ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

A Tunable Forced Alignment System Based on Deep Learning: Applications to Child Speech

Prad Kadambi,[a,b] Tristan J. Mahr,[c] Katherine C. Hustad,[d] and Visar Berisha[a,b]

[a] School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe [b] College of Health Solutions, Arizona State University, Tempe [c] Waisman Center, University of Wisconsin–Madison [d] Department of Communication Sciences and Disorders, University of Wisconsin–Madison

ARTICLE INFO

Article History:
Received May 23, 2024
Revision received October 25, 2024
Accepted December 3, 2024

Editor-in-Chief: Maria Grigos
Editor: Marisha Speights

https://doi.org/10.1044/2024_JSLHR-24-00347

ABSTRACT

Purpose: Phonetic forced alignment has a multitude of applications in automated analysis of speech, particularly in studying nonstandard speech such as children's speech. Manual alignment is tedious but serves as the gold standard for clinical-grade alignment. Current tools do not support direct training on manual alignments. Thus, a trainable speaker adaptive phonetic forced alignment system, Wav2TextGrid, was developed for children's speech. The source code for the method is publicly available along with a graphical user interface at https://github.com/pkadambi/Wav2TextGrid.
Method: We propose a trainable, speaker-adaptive, neural forced aligner developed using a corpus of 42 neurotypical children from 3 to 6 years of age. Evaluation on both child speech and on the TIMIT corpus was performed to demonstrate aligner performance across age and dialectal variations.
Results: The trainable alignment tool markedly improved accuracy over baseline for several alignment quality metrics, for all phoneme categories. Accuracy for plosives and affricates in children's speech improved more than 40% over baseline. Performance matched existing methods using approximately 13 min of labeled data, while approximately 45–60 min of labeled alignments yielded significant improvement.
Conclusion: The Wav2TextGrid tool allows alternate alignment workflows where the forced alignments, via training, are directly tailored to match clinical-grade, manually provided alignments.
Supplemental Material: https://doi.org/10.23641/asha.28593971

Technologies that automatically align a given speech sample and corresponding orthographic transcription such that individual phonemes and pauses are marked and labeled have the potential to transform clinical assessment in speech language pathology. Such audio-to-phoneme alignment is a fundamental step in many speech processing tasks such as calculation of vocal quality measures (Lin & Wang, 2011; Sonderegger & Keshet, 2012), text-to-speech synthesis (Okamoto et al., 2019; Peng et al., 2024), pathological speech analysis (Fontan et al., 2015; Stegmann et al., 2020), and pronunciation scoring for children's speech (Mathad et al., 2021; Oppelstrup et al., 2005) for L2 language learning (Witt & Young, 1997). Phonetic alignments are valuable for their use in service of a downstream task such as tracking phoneme acquisition, voice onset time computation, pronunciation scoring, and use in a speech synthesis system.

Despite the many use cases of alignment, the generation of clinical-grade alignment accuracy requires substantial effort using manual labeling. Manual alignment annotation is a time-intensive process that can take up to 400 times as long as the original duration of speech (Godfrey et al., 1992). One alternative to manual alignment is manual correction, which requires that phoneme boundaries generated by an automated tool are inspected and manually adjusted to their appropriate location if necessary. Although this significantly reduces the workload on annotators, manual correction can still take 10–30 times as

Correspondence to Prad Kadambi: pkadambi@asu.edu. **Publisher Note:** This article is part of the Special Issue: Select Papers From the 2024 Conference on Motor Speech—Basic Science and Clinical Innovation. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

long as the length of the original speech (Mahr, Berisha, et al., 2021). Therefore, many speech analysis pipelines that require alignments rely wholly on automated tools, referred to as forced aligners, for phonetically aligning speech.

Automated phonetic alignment systems, or forced alignment systems, typically use acoustic models from automatic speech recognition (ASR) pipelines to assign phoneme boundaries (Yuan et al., 2013). Many existing tools such as EasyAlign (Goldman, 2011), Prosodylab (Gorman et al., 2011), Kaldi (Povey et al., 2011), and the Montreal Forced Aligner (MFA; McAuliffe et al., 2017) address the forced alignment problem. These commonly used tools use hidden Markov models (HMMs) and acoustic models that are either monophone (Goldman, 2011; Gorman et al., 2011) or triphone (McAuliffe et al., 2017; Povey et al., 2011). In addition to conventional HMM systems, neural network forced alignment systems exploit the performance increases afforded by the use of pretrained deep learning models and large data sets. For example, NeuFA (Li et al., 2022) augments a Tacotron-inspired (Y. Wang et al., 2017) text encoder with a bidirectional attention mechanism and phoneme boundary detector. Zhu et al. (2022) adapted a pretrained Wav2-Vec2 (W2V2) model (Baevski et al., 2020) with a contrastive loss to learn alignments in a semisupervised fashion.

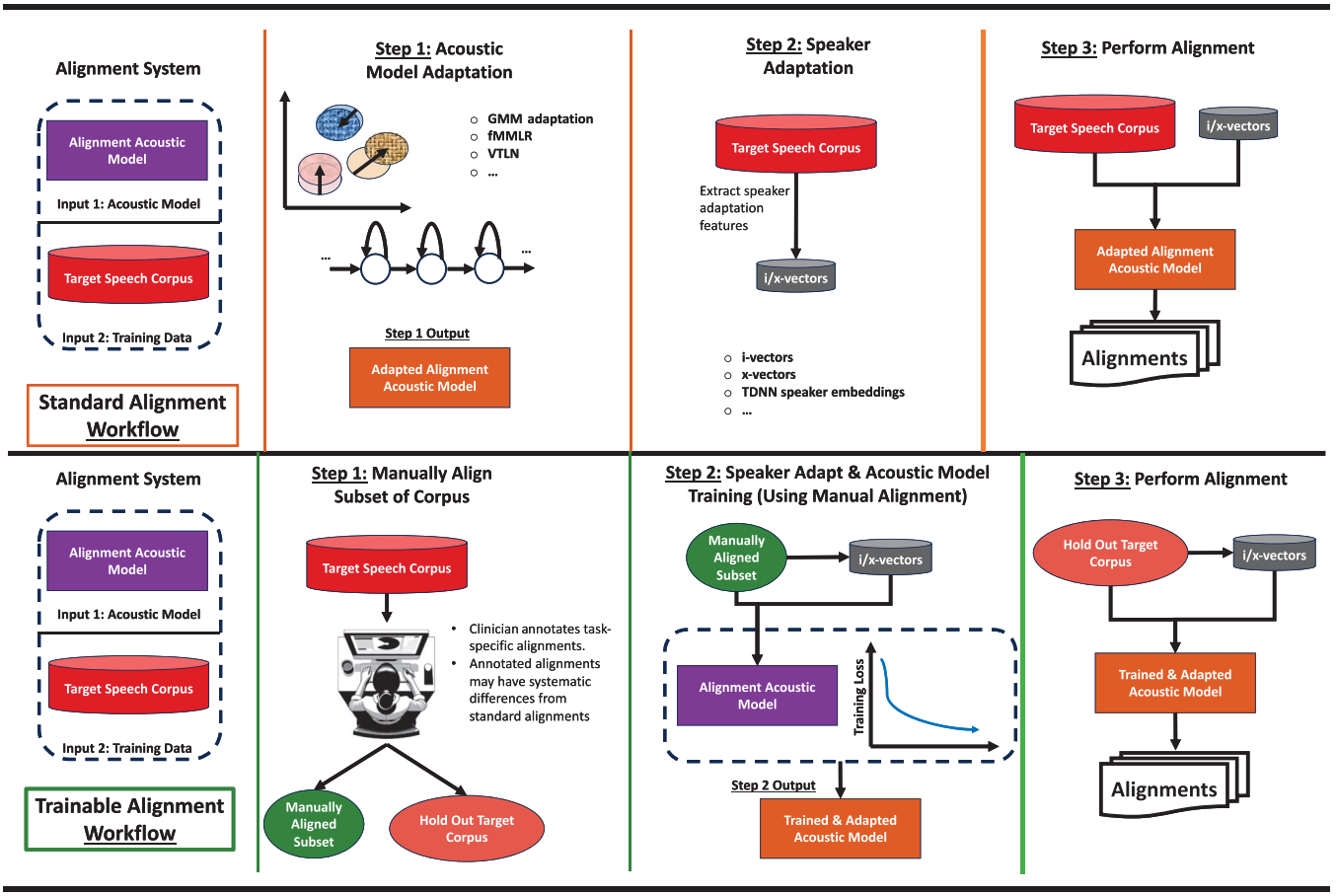
High-throughput, large-scale analysis of speech is only useful if the aforementioned systems can produce accurate forced alignments that can closely match gold standard, clinical-grade alignments without requiring manual correction. The problem of inaccurate forced alignments is compounded for child speech as the acoustic models that underlie the most popular forced alignment algorithms are developed and evaluated on adult speech corpora. Speech development in children poses yet another challenge as there are age-dependent increases in WER when children's speech is processed by acoustic models trained on adult speech; the WER for child speech can be up to 2–5 times the error rate for adult speech (Yeung & Alwan, 2018). As children exhibit far higher intra- and interspeaker variability than adults (Beckman et al., 2017), several approaches to modifying ASR acoustic models or acoustic pipelines for child speech have been proposed. These include feature augmentation techniques (Leggetter & Woodland, 1995) such as vocal tract length normalization (VTLN; Gerosa et al., 2007), speaker adaptation techniques (Gerosa et al., 2009) such as speaker-specific pronunciation dictionaries (Shivakumar et al., 2014), speaker-specific pitch vectors (Tai et al., 2022), and factor analysis methods (Dehak et al., 2010). The performance of forced alignment algorithms on child speech and the impact of speaker adaptation methods for children's speech remains understudied.

Two studies (Knowles et al., 2018; Mahr, Berisha, et al., 2021) analyzed the performance of state-of-the-art forced alignment algorithms (based on HMMs) on child speech. Knowles et al. (2018) observed age-dependent changes in alignment accuracy when using the Prosodylab aligner. In the study of Mahr, Berisha, et al. (2021), off-the-shelf forced alignment methods including Prosodylab, MFA, and Kaldi were evaluated. When compared with interrater alignment accuracy, they found that the MFA with speaker adaptive training (SAT) performed the best, within 5 percentage points of human interrater alignment accuracy. Both studies identify the training of the acoustic model on child data as a key factor in improved alignment accuracy, and Mahr, Soriano, et al. (2021) identified speaker adaptation as crucial for generating high-quality alignments.

Figure 1 shows a standard alignment workflow for most existing methods. First, aligner acoustic models are adapted to the target corpus. Subsequently, speaker adaptation is performed. Finally, the resultant model with both acoustic and speaker adaptation is used to generate alignments. We note that some earlier alignment tools do not follow this precise workflow; some may not even support acoustic model adaptation (Goldman, 2011) or may not support speaker adaptation (Goldman, 2011; Gorman et al., 2011), but the general workflow of existing tools is based on adapting an acoustic model to the target corpus. Another critical constraint with this workflow is that it is not possible to provide gold standard manual labels for training the aligners. This is a significant limitation as, in many cases, it is useful for researchers and clinicians to define specific alignment rules customized for some downstream task. For example, the forced alignment rules for speech from adult patients with Parkinson's disease are not likely to be the same as those for typically developing children. Notably, developmental features such as phonological processes or differences in speech rate and fluency require specialized alignment consideration. In this work, we propose an alternative workflow, one that allows for training of forced alignment systems on clinical-grade child speech alignments provided by experts to train and adapt alignment models jointly. This alternative trainable alignment workflow is also shown in Figure 1. A subset of the corpus is manually aligned following a well-specified labeling protocol. These labels are then used to train the acoustic model (with speaker adaptation) to directly adopt and mimic the rules of the manual forced alignments. Finally, the resultant alignment acoustic model can be used to perform alignment at scale on large quantities of data.

The key goals of this work are as follows. First, we operationalize this trainable alignment workflow with our algorithm, Wav2TextGrid, a general-purpose W2V2 forced alignment system with x-vector speaker adaptation. We

Figure 1. Comparison of a standard alignment workflow (top row) to our trainable alignment workflow (bottom row). In the standard alignment workflow, the acoustic model is first adapted to the target speech corpus, speaker adaptation is applied, and alignments are generated. In contrast, a trainable alignment workflow allows for gold-standard task-specific alignment rules to be learned by the acoustic model. A manual annotator provides task-specific alignments for a subset of the corpus. Speaker adaptation and acoustic model adaptation can still be performed, but the model is also trained using the manual alignments to produce a model that has learned task-specific alignment rules. Thus, final generated forced alignments are generated with a model that has learned the gold-standard alignment rules. GMM = Gaussian mixture model; VTLN = vocal tract length normalization; fMLLR = feature space maximum likelihood linear regression.



outline our algorithm, which allows for training and speaker adaptation of a forced alignment acoustic model on expert annotated alignments. Second, the workflow and algorithm are thoroughly validated using a multitude of alignment accuracy measures and across several phoneme categories to show that it matches gold standard manual alignments more closely than existing tools (which cannot be trained on manual alignments). Both TIMIT and a custom child speech corpus are used to show the applicability of Wav2TextGrid in populations across the age range. Third, as manual alignment annotation is arduous, we also quantify the amount of labeled data (i.e., minutes of audio that must be manually aligned) required by our workflow (a) to achieve equivalent performance to existing baseline tools and (b) to outperform this baseline. Finally, we demonstrate the practical value of our system on a target application of forced alignment: pronunciation evaluation. Pronunciation scores are calculated with both forced alignment methods and manual alignments, demonstrating that forced alignment methods that match manual annotations more closely also yield pronunciation scores with improved correlation with the gold standard pronunciation score derived from manual alignments. The code for the method is shared, and the method is installable as a command line utility or a graphical user interface.

Data

Child Speech Corpus

A speech corpus consisting of 42 typically developing children was used. By age group, 10 children were 3–4 years old, 10 children were 4–5 years old, 12 children were 5–6 years old, and 10 children were 6–7 years old.

Each age range was also gender matched. The children were native English speakers and spoke standard American English in the home. Standardized articulation and language tests and a hearing screening (parent report and pure-tone hearing screening) confirmed that the children had typically developing speech, language, and hearing abilities. For additional details regarding tests conducted to ensure typical development, the reader is directed to Mahr, Berisha, et al. (2021). Approval was granted for the study by the University of Wisconsin–Madison institutional review board (Social and Behavioral Sciences, MRR IRB 2016–0574). An informed consent form was obtained on behalf of all participants.

Speech Elicitation Task

Speech elicitations were selected from the set of stimuli provided in the Test of Children's Speech (Hodge & Daniels, 2007). Stimuli included single words, multiword phrases, and sentences. In a sound-attenuated room, speech-language pathologists (SLPs) collected speech with a picture prompted repetition task. Children heard a prerecorded version of the prompt along with an appropriate image/scene and had to repeat the prompt. Speech was recorded at a 44.1 kHz sampling rate using an Audio-Technica AT4040 microphone. A total of 3,764 files (total duration 128 min) were collected. An average of 89 utterances were elicited per child.

Manual Alignment Annotation

Two trained manual annotators, a certified SLP and an SLP graduate student, provided manual annotations of phonetic boundaries (labeling 2,801 and 963 files, respectively). Alignments were used as the ground truth labels for training the alignment system. Forced alignments were first generated using Prosodylab, and subsequently corrected manually by adjusting phoneme time boundaries in Praat. The alignment procedure consisted of (a) repairing gross alignment errors specific to the Prosodylab aligner (such as the insertion of false pause intervals) and (b) moving phoneme boundaries to their appropriate location using conventional spectral landmarks. Annotators did not remove phoneme intervals or relabel them. For example, despite the many interesting ways a young child can say the word animal, every animal token was annotated using the phoneme intervals /æ/, /n/, /ə/, /m/, /ə/, /l/. The two manual annotators calibrated their alignments by manually aligning two "training" speakers (a 3-year-old and a 6-year-old), and discrepancies were reviewed with the second author. In order to characterize interrater variability, both annotators labeled all files for four of 42 children (406 files; the four speakers in this subset were age-matched with the rest of the corpus). For a detailed overview of the annotation process used, the reader is directed to Mahr, Soriano, et al. (2021).

TIMIT

In addition to the child speech corpus, the TIMIT corpus (Garofalo et al., 1993) was used to evaluate the aligner performance on adult speech. Annotated phoneme boundaries in TIMIT use a phoneme set heavily inspired by ARPABET, an inexact match for the phoneme set used in this work. Our aligner used the CMU (Carnegie Mellon University) pronunciation dictionary (CMU, 1998). Thus, to map from the 61 phoneme classes in TIMIT to the 39 classes in the CMU dictionary, we follow the approach suggested by Lee and Hon (1989). Because the mapping of the phoneme /DX/ (flap [ɾ]), was ambiguous and context dependent, it was unmodified. During training, /DX/ frames were ignored, and during testing they were retained as they reduced the performance of all methods equally.

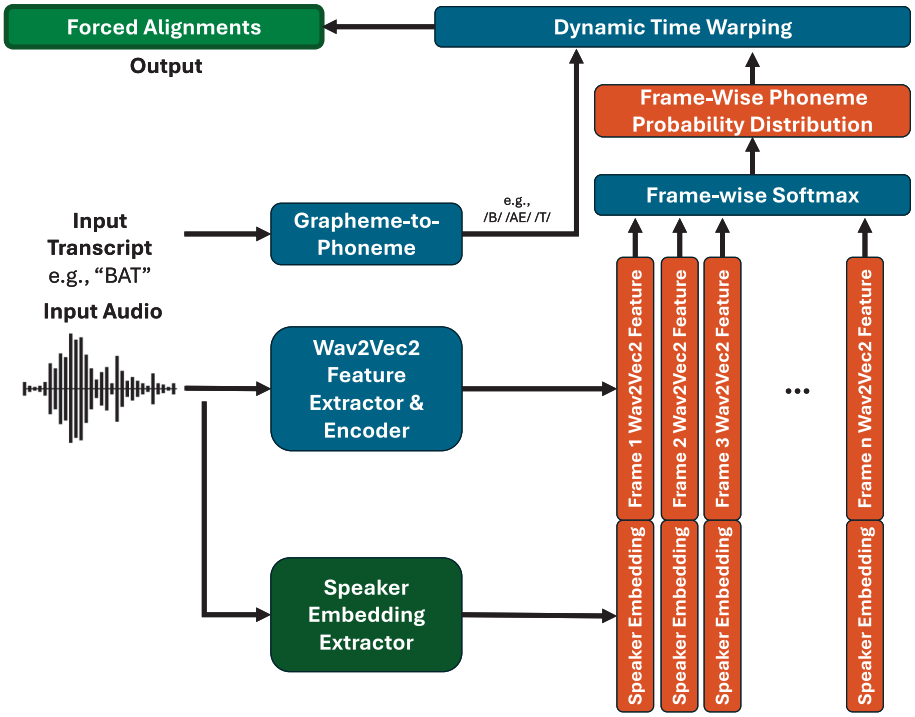
Method

W2V2 for Deep Forced Alignment

Figure 2 provides a block diagram of the alignment system used in this study. The method follows the similar core approach found in Zhu et al. (2022). In their work, a semisupervised W2V2 forced alignment system was trained on Librispeech (Panayotov et al., 2015) and CommonVoice 6.1 (Ardila et al., 2019). First, an initial alignment model was learned in unsupervised fashion using W2V2 with a BERT (Devlin et al., 2016) inspired phoneme sequence encoder. Their model was trained using a doublet loss including a contrastive term and a forward sum alignment loss commonly used in HMM training. For the final alignment model, they train a freshly initialized pretrained W2V2 using the output of the resultant model from the previous step previous model as ground truth. We use this final alignment model to initialize the weights of our system (W2V2 feature extractor, W2V2 feature encoder, and W2V2 classification head in Figure 2).

The system takes an audio file and its transcript as input. First, a grapheme-to-phoneme (G2P) conversion is applied on the transcript to obtain the expected phoneme sequence. G2P conversion was performed using the g2p_en python package (Park & Kim, 2019). The g2p_en package uses the CMU Pronunciation Dictionary (CMUdict; CMU, 1998) for in-vocabulary words. For out-of-vocabulary words, the package employs a neural network model to predict the phoneme sequence. For heteronyms, it uses part-of-speech tagging and context based parsing to assign the appropriate phoneme sequence. The input audio is passed into the aforementioned W2V2 network to generate per-frame feature vectors. Typically, W2V2

Figure 2. Block diagram of speaker adaptive Wav2Vec2 forced alignment system. The system requires an audio file and its corresponding transcript. A grapheme-to-phoneme conversion is used to generate the expected phoneme sequence for alignment. From each input audio file, a speaker embedding is extracted, frame-wise Wav2Vec2 features are calculated, and the speaker embedding is appended to each per-frame feature vector. A per-frame softmax provides a probability distribution over all phonemes for each frame. To produce the final alignment, dynamic time warping uses the per-frame phoneme probabilities to assign each phoneme in the expected phoneme sequence to a frame.



produces output feature vector frames at an interval of 20 ms. However, for W2V2 trained for forced alignment, the stride of the final convolution layer was reduced from 2 to 1 to generate feature vectors every 10 ms. This improved the granularity of the predicted phoneme boundaries.

Subsequently, a W2V2 classification head (frame-wise softmax) was applied on a frame-by-frame basis on the feature vectors to obtain the per-frame phoneme probabilities. Finally, from both the per-frame phoneme probability distributions and the expected phoneme sequence from G2P conversion, a Viterbi decoding step generated the final forced alignments. Given the expected phoneme sequence for the utterance, the most likely monotonic trajectory for the expected phoneme sequence was found across frames with Viterbi decoding.

A Trainable Alignment System

Three variants of the W2V2 alignment model were studied. The first model, which we designate *W2V2Base*, was identical to the frame-wise classification model in Zhu et al. (2022), termed "W2V2-FC-10 ms" in their work. No additional training, modification, or adaptation was performed for *W2V2Base*. The second variant, *W2V2Trained*, refers to a version of the aligner fine-tuned using manually annotated alignments as ground truth. The model was trained using a per-frame cross-entropy loss between the aligner predicted phoneme label for a frame, and the frame's ground-truth, manually annotated phoneme label. Finally, a third variant, W2V2TrainedSAT, was trained similar to W2V2Trained, but x-vector speaker adaptation (Snyder et al., 2018) was added to the model.

Speaker Adaptation

Speaker adaptation was performed by extracting speaker embeddings from each audio file. Given an input speech segment, a single speaker embedding vector was extracted for the entire segment. As shown in Figure 2, the extracted embedding vector was simply appended to the per-frame calculated W2V2 features, and the frame-wise softmax is performed over the concatenated W2V2 and speaker embedding vector. This allows flexibility in the choice of type of speaker embedding. An off-the-shelf, trained x-vector extractor network trained on the Voxceleb 1 (Nagrani et al., 2017) and Voxceleb 2 (Chung et al., 2018) data sets was used to extract a 128-dimensional x-vector for each utterance. The x-vector

extractor was not trained on the data sets used in this study.

Model Training

All W2V2 models were trained and implemented using the Huggingface transformers package in python (Wolf, 2019). The models were trained for 25 epochs with a learning rate of 2e−4, batch size of 64, and cosine learning rate decay. The training loss was a per-frame cross entropy loss between the aligner predicted phoneme class for the frame and the manually annotated phoneme label for the frame.

Baseline Comparison

The widely used MFA served as a baseline. The MFA was chosen as a baseline following previous observations by Mahr, Berisha, et al. (2021) that the MFA soundly outperformed other off-the-shelf alignment algorithms when evaluated on this child speech corpus. The acoustic model used with the MFA was the english_us_arpa model, consisting of a joint Gaussian mixture model (GMM) and HMM trained on Librispeech 960 h. Several configurations of the MFA were assayed. Alignments were generated without speaker adaptation (MFA_NoSAT), with speaker adaptation (MFA_Align), and by adapting the acoustic model's GMM means to the target corpus (MFA_Adapt). Additionally, in a fourth configuration, MFA_Train, a fresh HMM-GMM model was trained from scratch on the target corpus prior to alignment generation.

Model Training and Evaluation Method

Child Data Set

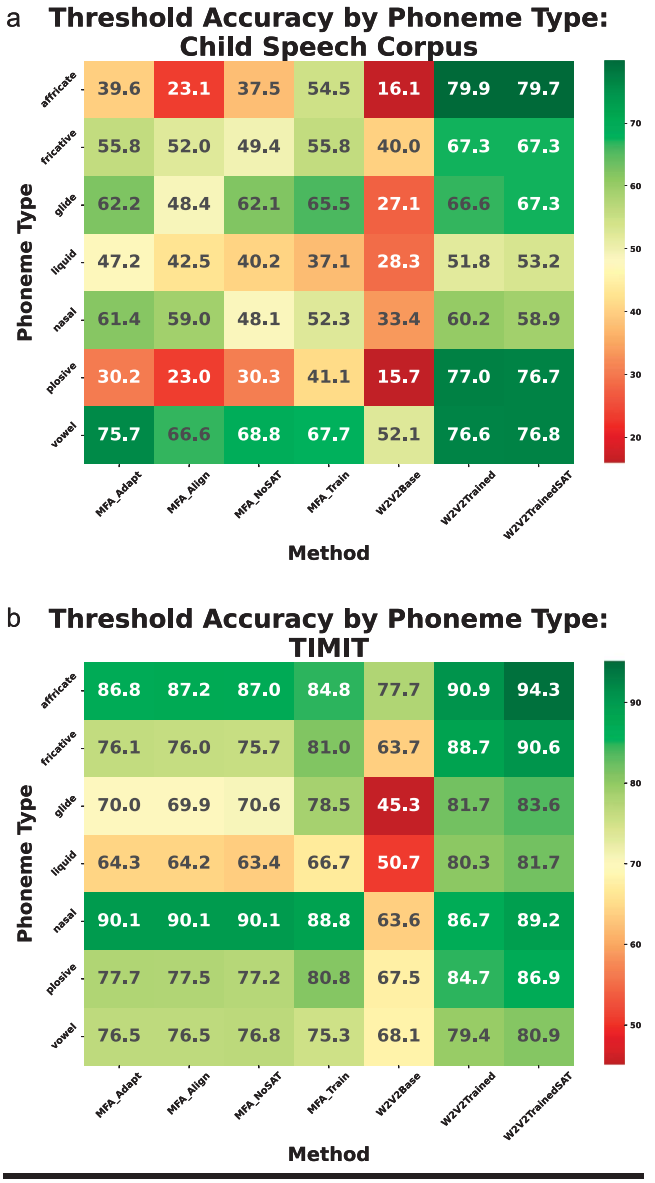
For Figure 3 and Tables 1 and 4, leave-one-speaker-out cross-validation (LOSOCV) was used to generate alignments. In LOSOCV, data from a speaker are held out, and the alignment model is trained on data from all remaining speakers and evaluated on the held out speaker's data. This process is repeated for each speaker in the corpus. Thus, in LOSOCV, we train 42 alignment models for the 42 speakers in our child speech corpus.

To evaluate alignment performance relative to inter-rater accuracy, a standard train/test split was used. The four speakers with two sets of manual annotations were used as the holdout set while the remaining 38 speakers comprised the training set. This train/test split was also used in generating Figure 4.

TIMIT

Results reported on TIMIT did not use LOSOCV; rather, the standard TIMIT train/test split was used. This is because the TIMIT corpus is much larger than the manually labeled child speech corpus.

Figure 3. (A) Heatmap of threshold accuracy by phoneme type and alignment method for the child speech corpus used in this study. Threshold accuracy is calculated in accordance with Kreuk et al., 2020, and Zhu et al., 2022, measuring the percentage of onset boundary errors smaller than 20 ms threshold. (B) Heatmap of threshold accuracy (percentage of onset errors less than a 20 ms threshold) by phoneme type and alignment method for the TIMIT data set.



Statistical Evaluation and Alignment Quality Metrics

To ensure thorough evaluation of forced alignment quality against manual annotation, several alignment quality metrics were calculated in accordance with

Table 1. Alignment accuracy metrics for the child speech data set analyzed in this study.

Method	Midpoint accuracy	Overlap percentage	Threshold accuracy (onset error < 20 ms; %)	Median onset error (ms) [IQR]	Median percent onset error (%)	Median offset error (ms) [IQR]	Median percent offset error (%)
MFA_NoSAT	78.9	73.1	54.8	16.6 [42.5]	13.4	20 [41]	14.9
MFA_Align	82	75.5	54.7	16.8 [29.7]	12.5	20 [28.5]	14.2
MFA_Adapt	86.8	79.5	59.5	13.5 [25.9]	10.5	18.8 [23.3]	12.1
MFA_Train	83.2	76.5	58.7	13.5 [32.6]	11.2	18.8 [30.6]	12.7
W2V2Base	74.4	68	42.5	25.5 [47.2]	18.7	27.5 [50]	20.3
W2V2Trained	93.9	85.6	73.2	*10 [17.3]*	7.4	*10 [17.1]*	8.1
W2V2TrainedSAT	*95.1*	*86.6*	*75.4*	*10 [17.3]*	*7.1*	*10 [16.9]*	*7.7*

Note. The best performance is denoted by italics. When more than one method performs best, all the methods are marked with italics. However, if only one method achieves the best performance, that method is highlighted using both italics and underline. IQR = interquartile range; MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.

previous studies on evaluating alignment. The onset and offset boundaries of a ground truth interval are denoted t_{on} and t_{off}, and the onset and offset boundaries of an aligner predicted interval are denoted a_{on} and a_{off}, respectively. Alignments were evaluated against ground truth annotations across seven phoneme categories and across all phonemes. The seven phoneme categories were vowels (in ARPABET: AA, AE, AH, AO, AW, AY, EH, ER, EY, IH, IY, OW, OY, UH, UW), nasals (M and N), plosives (P, B, T, D, K, G), affricates (CH, JH), fricatives (F, V, TH, DH, S, SH, Z, HH), liquids (L, R), and glides (W, Y). The following metrics were calculated.

Onset error and offset error. The onset error and offset error were calculated as the absolute difference between the ground truth boundaries and the aligner predicted boundary. Thus, for a given interval, onset error was calculated as $|t_{on} - a_{on}|$ and offset error was calculated as $|t_{off} - a_{off}|$. Cumulative distribution plots were generated to help visualize onset and offset errors.

Percentage onset error and offset error. Phoneme intervals can greatly vary in duration depending on phoneme type (see Supplemental Material S2). For example, a predicted interval with 20 ms boundary error for a long vowel may still contain most of the ground truth interval. However, such an error for a short burst consonant may result in most of the ground truth interval being missed. Thus, percentage onset and offset error is reported as well by normalizing by the length of the ground truth interval. Percentage onset error is calculated as $\frac{100 * |t_{on} - a_{on}|}{t_{off} - t_{on}}$.

Midpoint accuracy. Previous studies on alignment for child speech have used a midpoint containment accuracy metric (Knowles et al., 2018; Mahr, Berisha, et al., 2021). This metric calculated the percentage of predicted phoneme intervals that contained the midpoint of the ground truth, manually annotated phoneme interval. Age dependence of alignment error was analyzed by plotting the average alignment accuracy for each speaker versus age.

Threshold accuracy. Another previously used (Kreuk et al., 2020; Michel et al., 2016; Y.-H. Wang et al., 2017; Zhu et al., 2022) metric evaluates threshold accuracy of onset boundaries or offset boundaries. Intervals with onset error smaller than a threshold, τ, are considered correct, while intervals with onset error larger than τ are considered incorrect. Following Zhu et al. (2022), we select τ = 20 ms. As frame size is 10 ms, threshold accuracy is essentially a binary outcome variable that measures whether the predicted boundary was within two frames of the ground truth boundary.

Overlap percentage. Overlap percentage for a given interval was calculated as $\frac{100 * Overlap(s)}{t_{off} - t_{on}}$. The overlap between the manual interval and predicted interval is calculated and then divided by the length of the ground truth interval.

Interrater Agreement Evaluation

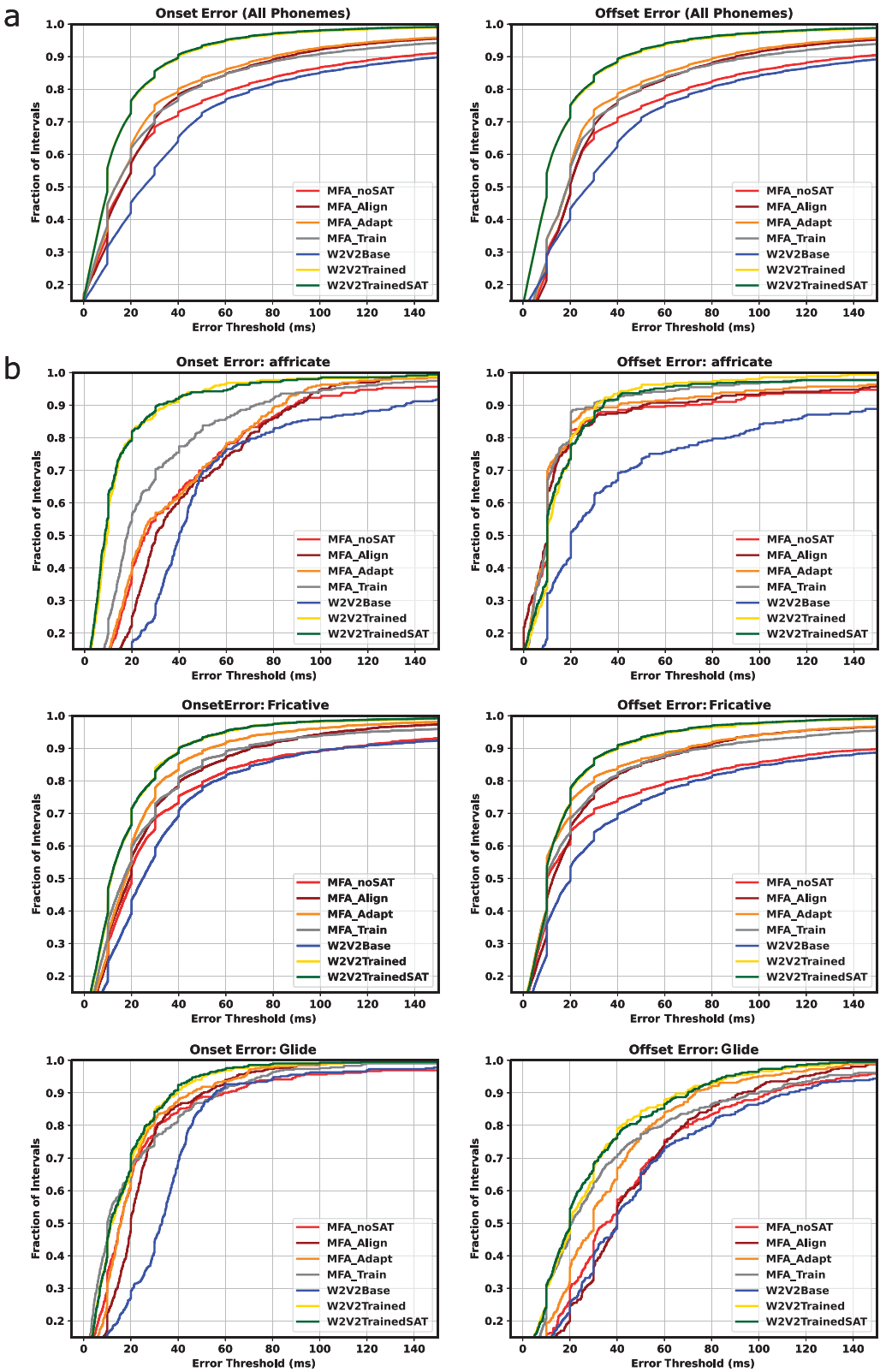
Often, manual segmentation of phoneme intervals involves subjective judgments in boundary placement, prompting the need for evaluating interrater agreement. Interrater accuracy metrics were calculated in the following manner. As described in the Manual Alignment Annotation section, of 42 speakers, four speakers contained annotations from two human annotators. One of the two sets of manual annotations was arbitrarily chosen as the reference alignments, and the accuracy of the alignments from the other annotator was evaluated against this reference. Interrater accuracy was also used to provide an upper bound or human level accuracy.

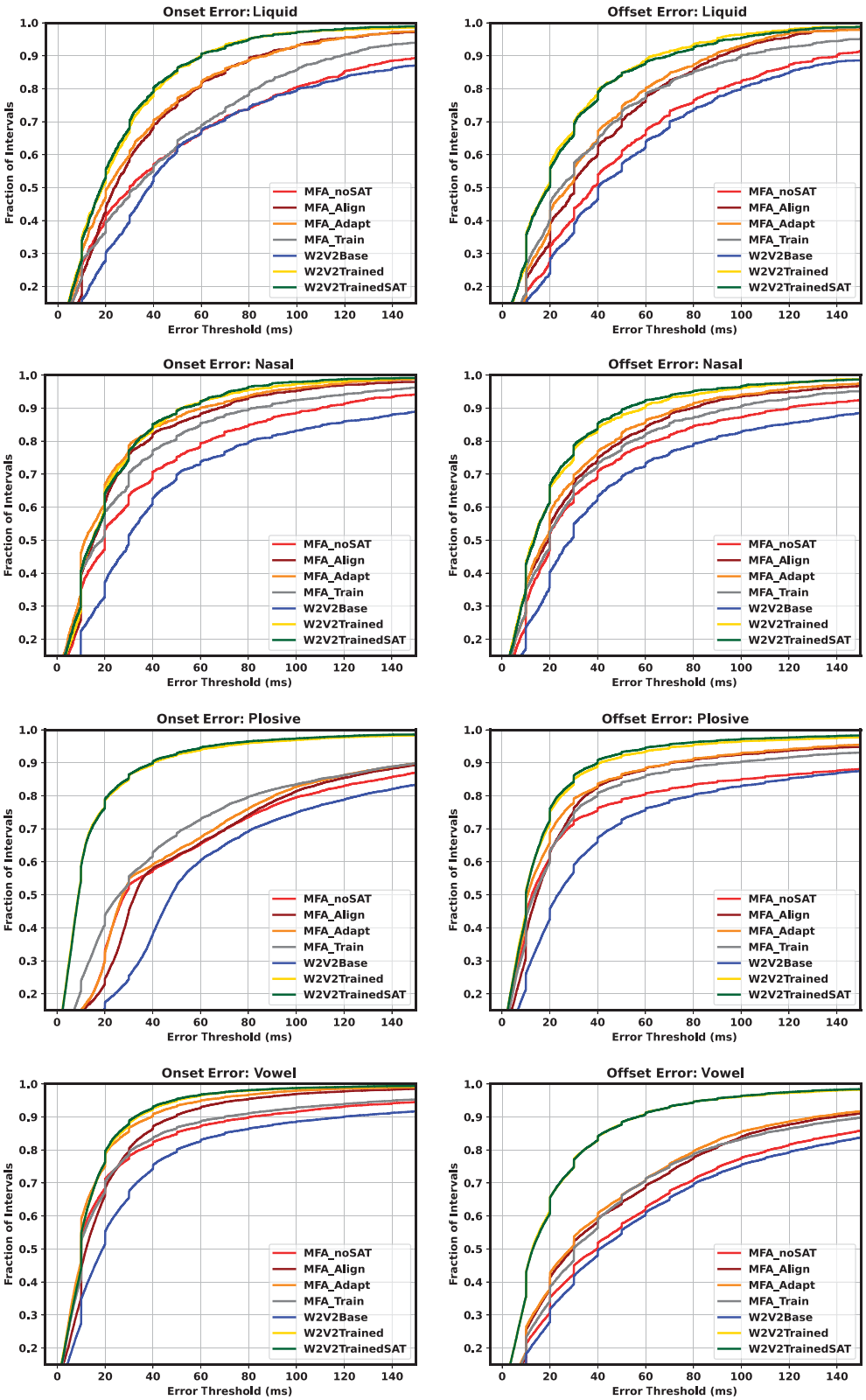
Out of Sample and Out of Utterance Evaluation

Child Speech

It is also necessary to evaluate the performance of the system for utterances not seen during training. For this purpose, we begin with the same data subsetting described in the Manual Alignment Annotation section.

Figure 4. (A) Empirical cumulative distribution function of onset error and offset error for all phoneme intervals in the child speech corpus. The *x*-axis is the error threshold in milliseconds. The *y*-axis denotes the fraction of phoneme intervals that have onset or offset error less than the threshold. (B) Empirical cumulative distribution of onset error and offset errors for by phoneme subtype. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.





As described in the Manual Alignment Annotation section, four of 42 speakers were held out for testing. This data set contained 3,358 files in the training set and 406 files in the test set. To create a test set with utterances not seen in the training set, we observed that the data set contained 98 unique utterances and randomly selected 30 of these utterances. Only files containing these 30 utterances were retained in the test set, while any file containing those utterances was removed from the training set. This resulted in a training data set with 2,554 files and a testing data set with 80 files (the test data set still consisted of four speakers). Then, the same training procedure in the Model Training section was used.

TIMIT

The TIMIT train data set consisted of 4,620 files, while the test data set consisted of 1,680 files. The training set contained 1,733 unique utterances, and the test set contained 632 unique utterances, with only two utterances overlapping between them. Therefore, no additional out of utterance evaluation was performed for TIMIT since the training and testing set were already sufficiently isolated by utterance.

Aligner Performance as a Function of Training Data Set Size

In many cases, it may not be practical to manually annotate a large set of files. However, both corpora used in this study contain a relatively substantial number of manually aligned files. Understanding of the W2V2 aligner's performance and the amount of labeled manual alignment necessary for the trainable alignment workflow requires characterizing the curve of alignment accuracy as a function of training data set size. We focus on the child speech corpus for this experiment as it is the largest deviation from the healthy adult speech used to train the aligner's acoustic model and likely requires the most labeled training data.

First, a randomly sampled subset of the training data was generated. Then the aligner (W2V2TrainedSAT) was trained on the training data subset and evaluated on the full test set. The size of the randomly generated subsets varied from 2% to 50% the size of the full training data set (2.5–64 min). The train/test split used in the Interrater Agreement Evaluation section was also used in this section. Performance was compared against the three references: the MFA baseline (a method that does not need manual alignments), W2V2TrainedSAT given the full training data set, and interrater accuracy.

Comparison of Downstream Metrics Calculated Using Manual Versus Forced Alignment

Intervals produced by alignment (automatic or manual) are often subject to downstream acoustic measurements such as measuring a phoneme's intensity, vowel's formant values, a speaker's rate of articulation, and so on. If forced alignment boundaries can provide a reliable replacement for gold standard manual alignments, then downstream measurements made on forced-aligned intervals should likewise be reliable compared to measurements made on manual intervals. We assess this criterion by comparing pronunciation score measurements on intervals produced by both alignment systems.

Pronunciation Scoring

A common application of forced alignments is automatic pronunciation scoring. Forced alignments are used to calculate goodness of pronunciation (GOP) scores (Witt & Young, 2000). Tu et al. (2018) describe one method of calculating GOP score using the phoneme log-likelihood ratio (PLLR). PLLR is the per-frame average of the ratio between the log probability of the target phoneme and the log probability of the acoustically most likely phoneme; in other words, the probability of the expected sound (the target) is compared to the probability of the produced sound. Alignments inform the expected phoneme for each frame, and if two alignment systems predict identical intervals, PLLR scores should also be identical. The acoustic model and code provided by Vidal et al. (2021) was used to compute PLLR. Average PLLR was calculated for each file in the corpus. The Pearson correlation between PLLR score generated using manual alignment and PLLR score for each forced alignment method was calculated. This analysis was conducted for the child speech corpus.

Results

Alignment Evaluation for Child Speech

Aggregate Alignment Accuracy

Table 1 summarizes midpoint accuracy, overlap percentage, threshold accuracy, and median onset/offset error metrics for the forced alignment methods. We find that these alignment accuracy metrics are highly correlated and relative ranking between alignment methods is preserved across accuracy metrics.

Out of all MFA methods, MFA_Adapt performed the best, as shown by the highest midpoint accuracy (86.8%), highest overlap percentage between manual and predicted intervals (79.5%), and highest threshold accuracy (59.5%) among the other MFA methods. In MFA_Adapt, the en_us_arpa model's GMM parameters were adapted to child speech corpus. As it was the best performing MFA method, it served as a benchmark for comparison against our W2V2 methods.

W2V2 models that were trained the manual alignments (W2V2Trained, W2V2TrainedSAT) outperformed all MFA methods. However, W2V2Base was not trained on manual alignments, performed worse than all MFA methods, and performed the worst overall. Over the MFA_Adapt baseline, W2V2TrainedSAT achieves improvements of 8.3 percentage points (pp), 7.1 pp, and 15.9 pp in midpoint accuracy, overlap percentage, and threshold accuracy respectively (see Table 1). Speaker adaptation improved midpoint accuracy for the MFA by 3.1 pp (MFA_noSAT vs. MFA_Align), and speaker adaptation improved midpoint accuracy by 1.2 pp for W2V2 (W2V2Trained vs. W2V2TrainedSAT).

Aligner Performance by Phoneme Type

A heatmap of the threshold accuracy is shown for each phoneme subtype in Figure 3a. W2V2Base performed worst out of all methods for and for all phoneme categories since it was not trained on any alignment data. Excluding W2V2Base from our discussion, it is clear that W2V2 methods outperformed the MFA on all phoneme subtypes. For both W2V2 and MFA, the phoneme subclass with the highest threshold accuracy was vowels. However, for plosives, W2V2 methods trained on manual alignments significantly improved accuracy from approximately 30% for the MFA to above 75% for W2V2 and also improved affricate accuracy from the MFA's approximately 40%–50% to nearly 80% for W2V2. The liquids, /R/ and /L/, were the most challenging subclass to align for W2V2 methods, resulting in approximately 50% threshold accuracy for the class. Despite this lower accuracy for liquids, both W2V2TrainedSAT and W2V2Trained both outperformed the MFA on liquids. Accuracy for glides and nasals was roughly comparable between MFA_Adapt and W2V2 methods.

Onset/Offset Error

In the previous discussion, phoneme boundary error was characterized as a binary outcome variable with respect to a 20-ms threshold. A complete treatment of alignment errors requires understanding the entire distribution onset and offset errors. In Figure 4a, the empirical cumulative distribution functions (eCDFs) of the onset error and offset error are shown for the child speech corpus for all phoneme intervals. In eCDF plots of error by phoneme type (see Figure 4b), the relative ranking of the alignment methods are consistent with Table 1 and are similar across levels of onset/offset errors and across phoneme type. Ordered from largest to smallest onset/offset error (rightmost to leftmost on the figure), W2V2Base had the largest boundary errors followed by MFA-based methods. Wav2Vec2Trained performed better still, and Wav2Vec2TrainedSAT performed the best.

Of particular note in Figure 4b were the following boundary types which had the largest error in absolute terms (in milliseconds) for all methods: offset error for liquids and offset error for glides. For the following boundary types, W2V2 achieved the largest improvement over MFA: vowel offset errors, plosive onset errors, and affricate onset error. These improvements are apparent in the horizontal gaps between the two W2V2 eCDF lines and the next MFA eCDF line. Discontinuities or jumps in the eCDF curves are due to the 10-ms granularity of the Prosodylab, MFA, and W2V2 aligners. (Hand alignments for the child speech corpus were generated by manually moving phoneme boundaries assigned by Prosodylab. Boundaries were only modified if corrections were necessary.) In Supplemental Material S1, eCDFs are also shown for percentage onset/offset error to show the distribution of error normalized to the length of the ground truth interval.

Interrater Accuracy Evaluation

Alignment accuracy was computed for all methods on the subset of four speakers with two sets of manual annotations (see Table 2). Wav2Vec2TrainedSAT and

Table 2. Interrater accuracy versus forced alignment accuracy.

Method	Midpoint accuracy	Overlap percentage	Threshold accuracy (onset error < 20 ms; %)	Median onset error (ms)	Median onset error (%)	Median offset error (ms)	Median offset error (%)	Num intervals
MFA_NoSAT	73.7	69.5	51.8	18.8	15.1	21.9	16.5	2149
MFA_Align	80.3	74.4	52.4	18.5	12.4	20.8	13.7	2220
MFA_Adapt	83.8	78	58.4	15	10.7	19.7	11.8	2285
MFA_Train	81	75.2	57.9	14.5	11.1	20	12.3	2284
W2V2Base	68.8	64	38.6	30	20.3	32.2	22.7	2296
W2V2Trained	92.8	85	70.4	10	6.9	10	7.6	2418
W2V2TrainedSAT	93.2	85.2	*71.2*	*10*	6.9	*10*	7.8	2428
Interrater	**93.6**	**86.1**	**75.5**	**6.8**	**4.9**	**7.6**	**5.9**	**2482**

Note. Alignments for a subset of four of 42 speakers in the child speech corpus were annotated by two manual annotators. For these four speakers, interrater accuracy and forced alignment accuracies are reported. The baseline interrater agreement is marked in bold. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.

W2V2Trained nearly matched interrater accuracy on all three alignment accuracy metrics reported. While threshold accuracy for W2V2TrainedSAT was 4.1 pp lower than interrater accuracy, midpoint accuracy and overlap percentage were within 1 pp of interrater accuracy.

Alignment Accuracy Across the Age Span

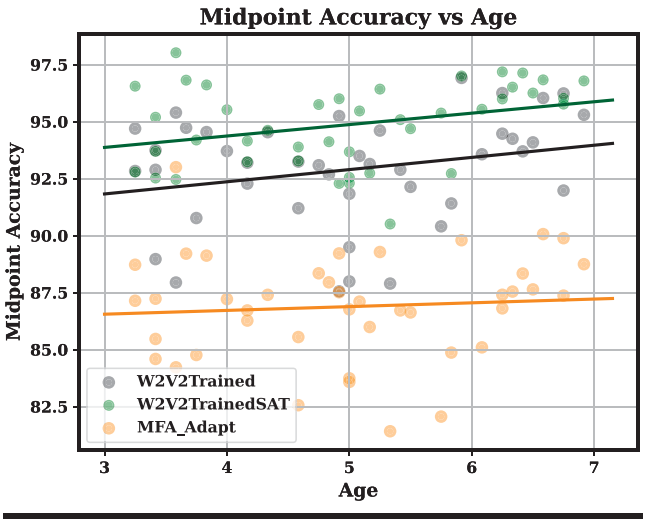
Figure 5 shows the effect of age on alignment accuracy. Alignment accuracy at any age was lower for MFA when compared to W2V2Trained and W2V2TrainedSAT. Expected midpoint accuracy for W2V2TrainedSAT varied from 94% for 3-year-olds to nearly 96% for 7-year-olds. Age only had a minor positive impact on accuracy for W2V2TrainedSAT, but for each method, the linear effect of age on alignment accuracy was statistically significant ($p < .05$). Additionally, on a per-speaker level, there were no outliers, that is, there were no individual speakers for whom alignment performance was significantly degraded.

Aligner Evaluation on Adult Speech: TIMIT

Aggregate Alignment Accuracy

To demonstrate the W2V2TrainedSAT system on adult speech as well as child speech, the model was trained and evaluated on the TIMIT data set. The standard train/test split was used. All MFA methods performed roughly on par with each other (see Table 3). The ranking of the methods by alignment accuracy on TIMIT was similar to the ranking on the child speech data set. W2V2Base performed the worst, achieving 52.4% threshold accuracy and 75% midpoint containment accuracy. W2V2TrainedSAT again outperformed all other methods.

Figure 5. Midpoint accuracy as a function of age for the child speech corpus plotted for MFA_Adapt, W2V2TrainedSAT, and W2V2Trained. A linear fit for each method shows the average accuracy at a given age. Each point represents one speaker. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.



Between MFA methods, adapting the acoustic model (MFA_Align vs. MFA_Adapt) had little effect on the accuracy—likely because TIMIT is similar to the training data used for the MFA's acoustic models. Additionally, speaker adaptation afforded much smaller increases in accuracy: between MFA_NoSAT and MFA_Align, midpoint accuracy and overlap percentage only improved by 0.2 pp and 0.1 pp, respectively. Speaker adaptation also only provided minor for W2V2. Between W2V2Trained and W2V2TrainedSAT, midpoint accuracy and overlap percentage improved by 0.7 pp and 0.2 pp, respectively.

Speaker Adaptation

The use of speaker adaptation increased midpoint containment accuracy by 3.1 pp for the MFA and increased the accuracy of W2V2 by 1.2 pp. However, on the TIMIT data set, this increase in accuracy was only 0.2 pp for MFA and 0.7 pp for W2V2. Studies on the effect of speaker adaptation in improving ASR word error rate (WER) have shown that speaker adaptation can provide a greater improvement to WER when the domain shift relative to healthy adult speech is greater (Geng et al., 2022).

Accuracy/Error by Phoneme Type

A heatmap of threshold accuracy by phoneme subclass was also generated for TIMIT (see Figure 3b). Similar to aggregate alignment accuracy, accuracy per phoneme subclass was also similar across all MFA methods. Nasals and affricates were the most accurately aligned phoneme subclass for all MFA configurations. Similar to results observed for child speech (see Figure 3a), liquids were the most challenging phonemes to align accurately. Comparison between W2V2TrainedSAT and the MFA reveals that the two alignment methods achieve similarly high accuracy for nasal phonemes, but W2V2TrainedSAT improves threshold accuracy by more than 9 pp for fricatives, 7 pp for affricates, 5 pp for glides, and 15 pp for liquids.

Out of Sample and Utterance Evaluation

Out of sample utterance performance of the alignment system was performed on a reduced version of the corpus using the method described in the Child Speech section. In Table 4, the performance of each assayed method is shown on this utterance isolated test set. Among the MFA-based methods, MFA_Train achieved the lowest median onset/offset error in milliseconds with 12.3 ms median onset error and 19.1 ms median offset error. Wav2Vec2TrainedSAT performed the best among all methods with 11.7 ms onset error and 12.8 offset error.

Although trained W2V2 methods outperformed MFA methods in onset/offset error, midpoint containment accuracy was lower. Specifically, midpoint containment accuracy for

Table 3. Alignment accuracy metrics for the TIMIT data set.

Method	Midpoint accuracy	Overlap percentage	Threshold accuracy (onset error < 20 ms; %)	Median onset error (ms) [IQR]	Median percent onset error (%)	Median offset error (ms) [IQR]	Median percent offset error (%)
MFA_NoSAT	78.4	82.1	64.6	8.4 [14.9]	14.9	7.8 [14.1]	14.1
MFA_Align	78.6	82.2	64.8	8.4 [14.8]	14.8	7.8 [14.2]	14.2
MFA_Adapt	78.8	82.3	64.9	8.4 [14.8]	14.8	7.7 [14.2]	14.2
MFA_Train	77.7	81.7	65.8	8.2 [13.9]	13.9	8.1 [14.1]	14.1
W2V2Base	75	78.6	52.4	14.1 [20.3]	20.3	12.6 [16.6]	16.6
W2V2Trained	78.5	*84.9*	70.5	*7.5 [11.0]*	11	7.5 [12.0]	12
W2V2TrainedSAT	*79.2*	84.7	*72.4*	*7.4 [10.0]*	*10*	*7.5* [10.9]	*10.9*

Note. The best performance is denoted by italics. When more than one method performs best, all the methods are marked with italics. However, if only one method achieves the best performance, that method is highlighted using both italics and underline. IQR = interquartile range; MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.

W2V2Trained and W2V2TrainedSAT was 80.3% and 80.8%, respectively, while accuracy for MFA_Adapt and MFA_Train was 85.4% and 83.1%, respectively. This is attributed to the higher interquartile range (IQR) in onset and offset error for W2V2Trained and W2V2TrainedSAT (onset error IQR 39.2 ms and 33.5 ms) versus MFA_Adapt and MFA_Trained (28.6 ms and 28.7 ms). Although trained W2V2 alignments had smaller median boundary error than MFA, larger alignment errors were slightly more frequent for W2V2 aligned intervals. This also explains the modestly lower overlap percentage for W2V2Trained (74.5%) and W2V2TrainedSAT (74.9%) versus MFA_Adapt (78.4%) and MFA_Align (77.6%). As the training and testing splits of the TIMIT data set only have two unique utterances in common, no additional out of utterance evaluation was performed for an adult speech data set.

Dependence of Aligner Performance of Training Data Size

The performance of the W2V2 aligner is sensitive to the size of the training data set. It is important to characterize the amount of labeled training data required for the model to produce usable, high-quality alignments. On the child speech corpus, W2V2TrainedSAT was trained on randomly sampled subsets of the training data set (see Figures 6a–6c). These results show that if using 10%–15% of the full training data set, threshold accuracy (see Figure 6a), interval overlap percentage (see Figure 6b), and midpoint accuracy (see Figure 6c) match the MFA_Adapt baseline. Additionally, the figures show that all accuracy metrics behave similarly as a function of data set size.

This provides a rough estimate of the amount of labeled alignments required to match the performance of the MFA on child speech (10%–15% of data is equivalent to 12 min of audio with manually annotated alignments). Figures 5a–5c show that when 40%–50% of the data set was used (roughly 45–60 min of manually aligned audio), accuracy was nearly equivalent to using the entire data set.

Pronunciation Score Calculated Using Forced Alignments

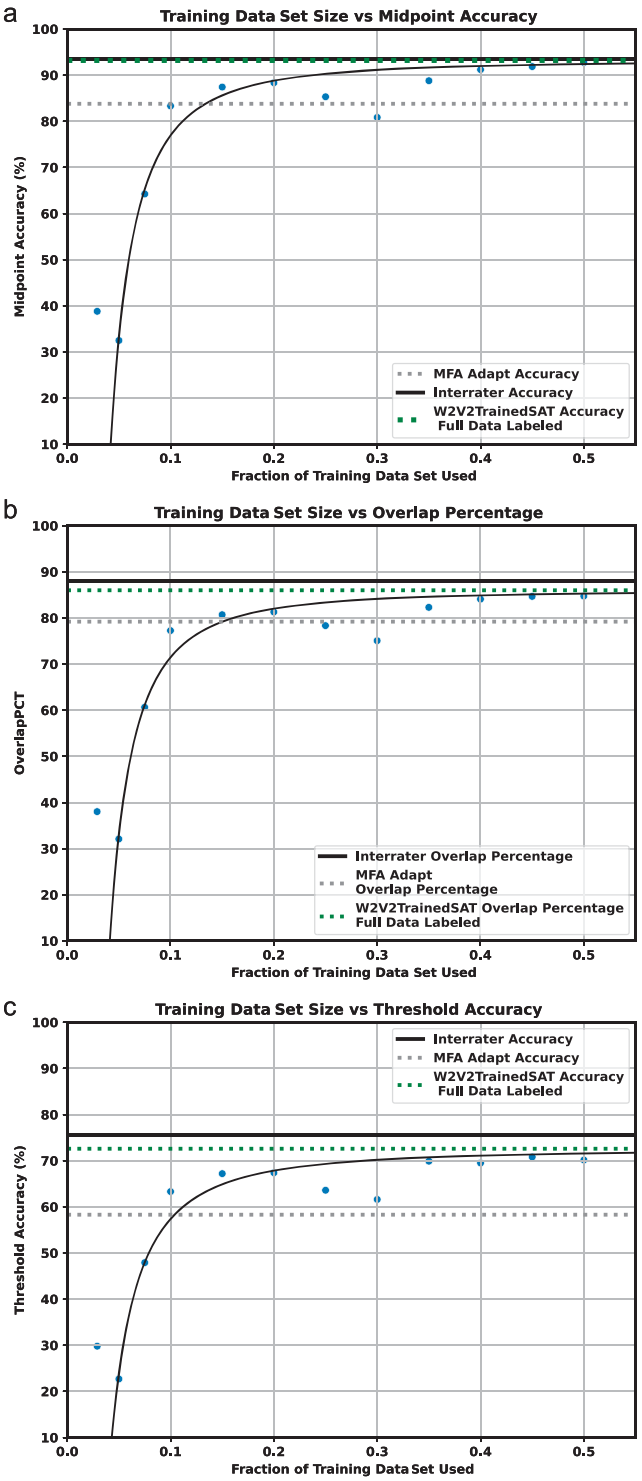
In many cases, forced alignment is performed in service of further downstream analysis. One example use case

Table 4. Alignment accuracy for held out utterances.

Method	Midpoint accuracy (%)	Overlap percentage	Threshold accuracy (onset error < 20 ms; %)	Median onset error (ms)	Onset error IQR (ms)	Median onset error (%)	Median offset error (ms)	Offset error IQR (ms)	Median offset error (%)	Num intervals
MFA_NoSAT	74.3	70.2	51.9	18.6	56.6	15.1	22.2	59.9	17.3	540
MFA_Align	79.9	75.1	51.2	19.4	32.2	17.1	21.1	29.6	16.9	562
MFA_Adapt	85.4	78.4	59	15.2	28.6	11.2	19.7	24	17.7	602
MFA_Train	83.1	77.6	61.6	12.3	28.7	8.6	19.1	25.5	14.2	575
W2V2Base	66.9	63	36.7	30	68.5	20.9	34	72.6	22.2	589
W2V2Trained	80.3	74.5	56.1	12.7	39.2	10.2	14.6	43.3	12.8	610
W2V2TrainedSAT	80.8	74.9	60.4	11.7	33.5	11.9	12.8	35.9	11.3	629

Note. Using the same data set from Table 2, 20% of the unique utterances in the training data set were removed. The procedure is outlined in the Out of Sample and Utterance Evaluation section. These removed utterances were the only utterances retained in the test set (still composed of four of 42 speakers). IQR = interquartile range; MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.

Figure 6. (A) W2V2TrainedSAT may be susceptible to poor performance if the training data set size is small. Thus, W2V2TrainedSAT was trained from scratch on smaller subsets of the child speech corpus to characterize performance given only a limited corpus size to train the alignment system. The figure above shows that the threshold accuracy achieved by the Montreal Forced Aligner (MFA) is matched by W2V2TrainedSAT using only ~10% of the training data (approximately 12 min of speech). Threshold accuracy using 35% of the training data nearly matched the accuracy achieved using all of the training data. (B) To match the interval overlap percentage achieved by the MFA, W2V2TrainedSAT requires 10%–15% of the training data. Improvements in overlap percentage diminish beyond data set size of 40% of the original training data. (C) Midpoint accuracy performance matches the MFA baseline using 10%–15% of the training data. Improvements are marginal after increasing the data set size beyond 45% of the original training data. SAT = speaker adaptive training; W2V2 = Wav2Vec2.



is automatic assessment of pronunciation (Witt & Young 1997, 2000). To evaluate the impact of forced alignment error on this application, we compute the GOP scores via the PLLR using all alignment methods. The average pronunciation score was computed for each file for each alignment method, and Table 5 shows the correlation of pronunciation scores from forced aligned files and pronunciation scores from manually aligned files. Correlations between pronunciation scores ranged from .74 (W2V2Base) to .87 (W2V2TrainedSAT). Given that the same acoustic model was used to evaluate the phoneme likelihoods, the only factor driving differences in pronunciation score for a given utterance was alignment errors. It is therefore unsurprising that the rank order of methods in terms of alignment accuracy also matches the rank order of methods by the strength of correlation between force aligned and manually aligned pronunciation scores.

Discussion

Numerous ready-to-use forced alignment systems are available (Goldman, 2011; Gorman et al., 2011; Kreuk et al., 2020; McAuliffe et al., 2017; Povey et al., 2011; Zhu et al., 2022), but few leverage recent advances from neural acoustic models or include speaker adaptation. Fewer still allow training on manually prepared, clinical-grade alignments. In this study, we introduce a speaker adaptive forced alignment tool based on a neural transformer acoustic model that can be trained on gold standard alignments. The utility of the method and of training on labeled alignments is demonstrated on both standard adult speech and a child speech corpus.

Forced alignment is often used to analyze speech that is not healthy adult speech (which we term NAH speech). However, acoustic properties of each phoneme interval in NAH speech can be markedly different from adult speech. Existing models, trained only on healthy adult speech, may be unable to segment these intervals accurately despite acoustic model adaptation as properties of phoneme intervals may have changed in NAH speech. For example, *MFA_Adapt* only achieved a threshold accuracy of 59% on the child speech corpus, while W2V2TrainedSAT, which included additional training using manual alignments, achieved 75.4%. We attribute this discrepancy between the two to the benefit of training on manual annotations.

Table 5. Correlation of pronunciation score calculated manual and forced alignments.

Alignment method	Pearson correlation between manual pronunciation score and force aligned pronunciation score
MFA_NoSAT	.77
MFA_Align	.84
MFA_Adapt	.85
MFA_Train	.75
W2V2Base	.74
W2V2Trained	.85
W2V2TrainedSAT	.87

Note. Goodness of pronunciation score was calculated using automated forced alignments from each method on the child speech corpus. Following a per-speaker average, automated scores were correlated with manual aligned scores. Rank order of the strength of the correlation matches rank order by alignment accuracy. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; W2V2 = Wav2Vec2.

Alignment Accuracy: Child Speech Versus Adult Speech

Comparison of Tables 1 and 3 reveal that alignment accuracy was actually higher for child speech data set compared to TIMIT. This may seem unexpected as the original training data for all acoustic models used in this study (other than MFA_Train, which was only trained on the data set in question) was healthy adult speech. Several factors contribute to the comparatively lower accuracy on TIMIT: the TIMIT manual annotation process, the greater number of unique utterances in TIMIT, and the complexity of the elicitations in TIMIT. First, the TIMIT manual alignments were annotated with 61 unique phoneme tokens. These 61 phonemes had to be mapped to the 39 phonemes in the CMU pronunciation dictionary. This process was inexact. For example, the flap /DX/could not be converted since a one-to-one mapping to the CMUdict did not exist. Rather, it was simply retained and reduced the accuracy of all methods equally. Second, TIMIT contained greater annotator heterogeneity. The train/test splits of the child speech corpus using LOSOCV contained alignments from the same annotator in most cases, while for TIMIT, a far larger number of annotators were used.

Another factor contributing to the higher alignment accuracy observed in the child speech data set was elicitation complexity: 38 of the 98 unique utterances in the child speech corpus were single-word utterances, which are easier to align than full sentences. Also, the low number of unique stimuli ensures that for nearly every stimulus in the test set, the aligner is trained against multiple versions of the same stimulus among the other speakers in the training set. TIMIT stimuli were far more diverse than the stimuli used for the child speech corpus. TIMIT consists of 6,300 total utterances with 2,342 of them being unique. The TIMIT training data set contained 1,733 unique elicitations out of 4,620 audio files, and the testing data set contained 632 unique elicitations out of 1,680 total files. Only two elicitations were common across the training and testing data sets for TIMIT.

This comparative analysis highlights two important considerations when using this system. First, it is difficult to compare accuracy results across corpora as each corpus has different alignment rules that the aligner adopts internally post training. Second, one of the ways in which the proposed alignment system attains higher accuracy for the child speech is by taking advantage of the fact that most of the speech elicitations are common across participants. That is, the aligner learns the relationship between the alignments and the specific speech stimuli used in the study. Although this is useful for certain downstream tasks (e.g., pronunciation evaluation), it may not be for others where there is more variability in speech elicitations. Future work should evaluate the effectiveness of this alignment system on more spontaneous speech.

How Accurate Do Alignments Need to Be?

In several applications, forced alignment is performed to study subclasses of phonemes or particular phonemes of interest. For example, children with cerebral palsy may have a limited consonant inventory and struggle with consonant clusters (e.g., /sp/, /st/, /bl/) or sounds that require coordination of multiple articulators such as /ʃ/ or /dʒ/ (Mei et al., 2014). Children with cleft palate may also struggle to produce plosives due to altered oral anatomy (Kummer, 2013). In these use cases, researchers may wish to isolate particular phonemes, measure acoustic properties of the sounds, or even extract biomarkers from the isolated sounds for comparison against healthy baselines. However, the amount of permissible alignment error is likely application dependent and unknown a priori for most use cases.

Some studies have found that substantial forced alignment error did not impede the validity of acoustic metric extraction pipelines. Knowles et al. (2018) studied center of gravity (CoG) extracted on /s/ intervals on children aged 2- to 5-year-olds using the Prosodylab aligner. They found that even with extremely low alignment accuracy (midpoint containment accuracy < 25%–75%), alignment error had no impact ($p > .5$) on CoG measurements.

Similarly, Mathad et al. (2021), studied the impact of forced alignment errors using MFA generated alignments to calculate pronunciation scores (PLLR) for typically developing children and children with cerebral palsy. Modeling pronunciation score as a linear function of alignment error, they found that despite boundary prediction errors on the order of 30–60 ms, alignment error was a weak predictor of pronunciation score. Alignment error was only a significant predictor for plosives and fricatives, and its effect size was tiny in both cases. Incidentally, plosives, fricatives and affricates were the phoneme subclasses for which W2V2TrainedSAT showed the greatest performance improvement over the MFA (see Figure 3a).

Despite the two previous studies showing the applicability of forced alignments in cases of moderate to high alignment accuracy, it is unclear whether their conclusions can be generalized outside the narrow context of the studies. For example, measuring an acoustic feature that changes over time in a segment (i.e., formant trajectories) might not be as robust to imprecise alignments as ones that average over many slices of time in a segment (i.e., mean formant values). For such time-varying cases, the system proposed in this work serves to improve both forced aligner quality and the fidelity of results obtained from these alignments.

Moreover, under high alignment error, extracted downstream measurements are invalid. For example, consider the plosives from the child speech corpus studied in this work. The median percent onset error for MFA_Adapt was nearly 60%, training the acoustic model on the manual alignments in the W2V2TrainedSAT method greatly reduced the percent onset error for plosives to just 17% (see Supplemental Material S1). This means that for most plosive intervals aligned by MFA, the size of the onset error was larger than half the interval length itself. If more than half of the predicted interval contains sounds other than the predicted sound (as is the case with MFA and plosives), metrics such as PLLR can no longer be directly interpreted as a pronunciation score which measures a ratio of expected phoneme probability over maximum phoneme probability.

Quantifying Required Manually Alignment Effort

Our results have shown that labeling only 10%–15% (13 min) of the data yields performance exceeding the MFA baseline. A significant increase in accuracy over baseline is achieved by labeling 45–60 min of training data. Gains in alignment accuracy when beyond 35%–45% (45–60 min) of training data was used were marginal. This is consistent with WER performance of W2V2 on an ASR task as a function of available training data. Baevski et al. (2020) showed that for Librispeech ASR, test WER when using 10 min, 1 hr, 10 hr, or 960 hr of the Librispeech was 6.3%, 3.8%, 2.9%, and 1.7%, respectively.

Limitations and Future Work

The use of manual alignments from a single annotator to both train and evaluate the model could lead to over optimistic evaluation metrics or to overfitting to annotator-specific artifacts present in the labels. Two annotators were used to prevent this issue, but labels from two distinct annotators were only collected for a small subset of the corpus due to the time-intensive nature of manual alignment. For each

speaker, the time required to manually correct alignments ranged from 3.5 hr to more than 8 hr. Thus, the model was not evaluated on alignments from an annotator that did not also provide training alignments.

When comparing neural end-to-end acoustic models to classical approaches, a key difference is apparent in available acoustic model adaptation methods. DNN-GMM, HMM-GMM use MFCC inputs, which have a long history of physiologically motivated adaptation methods based on established theories of speech production. These include techniques such as VTLN and cepstral mean and variance normalization (CMVN). Many of these have been used in child ASR, and for disordered speech ASR (Giuliani & Gerosa, 2003; Serizel & Guiliani, 2017). End-to-end systems like W2V2 typically rely on data driven adaptation methods that may not be physiologically interpretable.

Perhaps the largest limitation of our system is that manual labeling of alignments on approximately 10–15 min of audio is required at all to match the performance of the MFA baseline and 45–60 min of labeled data is required to even approach the level at which alignment accuracy begins to saturate. Given that alignment can take 1–5 min per utterance (i.e., per sentence) alternatives to manual alignment are necessary. One future avenue of research could be in bootstrapping alignment following Zhu et al. (2022). First, alignments can be generated with another forced alignment system such as the MFA and be used as noisy ground truth. Then, fine-tuning can be performed with the forward-sum loss used to train HMMs to learn improved alignments.

Alternatively labeling effort could directly be reduced by clever prioritization of the files to manually annotate first. Such methods, termed "active learning" methods, have been used successfully in speech recognition (Huang et al., 2016). To achieve similar accuracy as randomly choosing training samples for manual alignment, active learning could reduce the amount of manually aligned training data required by an order of magnitude.

Conclusions

In this work, we devised a novel, speaker adaptive, W2V2-based neural forced alignment system as part of a trainable alignment workflow: the forced aligner was directly trained to mimic clinical-grade manual alignments. The system was evaluated on a corpus consisting of 42 children from the ages of 3- to 7-year-olds. In aggregate, and across every phoneme subclass, the method improves an array of alignment evaluation metrics compared to the best performing off-the-shelf alignment methods. Evaluation was also performed on adult speech to demonstrate the method is useful across various types of speech. This method, Wav2TextGrid, could greatly increase throughput in analysis of child speech by eliminating the need for time consuming hand correction of forced alignments.

Data Availability Statement

We are unable to share the data used in this study due to privacy and Health Insurance Portability and Accountability Act concerns.

Acknowledgments

This work was funded by National Institute on Deafness and Other Communication Disorders Grant R01DC019645-03, awarded to Katherine C. Hustad.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). *Common voice: A massively-multilingual speech corpus.* arXiv. https://doi.org/10.48550/arXiv.1912.06670

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 & Language, 45,* 278–299. https://doi.org/10.1016/j.csl.2017.02.010

Carnegie Mellon University. (1998). *The Carnegie Mellon pronouncing dictionary* (Version 0.6). http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). *VoxCeleb2: Deep speaker recognition.* arXiv. https://doi.org/10.48550/arXiv.1806.05622

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2016). *BERT: Bidirectional encoder representations from transformers.*

Fontan, L., Pellegrini, T., Olcoz, J., & Abad, A. (2015). Predicting disordered speech comprehensibility from goodness of pronunciation scores. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies,* 42–46. https://doi.org/10.18653/v1/W15-5108

Garofalo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *The DARPA TIMIT acoustic-phonetic continuous speech corpus LDC93S1* [Web download]. Linguistic Data Consortium.

Geng, M., Xie, X., Ye, Z., Wang, T., Li, G., Hu, S., Liu, X., & Meng, H. (2022). Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30,* 2597–2611. https://doi.org/10.1109/TASLP.2022.3195113

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication, 49*(10–11), 847–860. https://doi.org/10.1016/j.specom.2007.01.002

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, Article 7. https://doi.org/10.1145/1640377.1640384

Giuliani, D., & Gerosa, M. (2003, April). Investigating recognition of children's speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)* (Vol. 2, pp. II–137). IEEE.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Goldman, J.-P. (2011). EasyAlign: An automatic phonetic alignment tool under Praat. *Proceedings of Interspeech 2011,* 3233–3236. https://doi.org/10.21437/Interspeech.2011-815

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics, 39*(3), 192–193.

Hodge, M. M., & Daniels, J. (2007). *TOCS+ intelligibility measures* [Computer sofware]. University of Alberta.

Huang, J., Child, R., Rao, V., Liu, H., Satheesh, S., & Coates, A. (2016). *Active learning for speech recognition: The power of gradients*. arXiv. https://doi.org/10.48550/arXiv.1612.03226

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining factors influencing the viability of automatic acoustic analysis of child speech. *Journal of Speech, Language, and Hearing Research, 61*(10), 2487–2501. https://doi.org/10.1044/2018_JSLHR-S-17-0275

Kreuk, F., Keshet, J., & Adi, Y. (2020). *Self-supervised contrastive learning for unsupervised phoneme segmentation*. arXiv. https://doi.org/10.48550/arXiv.2007.13465

Kummer, A. W. (2013). *Cleft palate & craniofacial anomalies: Effects on speech and resonance* (3rd ed.). Cengage Learning.

Lee, K.-F., & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*(11), 1641–1648. https://doi.org/10.1109/29.46546

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language, 9*(2), 171–185. https://doi.org/10.1006/csla.1995.0010

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y., & Wang, Y. (2022). NeuFA: Neural network based end-to-end forced alignment with bidirectional attention mechanism. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 8007–8011. https://doi.org/10.1109/ICASSP43922.2022.9747085

Lin, C.-Y., & Wang, H.-C. (2011). Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection. *The Journal of the Acoustical Society of America, 130*(1), 514–525. https://doi.org/10.1121/1.3592233

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., & Hustad, K. C. (2021). Performance of forced-alignment algorithms on children's speech. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2213–2222. https://doi.org/10.1044/2020_JSLHR-20-00268

Mahr, T. J., Soriano, J. U., Rathouz, P. J., & Hustad, K. C. (2021). Speech development between 30 and 119 months in typical children II: Articulation rate growth curves. *Journal of Speech, Language, and Hearing Research, 64*(11), 4057–4070. https://doi.org/10.1044/2021_JSLHR-21-00206

Mathad, V. C., Mahr, T. J., Scherer, N., Chapman, K., Hustad, K. C., Liss, J., & Berisha, V. (2021). The impact of forced-alignment errors on automatic pronunciation evaluation. *Proceedings of Interspeech 2021,* 1922–1926. https://doi.org/10.21437/Interspeech.2021-1403

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017,* 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Mei, C., Reilly, S., Reddihough, D., Mensah, F., & Morgan, A. (2014). Motor speech impairment, activity, and participation in children with cerebral palsy. *International Journal of Speech-Language Pathology, 16*(4), 427–435. https://doi.org/10.3109/17549507.2014.917439

Michel, P., Räsänen, O., Thiolliere, R., & Dupoux, E. (2016). *Blind phoneme segmentation with temporal prediction errors*. arXiv. https://doi.org/10.48550/arXiv.1608.00508

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). *VoxCeleb: A large-scale speaker identification dataset*. arXiv. https://doi.org/10.48550/arXiv.1706.08612

Okamoto, T., Toda, T., Shiga, Y., & Kawai, H. (2019). Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 214–221. https://doi.org/10.1109/ASRU46091.2019.9003956

Oppelstrup, L., Blomberg, M., & Elenius, D. (2005). Scoring children's foreign language pronunciation. *Proceedings, FONETIK*, 51–54.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Park, K., & Kim, J. (2019). *g2pE* [Source code]. https://pypi.org/project/g2p-en/

Peng, P., Huang, P. Y., Li, S. W., Mohamed, A., & Harwath, D. (2024). *Voicecraft: Zero-shot speech editing and text-to-speech in the wild*. arXiv. https://doi.org/10.48550/arXiv.2403.16973

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesel, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Serizel, R., & Giuliani, D. (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering, 23*(3), 325–350. https://doi.org/10.1017/S135132491600005X

Shivakumar, P. G., Potamianos, A., Lee, S., & Narayanan, S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. *Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI 2014),* 15–19. https://www.isca-archive.org/wocci_2014/shivakumar14_wocci.pdf [PDF]

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Sonderegger, M., & Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured prediction. *The Journal of the Acoustical Society of America, 132*(6), 3965–3979. https://doi.org/10.1121/1.4763995

Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S., Shelton, K., Duncan, C. J., & Berisha, V. (2020). Early

detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *NPJ Digital Medicine, 3*(1), Article 132. https://doi.org/10.1038/s41746-020-00335-x

Tai, C.-L., Lee, H.-S., Tsao, Y., & Wang, H.-M. (2022). *Filter-based discriminative autoencoders for children speech recognition*. arXiv. https://doi.org/10.48550/arXiv.2204.00164

Tu, M., Grabek, A., Liss, J., & Berisha, V. (2018). *Investigating the role of L1 in automatic pronunciation evaluation of L2 speech*. arXiv. https://doi.org/10.48550/arXiv.1807.01738

Vidal, J., Bonomi, C., Sancinetti, M., & Ferrer, L. (2021). Phone-level pronunciation scoring for Spanish speakers learning English using a GOP-DNN system. In *Interspeech* (pp. 4423–4427).

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). *Tacotron: Towards end-to-end text-to-speech synthesis*. arXiv. https://doi.org/10.48550/arXiv.1703.10135

Wang, Y.-H., Chung, C.-T., & Lee, H.-Y. (2017). *Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries*. arXiv. https://doi.org/10.48550/arXiv.1703.07588

Witt, S. M., & Young, S. J. (1997, September). Language learning based on non-native speech recognition. In *Eurospeech* (pp. 633–636).

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30*(2–3), 95–108. https://doi.org/10.1016/S0167-6393(99)00044-8

Wolf, T. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv. https://doi.org/10.48550/arXiv.1910.03771

Yeung, G., & Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. *Proceedings of Interspeech 2018,* 1661–1665. https://doi.org/10.21437/Interspeech.2018-2297

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., & Wang, W. (2013). Automatic phonetic segmentation using boundary models. *Proceedings of Interspeech 2013,* 2306–2310. https://doi.org/10.21437/Interspeech.2013-540

Zhu, J., Zhang, C., & Jurgens, D. (2022). Phone-to-audio alignment without text: A semi-supervised approach. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 8167–8171. https://doi.org/10.1109/ICASSP43922.2022.9746112