Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling

Theodoros Kouzelis[1], *Georgios Paraskevopoulos,*[1,2], *Athanasios Katsamanis*[1], *Vassilis Katsouros*[1]

[1] Institute for Language and Speech Processing, Athena Research Center,
[2]School of ECE, National Technical University of Athens, Athens, Greece

theodoros.kouzelis@athenarc.gr, g.paraskevopoulos@athenarc.gr, nkatsam@athenarc.gr,
vsk@athenarc.gr

Abstract

The study of speech disorders can benefit greatly from time-aligned data. However, audio-text mismatches in disfluent speech cause rapid performance degradation for modern speech aligners, hindering the use of automatic approaches. In this work, we propose a simple and effective modification of alignment graph construction of CTC-based models using Weighted Finite State Transducers. The proposed weakly-supervised approach alleviates the need for verbatim transcription of speech disfluencies for forced alignment. During the graph construction, we allow the modeling of common speech disfluencies, i.e. repetitions and omissions. Further, we show that by assessing the degree of audio-text mismatch through the use of Oracle Error Rate, our method can be effectively used in the wild. Our evaluation on a corrupted version of the TIMIT test set and the UCLASS dataset shows significant improvements, particularly for recall, achieving a 23-25% relative improvement over our baselines.

**Index Terms**: fluency disorder, phonetic alignment, WFST

1. Introduction

Speech disfluencies are common for both children and adults [1]. Stuttering when reading aloud or speaking spontaneously is an example of a disfluency that can have a profound impact on an individual's ability to communicate effectively [2]. Stuttering typically begins in childhood [3] and manifests itself as interruptions in the flow of speech such as repeating words or sounds and committing false starts. Diagnosis and assessment of such disfluencies have traditionally been done by clinicians, who manually count stuttering events after having transcribed and aligned text and audio of recorded sessions to classify disfluencies [4]. Disfluencies in both spontaneous speech and aloud reading are not prevalent only in the speech of individuals with speech disorders. In [5], Shriberg demonstrates that there is a 50% probability for a sentence of 10–13 words to include a disfluency and that the probability increases with sentence length.

In studies of stuttering or other disfluencies, the analysis of large corpora is essential due to the rich inherent variability in speech production. However, the traditional method of manually annotating and aligning words and phonemes is an extremely time-consuming process and can take up to 130 times real-time [6]. Automatic forced alignment, which detects time boundaries for each phoneme or word of already transcribed speech, can replace or speed up manual alignment depending on the required level of accuracy [7].

Manual transcription processes, however, typically correct or remove any disfluency events, resulting in transcripts that are considerably cleaner than the actual speech [8]. Further,

Figure 1: *An overview of our proposed approach. Given an approximate transcription, the disfluent verbatim utterance is aligned with the corresponding sections of the audio.*
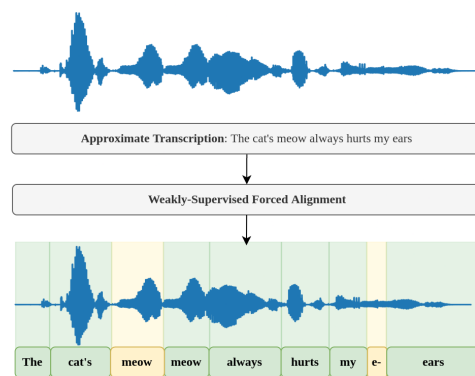

when speech contains disfluencies, it is necessary to manually transcribe any corresponding events before the alignment since the performance of conventional forced alignment approaches can be significantly affected by mismatches between speech and text [9]. Therefore, it is crucial for robust speech-to-text alignment to be able to detect disfluencies and align them accurately with the appropriate parts of the utterance.

Conventionally, speech-to-text alignment is performed using Hidden Markov Models - Gaussian Mixture Model (HMM-GMM) systems by application of the standard Viterbi-based forced alignment. However, very long audio and potential mismatches between speech and text pose a challenge to these systems as they are susceptible to error propagation. In the literature these issues are usually mitigated by iterative approaches [9, 10, 11]. In [11], the alignment and correction of approximate transcriptions are addressed with an iterative approach and the use of a modified lattice that allows word insertions and deletions. A similar approach with a grammar-based decoder has been used for the alignment of dysarthric speech [12]. Prior research has explored the effectiveness of using an augmented decoding grammar based on the original transcription to identify disfluencies, including word and syllable repetitions, in both stuttering detection [13] and disfluency detection during oral reading [14, 15]. However, these studies have not addressed the issue of performance degradation of forced alignment when faced with untranscribed disfluencies.

While HMM-GMM based techniques have been thoroughly researched for the alignment of speech and text, recent works are exploring the use of neural networks with promising results [16, 17, 18, 19]. In most recent neural forced aligners the final alignment is extracted from the emission probabilities by Dynamic Time Wrapping (DTW) [16] or by following

Figure 2: *The composing parts of the alignment graph $\mathcal{A}$. (a), (b) present the linear FSA and the Modified FSA constructed on the sentence "don't ask" with CMU phonetic transcription [D AA N T AE S K]. Part-word repetition arcs (PW) are illustrated in purple, word repetition arcs (W) in blue, and word deletion arcs (D) in red. (b), (d) are an example Emissions Graph and the standard CTC Topology Graph constructed over the CMU phone set. For clarity, we illustrate the CTC topology over two phones {AA, AE}.*
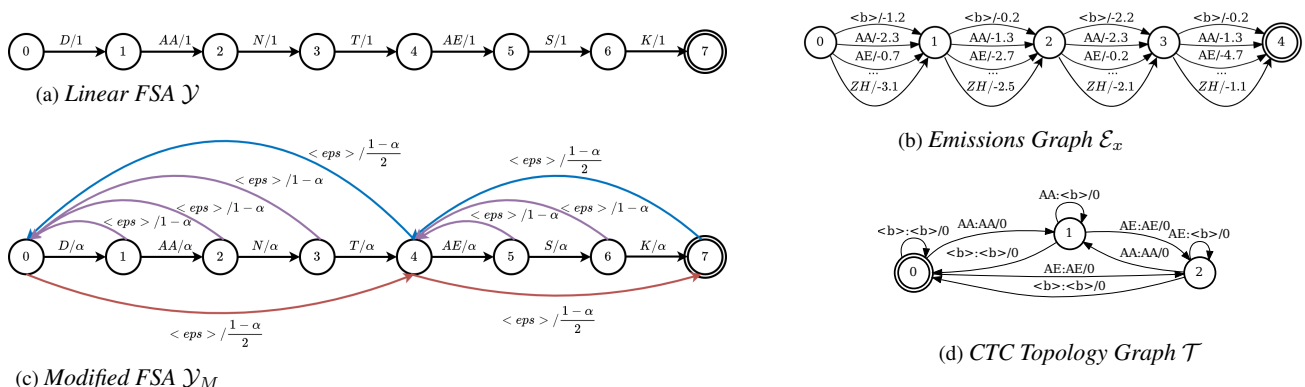
the procedure in [18] if the blank token of Connectionist Temporal Classification (CTC) training is present i.e., cumulative score, blank distribution, and beam search decoding [20]. These methods limit the modifiability made possible by decoding with Weighted Finite State Transducers (WFSTs). WFST representations enable the efficient incorporation of lexicons and language models into CTC decoding [21], memory-efficient variations of the standard CTC [22], and even training with partially labeled data [23].

In this work, we investigate the performance degradation of state-of-the-art forced aligners under untranscribed disfluencies. We demonstrate that given the emission probabilities of a frame classification or CTC-based model, the forced alignment can be achieved with WFST operations. Building on this, we propose a weakly-supervised forced alignment that aims at aligning the full verbatim disfluent utterance without the need for manual transcription. An overview of our proposed system is presented in Figure 1. To address an in the wild scenario, where the severity of audio-text mismatch is unknown we constrain the degrees of freedom of the weakly-supervised forced alignment leveraging Oracle Error Rate [24]. To test our method we employ a *state-of-the-art* neural forced aligner [16] that is based on Wav2Vec2 [25]. We evaluate our approach on the UCLASS dataset [4] and a corrupted version of the TIMIT test set with synthesized disfluencies.

Our key contributions can be summarized as follows: (a) We address the performance degradation of forced alignment under untranscribed disfluencies, (b) we introduce a weakly supervised forced alignment that is applicable to modern CTC-based neural forced aligners, (c) we construct and provide DisfluenTIMIT, a testbed for the evaluation of forced alignment under untranscribed disfluencies which we hope will inspire further research on the topic. Our code is available as open-source[1].

2. Method

For the methods described in this Section, we assume: (a) an input sequence $X = [x_0, ..., x_{T-1}]$ corresponding to the speech signal, (b) a phonetic transcription $Y = [y_0, ..., y_{N-1}]$, $y_n \in \mathcal{P}$, where $\mathcal{P}$ is the phonetic vocabulary, and (c) a net-

work that given X generates frame-based phone posteriors in the log-domain $\log P(p_t|X)$ with $p_t \in \mathcal{P}$ and $t \in T$.

Weakly supervised forced alignment with WFSTs: Given the phonetic transcription Y, a label graph $\mathcal{L}_Y$ that represents all the valid alignment paths can be constructed. Further, using the phone posterior log probabilities, an emission graph $\mathcal{E}_x$ (Figure 2b) can be created, where arc weights correspond to $\log P(p_t|X)$. The alignment graph is derived by composing the emission and label graphs, $\mathcal{A} = \mathcal{L}_Y \circ \mathcal{E}_x$. The shortest path in $\mathcal{A}$ is the final frame-level alignment. The label graph can be further deconstructed to its composing parts $\mathcal{L}_Y = \mathcal{T} \circ \mathcal{Y}$, a topology graph $\mathcal{T}$ (Figure 2d) and a linear Finite State Acceptor (FSA) $\mathcal{Y}$ (Figure 2a), that accepts only the input phonetic transcription Y. Thus the alignment graph can be written as:
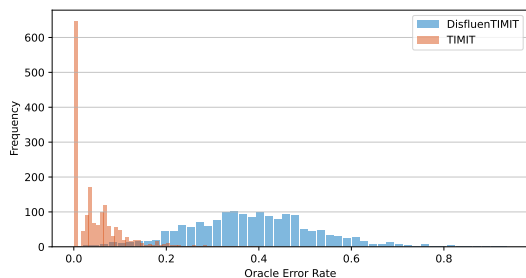
$$\mathcal{A} = \mathcal{T} \circ \mathcal{Y} \circ \mathcal{E}_x \qquad (1)$$

To enable the automated recognition and alignment of disfluencies, we suggest a straightforward adaptation to the linear FSA. Given an approximate transcription, a Modified FSA $\mathcal{Y}_M$ incorporating non-consuming $< eps >$ arcs is constructed. Figure 2c shows an example Modified FSA for the sentence "don't ask" [D AA N T AE S K]. Following the horizontal arcs, the original transcription is obtained. Traversing the blue arcs allows both word and phrase repetitions, such as "don't ask don't ask". Red arcs allow the deletion of a word or phrase. Omissions of words are frequently observed during oral reading, particularly in children [26]. Part-word repetitions are allowed by the purple arcs. Unlike the word lattices described in [13] and [14], our approach entails modeling of part-word repetitions at the phoneme level rather than the syllable level, alleviating the need for manual segmentation of words into syllables. Intra-word repetitions are not modeled since they are not as common in disfluent speech [13].

The arc weights of $\mathcal{Y}_M$ are constrained by a prior probability α which restraints excessive changes to the input phonetic transcription. For each state, there is a probability of α assigned to traversing a horizontal arc that follows the path outlined by the transcription. Conversely, the probability of traversing a non-consuming $< eps >$ arc is $1 - \alpha$. In the case of the start and end word states, the probability is equally divided between the deletion (red) and the repetition (blue) arc. When $\alpha = 1$ the Modified FSA degenerates to the linear FSA. As noted by

Figure 3: *Oracle Error Rates for TIMIT and DisfluenTIMIT test sets. DisfluenTIMIT is presented in Section 3.*

fluencies are labeled accordingly, and the ground truth disfluent phonetic transcription is acquired.

Table 2: *Types of disfluencies in the samples from UCLASS and DisfluenTimit. The 3^d and 4th row present average percentage of each type.*

Label	Disfluency Type	DisfluenTimit	UCLASS
PW	Part-word repetitions	4.7%	0.4%
W	Word repetitions	4.8%	7.0%
PH	Phrase repetitions	4.7%	0.7%
D	Word deletions	4.7%	0%
I	Injections	0%	2.3%

More specifically, to generate word repetitions, we repeat a randomly selected word one, two, or three times before the original utterance. For phrase repetitions, we repeat a phrase consisting of two to three words. For part-word repetitions, we copy and append a random number of phones before the selected word. For word deletions, we delete one, two, or three consecutive words from the audio. To ensure a smoother transition between the added audio and the existing clip, we use interpolation as described in [29]. The original TIMIT 61 phones are collapsed into the 39 CMU phone set as in [16].

prior studies [13, 14, 27] the weights of $\mathcal{Y}_M$ are effective when $\alpha \geq 0.9$ as the log probabilities of graph $\mathcal{E}_x$ are of larger orders of magnitude. Thus we will denote $\alpha = 1 - 10^{-\beta}$ with $\beta \in [0, +\inf)$ and consider values of β.

Adaptively choosing a value for β: The modified FSA adds degrees of freedom in the forced alignment process allowing the recognition of untranscribed events, but also the possibility of false alarms. Thus, the value of β must balance this trade-off. In previous works [13, 14, 27], the arc weights are empirically decided to fit the target data. But in a real word scenario, the mismatch between audio and transcription as well as the severity of speech disfluencies are unknown. So, the value of β should be proportional to the mismatch between the spoken utterance and the transcription. A simple method to quantify this mismatch is Oracle Error Rate (OER) [24]. A low OER indicates that the transcription of the utterance is mostly correct. We calculate OER as in [24, 28], with a biased bi-gram language model.

Figure 3 demonstrates that utilizing OER is an effective means of distinguishing between clean and disfluent samples. Building on this, we derive an intuitive formula to condition β, and consequently the degrees of freedom of the weakly-supervised forced alignment, on OER. For a sampled audio-transcription pair S we set $\beta_S = 10^{(1-O_S)}$, where O_S is the OER of sample S. Since OER can be greater than 1, we explicitly restrict $O_S \in [0, 1]$ and, thus, $\beta \in [1, 10]$.

3. Data

The scarcity of speech disfluency datasets and the need for manual time-aligned data pose significant challenges for evaluating speech alignment methods. To overcome these challenges, we create the DisfluenTimit dataset by introducing speech disfluencies into the standard Timit test set. Table 2 displays the disfluency types and their percentage of occurrence calculated over the total number of words in the two datasets.

UCLASS is a commonly used dataset in disfluency-related machine learning studies, consisting of 457 audio recordings of children with known stuttering issues. UCLASS Release Two contains five recordings with manually aligned transcriptions in orthographic format, where only the onset of each word or disfluency is labeled. Since disfluencies are flagged we are able to extract the approximate transcription for our experiments.

DisfluenTIMIT is a synthetic dataset that includes various types of disfluencies, including part-word repetitions, word repetitions, phrase repetitions, and word deletions. To generate these disfluencies, we followed a very similar approach to [29], sampling and copying the audio or removing it in the case of deletions. For an audio file with n words, we randomly sample a disfluency percentage $p \in [0.1, 0.2, 0.3]$ and insert $\lceil p \cdot n \rceil$ random disfluencies of types $d \in [PW, W, PH, D]$. The dis-

4. Experimental Setup

To evaluate the methods described in Section 3 we utilized a pre-trained *state-of-the-art* frame classification model called `W2V2-FC`[2][16] that has been trained on Librispeech [30]. We compare our proposed method against two strong baselines: `Montreal Forced Aligner (MFA)` and `W2V2-FC` with `DTW` as in [16]. All WFST operations described in Section 3 are executed using the k2 framework[3]. For each audio, we extract the emission probabilities using `W2V2-FC`, resulting in a 2D tensor $E \in \mathbb{R}^{N \times T}$ where T is the number of frames and N=42, including the 39 CMU phones and special tokens such as [SIL] for silence, [UNK] for unknown phones, and [PAD] for padding. As samples in the UCLASS dataset are longer, we divide the audio into 10-second chunks, feed them through the model, and concatenate the output emission probabilities in the time dimension. The [SIL] token and the *blank* token $< b >$ of the topology graph $\mathcal{T}$ are mapped in the same token id, as `W2V2-FC` has been fine-tuned with the cross-entropy loss instead of CTC, and thus the blank token is absent. In all our experiments the transition weights of the Modified FSA in the weakly-supervised approach are adaptively chosen for each sample as described in Section 2.

We evaluate the alignment accuracy using precision, recall, F1 score, and R-value [31]. A hit is counted for each predicted phone boundary that falls within a tolerance of τ and matches the predicted phone. For DisfluenTIMIT, we only evaluate the phone onsets since each boundary marks the onset and offset of consecutive phones, while for UCLASS, we evaluate only the word onsets. We use a tolerance of 40ms and 100ms for DisfluenTIMIT and UCLASS respectively. Additionally, we measure the percentage of accurately predicted frame labels at a 10ms timescale for DisfluenTIMIT (Overlap metric in Table 1). The relative performance reduction for each metric M is calculated as:

$$M_{\downarrow\%} = \frac{M_{verbatim} - M_{approximate}}{M_{verbatim}} \times 100 \quad (2)$$

---

[2]https://huggingface.co/charsiu/en_w2v2_fs_10ms
[3]https://github.com/k2-fsa/k2

Table 1: *Evaluation results on UCLASS and DisfluenTIMIT. WFST and WFST_WS denote the standard and weakly-supervised force alignment presented in Section 2.*

Dataset	Method	P	$P_{\downarrow\%}$	R	$R_{\downarrow\%}$	F1	$F1_{\downarrow\%}$	R-val	$R\text{-val}_{\downarrow\%}$	Overlap	$\text{Overlap}_{\downarrow\%}$
DisfluenTIMIT	MFA	0.56	14.9	0.48	29.4	0.51	23.4	0.58	17.9	0.50	24.4
	DTW	0.59	9.0	0.47	22.4	0.52	17.4	0.59	14.1	0.57	15.4
	WFST	0.57	12.3	0.47	21.9	0.51	18.3	0.584	14.4	0.57	15.3
	WFST_WS(Ours)	**0.61**	**5.8**	**0.60**	**1.7**	**0.60**	**3.8**	**0.66**	**3.4**	**0.67**	**1.1**
UCLASS	MFA	0.36	28.4	0.37	33.1	0.36	30.8	0.42	23.4	-	-
	DTW	0.54	**1.67**	0.52	4.8	0.53	3.3	0.57	2.36	-	-
	WFST	0.55	2.2	0.52	6.0	0.54	4.1	**0.58**	3.1	-	-
	WFST_WS(Ours)	**0.56**	1.9	**0.54**	3.8	**0.54**	2.9	**0.58**	2.3	-	-

Table 3: *The effect of word repetition W, word deletions D and part-word repetition PW arcs on performance metrics for DisfluenTIMIT dataset. Minus (-) denotes without.*

Method	P	R	F1	R-val	Overlap
WFST_WS	0.61	0.60	0.60	0.66	0.67
-PW	0.59	0.58	0.58	0.64	0.65
-D	0.59	0.59	0.58	0.63	0.66
-W	0.59	0.47	0.52	0.60	0.57
-W-D	0.57	0.47	0.51	0.59	0.56
-W-PW	0.59	0.47	0.52	0.59	0.57
-PW-D	0.57	0.58	0.58	0.62	0.64

Where, $M_{approximate}$ and $M_{verbatim}$ denote the performance metrics calculated with untranscribed and transcribed disfluencies, respectively.

To evaluate the effectiveness of the adaptive approach presented in Section 2 for adjusting the Modified FSA weights, we conduct an experiment that simulates a real-world scenario where a combination of disfluent and clear speech samples are encountered without prior knowledge of the speech-text mismatch. We create a mixed test set, referred to as MixTIMIT, by randomly selecting the disfluent or fluent version of a sample from the DisfluenTIMIT and TIMIT test sets, respectively.

5. Results and Discussion

Table 1 displays the alignment performance of all methods on both datasets. Weakly-supervised forced alignment with WFSTs demonstrates superior performance across all metrics, with the smallest performance reduction when disfluencies are not transcribed. However, it is worth noting that while recall and frame overlap are significantly improved, precision is only marginally affected due to the presence of false alarms. Interestingly, DTW outperforms MFA in both datasets indicating that it is less affected by mismatches between audio and transcription.

To investigate the impact of the number of untranscribed disfluencies on the performance, we divide uniformly DisfluenTIMIT into three categories based on the percentage of disfluencies in relation to the total number of words. The results are presented in Figure 4, which clearly demonstrates that our proposed approach is robust to untranscribed disfluencies. In contrast, all other methods and especially MFA, experience significant performance degradation, particularly in severe cases where the percentage of disfluencies is high.

Additional arcs: In Table 3 we present the performance gains associated with each of the additional arcs to gain insight into the relative importance of these components. Our results indicate that omitting repetition arcs W has the most significant impact on alignment performance, suggesting that untranscribed word or phrase repetitions are a primary cause of performance degradation of forced alignment.
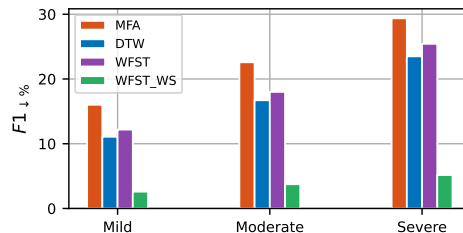

Figure 4: *Relative performance reduction of F1 metric under mild, moderate and severe untranscribed disfluencies.*

Effects of adaptive β: Using the MixTIMIT test set we compared the impact of adaptively choosing β for each sample versus globally setting β to an a priori value for all test samples. The results are presented in Table 4. While the performance isn't improved substantially, our adaptive approach succeeded at mitigating the necessity of prior knowledge about the severity of untranscribed speech disfluencies in target data or an extensive grid search.

Table 4: *Global (G) vs adaptive (A) β ablation on MixTIMIT.*

β	P	R	F1	R-val	Overlap
1 (G)	0.69	0.67	0.68	0.72	0.63
10 (G)	0.69	0.67	0.68	0.73	0.63
100 (G)	0.68	0.65	0.66	0.71	0.62
1000 (G)	0.67	0.61	0.64	0.69	0.58
$10^{(1-O_S)}$ (A)	0.70	0.68	0.69	0.74	0.63

6. Conclusions and Future Work

In this paper, we have explored the performance decrease of state-of-the-art forced aligners under untranscribed disfluencies. We proposed a weakly-supervised forced alignment, by a straightforward modification of the alignment graph construction of CTC-based models, utilizing WFSTs. Our method enables the alignment of common disfluencies without the need for manual transcription while incurring minimal performance decrease when disfluencies are not transcribed. Utilizing OER to quantify the severity of untranscribed disfluencies before alignment enables the effective application of our approach in the wild. In feature work, we plan to further explore the use of WFSTs to model pronunciation variants and other disfluency types such as injections, and also make the Modified FSA's weights learnable through Expectation Maximization.

7. Acknowledgements

8. References

[1] A. L. Choo, D. Greenberg, H. Li, and A. Talwar, "Rate of stuttering and factors associated with speech fluency characteristics in adult struggling readers," *Journal of Learning Disabilities*, vol. 56, no. 1, pp. 7–24, 2023, pMID: 35549596. [Online]. Available: https://doi.org/10.1177/00222194221095265

[2] R. Lowe, R. Menzies, M. Onslow, A. Packman, and S. O'Brian, "Speech and anxiety management with persistent stuttering: Current status and essential research," *Journal of Speech, Language, and Hearing Research*, vol. 64, pp. 1–16, 01 2021.

[3] E. Yairi and N. Ambrose, "Epidemiology of stuttering: 21st century advances," *Journal of Fluency Disorders*, vol. 38, no. 2, pp. 66–87, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094730X12001052

[4] P. Howell, S. Davis, and J. Bartrip, "The uclass archive of stuttered speech." *Journal of speech, language, and hearing research : JSLHR*, 03 2009.

[5] E. Shriberg, "Preliminaries to a theory of speech disfluencies," 1994.

[6] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–677, 2004.

[7] N. J. Young and M. McGarrah, "Forced alignment for nordic languages: Rapidly constructing a high-quality prototype," *Nordic Journal of Linguistics*, p. 1–27, 2021.

[8] V. Zayats, T. Tran, R. A. Wright, C. Mansfield, and M. Ostendorf, "Disfluencies and human speech transcription errors," in *Interspeech*, 2019.

[9] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[10] P. J. Moreno, C. F. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," *5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998.

[11] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Interspeech*, 2006.

[12] Y. T. Yeung, K. H. Wong, and H. Meng, "Improving automatic forced alignment for dysarthric speech transcription," in *Proc. Interspeech 2015*, 2015, pp. 2991–2995.

[13] S. Alharbi, M. Hasan, A. J. H. Simons, S. Brumfitt, and P. D. Green, "A lightly supervised approach to detect stuttering in children's speech," in *Interspeech*, 2018.

[14] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, "Mispronunciation detection in children's reading of sentences," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1207–1219, 2018.

[15] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst, and H. V. hamme, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, 2009, spoken Language Technology for Education. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639309000697

[16] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8167–8171, 2021.

[17] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7274–7278.

[18] Y. Teytaut and A. Roebel, "Phoneme-to-Audio Alignment with Recurrent Neural Networks for Speaking and Singing Voice," in *Proceedings of Interspeech 2021*, ser. Proceedings of Interspeech 2021, International Speech Communication Association. Brno, Czech Republic: International Speech Communication Association, Aug. 2021, pp. 61–65. [Online]. Available: https://hal.science/hal-03552964

[19] M. C. Kelley and B. V. Tucker, "A comparison of input types to a deep neural network-based forced aligner," in *Interspeech*, 2018.

[20] Y. TEYTAUT, B. Bouvier, and A. Roebel, "A study on constraining Connectionist Temporal Classification for temporal audio alignment," in *Proc. Interspeech 2022*, 2022, pp. 5015–5019.

[21] Y. Miao, M. A. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, 2015.

[22] A. Laptev, S. Majumdar, and B. Ginsburg, "Ctc variations through new wfst topologies," in *Interspeech*, 2021.

[23] V. Pratap, A. Hannun, G. Synnaeve, and R. Collobert, "Star temporal classification: Sequence classification with partially labeled data," *arXiv preprint arXiv:2201.12208*, 2022.

[24] V. Peddinti, V. Manohar, Y. Wang, D. Povey, and S. Khudanpur, "Far-field asr without parallel data," 09 2016, pp. 1996–2000.

[25] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[26] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, "Automatic evaluation of reading aloud performance in children," *Speech Communication*, vol. 94, pp. 1–14, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639316303739

[27] P. Bell and S. Renals, "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 675–680.

[28] J. Yang, L. Ondel, V. Manohar, and H. Hermansky, "Towards automatic methods to detect errors in transcriptions of speech recordings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3747–3751.

[29] T. Kourkounakis, A. Hajavi, and A. Etemad, "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2986–2999, 2021.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

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