
Proceedings Paper:
Alharbi, S., Hasan, M., Simons, A.J.H. orcid.org/0000-0002-5925-7148 et al. (2 more authors) (2018) A lightly supervised approach to detect stuttering in children's speech. In: Proceedings of Interspeech 2018. Interspeech 2018, 02-06 Sep 2018, Hyderabad, India. ISCA , pp. 3433-3437.

10.21437/Interspeech.2018-2155

Interspeech 2018
2-6 September 2018, Hyderabad

A Lightly Supervised Approach to Detect Stuttering in Children's Speech

Sadeen Alharbi[1], *Madina Hasan*[1], *Anthony J H Simons*[1], *Shelagh Brumfitt*[2], *Phil Green*[1]

[1]Computer Science Department, The University of Sheffield,
[2]Human Communication Sciences Department, The University of Sheffield,
Sheffield, United Kingdom.

[ssmalharbi1,m.hasan,a.j.simons,s.m.brumfitt,p.green]@sheffield.ac.uk

Abstract

In speech pathology, new assistive technologies using ASR and machine learning approaches are being developed for detecting speech disorder events. Classically-trained ASR model tends to remove disfluencies from spoken utterances, due to its focus on producing clean and readable text output. However, diagnostic systems need to be able to track speech disfluencies, such as stuttering events, in order to determine the severity level of stuttering. To achieve this, ASR systems must be adapted to recognise full verbatim utterances, including pseudo-words and non-meaningful part-words. This work proposes a training regime to address this problem, and preserve a full verbatim output of stuttering speech. We use a lightly-supervised approach using task-oriented lattices to recognise the stuttering speech of children performing a standard reading task. This approach improved the WER by 27.8% relative to a baseline that uses word-lattices generated from the original prompt. The improved results preserved 63% of stuttering events (including sound, word, part-word and phrase repetition, and revision). This work also proposes a separate correction layer on top of the ASR that detects prolongation events (which are poorly recognised by the ASR). This increases the percentage of preserved stuttering events to 70%.

Index Terms: fluency disorder, speech language pathology, children's speech, stuttering detection

1. Introduction

Stuttering is a speech disfluency disorder that typically begins in childhood. Stuttering manifests itself by four years of age in 95% of sufferers [1], just as the child is learning to talk. The prognosis for full recovery dramatically reduces if stuttering persists into adolescence; therefore it is critical to address speech disorder problems in early childhood, since delays in medical interventions can result in wide-ranging social and mental difficulties [2, 3].

Clinicians usually diagnose stuttering by counting the number of stuttering events, to determine the severity of the condition [4, 5, 6]. This may done in real time, using a pen and paper, while the child is reading a set passage, which in this work we refer to as the original prompt (OP). The effectiveness of this approach greatly depends on experience [7]. Alternatively, clinicians may first transcribe a recorded session and then classify each stuttering event into one of several categories (including different kinds of repetition, prolongation, blocks and interjections). This provides a more accurate diagnosis [8]. The process of transcribing each spoken word takes time, effort and knowledge of the relevant categories.

In this work, we explore the use of automatic speech recognition (ASR) to create accurate transcriptions of stuttered speech, which may be used for diagnosis and also may be archived for further investigative research into the condition. ASR is already widely used in speech pathology as an assistive technology [9, 10]. However, it is well known that children's speech poses problems for ASR. Previous research has reported poor performance of ASR systems when recognising children's speech [11, 12]. This is caused by factors such as variable speech rate, or small vocal tract length [11]. Despite substantial reported efforts directed towards improving ASR for children's speech, progress in this area is still limited. [13].

Detecting stuttering events in children's speech is even harder. We seek to address this problem using a two-fold approach that targets the most common stuttering events in recordings of children performing a reading task. The first part is an ASR that uses task-oriented word (and sub-word) lattices to improve the detection of certain classes of stuttering event (sound, word, part-word and phrase repetition, and revision). The second part is a prolongation detector, based on the correlation of successive voiced frames, which identifies segments that correspond to prolongation events. The output of this stage is added as a correction layer to the ASR system.

The rest of the paper is organised as follows. Section 2 presents an overview of related studies that use various techniques to detect stuttering. The speech corpus, annotation guidelines and development methodology for the ASR used in our approach are described in Section 3. The lattice re-scoring approach is described in Section 4 and Section 5 explains the prolongation detection system. The integrated diagnostic system is described in Section 6. Section 7 presents the experiments used in this study. Finally, Section 8 summarises our findings and future work plans.

2. Literature review

It is commonly known that children have smaller vocal tracts, which creates challenges in recognising their speech [14, 12]. Attempting to distinguish stuttering events adds to the complexity of this task.

Early studies in disfluency detection that appeared to report positive results [15, 16] lacked complete statistical findings; therefore, their significance cannot be determined. Other studies presented stuttering events in isolated speech segments to an artificial neural network (ANN), such that the ANN was actually performing a classification task, rather than recognition in continuous speech [17, 18]. Further studies used a hybrid approach to detect stuttering in childrens reading tasks: Heeman et al. [19, 20] merged ASR outputs with the clinician's own manual annotations to produce corrected transcripts of the stuttering speech; this approach could not be described as fully automatic.

Two main approaches to prolongation detection are reported in the literature. In the supervisory approach, the sound signal is manually segmented into normal or prolongation seg-

10.21437/Interspeech.2018-2155

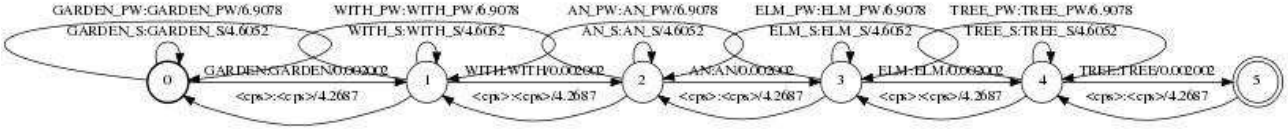

Figure 1: *Corresponding FST graph for the prompt 'GARDEN WITH AN ELM TREE'. The arc* ***GARDEN_PW*** *allows part-word repetition and it could be [gar] or [den] while* ***GARDEN_S*** *is sound repetition and it could be [ga]. The go-back transition allows word/phrase repetition and revision.*

ments and the labelled data is then used to train classifiers, such as SVM [15] or ANN [21]. In the unsupervised approach, quasi-silent areas of the signal are first removed using an automatic speech detection model and then the similarity between successive frames is used to produce initial estimates of possible prolongation segments. If the duration of the detected segment is found to be greater than a predefined threshold, the segment is labelled as a prolongation event; otherwise, it is considered a normal segment [22].

3. Data

3.1. Corpora of children's speech

The present study is based on a standard reading task that is used by clinicians to diagnose stuttering in children. For training purposes, we obtained all of the recordings of children's read speech from the UCLASS stuttering corpus, UCLASS Release Two [23]. It contained 107 recordings of readings contributed by 40 different speakers. We transcribed 48 speech samples (totalling 120 minutes) to use in the current research. The data was divided into 40 samples used as a training set, 4 samples used as a test set and 4 samples used for a development set. Each speaker read a passage once. To improve the ASR model, 7 hours from the PF-Star [24] corpus of children's read speech were added to the training set.

3.2. Data transcription and annotation

Table 1: *Stuttering types that considered in this study with their corresponding abbreviations.*

Label	Stuttering Type
S	Sound repetitions
PW	Part-word repetitions
W	Word repetitions
PH	Phrase repetitions
R	Revision repetitions
P	Prolongation

These 48 recordings did not have any associated transcriptions. However, UCLASS Release One [23] contains another dataset of spontaneous stuttering speech, for which 31 recordings had transcriptions. We have made a full verbatim transcription of the read speech dataset following the same conventions used for this subset. Transcriptions were orthographic, and included conventional forms to represent stuttering dysfluencies, for example: *This is a a a amazing*.

The annotation approach followed in this study is the same approach followed in our previous work [25] and it is the one proposed by Yairi and Ambrose [26]. In this study, all types of stuttering were considered except the interjection and block types. All stuttering types examined in the study are listed with their corresponding abbreviations in Table 1.

4. Lightly-supervised lattice decoding

Given an approximate transcription (close to a manual transcript but not exact), a more accurate transcription can be generated using a biased language model (LM). This approach is known as a lightly supervised approach. It has been used successfully to generate improved transcriptions for acoustic model training, so avoiding the need for expensive manual human transcription [27, 28, 29]. It has also been used to align and correct approximate transcriptions of long audio recordings [30], and for audio indexing and displaying subtitles. In a tutoring application, this approach was used by [31] to track and align a child's read passage.

The current work used a lightly supervised approach to track and identify stuttering events in a reading task. We used the original clean prompt (OP) as an approximation of a human manual transcript (that should include stuttering events). An initial decoding run used the draft ASR hypothesis to automatically align with the OP by dynamic programming. This step eased the merging of several hypothesised segments into the corresponding original prompt utterance. In the second stage, we used task-oriented finite state transducers (FST) for second-pass decoding. These task-oriented lattices were automatically generated from the OP and weights were tuned to allow for possible stuttering events. Figure 1 demonstrates a lattice that was generated for the OP '*garden with an elm tree*'. This can be considered a forced alignment with additional features. A set of arcs were added to each word in the generated lattice. The first two arcs allowed multiple occurrences of sound and part-word repetitions. Sound repetitions are represented by the suffix **_S** and include all repeated sounds that could occur in the beginning of the word, such as *ga ga* in *garden*. Part-word repetitions are represented by the suffix **_PW** and include all repeated syllable that could occur in the word *garden*. Traversing go-back arcs allowed for the possibility of both word/phrase repetitions and revisions, such as '*garden with garden with*'.

5. Prolongation detection system

Whereas ASR is highly effective in identifying segments classified by frequency-based features, it is less successful in identifying segments classified by time-based features, such as prolongation events, e.g. 'mmmommy'. To handle this, we used a separate autocorrelation algorithm, which measured the similarity between successive speech frames and proposed prolongation events as a correction to the ASR word lattice. Stutterers usually have a lower speaking rate than normal speakers, so thresholds for detecting prolongations had to account for natural variations in fluency and speaking rate on different occasions [32]. We followed the unsupervised approach of [22, 33], which uses two thresholds to decide whether two successive frames

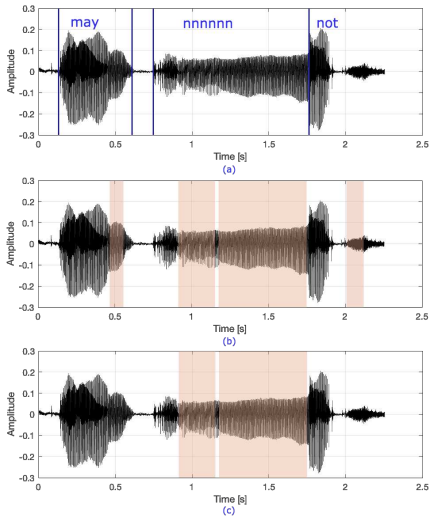

Figure 2: *Prolongation detection. (a) Speech sample 'may not' with prolongation in letter 'n' in word 'not' (b) highly similar segments, and (c) detected prolonged segment which is longer than threshold.*

were similar, and whether the duration of similar frames was sufficient to count as a prolongation. We found empirically that 0.9 was the best value for the first; but the second threshold had to be set dynamically, according to speaking rate.

Our prolongation detector used a prior filter to mark frames judged to be silence (the 'vadsohn' voice activity detector from the 'voicebox' toolbox [34]), in order to decrease the incidence of false alarms. It then used autocorrelation to measure the similarity between speech frames as a function of the lag, to detect prolongations. The length threshold for accepting candidate prolongation events was then normalised by the speaking rate, in order to discount false alarms, whose length fell below the threshold. The speaking rate detector used smoothed short-term energy and zero-crossing rates to detect a syllable when the energy reached a maximum in the absence of a peak in the zero-crossing rate. This gave accurate enough estimates of syllables and informed the speaking rate.

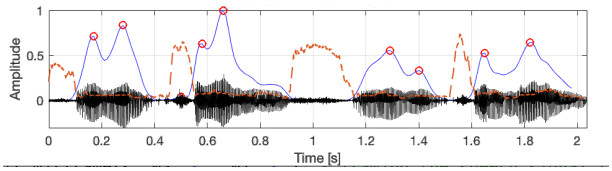

Figure 3: *An example of syllable counting method. Solid line, and dashed line are energy signal, and zero-crossing rate, respectively. Circle marks are considered in the syllable counting process.*

6. Integrated system

The proposed integrated system is presented in Figure 4. The speech signals were initially parameterised to mel-frequency cepstral coefficient (MFCC) feature sets. Subsequently, the MFCC frames were analyzed in parallel by a feature-based ASR trained on stuttering speech and by an autocorrelation-based prolongation detector. The ASR produced a transcription that

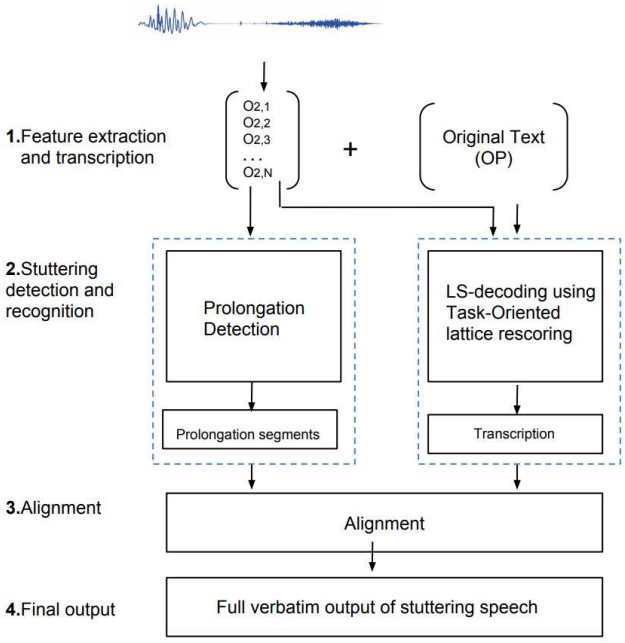

Figure 4: *Main stages of the proposed integrated system*

contained most types of stuttering event, such as sound, part-word, word or phrase repetitions and revisions, but not prolongations. The prolongation detector acted in parallel to detect prolongation events and served as a correction layer for the ASR. Thereafter, all detected prolonged segments were aligned with the ASR-output to produce a detailed verbatim transcript containing stuttering events.

7. Experiments

The following section presents our experiments on UCLASS data for detecting stuttering events using the method discussed in Sections 4 and 6. We used the Kaldi ASR toolkit [35] to train the acoustic model, following the WSJ recipe. In the following experiments, we evaluated the systems performance on sound and word repetitions, revision and prolongation. While the system could have detected part-word and phrase repetitions, these types of stuttering event were not present in the evaluation data. Below, we use the conventional $WordErrorRate$ to measure decoded output against an accurate manual transcription (including stuttering events), $MissRate$ to measure the count of missed stuttering events as a fraction of all events, and $FalsePositiveRate$ to measure the count of false alarms as a fraction of the number of original words.

7.1. Baseline experiments

In normal ASR systems, a statistically-trained language model (LM) always outperforms a specially-constructed lattice, due to the flexibility given to the ASR system to determine the best-matched word sequences. However, our experiments show that task-specific lattices perform better when seeking to detect rare stuttering events. Our initial experiments evaluated the performance of the ASR when the LM relied on the original prompt (OP) only. In the first experiment, we evaluated the ASRs performance when applying a statistical LM (tri-gram) that had

Table 2: *WER, miss rate (misses) and false positive rate (FPR) on the test set*

Measures	% WER	misses	FPR	misses	FPR	misses	FPR	misses	FPR
Stuttering-type		Sound repetition		Word repetition		Revision		AVG	
Statistical LM from stuttered training data	7.25	90%	0%	40%	0.29%	50%	0%	68%	0.29%
Task-oriented lattices	5.7	60%	0.59%	0%	0.29%	25%	0%	37%	0.89%

been built from the OP. The obtained WER was 8%. As expected, the results confirmed that the ASR, when built from the OP, tended to delete all stuttering events. Thus, the miss rate was 100%. In a second experiment, the lattices were rescored using a deterministic LM that was created from the OP with no additional features. The obtained WER was 7.9% without any stuttering recognition.

7.2. Decoding with task-oriented lattices

To evaluate the performance of the ASR system in detecting stuttering events, we used similar criteria to those applied in the NIST scoring tool [36]: insertions of stuttering events were considered false alarms, deletions were considered misses and the substitution of detected events by events from another stuttering category were also misses.

The results presented in Table 2 compare the ability of an ASR with a statistical LM that was trained on stuttering data, versus an ASR using task-specific lattices. As expected, a slight improvement (7.25%) was seen after training the statistical LM on stuttering data, as opposed to OP data. Although the ASR, after being retrained on stuttering data, was able to detect some stuttering events, the miss rate was 68%, which is still high. However, after re-scoring the word lattices using the task-oriented approach, the WER improved to 5.7, with a 45.5% improvement in the miss-rate compared to the statistical LM. Using the re-scoring approach, we preserved 63% of the stuttering events.

The least-well detected events were sound repetitions, where the miss rate reached 60%. A detailed investigation showed that these deleted sounds came from low-quality recordings. Additionally, these sounds were judged to be barely recognisable even by humans. Word repetitions were all successfully detected, with no misses; whereas the system missed 25% of the revision type. As shown in the revision example in Figure 5 (a), the ASR detected revisions, even when a word from the revised phrase was deleted. By contrast, the average false alarm rate was 0.89%, which is considered relatively low. This was mainly due to the constraints applied in each task-specific lattice.

Figure 5 contrasts the decoding behaviour of a task-specific lattice with stuttering arcs (a) against the deterministic lattice built from the OP (b), to show how adding the stuttering arcs allows successful detection of stuttering, versus no detection. Task-specific lattices are possibly over-constrained, in that they will not detect word insertions or substitutions, such as 'I' instead of 'he', which will be deleted in the output. While this worsens the WER (compared to a statistically-trained LM), it does not impact on the detection of stuttering events. Word-substitution happens naturally in children's speech [37], because the child often anticipates the next word. This behaviour is neither considered a revision-incomplete phrase, nor another category of stuttering event.

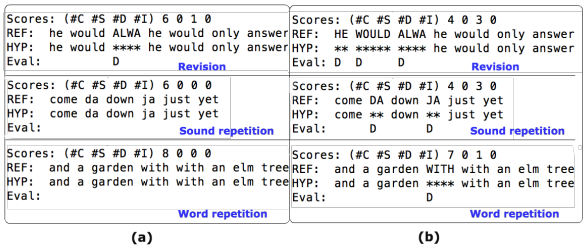

Figure 5: *Decoding, (a) example showing the ability of the ASR to detect stuttering events after applying task-oriented lattices; (b) example showing the deletion of stuttering events on the baseline ASR.*

7.3. Integration with prolongation detection system

The prolongation detector was evaluated on the test set. It successfully detected all prolongation events and had a 0% miss rate. From the results, it was clear that the detection of similarly correlated successive frames resulted in the effective identification of prolongation during continuous speech. However, artefact noises, such as background noises during the recording and heavy breathing, were still erroneously identified as prolongation. Using this approach resulted in a 5.3% false alarm rate. In the future, all detected false alarms could be minimised by applying a better silence remover.

A full verbatim translation, which was obtained after aligning the prolonged segment with the transcription that was produced from the ASR, preserved a total of 70% of stuttering events with a 6.25% false alarm rate. Although the false alarm rate increased after integrating systems, applying a better noise remover in the future could reduce the effects of this problem.

8. Conclusions and future work

Identifying stuttering events in children's speech is a hard task, due to the lack of available data needed for conventional statistical training methods. The rarity of particular stuttering events meant that these would most likely be deleted in the transcript. We addressed this problem initially using a lightly-supervised ASR with task-specific lattice re-scoring, which greatly increased the detection of many classes of stuttering event, apart from prolongations. A different approach was needed to detect prolongations, based on the correlation of successive voiced frames. The output of this stage was added as a correction layer to the ASR system. For future work, we could minimise false alarms in the prolongation detector by applying a better silence remover. Also, we plan to add more features to the task-oriented lattice, in order to attempt to recognise interjection events.

9. Acknowledgements

This research has been supported by the Saudi Ministry of Education, King Saud University

10. References

[1] E. Yairi and N. Ambrose, "Epidemiology of stuttering: 21st century advances," *Journal of fluency disorders*, vol. 38, no. 2, pp. 66–87, 2013.

[2] L. Iverach, S. OBrian, M. Jones, S. Block, M. Lincoln, E. Harrison, S. Hewat, R. G. Menzies, A. Packman, and M. Onslow, "Prevalence of anxiety disorders among adults seeking speech therapy for stuttering," *Journal of anxiety disorders*, vol. 23, no. 7, pp. 928–934, 2009.

[3] Y. Tran, E. Blumgart, and A. Craig, "Subjective distress associated with chronic stuttering," *Journal of fluency disorders*, vol. 36, no. 1, pp. 17–26, 2011.

[4] J. Yaruss, "Clinical measurement of stuttering behaviors," *Contemporary Issues in Communication Science and Disorders*, vol. 24, no. 24, pp. 33–44, 1997.

[5] J. S. Yaruss, "Real-time analysis of speech fluency: Procedures and reliability training," *American Journal of Speech-Language Pathology*, vol. 7, no. 2, pp. 25–37, 1998.

[6] E. G. Conture, *Stuttering: Its nature, diagnosis, and treatment*. Allyn & Bacon, 2001.

[7] S. B. Brundage, A. K. Bothe, A. N. Lengeling, and J. J. Evans, "Comparing judgments of stuttering made by students, clinicians, and highly experienced judges," *Journal of Fluency Disorders*, vol. 31, no. 4, pp. 271–283, 2006.

[8] H. H. Gregory, J. H. Campbell, C. B. Gregory, and D. G. Hill, *Stuttering therapy: Rationale and procedures*. Allyn & Bacon, 2003.

[9] P. Howell, S. Sackin, and K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: Ii. ann recognition of repetitions and prolongations with supplied word segment markers," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 5, pp. 1085–1096, 1997.

[10] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia)," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005)*. ISCA, 2005, pp. 149–152.

[11] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

[12] M. Russell and S. DArcy, "Challenges for computer recognition of childrens speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[13] S. Chandrakala and N. Rajeswari, "Representation learning based speech assistive system for persons with dysarthria," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1510–1517, 2017.

[14] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński, "Automatic detection of disorders in a continuous speech with the hidden markov models approach," in *Computer Recognition Systems 2*. Springer, 2007, pp. 445–453.

[16] ——, "Automatic detection of prolonged fricative phonemes with the hidden markov models approach," *Journal of Medical Informatics & Technologies*, vol. 11, p. 2007, 2007.

[17] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and E. Smołka, "Artificial neural networks in the disabled speech analysis," in *Computer Recognition Systems 3*. Springer, 2009, pp. 347–354.

[18] ——, "Hierarchical ann system for stuttering identification," *Computer Speech & Language*, vol. 27, no. 1, pp. 228–242, 2013.

[19] P. A. Heeman, A. McMillin, and J. S. Yaruss, "Computer-assisted disfluency counts for stuttered speech." in *INTERSPEECH*, 2011, pp. 3013–3016.

[20] P. A. Heeman, R. Lunsford, A. McMillin, and J. S. Yaruss, "Using clinician annotations to improve automatic speech recognition of stuttered speech," *Interspeech 2016*, pp. 2651–2655, 2016.

[21] P. Howell and S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech," in *Proceedings of the first World Congress on fluency disorders*, vol. 2, 1995, pp. 372–374.

[22] I. Esmaili, N. J. Dabanloo, and M. Vali, "An automatic prolongation detection approach in continuous speech with robustness against speaking rate variations," *Journal of medical signals and sensors*, vol. 7, no. 1, p. 1, 2017.

[23] P. Howell, S. Davis, and J. Bartrip, "The university college london archive of stuttered speech (uclass)," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 2, pp. 556–569, 2009.

[24] M. Russell, "The pf-star british english childrens speech corpus," *The Speech Ark Limited*, 2006.

[25] S. Alharbi, M. Hasan, A. J. Simons, S. Brumfitt, and P. Green, "Detecting stuttering events in transcripts of childrens speech," in *International Conference on Statistical Language and Speech Processing*. Springer, 2017, pp. 217–228.

[26] E. Yairi and N. G. Ambrose, "Early childhood stuttering ipersistency and recovery rates," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 5, pp. 1097–1112, 1999.

[27] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech Language*, vol. 16, no. 1, pp. 115 – 129, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523080190186X

[28] L. Chen, L. Lamel, and J. L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. I–189–92 vol.1.

[29] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–737.

[30] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Ninth International Conference on Spoken Language Processing*, 2006.

[31] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, "Detection of mispronunciations and disfluencies in children reading aloud," *Proc. Interspeech 2017*, pp. 1437–1441, 2017.

[32] C. R. F. d. Andrade, L. M. Cervone, and F. C. Sassi, "Relationship between the stuttering severity index and speech rate," *Sao Paulo Medical Journal*, vol. 121, no. 2, pp. 81–84, 2003.

[33] W. Suszyński, W. Kuniszyk-Jóźkowiak, E. Smołka, and M. Dzieńkowski, "Prolongation detection with application of fuzzy logic," *Annales Universitatis Mariae Curie-Sklodowska, sectio AI–Informatica*, vol. 1, no. 1, pp. 1–8, 2015.

[34] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[36] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. sigir*, vol. 7, 2007, pp. 51–57.

[37] C. E. Westby, "Language performance of stuttering and nonstuttering children," *Journal of Communication Disorders*, vol. 12, no. 2, pp. 133–145, 1979.