WOCCI 2017: 6th International Workshop on Child Computer Interaction
13 November 2017, Glasgow, Scotland, UK

Automatic recognition of children's read speech for stuttering application

Sadeen Alharbi, Anthony J H Simons, Shelagh Brumfitt, Phil Green

The University of Sheffield, Sheffield, United Kingdom.
[ssmalharbi1,a.j.simons,s.m.brumfitt,p.green]@sheffield.ac.uk

Abstract

Stuttering is a common speech disfluency that may persist into adulthood if not treated in its early stages. Techniques from spoken language understanding may be applied to provide automated diagnoses of stuttering from voice recordings; however, there are several difficulties, including the lack of training data involving young children and the high dimensionality of these data. This study investigates how automatic speech recognition (ASR) could help clinicians by providing a tool that automatically recognises stuttering events and provides a useful written transcription of what was said. In addition, to enhance the performance of ASR and to alleviate the lack of stuttering data, this study examines the effect of augmenting the language model with artificially generated data. The performance of the ASR tool with and without language model augmentation is compared. Following language model augmentation, the ASR tool's performance improved recall from 38% to 62.2% and precision from 56.58% to 71%. When mis-recognised events are more coarsely classified as stuttering/ non-stuttering events, the performance improves up to 73% in recall and 84% in precision. Although the obtained results are not perfect, they map to fairly robust stutter/ non-stutter decision boundaries.

Index Terms: speech recognition, human-computer interaction, stuttering recognition, ASR for children

1. Introduction

Stuttering is a complex disorder, an uncontrolled disfluency, which can cause a wide range of social and mental problems [1, 2]. Late intervention during childhood leads the disorder be considered as a chronic condition during adulthood because it is associated with several aspects of life such as disruptions in life quality [3], communication difficulties [4], job performance measurement [5]. In addition, The risk of mental difficulties is increased for people who stutter such as social anxiety [6]. The risks are increased during adulthood rather than childhood because most young children are not fully aware of their disfluency yet. Thus, clinician intervention should take place as early as the preschool years because later intervention does not help. Furthermore, it is not possible to determine a child's chance of naturally recovering, and children are less tractable as they get older due to the reduction of neural plasticity [7]. During the assessment phase, clinicians need to carefully measure the stuttering events to determine the severity of stuttering. This measurement is usually conducted by counting the number of stuttering events in the child's speech. This process is extremely dependent on the clinician's experience [8]. In another approach, the clinician transcribes a recorded session and classifies each spoken term into one of several normal, disfluent or stuttering categories [9]. This process takes a long time because of the need to write every spoken word, which takes time and effort and requires knowledge of the relevant categories. An automated speech transcription of the recorded speech using automatic speech recognition (ASR) could help clinicians speed up the assessment process and store the data for further investigations.

However, understanding children's speech is well known to be a challenge even for humans, due to several factors, such as speech spontaneity, a slow rate of speech and variability in the vocal effort [10]. Therefore, a large amount of data is required to train an ASR with an acceptable word error rate (WER) and to process the ASR output to automatically identify the stuttering events in the transcription. Researchers in this area have investigated three main approaches to detect stuttering events. The first area of study involves attempts to detect stuttering events from recorded speech signals. Howell and Sackin [11], for example, proposed the first attempt at stuttering recognition. Their study applied artificial neural networks (ANNs) and focused on identifying repetitions and prolongations. Their model correctly identified only 43% of the repetitions and 58% of the prolongations, but correctly identified 80% of the prolongations and repetitions combined. Geetha et al.[12] presented an objective method of differentiating stuttering disfluencies. They used ANN techniques on two groups of disfluent children. Several features were chosen to discriminate between normal and stuttering speech. They reported that ANN classifiers could predict the classifications of normal speech and stuttering with 92% accuracy. Another approach has been used to detect stuttering events from transcriptions. Mahesha and Vinod [13] used a lexical rule-based algorithm to detect and estimate the severity of 4 types of stuttering events: Interjection, word repetition, syllable repetition and prolongation, in orthographic transcripts from University College London's Archive of Stuttered Speech (UCLASS) [14].

The third approach is a combination of the previous two approaches aiming to apply ASR to recognise stuttering events. The first speech recognition attempt was made by Nöth et al [15]. The system modelled a grammar that consider positional regularities of the stutterers' disfluencies to recognize different stuttering events. However, the authors did not provide many details regarding the system evaluation. In addition, no information was included about the location and classification of the stuttering events. Another study employing an ASR approach was proposed by Heeman et al. [16] in an attempt to merge a clinician's annotations with an ASR transcript to produce an annotated transcript of audio files (between one and two minutes in duration) of read speech. Three types of stuttering were considered in [16]: revisions, interjection and repetitions. However, the proposed system relied on the availability of the clinician's annotations of the read recordings. This work investigates the ability to build an ASR that is able to recognise different stuttering events in children's read speech and produce

This project is part of a CloudCAST project. CloudCAST is an International Network (IN-2014-003) funded by the Leverhulme Trust

10.21437/WOCCI.2017-1

a useful word transcription of what was said by augmenting the language model with artificially generated data. Moreover, it is well known that the main limitation in children's speech-related research is the lack of large publicly available corpora for training purposes. The amount of data for stuttering children is even smaller and not always transcribed which is not ready for machine training. This research explores an approach which makes best use for existing stuttering data and augment a larger database of normal child speech to confessor for lack of stuttering data.

The remainder of the paper is organised as follows. The data augmentation design is presented in Section 2. The ASR setup involving guidelines and methodology used for preparing the stuttering data transcriptions from UCLASS data [14] are described in Section 3. Section 4 presents the experiments used in this study. Finally, the conclusion and future work are discussed in Section 5.

2. Data Augmentation

The essential components of the ASR system are the acoustic model and the statistical language model (LM). Most existing ASRs generate final word hypotheses based on the probability distributions of each word available in training text corpora provided by the LM. The performance of the ASR is highly dependent on the amount and style of the text seen in training corpora. In general, rich text leads to a better model. However, the text that has been used to train the model needs to match the language style used in the ASR application. Thus, many problems have been raised in some ASR applications due to the difficulty of providing a good match, in-domain and sufficient text to reach a satisfactory level of performance in the ASR. Several solutions have been proposed.

The first solution is LM adaptation. Many approaches to LM adaptation have been proposed, such as dynamic adaptation, consisting of continuously updating in time the LM probability distributions [17]. In the previous work, Tatsuya et al. [18] studied the LM adaptation of automatic real-time lecture transcription using information provided in presentation slides used in a lecture. However, due to the small amount and fragmentary text content provided in the presentation slides, the author applied a global and local adaption scheme. The adaptation of global topics is used based on probabilistic latent semantic analysis (PLSA) by adding keywords shown in all the slides. For local adaption, they also used a cache model to store each word mentioned in the slide used during each utterance, and the occurrence probabilities of these words are heightened as they are more likely to be re-used. They reported that the proposed approach achieves a significant enhancement of recognition accuracy, particularly in the detection rate of content keywords.

Another solution could be applying data augmentation which somehow generates extra artificial data for events that are not commonly observed in the available training data. The artificial data have to be generated from some other source of knowledge which provides some relative frequencies of the artificial events that you want to generate. So, the revised probabilities in the LM correspond more closely to the target language that wish to recognise.

The method suggested in this paper is inspired by Tatsuya et al. [18] and begins with the speculation that if the UCLASS read corpus (the corpus that including stuttering events) is augmented, then the probability of stuttering events will increase in general and be considered during recognition.

According to Howell and Vause, the vowel that occurs often sounds like schwa in stuttered repetitions of a syllable even when schwa is not intended [19]. The commonness of the schwa is probably on the basis of its being the most easily and readily articulated of the vowels. The occurrence of schwa in syllable repetition usually happen when the person begins to vocalize while breathing. In rapid repetitive blocks there is insufficient time to make the articulatory movements necessary to produce the appropriate vowel of the syllable being repeated. Hence the more readily formed schwa tends to precede the appropriate vowel [20].

In the system, the child will read a known passage. Thus, the child in the test recording may stutter in any word of the given passage. Thus, adding a schwa sound to each word per utterance will increase the probability of stuttering in that word, which may lead to improving the recognition of stuttering events overall. For this study, each utterance in the training corpus, such as 'come down', we automatically generated a stuttering event by taking the first letter of the first word and adding a vowel after it, 'ca come' and adding the whole utterance with the new stuttering event to the augmented corpus 'ca come down'. Then, we took the first letter of the second word with a vowel and add it as another new utterance 'come da down'. Using this approach augmented the probability of stuttering events in the corpus used in the LM, maintaining the balance for the clean words (non-stuttered) in each utterance.

In this study, we trained an acoustic model using 9 hours of speech of which 7 hours came from PF-Star corpus [21] of normal child speech and 2 hours initially from the UCLASS corpus. We needed to train on both datasets because training on just UCLASS data alone was insufficient. In both corpora, children reading from simple stories such as 'Arthur the Rat' and 'Poor Fisherman'. In the combined training set, the probability of stuttering events is small because of the dominant of PF-star.

Algorithm 1 Augmentation Algorithm

```
$N$: Number of utterances $\rightarrow U_n$
$W_n$: Number of words in $U_n$
$S$: stuttering event
for $W_1$ to $W_n$ do
    $\hat{W_n} = S@W_n$
    $\hat{U_n^w} = \left\{ W_{n-1}, \hat{W_n}, W_{n+1} \right\}$
end for
```

3. Kaldi-ASR Setup

We used the Kaldi ASR toolkit [22] both to build the acoustic model and to train an ASR system to recognise stuttering events.

3.1. Training Data Preparation

The ASR system needs to recognise children's with stuttering speech, we started off with UCLASS [14], Release 2 (read speech) and we add to this a much larger corpus of children's read speech from PF-Star [21] to satisfy training requirements. The PF-Star sentences were all transcribed but the UCLASS, Release 2, sentences were not transcribed. However, we used the transcription convention of an earlier transcription's of UCLASS, Release 1, in order to transcribe release 2 except in sound repetition. To transcribe sound repetition, we inserted orthographic vowels which would aid the pronunciation model in the ASR after each repeated sound such as 'wa what'. This transcription approach were applied for all experiments. We experimented the ASR performance with transcription of repeated sound with no vowel after repeated sound following the exact transcription approach proposed by the UCLASS dataset, however, the ASR failed to recognise any repeated sound. So, we add a vowel during the transcription process after each repeated sound and apply it for the experiments proposed in our study.

The PF-Star corpus includes samples from 158 children aged 4 to 14 years. Most of the children recorded 20 SCRIBE sentences, a list of 40 isolated words, a list of 10 phonetically rich sentences, 20 generic phrases, an accent diagnostic passage (the sailor passage) and a list of 20 digit triples. In the beginning, the system were trained with the designated training set (86 speakers, approximately 7 hours and 30 mins) and tested it with the evaluation test set (60 speakers, approximately 5 hours and 50 minutes). This corpus contains simultaneous recordings from both a headset microphone and a desk microphone. Recordings from the desk microphone were used for training and testing to evaluate the results while taking into account the domestic background. We prepared all the data preparation scripts to provide the necessary files for Kaldi.

The speech samples obtained from UCLASS, Release 2, are much smaller compared with the PF-Star corpus. The complete database consists of recordings of monologues, readings and conversations without transcriptions provided. It contains 107 reading recordings contributed by 40 different speakers. In this work, only 48 read speech samples (about 2 hours) from 35 different speakers were used. As is usual with small datasets a cross-validation (CV) technique was used to partition the stuttering data and using partitioning turn as test set.

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

Table 1: *Data Annotation*

3.2. Data Transcription and Normalisation

In this study, we used the 48 publicly available audio recordings of children's read speech in Release Two of UCLASS [14]. The transcription approach for these files followed in this study is the one proposed by Yairi and Ambrose [23] and used by Fabiola and Claudia [24]. Their approach considered eight types of stuttering: 1) sound repetitions, which include phoneme repetitions (e.g. 'fa face'); 2) part-word repetitions, which consider a repetition of less than a word and more than a sound (e.g. 'any anymore'); 3) word repetitions that count the whole word repeated (e.g. 'mommy mommy'); 4) prolongations, which involve an inappropriate duration of a phoneme sound (e.g. 'mmmay'); 5) phrase repetitions that repeat at least two complete words (e.g. 'it is it is') 6) interjections, which involve the inclusion of meaningless words (e.g. 'ah', 'um'); 7) revisions that attempt to fix grammar or pronunciation mistakes (e.g. 'I ate I prepared dinner'); 8) The block type includes inappropriate breaks in different parts of the sentence in between or within words. In this study, all types of stuttering were considered, except the prolongation and block types because these events are better recognised by special HMM with explicit duration modeling and we used a standard HMM in current study. All stuttering types examined in the study are listed with their corresponding abbreviations in Table 1. The transcription methodology was reviewed by a speech language pathologist (SLP; co-author Brumfitt). Moreover, the complete transcribed text has been normalised. Text normalisation considered to be a prerequisite step for lots of downstream speech and language processing tasks. Text normalisation categorises text entities like dates, numbers, times and currency amounts, and transforms those entities into words.

3.3. Kaldi Acoustic Modeling

The Kaldi toolkit has standard recipes. In this study, the Wall Street Journal (WSJ) recipe was followed. The WSJ recipe began with training a monophone system that used standard 13-dimensional MFCCs. To reduce the channel effect, cepstral mean normalisation was applied. Then, using the information obtained from the monophone system, a triphone system was built using speaker-independent alignments. Next, a linear discriminant analysis (LDA) transformation was used to select the most discriminative dimensions from a large context. This included taking five frames to the left and five frames to the right. A more refined step was developed using a maximum likelihood linear transform (MLLT) on top of the LDA feature. Finally, the final GMM acoustic model was ready for use in offline decoding. All steps for developing a GMM acoustic model were applied to develop a GMM acoustic model that trained using the PF-Star corpus with the stuttering corpus obtained from UCLASS.

3.4. Language Model Augmentation and Pronunciation Dictionary

The SRILM n-gram LM toolkit [25] was used in this study. To create an LM properly, the text corpus and the word list for all the training data should be used. Then, an n-gram should be generated to be used as an input to Kaldi. For this work, trigram LMs using the SRILM tool were chosen to build our LMs for the experiments. In trigram LMs, for story text, for example, each word in the story is viewed as a different word according to its position. For instance, in 'the rats heard a great noise in the loft;' the two examples of 'the' are addressed as two different words. The trigram model forces the ASR system to recognise words in the order in which they occur in the story by placing the probability for a word on the word that follows it in the story. The first LM is used for the baseline experiment. The corpus in this LM essentially contains the text from the training data, which contains the text from training set of PF-star and the text from the training set of the UCLASS corpus. The other LM is augmented using the technique mentioned in Section 2 and used for same acoustic model that been used in the baseline experiment. The LM creation process also requires a pronunciation dictionary that consists of the phonetic sequence(s) for each word in the dictionary. The British English Example Pronunciation (BEEP) dictionary was used for this purpose [26]. For words that are not in the dictionary, such as the stuttered words, the Sequitur tool [27] was applied to estimate the phonetic sequences given the letters of the word. This provided an estimation of the pronunciation of the unavailable word. The pronunciation process for some stuttering vocabularies also had to be manually checked to ensure they are correct.

4. Experiments

4.1. Baseline Experiments

Baseline experiments were conducted to determine how the ASR behaved when trained on mostly fluent speech (from PF-Star) with only a small amount of stuttering data (from

	Ground Truth							ASR output						
Test Set	I	W	S	PW	PH	R	Total	I	W	S	PW	PH	R	Total
Set 1	3	4	8	0	1	2	18	1	4	3	0	0	2	10
Set 2	8	10	13	2	4	4	41	1	5	2	1	1	3	13
Set 3	3	2	0	0	7	3	15	0	0	0	0	3	1	4
Set 4	10	0	5	1	0	0	16	1	0	3	1	0	0	5
Set 5	10	3	5	0	1	2	21	2	0	2	0	0	2	6
Set 6	6	9	14	1	3	3	36	1	4	4	0	0	3	12
Set 7	3	5	1	0	1	0	10	0	3	1	0	1	0	5
Set 8	11	6	12	0	1	2	32	1	4	3	0	0	1	9
Set 9	2	2	4	0	1	5	14	0	1	1	0	0	4	6
Set 10	17	13	32	1	4	1	68	3	4	6	1	1	1	16
Set 11	17	7	27	3	4	3	61	3	6	8	1	3	1	22
Set 12	6	4	4	0	7	0	21	3	3	1	0	7	0	14

Table 2: *Baseline detailed results*

UCLASS). As mentioned before, the baseline ASR acoustic model was trained on the PF-star corpus with the UCLASS corpus. The UCLASS corpus transcriptions included stuttering events. The LM of the baseline ASR depended on the text corpora available for the training data. These initial experiments were performed using 12-fold cross-validation (CV) sets to verify the reliability of the model's performance. The 12-fold cross-validation (CV) sets were determined after dividing the 48 stuttering recordings from the UCLASS corpus by 4. Thus, each partition included four speech recordings, representing 10% of the complete stuttering data. Table 2 shows the CV results of the ASR baseline output regarding the detailed number of stuttering events detected compared to the ground truth stuttering events. Table 3 demonstrates the results of evaluating the performance of the baseline ASR. The results clearly suggest that the model's performance is very poor. For a diagnosis system that determines whether patients should start receiving a treatment, false negatives (FNs) are more important than false positive (FPs), which would fail to diagnose patients who genuinely require treatment. [28][29]. The current baseline model can only detect 38% of the total number of stuttering events, which is not sufficient for helping a therapist to diagnose a child who stutters. The poor performance of the ASR system for detecting stuttering events can be explained by the high perplexity of the LM, as the probability you are describing is of stuttering events in the combined corpus is small because the dominate of PF-Star. We also found in the baseline experiments that generated pronunciation in the pronunciation dictionary for 'some' stuttering events which considered as an OOV by the Sequitur tool is sometimes faulty which penalizes both the LM and WER. Then, we fix that manually in the pronunciation dictionary. Furthermore, the model reports 56.58% in precision, and almost half of the stuttering events produced in the ASR output transcription are FPs. In terms of the WER, the system performed well in most test sets, with an average of 19.82%. This is primarily because the task is a reading task and most of the words context already exists in the LM. However, the ASR reports a high WER in set numbers 10 and 11 due to the high number of stuttering events in these two test sets. The total numbers of stuttering events are 68 and 61 for set 10 and set 11, respectively.

Test Set	Recall	Precision	F-measure	WER
Set 1	55.5	77	64.5	6.5
Set 2	32	65	43	18.2
Set 3	26.6	40	32	10.4
Set 4	31.2	50	39	14.4
Set 5	28.5	54.5	37.5	13.2
Set 6	33.3	57.1	41	24.3
Set 7	50	71.4	59	6.8
Set 8	28	33.3	30.5	25.1
Set 9	42.8	60	50	7.8
Set 10	23.5	37.2	29	46.9
Set 11	36	63	46	42.9
Set 12	71.4	70.5	70.9	20.8
Average	**38%**	**56.58%**	**46%**	**19.82%**

Table 3: *ASR performance in baseline experiments*

4.2. Effect of LM Augmentation

After creating a LM with an augmented corpus, as explained in Section 2, we repeated the training with the augmented LM and analysed the impact on the ASR performance. Table 4 shows the performance obtained with the different test sets (CV) of the ASR output and illustrate the detailed number of stuttering events detected compared to the ground truth stuttering events.

Word/ phrase/ revision repetitions do not cause a serious issue in the ASR recognition performance, as the child mainly will repeat a word, phrase or revise a sentence using the same words that already exists in the LM of the ASR. Thus, the ASR is more likely to recognise them. Augmentation inserts more word repetition and revisions into the LM, so that these are more likely to be detected in recognition. For the part word repetition events, this event occurred only 8 times in all 12 sets as shown in table 2 and table 4 which represents only 2% of other stuttering events. All 8 times of PW occurred mostly in a compound words which were formed of two words that were put together such as 'anywhere' and 'anymore'. For example, the child who stutter is more likely to say 'any anymore' which considered as a PW event. Although, 50% of these events have been successfully recognised in baseline experiments, the recognition of this events couldn't be improved after augmentation process. The augmentation model not used to augment this type of stutter-

	Ground Truth							ASR output						
Test Set	I	W	S	PW	PH	R	Total	I	W	S	PW	PH	R	Total
Set 1	3	4	8	0	1	2	18	1	5	3	0	0	5	14
Set 2	8	10	13	2	4	4	41	0	7	6	1	3	4	21
Set 3	3	2	0	0	7	3	15	2	1	0	0	4	1	8
Set 4	10	0	5	1	0	0	16	8	0	4	1	0	0	13
Set 5	10	3	5	0	1	2	21	6	1	3	0	0	2	12
Set 6	6	9	14	1	3	3	36	4	6	6	0	3	3	22
Set 7	3	5	1	0	1	0	10	0	3	0	0	1	0	4
Set 8	11	6	12	0	1	2	32	6	4	4	0	0	2	16
Set 9	2	2	4	0	1	5	14	2	1	2	0	1	5	11
Set 10	17	13	32	1	4	1	68	13	8	10	0	3	1	35
Set 11	17	7	27	3	4	3	61	7	7	16	2	4	3	43
Set 12	6	4	4	0	7	0	21	4	4	3	0	6	0	18

Table 4: *AUG detailed results*

Test Set	**Recall**	**Precision**	**F-measure**	**WER**
Set 1	78	93.3	85	5.34
Set 2	51	64	56.8	13.23
Set 3	53.3	53.3	53.3	7.71
Set 4	72.2	62	66.7	10.66
Set 5	57	67	61.6	10.56
Set 6	61	81.5	69.8	19.25
Set 7	40	67	50	5.46
Set 8	50	55	53.4	21.62
Set 9	79	79	79	6.35
Set 10	51.5	62.5	56.5	40.86
Set 11	67.2	81.2	73.5	33.71
Set 12	86	82.6	84.3	16.89
Average	**62.2%**	**71%**	**66%**	**15.9%**

Table 5: *ASR performance after LM augmentation*

Test Set	**Recall**	**Precision**	**F-measure**	**WER**
Set 1	83.33	100	90.9	5.34
Set 2	66	82	73	13.23
Set 3	80	80	80	7.71
Set 4	72.2	62	66.7	10.66
Set 5	62	72	66.6	10.56
Set 6	64	85	73	19.25
Set 7	60	100	75	5.46
Set 8	68	74.41	71	21.62
Set 9	86	86	86	6.35
Set 10	70.5	88	78.3	40.86
Set 11	78	96	86	33.71
Set 12	86.4	82.6	84.5	16.89
Average	**73%**	**84%**	**77.6%**	**15.9%**

Table 6: *After coarsely reclassifying all stuttering events as belonging to either the SLD or other disfluency group*

ing due to the small amount of examples of PW events in the used corpus. The augmentation model mainly focused to enhanced the recognition of sound repetitions event and this is clearly achieved as the ASR can only recognise 34 repeated sound in the baseline which increased to 57 after augmentation process. Moreover, using this augmentation model will increase the probability of all other stuttering events in general as event such as interjection will be repeated many time in the corpus through augmentation process.

Table 5 shows the performance evaluation of the ASR after LM augmentation. As illustrated, there is a significant improvement in the results. Compared to the baseline presented in Table 3, we obtained an absolute gain of 24.2% in recall and 14.42% in precision. Moreover, the WER of the ASR is improved for all test sets (CV), with an average of 15.9%. One explanation might be that the correction of the pronunciation of stuttering events in the pronunciation dictionary positively affected both the LM and WER. Furthermore, the results prove that WER is relatively independent of the stuttering recognition task. As shown in Set 10 and Set 11 in table 5, the WER is high comparing to other testing sets, however, the recall percentage is higher than Set 7 which is 5% WER. These results indicate that the performance of the ASR for stuttering recognition mainly depend on the type of stuttering events not depend on the WER. One observation during the analysis of the results was that many FP events were reported in the same utterance containing real stuttering events (TP). The ASR misrecognised the correct stuttering event (TP) and produced another close stuttering, but it was considered an FP event. For example, in the ground truth transcription, there is a sound repetition in 'fo for', and the pronunciation of for in the pronunciation dictionary of the ASR could be 'f ao r' or 'f ao'. Thus, the ASR recognised this case as 'for for', which is word repetition and counted as (FP). According to Yairi and Ambrose [23], stuttering disfluency types are divided into two groups: 'stuttering-like disfluencies' (SLDs), which include part-word repetition, sound repetition, word repetition and prolongation and other disfluencies, which include interjection, revision and phrase repetition. Normally, during the analysis and stuttering counts, clinicians mark all stuttering events on the transcript and add the numbers of the types that belong to SLDs and add those belonging to 'other disfluencies'.

Then, they add up totals withen each categories. The Yairi and Ambrose [23] study leads to the conclusion that there is no effective difference between types of stuttering event within the SLD group, when these are aggregated for diagnosis purposes. For example, the miscount sound repetition as a word repetition has no real effect on the quality of the final decision, as they are in the same disfluency group (SLDs) and they will sum up at the end. The reclassification happens to treat some FPs from the earlier experiments as TPs in the new experiments, which reflects a coarser classification, relevant to diagnosis. Table 6 presents the results. The recall improved by 10% while precision improved by 13%.

5. Conclusion

In this work, we studied how speech technology could help clinicians in speeding up the diagnosing process for children who stutter. We built an ASR with augmented LM to recognise stuttering in audio files recorded of children reading stories, to obtain orthographic transcriptions of what was said. Our method leads to better LM, with which lower WERs are obtained and a greater ability to recognise stuttering events. The results show that the performance of the ASR on detecting stuttering improved to reach 73% in recall and 84% in precision. We can make further improvements to this model, which must start by detecting prolongation and blocking events. However, even with this current level of accuracy, we may able to make some judgment as to how reliable the current ASR would be for the automated diagnosis of stuttering. We believe that even with this level of recall (73%), we can map this to fairly robust reasonable stutter/non-stutter decision boundaries.

6. Acknowledgements

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

7. References

[1] 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.

[2] 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.

[3] A. Craig, E. Blumgart, and Y. Tran, "The impact of stuttering on the quality of life in adults who stutter," *Journal of fluency disorders*, vol. 34, no. 2, pp. 61–71, 2009.

[4] A. Craig and Y. Tran, "Fear of speaking: chronic anxiety and stammering," *Advances in Psychiatric Treatment*, vol. 12, no. 1, pp. 63–68, 2006.

[5] J. F. Klein and S. B. Hood, "The impact of stuttering on employment opportunities and job performance," *Journal of fluency disorders*, vol. 29, no. 4, pp. 255–273, 2004.

[6] E. Blumgart, Y. Tran, and A. Craig, "Social anxiety disorder in adults who stutter," *Depression and Anxiety*, vol. 27, no. 7, pp. 687–692, 2010.

[7] M. Jones, M. Onslow, A. Packman, S. Williams, T. Ormond, I. Schwarz, and V. Gebski, "Randomised controlled trial of the lidcombe programme of early stuttering intervention," *BMJ*, vol. 331, no. 7518, p. 659, 2005. [Online]. Available: http://www.bmj.com/content/331/7518/659

[8] 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.

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

[10] 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.

[11] 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.

[12] Y. Geetha, K. Pratibha, R. Ashok, and S. K. Ravindra, "Classification of childhood disfluencies using neural networks," *Journal of fluency disorders*, vol. 25, no. 2, pp. 99–117, 2000.

[13] P. Mahesha and D. Vinod, "Using orthographic transcripts for stuttering dysfluency recognition and severity estimation," in *Intelligent Computing, Communication and Devices*. Springer, 2015, pp. 613–621.

[14] 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.

[15] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, and T. Wittenberg, "Automatic stuttering recognition using hidden markov models," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] 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.

[17] M. Federico, R. De Mori, and K. Ponting, "Language model adaptation," 1999.

[18] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4929–4932.

[19] P. Howell and L. Vause, "Acoustic analysis and perception of vowels in stuttered speech," *The Journal of the Acoustical Society of America*, vol. 79, no. 5, pp. 1571–1579, 1986.

[20] J. G. Sheehan, "Stuttering behavior: A phonetic analysis," *Journal of Communication Disorders*, vol. 7, no. 3, pp. 193–212, 1974.

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

[22] 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.

[23] E. Yairi and N. Ambrose, *Early Childhood Stuttering for Clinicians by Clinicians*, ser. For clinicians by clinicians. PRO-ED, 2005. [Online]. Available: https://books.google.co.uk/books?id=9H8rAQAAMAAJ

[24] F. Staróbole Juste and C. R. Furquim de Andrade, "Speech disfluency types of fluent and stuttering individuals: age effects," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 2, pp. 57–64, 2010.

[25] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *Interspeech*, vol. 2002, 2002, p. 2002.

[26] A. Robinson, "The british english example pronunciation (beep) dictionary," *Retrieved from World Wide Web: ftp://svrftp. eng. cam. ac. uk/pub/comp. speech/dictionaries/beep. tar. gz*, 1996.

[27] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[28] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Classification evaluation," *Nature Methods*, vol. 13, no. 8, pp. 603–604, 2016.

[29] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.