Using AI to Automate Phonetic Transcription and Perform Forced Alignment for Clinical Application in the Assessment of Speech Sound Disorders

Ying Li[1], **Duc-Son Pham**[1], **Roslyn Ward**[2], **Neville Hennessey**[2], **Tele Tan**[1]

[1]School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth 6102, Australia
[2]School of Allied Health, Curtin University, Perth 6102, Australia
ying.li26@postgrad.curtin.edu.au, dspham@ieee.org, {r.ward, n.hennessey, t.tan} @curtin.edu.au

Abstract

Speech-language pathologists (S-LPs) routinely use phonetic transcription to profile and describe the characteristics of a child's speech in the assessment of speech sound disorders (SSDs). The literature identifies phonetic transcription as a demanding perceptual skill, with accuracy and reliability dependent on experience, available resources, and the nature of SSDs. Automatic speech recognition and segmentation techniques, which recognize, transcribe, and align audio file content, have been identified as a possible tool to improve the accuracy and efficiency of the auditory perceptual transcription undertaken by S-LPs. In this paper, we propose a model to automate phonetic transcriptions and perform forced alignment for childhood-disordered speech. Utilizing the state-of-the-art wav2vec 2.0 acoustic model and advanced post-processing algorithms, our model achieves a phoneme error rate of 0.15 and an F_1 Score of 82% on the UltraSuite dataset. These results suggest a level of accuracy greater than what has been reported for auditory-perceptual transcription in the clinical setting.

Introduction

Speech sound disorder (SSD) is an umbrella term that describes a heterogeneous group of individuals who have difficulties in producing speech, thereby interfering with communication (Carter, Paul, and Marchman 2014). It is the most prevalent communication disorder in young children, affecting approximately 3-6% of Australian preschoolers and accounting for up to 75% of a pediatric speech-language pathologist's (S-LP) caseload (Lewis et al. 2011). Early identification and intervention are crucial to minimizing long-term consequences, which may include lower academic success, decreased social interaction, and an increased risk of juvenile delinquency (Catts 1993).

Evidence-based practice guidelines recommend S-LPs use phonetic transcription, based on the International Phonetic Alphabet, to identify and classify a child's speech patterns. This process is fundamental to informing the diagnosis of the SSDs and subsequent selection of the most effective intervention approach and tracking therapy progress (Network 2017). However, the accuracy of phonetic transcription, as a profession-specific skill, is limited by S-LPs' experience, available time, and perceptual bias (Mallaband 2024). Whilst the transcriber's experience can be addressed through practice and training to some extent (Titterington and Bates 2021), there remain intrinsic challenges in using phonetic transcription to identify markers of speech impairments (Gibbon and Lee 2017).

With advances in artificial intelligence (AI), a growing body of literature has acknowledged its potential value in supporting S-LPs in generating **phonetic transcription** (Duffy 2016; Li et al. 2020; Naeini et al. 2024). This task is commonly facilitated by adapting automatic speech recognition (ASR) (Bhardwaj et al. 2022). Recent studies (Yi et al. 2020; Yue et al. 2020; Hernandez et al. 2022) have fine-tuned the latest transformer-based acoustic models and language models (Baevski et al. 2020; Radford et al. 2023) on disordered *adult* speech, achieving remarkable performance. Some studies have explored phoneme-level recognition (Shahin and Ahmed 2024) and reported good average performance for typically developing children and second-language learners. However, these models usually experience a significant drop in performance when applied to speech from children with SSDs (McKechnie et al. 2018). This drop is due to the high variability in the acoustic and linguistic characteristics of disordered childhood speech, as well as the limited availability of annotated speech corpora for this population (O'Shaughnessy 2024).

Forced alignment, which involves identifying the onset and offset time of phonemes, has been demonstrated as an effective approach to improve the quality and transparency of phoneme recognition (Shabber and Bansal 2024). Recent studies highlight that forced alignment allows phoneme labels to align more accurately with their corresponding acoustic features, enabling AI models to learn more phoneme-specific information and improve recognition accuracy (Graves and Schmidhuber 2005; Graves et al. 2006; Kalinli 2012).

In this paper, we present a model that simultaneously performs both phonetic transcription and forced alignment effectively on *childhood*-disordered speech, hereinafter referred to as the **phoneme segmenter**. To automate phonetic transcription, we adapt the advanced acoustic self-supervised learning model, wav2vec 2.0, in a novel way. While wav2vec 2.0 is traditionally used for word-level ASR with connectionist temporal classification (CTC), our adap-

tation focuses on phoneme-level transcription to capture subtle phonetic issues characteristic of children with speech sound disorders. We develop a novel algorithm for forced alignment. Furthermore, we develop and evaluate a transfer learning approach to address the challenge of data scarcity in disordered childhood speech corpora. This method facilitates the integration of AI into routine clinical speech analysis.

Methodology

Phoneme Segmenter

In this paper, we introduced a phoneme segmenter, as illustrated in Figure 1. The key innovation of this pipeline lies in adapting the state-of-the-art wav2vec 2.0 model, originally developed for ASR, for the task of generating phonetic transcriptions and performing forced alignment. Our code is publicly available online.

Code —
https://github.com/YingLi001/phoneme-segmenter

To automate *phonetic transcription* for childhood-disordered speech, we fine-tuned the `wav2vec2-xls-r` pre-trained model on an SSD dataset. The model was pre-trained on 436,000 hours of unlabeled speech data sampled at 16kHz, covering 128 different languages. During pre-training, the model acquired latent representations of multiple languages. However, these representations required additional specialization through fine-tuning on a specific "downstream" task to achieve effective performance.

In this study, we kept the feature extractor unchanged, fine-tuned the learned representations with labeled datasets, and added a randomly initialized fully connected (FC) layer on top of the Transformer architecture for phoneme prediction. To optimize the model, we minimized the CTC loss (Graves et al. 2006).

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \text{Align}(Y)} \prod_{t=1}^{T} P(\pi_t \mid X) \tag{1}$$

where,

- $\sum_{\pi \in \text{Align}(Y)}$ represents the summation over all possible alignment paths π for the target sequence Y;
- $\prod_{t=1}^{T}$ denotes the product of probabilities over all time steps t, where T is the length of the input sequence;
- $P(\pi_t \mid X)$ is the probability of observing the token π_t at time step t, given the input sequence X.

It is calculated based on the phoneme error rate (PER), which is derived from the Levenshtein distance, which quantifies the minimum number of single-character edits—substitutions, insertions, or deletions—needed to transform the predicted phoneme sequence into the actual sequence.

Algorithm 1 describes our proposed method for *forced alignment*. Given phoneme predictions from the model using CTC collapse, which merged consecutive identical labels that were not separated by a blank symbol, a bias factor β was introduced to adjust the boundaries of each phoneme. This is because the recognized phonemes might be located closer to either the start or the end of the true segment due to frame-based prediction and contextual information.

The rationale for not using the existing forced alignment models was based on several key considerations. First, the existing segmentation models (Kreuk et al. 2020; Kreuk, Keshet, and Adi 2020) trained on datasets of normal adult speakers tended to introduce higher errors when applied to disordered speech in children. Second, our work focused on phonemes, the smallest unit of speech that can differentiate one word element from another. As phoneme duration in children can be short, segmentation models tended to overlook segments with short duration during the aligning of boundaries. This oversight often results in forced alignment errors. Finally, we observed notable improvements in phoneme recognition and segmentation by incorporating an advanced wav2vec 2.0 model (`wav2vec2-xls-r`). Given these considerations and the need for sensitivity in detecting subtle yet diagnostically significant phonemes, we decided to develop an algorithm for disordered childhood speech.

Algorithm 1: Forced Alignment

procedure FORCEDALIGNMENT(timedTokenList, seconds, β)
    timedTokenList is a list of tuples of labels and their timings
    seconds is the duration of the speech sample in seconds
    β is the bias factor, which is positive and smaller than 1.
    FA ← new List
    for ii in range of length timedTokenList **do**
        if ii equals length of timedTokenList - 1 **then**
            upper ← seconds
            lower ← timedTokenList[ii - 1][time] * (1 - β) + timedTokenList[ii][time] * β
        else if ii equals 0 **then**
            upper ← timedTokenList[ii + 1][time] * β + timedTokenList[ii][time] * (1 - β)
            lower ← 0
        else
            upper ← timedTokenList[ii + 1][time] * β + timedTokenList[ii][time] * (1 - β)
            lower ← timedTokenList[ii - 1][time] * (1 - β) + timedTokenList[ii][time] * β
        end if
        Append tuple (timedTokenList[ii][label], lower, upper) to FA
    end for
    return FA
end procedure

We further developed a post-processing method, referred to as cleaning, to consolidate successive duplicate segments that often arise due to model overfitting. The final output of the phoneme segmenter model consisted of an aligned phonetic transcription accompanied by a spectrogram, as illus-

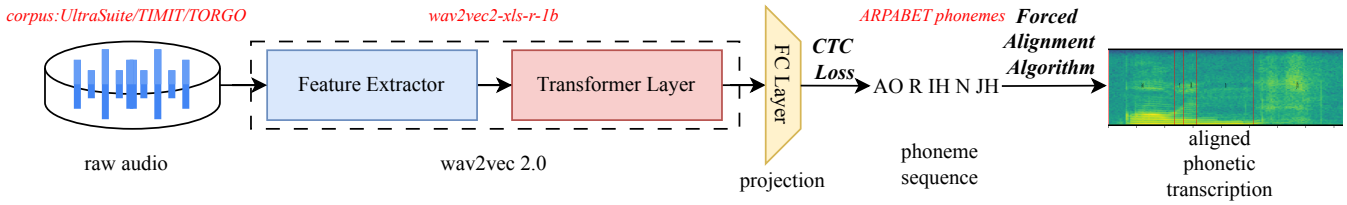

Figure 1: Architecture of the proposed phoneme segmenter.

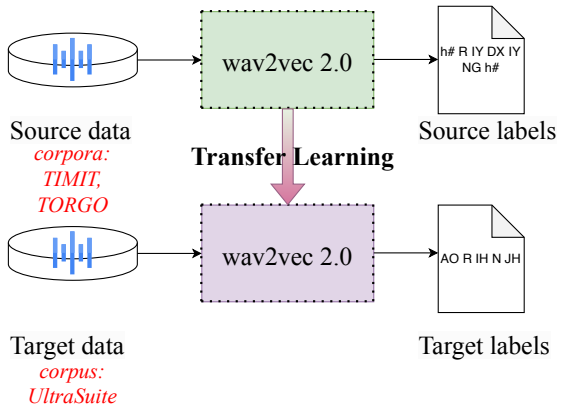

Figure 2: Transfer learning scheme.

trated in Figure 1. This visualization method was designed to enhance the explainability of the automated results, and providing S-LPs with clearer insights to support decision-making.

Transfer Learning

Transfer learning (TL) is a technique designed to leverage knowledge from a source domain to improve the learning of a predictive function in a target domain, as illustrated in Figure 2. Inspired by recent studies (Christensen et al. 2013; Fainberg et al. 2016; Smith et al. 2017), we applied TL to enhance the performance of the phoneme segmenter on a childhood-disordered speech corpus. Specifically, we transferred the knowledge encoded in the wav2vec 2.0 model, pre-trained on normal adult speech and fine-tuned on disordered adult speech, to a disordered childhood speech corpus.

Datasets

We used two datasets specifically designed for SSDs. The first dataset is UltraSuite (Eshky et al. 2019), an acoustic and ultrasound data collection featuring recordings from children aged 5 to 12 years with typically developing speech, as well as children aged 5 to 13 years with various SSDs, including childhood apraxia of speech, phonological delay, and articulation disorder. All speech files are sampled at 22.05 kHz with 16-bit sample resolution. In this study, we only used the SSD instances in the UltraSuite dataset for training and evaluating the proposed method.

The second dataset is TORGO (Rudzicz, Namasivayam, and Wolff 2012), an acoustic and articulatory speech corpus consisting 8 dysarthric speakers, aged 16 to 50 years, and 7 age-matched control speakers. It includes non-words, words, and sentences, of which words and sentences are used in this study. The individual wave file is encoded in the linear PCM format at 16kHz, which was used to evaluate the TL.

While our primary focus was on *child* speakers, we also utilized the popular TIMIT dataset (Garofolo et al. 1993) to demonstrate the benefits of TL. This is a standard acoustic-phonetic corpus used for the evaluation of speech-related tasks. It consists of 6,300 utterances produced by 630 healthy *adult* American speakers from 8 dialect regions. The corpus contains approximately 5 hours of speech recordings that are stored in 16-bit and 16kHz waveform files, associated orthographic transcriptions of the words the person said, and time-aligned phonetic transcriptions.

Dataset	TRAIN		TEST	
	SSD	TD	SSD	TD
UltraSuite	1870	-	804	-
TIMIT	-	4620	-	1280
TORGO	1588	1928	636	619

Table 1: Distribution of the instances included in this study from the preprocessed UltraSuite, TIMIT, and TORGO datasets.
Note: TD represents individuals with typical speech.

Data Pre-processing

To ensure compatibility with the wav2vec 2.0 model, we downsampled all raw audio files of SSD instances in the UltraSuite dataset from 22.05 kHz to 16 kHz. Additionally, since the UltraSuite dataset lacks a predefined training and test split, we allocated 70% of instances to the training set and 30% to the test set. As detailed in Table 1, only SSD instances from UltraSuite dataset were utilized to evaluate the performance of phoneme segmenter on childhood-disordered speech.

To facilitate efficient data loading, we used the *datasets* library in the Hugging Face. Since the UltraSuite dataset was not directly available within the library's offerings, we developed a custom data-loading script. In this script, each audio sample from the UltraSuite was treated as an individual instance with several attributes. The key attributes are outlined below, with non-essential attributes excluded during

Exp	Model	Dataset	PER	P	R	F_1	R-value
1	Zhu, Zhang, and Jurgens†	UltraSuite	0.20	0.45	0.73	0.55	0.35
2	Ribeiro et al.∗	UltraSuite	0.63	0.75	0.70	0.73	0.76
3	Phoneme Segmenter	UltraSuite TIMIT	0.15	0.82	0.82	0.82	0.85
4	Phoneme Segmenter TL	TORGO UltraSuite	0.12	0.85	0.86	0.86	0.88

Table 2: Results of phoneme recognition and alignment experiments, including comparisons with baseline methods. ∗ is measured by word error rate instead of PER. † indicates an evaluation by ourselves.

preprocessing to simplify data handling. During this phase, the necessary components such as the tokeniser, feature extractor, processor, and data collator were created, enabling efficient handling and analysis of complex speech datasets for model training and evaluation.

In total, the dictionary included 61 ARPABET phonemes (Seneff and Zue 1988) and three special tokens, "[UNK]", "[PAD]" and "|" for "unknown", "padding" and " ' " respectively.

- File: Path of the audio file.
- Text: The transcription for the audio file.
- Phonetic Detail: Phonetic transcription formatted as '<start_sample> <stop_sample> <phoneme> <new_line>', where 'start_sample' and 'stop_sample' are the integer sample numbers marking the start and stop of the phoneme segment, respectively, and 'phoneme' represents a single sound unit using ARPABET symbols.

Evaluation Measures

The proposed phoneme segmenter architecture was evaluated using two key aspects: phoneme recognition performance and forced alignment performance.

Phoneme Recognition Phoneme recognition performance was assessed using the PER, where lower values indicate better performance. The PER is calculated as:

$$\text{PER} = \frac{D + S + I}{N}, \tag{2}$$

where D represents deletions, S substitutions, I insertions, and N the total number of phonemes.

Forced Alignment The performance of temporal alignment was evaluated using precision (P), recall (R), F_1 Score and R-value (Räsänen, Laine, and Altosaar 2009), calculated via the midpoint method (Mahr et al. 2021), where higher values indicate better performance. The F_1 Score is defined as:

$$F_1 \text{ Score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}}, \tag{3}$$

where *Precision* represents the proportion of matched predictions correct; *Recall* the proportion of ground truths correctly classified.

Experimental Results

Phoneme Recognition

The phoneme recognition performance of SSD instances in the UltraSuite dataset is presented in Table 2. The proposed model demonstrated advanced performance on the UltraSuite test set, achieving a PER of 0.15. Fine-tuning on the UltraSuite dataset was completed in approximately 2 hours, 33 minutes, and 10 seconds on a Linux machine equipped with a 16-core CPU and an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

To demonstrate the superior performance of our proposed method, we compared it with several baseline alternatives. As shown in Table 2, we evaluated the model proposed by Zhu, Zhang, and Jurgens on the UltraSuite dataset, achieving a PER of 0.2 after collapsing the 61 TIMIT phonemes into the 39 CMU phoneme set. Since phoneme folding is known to improve performance (Lee and Hon 1989), we anticipate that our model would achieve a PER improvement of at least 0.05. Additionally, we compared our results with those reported by the UltraSuite team (Ribeiro et al. 2019), where their model's phoneme recognition performance was measured using Word Error Rate (WER), which is generally higher than PER. Despite the difference in evaluation metrics, our model demonstrated superior phoneme recognition performance.

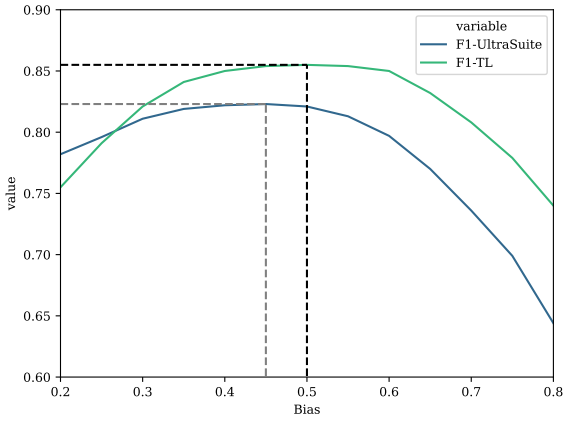

Figure 3: The F_1 value against bias factor β for the UltraSuite and TL experiments, respectively.

Forced Alignment

The forced alignment performance of SSD instances in the UltraSuite dataset was presented in Figure 3. The proposed model achieved the highest F_1 Score of 0.82 and an R-value of 0.85 with $\beta = 0.45$, significantly outperforming the existing solutions (Ribeiro et al. 2019; Zhu, Zhang, and Jurgens 2022). The bias factor $\beta = 0.45$ suggests that the boundaries originally closer to the end of the true segment have shifted farther from the uppermost segment.

Transfer Learning

With TL using TIMIT, TORGO and UltraSuite datasets, both phoneme recognition and forced alignment showed improvement. For phoneme recognition, as illustrated in Table 2, PER decreased from 0.15 to 0.12. For forced alignment, as shown in Figure 3, the proposed phoneme segmenter achieved the highest F_1 Score of 0.86 with $\beta = 0.5$, representing a 4% increase. The unbiased factor $\beta = 0.5$ further suggests that TL has enhanced alignment performance. These results align with the conclusions of studies that have demonstrated leveraging out-of-domain data can enhance segmentation performance for childhood-disordered speech (Christensen et al. 2013; Fainberg et al. 2016).

To better understand the improvements introduced by TL, we visualized a representative sample in Figure 4. The sample, "orange", was produced by an SSD child within the UltraSuite dataset. Subplot (a) illustrates the phoneme recognition and alignment results without TL, whereas subplot (b) presents the outcomes after applying TL. Subplot (c) displays the human-labeled phonetic transcription and timestamps. Notably, the alignment performance, especially for the onset and offset timestamps of the phonemes "ih" and "n" have been markedly improved with the application of TL.

Conclusion

In this paper, we introduce a novel phoneme segmenter designed to automate phonetic transcription and perform forced alignment, specifically tailored for clinical applications in assessing speech sound disorders in children. Our proposed method leverages `wav2vec2-xls-r` and we conduct a comprehensive evaluation of our architecture by assessing phoneme recognition and forced alignment with and without the application of TL technique. We benchmark our model against existing methods using the UltraSuite dataset, achieving a competitive performance with a PER of 0.15 and an F_1 Score of 82%. This result suggests that automatically generated phonetic transcriptions and phoneme boundaries for children's speech can be achieved with a high level of accuracy. This may help improve the accuracy currently reported in clinical settings, which rely solely on auditory-perceptual methods (Mallaband 2024).

Overall, this research introduces a comprehensive automated solution for future clinical application, focusing not only on transcribing what is said but also the precise timing of transitions between sounds. Future research will focus on enhancing the robustness of our model by incorporating clinical data and critical acoustic features, as well as including additional benchmarks and comparing our AI models with clinical assessments. This research is being conducted within a knowledge translation framework with the ultimate goal of applying the findings to clinical practice.

Acknowledgments

This work has been supported by the Western Australian Future Health Research and Innovation Fund (WANMA2023Ideas/1), which is an initiative of the WA State Government; and an Industry Research Grant (PROMPT 59592).

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv Neural Inf Process Syst*, 33: 12449–12460.

Bhardwaj, V.; Ben Othman, M. T.; Kukreja, V.; Belkhier, Y.; Bajaj, M.; Goud, B. S.; Rehman, A. U.; Shafiq, M.; and Hamam, H. 2022. Automatic speech recognition (asr) systems for children: A systematic literature review. *Appl Sci*, 12(9): 4419.

Carter, A. S.; Paul, R.; and Marchman, V. A. 2014. Diagnostic criteria for speech sound disorders in young children. *J Child Psychol Psychiatry*, 55(6): 619–626.

Catts, H. W. 1993. The relationship between speech-language impairments and reading disabilities. *J Speech Lang Hear Res*, 36(5): 948–958.

Christensen, H.; Aniol, M. B.; Bell, P.; Green, P. D.; Hain, T.; King, S.; and Swietojanski, P. 2013. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *Interspeech 2013*, 3642–3645.

Duffy, J. R. 2016. Motor speech disorders: where will we be in 10 years? In *Semin Speech Lang*, volume 37, 219–224. Thieme Medical Publishers.

Eshky, A.; Ribeiro, M. S.; Cleland, J.; Richmond, K.; Roxburgh, Z.; Scobbie, J.; and Wrench, A. 2019. UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions. *arXiv preprint arXiv:1907.00835*, 1888–1892.

Fainberg, J.; Bell, P.; Lincoln, M.; and Renals, S. 2016. Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation. In *Interspeech 2016*, 1598–1602.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; and Pallett, D. S. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93: 27403.

Gibbon, F. E.; and Lee, A. 2017. Electropalatographic (EPG) evidence of covert contrasts in disordered speech. *Clin Linguist Phon*, 31(1): 4–20.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML 2006*, 369–376.

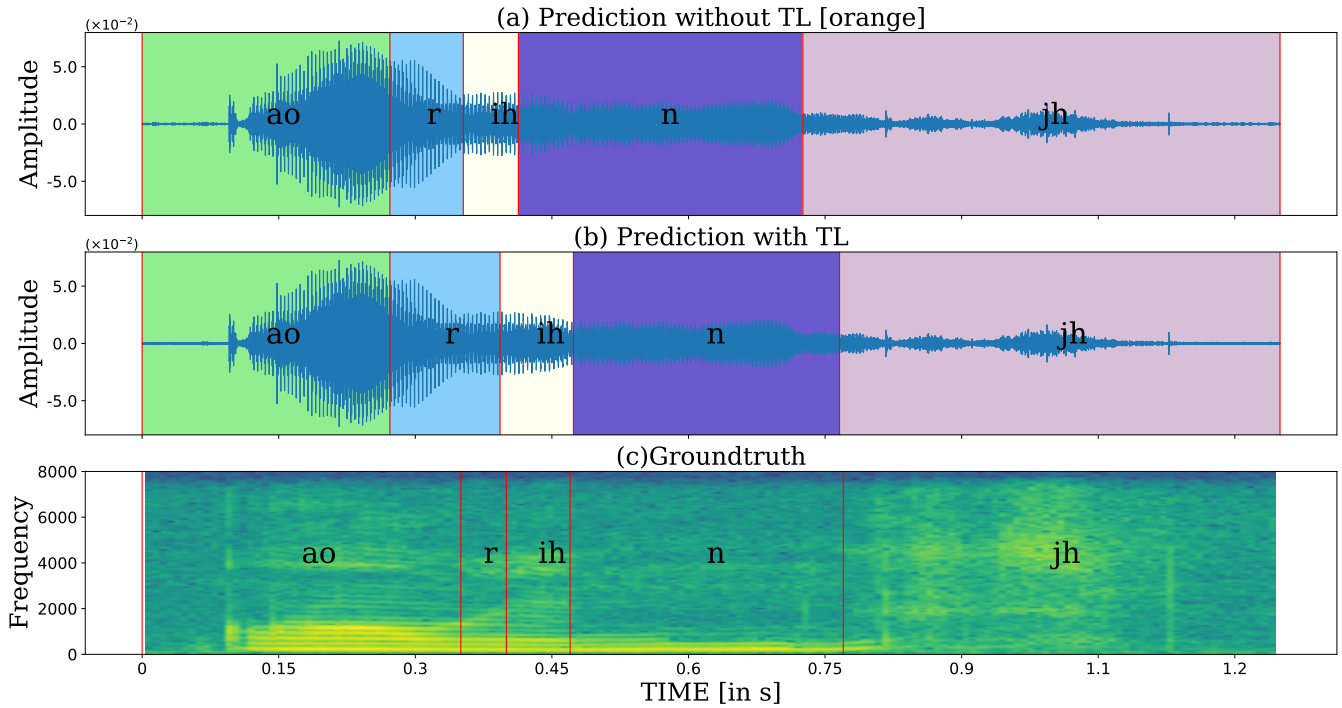

Figure 4: Qualitative visualization of the improvement achieved by employing TL for childhood-disordered speech using the word "orange".

Graves, A.; and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *J Neural Netw*, 18(5-6): 602–610.

Hernandez, A.; Pérez-Toro, P. A.; Nöth, E.; Orozco-Arroyave, J. R.; Maier, A.; and Yang, S. H. 2022. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv preprint arXiv:2204.01670*.

Kalinli, O. 2012. Automatic phoneme segmentation using auditory attention features. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Kreuk, F.; Keshet, J.; and Adi, Y. 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. *arXiv preprint arXiv:2007.13465*.

Kreuk, F.; Sheena, Y.; Keshet, J.; and Adi, Y. 2020. Phoneme boundary detection using learnable segmental features. In *ICASSP 2020*, 8089–8093. IEEE.

Lee, K.-F.; and Hon, H.-W. 1989. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans Acoust*, 37(11): 1641–1648.

Lewis, B. A.; Freebairn, L. A.; Hansen, A. J.; Iyengar, S. K.; and Taylor, H. G. 2011. Subtyping children with speech sound disorders by endophenotypes. *Top Lang Disord*, 31(2): 112–127.

Li, X.; Dalmia, S.; Mortensen, D.; Li, J.; Black, A.; and Metze, F. 2020. Towards zero-shot learning for automatic phonemic transcription. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8261–8268.

Mahr, T. J.; Berisha, V.; Kawabata, K.; Liss, J.; and Hustad, K. C. 2021. Performance of forced-alignment algorithms on children's speech. *J Speech Lang Hear Res*, 64(6S): 2213–2222.

Mallaband, L. J. 2024. The agreement of phonetic transcriptions between paediatric speech and language therapists transcribing a disordered speech sample. *Int J Lang Commun Disord*.

McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.; Monroe, P.; McCabe, P.; and Ballard, K. J. 2018. Automated speech analysis tools for children's speech production: A systematic literature review. *Int J Speech-Lang Pathol*, 20(6): 583–598.

Naeini, S. A.; Simmatis, L.; Jafari, D.; Yunusova, Y.; and Taati, B. 2024. Improving Dysarthric Speech Segmentation With Emulated and Synthetic Augmentation. *IEEE J Transl Eng Health Med*.

Network, C. S. D. R. 2017. Good practice guidelines for the transcription of children's speech in clinical practice and research. www.speechtherapy.org.uk. Accessed: 2024-07-03.

O'Shaughnessy, D. 2024. Trends and developments in automatic speech recognition research. *Comput Speech Lang*, 83: 101538.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, 28492–28518. PMLR.

Räsänen, O. J.; Laine, U. K.; and Altosaar, T. 2009. An improved speech segmentation quality measure: the r-value. In

Tenth Annual Conference of the International Speech Communication Association. Citeseer.

Ribeiro, M. S.; Eshky, A.; Richmond, K.; and Renals, S. 2019. Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions. *arXiv preprint arXiv:1907.00818*, 16–20.

Rudzicz, F.; Namasivayam, A. K.; and Wolff, T. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang Resour Eval*, 46: 523–541.

Seneff, S.; and Zue, V. 1988. Transcription and alignment of the TIMIT database. *TIMIT CD-ROM Documentation*.

Shabber, S. M.; and Bansal, M. 2024. Temporal feature-based approaches for enhancing phoneme boundary detection and masking in speech. *Int J Speech Technol*, 27(2): 425–436.

Shahin, M.; and Ahmed, B. 2024. Phonological-Level Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2024*, 307–311.

Smith, D. V.; Sneddon, A.; Ward, L.; Duenser, A.; Freyne, J.; Silvera-Tawil, D.; and Morgan, A. 2017. Improving Child Speech Disorder Assessment by Incorporating Out-of-Domain Adult Speech. In *Interspeech 2017*, 2690–2694.

Titterington, J.; and Bates, S. 2021. Teaching and learning clinical phonetic transcription. In *Manual of clinical phonetics*, 175–186. Routledge 2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN.

Yi, C.; Wang, J.; Cheng, N.; Zhou, S.; and Xu, B. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*, 1227–1241.

Yue, Z.; Xiong, F.; Christensen, H.; and Barker, J. 2020. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *ICASSP 2020*, 6094–6098. IEEE.

Zhu, J.; Zhang, C.; and Jurgens, D. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022*, 8167–8171. IEEE.