EXPLORING APPROPRIATE ACOUSTIC AND LANGUAGE MODELLING CHOICES FOR CONTINUOUS DYSARTHRIC SPEECH RECOGNITION

Zhengjun Yue[⋆] *Feifei Xiong*[⋆] *Heidi Christensen*[⋆†] *and Jon Barker*[⋆]

[⋆] Speech and Hearing Group (SPandH), Dept. of Computer Science, University of Sheffield, UK
[†] Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, UK

ABSTRACT

There has been much recent interest in building continuous speech recognition systems for people with severe speech impairments, e.g., dysarthria. However, the datasets that are commonly used are typically designed for tasks other than ASR development, or they contain only isolated words. As such, they contain much overlap in the prompts read by the speakers. Previous ASR evaluations have often neglected this, using language models (LMs) trained on non-disjoint training and test data, potentially producing unrealistically optimistic results. In this paper, we investigate the impact of LM design using the widely used TORGO database. We combine state-of-the-art acoustic models with LMs trained with data originating from LibriSpeech. Using LMs with varying vocabulary size, we examine the trade-off between the out-of-vocabulary rate and recognition confusions for speakers with varying degrees of dysarthria. It is found that the optimal LM complexity is highly speaker dependent, highlighting the need to design speaker-dependent LMs alongside speaker-dependent acoustic models when considering atypical speech.

Index Terms— Continuous dysarthric speech recognition, language modelling, out-of-domain data

1. INTRODUCTION

Dysarthria is a speech disorder caused by a disruption in the neuromotor interface [1] which impedes the physical production of speech. People with moderate to severe dysarthria are often only intelligible to close friends and family, and in general, communicating with others can be very challenging. This extends to communicating with machines, and although some progress has been made in recent years, dysarthric automatic speech recognition (ASR) remains a challenging research area that is lagging decades behind the sustained progress seen for mainstream automatic ASR tailored for *typical* voices. The large systematic differences between typical and dysarthric speech, coupled with the high degree of variability between speakers with dysarthria (depending on severity and type) means that large resources are required for training adequate acoustic models. However, very few dysarthric speech datasets are available. Further, the databases that do exist, have usually not been collected for the purpose of training ASR systems but instead for purposes such as diagnosis and impairment severity assessment. This means that researchers are faced with challenging choices when attempting to set up experimental frameworks aimed at facilitating meaningful research on improving continuous dysarthric speech recognition.

Early work on dysarthric speech has focused on isolated word recognition using the highly influential UASpeech corpus [2]. More recently, focus has moved on to continuous speech recognition. Here, the two widely available datasets are Nemours [3] and TORGO [4]. Research has focused on improving the acoustic modelling to better handle the mismatch to typical speech, e.g., the use of adaptation techniques to generate speaker dependent models using limited amounts of data [5], and demonstrating the benefit of adding articulatory information to improve traditional acoustic modeling of dysarthric speech [6]. Recently, further ASR performance improvements have been made by exploring neural network architectures such as DNNs, CNNs, TDNNs and LSTMs [7, 8].

However, unlike UASpeech and the smaller homeService corpus (recorded isolated command word interactions, [9]), neither Nemours nor TORGO were designed for ASR research. Nemours was motivated by intelligibility assessment and TORGO for comparative study of dysarthric and typical speech. As such, although they may allow a partitioning of speakers into training and testing sets, they do not provide a disjoint set of training and test sentences. For example TORGO features a lot of repeated prompts. This is sensible for assessment or across speaker comparisons, but not convenient for ASR. In fact, the standard approach of using a leave-one-speaker-out cross-validation setup with this dataset has encouraged previous researchers to train language models on training sets that is almost completely overlapping with the test set.

Working with TORGO, [4], [5] and [6] employed a back-off bigram LM while [7] and [8] applied a standard trigram LM with interpolated Kneser-Ney discounting [10] to the training data prompts – despite the overlap with the test data. Studies on the phrase-based dysarthria corpus Nemours [3] such as [11] simply utilized the bigram statistical LM trained on the whole corpus itself. [12] employed external text from the TIMIT dataset for LM training, but without providing any details of the LM setups. We believe that many of the reported ASR performances have been achieved with LMs unfairly biases towards the language specifics of that corpus (both the train and test part) and hence will have been overly optimistic, and less able to generalise to truly unseen utterances.

In this work, we aim to develop a reproducible benchmark for state-of-the-art continuous speech ASR using open tools and fairly designed language models. We re-evaluate the state-of-the-art for TORGO, building LMs over a range of vocabulary sizes for different utterance types by introducing the external, out-of-domain LibriSpeech corpus [13] as the source for LM estimation. Then, combined with state-of-the-art acoustic models (AMs), we analyze the influence of the LM vocabulary size on speakers with varying severity of dysarthria to find the trade-off between acoustic and language modelling constraints.

This work is supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for PAthological Speech processing; Grant Agreement No. 766287).

	Severe				**M/S**	**Moderate**	**Mild**	
	F01	M01	M02	M04	M05	F03	F04	M03
Number of utterances in the training set	16158	15647	15620	15735	15814	15314	15719	15586
Number of utterances in the test set	228	739	766	651	572	1072	667	800
% Prompt overlap between train and test set	100%	99.1%	98.2%	98.2%	98.9%	95.7%	98.6%	99.7%

Table 1: TORGO dataset statistics per (F)emale and (M)ale speaker. 'M/S': moderate to severe intelligibility.

2. CHALLENGES OF USING DYSARTHRIC SPEECH CORPORA FOR CONTINUOUS SPEECH RECOGNITION

Two corpora are commonly used for continuous dysarthric ASR: Nemours [3] and TORGO [4]. The Nemours dataset consists of 74 sentences spoken by each of 11 male speakers with different severity of dysarthria and 11 male typical speakers. The repetition in the dataset and the small amount of utterances make the Nemours not suitable for the ASR task. The TORGO dataset contains 21 hours of aligned acoustic and articulatory recordings collected from 15 speakers [4]. Eight of the speakers (5 males, 3 females) have different degrees of dysarthria, while the other seven are non-dysarthric typical speakers (4 males, 3 females). Compared with other existing American/Canadian English dysarthria datasets including Nemours [3] (continuous sentences) and UAspeech [2] (isolated words only), TORGO comprises both word and sentence prompts: 615 unique words and 354 unique sentences. The total vocabulary size is 1573, of which the vocabulary size for the sentence prompts on their own is 1083. Together with the articulatory recordings, this makes this dataset particularly interesting.

TORGO does not come with a pre-defined training and test partition. Instead researchers have used the leave-one-speaker-out approach to maximise the available training data. There is a large overlap between any given speaker's utterances (in response to word and sentence prompts) and those seen in their training set (provided by the remaining 14 speakers) as all speakers have had the same prompts and contributed very similar utterances. Table 1 summarises the number of utterances in the leave-one-speaker-out TORGO training and test sets per speaker (after excluding the recordings that are shorter than 25 ms and any wrongly annotated audio). Although the corpus contains from 15,314 to 16,158 recorded utterances per speaker, only a fraction of these (between 951 and 969) are in fact unique, indicating the high degree of repetition within and across speakers. The extremely high number of overlapping prompts between training and test sets means that any LM trained on any speaker's corresponding training part of the dataset will be highly tuned to the test set.

When setting up an evaluation framework within which to explore e.g., acoustic model improvements, it is essential that the chosen LM reflects a realistic scenario as best as possible. We propose to use out-of-domain (OOD) data to train LMs with a higher perplexity to allow for a more reasonable decoding space (in terms of WER). Note, this will evidently result in a worse baseline performance than previously assumed, but one which is more meaningful in terms of evaluating success of acoustic modelling strategies in general, not just fitting the (non-ASR) database available for research.

3. LANGUAGE MODELLING

Language models impose a syntactic and semantic constraint on the ASR decoding process by assigning probabilistic estimates for the occurrence of short word sequences ('n-grams') [14]. The LM is represented as a prior probability in the computation of the posterior estimates, which is typically trained using large amounts of natural language text data [15]. When it comes to low resource data, care has to be taken to not unfairly design the LM so as to give over-optimistic results by training it on within-corpora data.

To explore the effect of using different LMs we first evaluate the ASR system using the LM used in previous TORGO-based studies [7, 8] (from hereon referred to as TORGO LM). This LM covers both the word and the sentence prompts, but in order to assess the WER for each of those two separate ASR tasks separately, we further train *task-specific* TORGO LMs for the word and sentence recognition tasks separately, to see how these two distinct tasks are affected by the choice of LM. Finally, we build out-of-domain LMs originating from LibriSpeech to explore the optimal complexity of the language model.

The acoustic models (AMs) used in our experiments are GMM-HMMs and DNN-HMMs trained with both dysarthric and typical speech [5, 7, 8]. In particular, the GMM-HMM employs a triphone model with speaker adapted transformation, and the DNN-HMM uses a DNN network trained using cross-entropy following setups from [7]. The experiments are conducted in Kaldi [16] using the SRILM [17] toolkit for language modelling. Specific training details for the LMs are presented in Sections 3.1 - 3.3. Leave-one-speaker-out cross validation is employed to perform speaker-independent speech recognition, i.e., for each split, 14 speakers are used for training and testing is performed on a single held out speaker.

3.1. TORGO LM

The TORGO LM is reproduced from [7, 8]), which is a trigram LM built on prompts of the training stage data in TORGO. In addition to testing the whole test set, we also inspect the ASR performance by prompt type, i.e., considering word and sentence recognition as two different tasks.

Results summarised in Table 2, show that both GMM-HMM and DNN-HMM systems give a much better performance on sentence than on word tasks for each severity level, with a 26.0% higher performance on average. We believe that this is due to the strong LM. The DNN provides varied benefit across the tasks in comparison to the GMM-based AM. It is noticeable that for severely dysarthric speech, although the DNN decreases the overall WER on the full test set by 12%, when the results are reported per task (last two rows of Table 2), it is evident that this overall decrease is the result of a (modest) increase in the word task (2.2%), and a large decrease for the sentence task (13.7%). Overall, the results show that reporting task-specific results gives a more nuanced picture of the performance (and the confluence between the AM and LM), and that the severity level further affects how the AM and LM interact.

3.2. Task-specific TORGO LMs

The task-specific TORGO LM for the isolated word utterances is built as a standard unigram LM (TORGO unigram LM), whereas

TORGO LM						
	Severe		Moderate		Mild	
Task	GMM	DNN	GMM	DNN	GMM	DNN
Full Test set	69.6	**57.6**	35.9	**33.0**	15.1	**14.3**
Isolated words	**79.8**	82.0	66.3	**65.5**	22.4	**19.5**
Sentences	62.0	**48.3**	23.3	**22.7**	**11.2**	12.2

Table 2: ASR performance [WER] using different AMs and the TORGO LM for full, isolated words and sentences tasks and averaged for speakers with different dysarthria severity.

the sentence specific LM is a trigram model (TORGO trigram LM). The TORGO unigram LM is constructed on the 615 unique isolated words utterances in TORGO, and is a uniform word grammar network where all words in the corpora are in parallel and assigned the same log probability following similar setups for .e.g., the isolated word tasks in UASpeech [18].

The TORGO trigram LM is built on 313 to 354 (depending on speaker) unique training sentence prompts as defined by the speaker-specific TORGO training data split, applying Witten-Bell discounting [19]. The two LMs are evaluated on word and sentence utterances respectively as two specific tasks.

Task-specific TORGO LMs						
	Severe		Moderate		Mild	
LM	GMM	DNN	GMM	DNN	GMM	DNN
TORGO unigram LM	**61.5**	62.8	54.9	**48.2**	19.2	**15.9**
TORGO trigram LM	59.7	**41.8**	16.0	**12.8**	3.1	**2.0**

Table 3: ASR performance [WER] using different AMs and the task-specific TORGO LMs for isolated words (TORGO unigram LM) and sentences (TORGO trigram LM) tasks and averaged for speakers with different dysarthria severity.

Table 3 shows the results of testing with the task-specific LMs, and Figure 1 compares the results in Table 3 with Table 2 (the second and the third rows) and draws the improvement lines for both tasks. Not surprisingly, both of the two task-specific TORGO LMs give better results than the general TORGO LM evaluated on the corresponding utterance type subset. It is seen that as the dysarthric severity increases, the improvement made by the TORGO unigram LM on the isolated word task increases, while the opposite is the case for the TORGO trigram LM for the sentence task. The consistent improvement across speakers on the words performance is caused by the constraint made by the unigram LM, which forces the ASR system to output a single word. It can also eliminated some of the insertion errors caused by the slow speaking rate characterised by the moderate and severe group. The sentences performance of the mild speakers drops from 12.2% to an extremely optimistic value (2.0%), indicating that the constraint (e.g., reduction of training corpus) makes the trigram LMs stronger to result in overly optimistic evaluation.

3.3. Out-of-domain LibriSpeech LMs

To measure the impact of the biases introduced by the TORGO LMs, we compare the ASR performances to those obtained with LMs built from non-TORGO texts. For this purpose we introduce the LibriSpeech corpus[13] as the out-of-domain text corpus in our experiments. It is a read speech dataset based on LibriVox's audio books, containing 1000 hours of speech and around 803 million tokens from 14,500 public domain books used for LM training. A vocabulary size of the 200,000 most frequent words is selected to be in the lexicon.

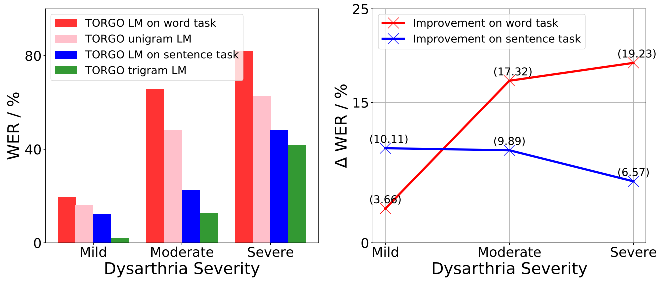


Fig. 1: Comparing task-specific and full TORGO LMs.

By gradually extending the vocabulary originating from LibriSpeech in line with the decreasing order of word frequency, we can build LibriSpeech unigram LMs (for the isolated word task) over a range of vocabulary sizes: {2k, 5k, 10k, 15k, 20k, 25k, 30k, 35k, 40k, 45k, 50k, 100k, 150k, 200k}. Likewise, a series of LibriSpeech trigram LMs is built for the sentence task by pruning the pre-trained official 3-gram LibriSpeech LM using the CHANGE-LM-VOCAB method in SRILM toolkit [17], which modifies the LM size by limiting the vocabulary to variously sizes subsets. The OOV words are converted to <UNK> tag in the unigram while any N-grams containing OOV words are removed, then the model is re-normalized. To achieve comparable OOV rates for the smaller vocabulary sized LM for both the word and the sentence tasks, vocabulary sizes starting from 0.1k are introduced for the sentence task.

4. ANALYSIS OF LIBRISPEECH LANGUAGE MODELS

In this section we further analyse the results obtained using the out-of-domain LibriSpeech LMs. In addition to the standard WER we also report the out-of-vocabulary rate (OOV rate), together with the recognition confusion to explore the LM complexity for different severity levels of dysarthria. Specifically, the recognition confusion measures how much confusion the system experiences when attempting to recognise words it is aware of, i.e., the in-vocabulary words. The recognition confusion, x can be calculated as follows

$$x = \frac{i - c}{i} = 1 - \frac{c}{i} \tag{1}$$

where c denotes the number of correctly recognized words, and i is the number of in-vocabulary words.

Figure 2a and 2b shows the WER, OOV rate and recognition confusion for the range of vocabulary sizes. It allows us to compare the relationship between the impaired speech severity and complexity of the LibriSpeech LMs, by means of WER, OOV rate. The colored circles on each line denotes the lowest WER of the LM with certain vocabulary size. Table 4 shows the results from these selected vocabulary sizes (indicated by 'optimal vocab size') for the LibriSpeech LM[1]. Comparing the results in Table 4 with the TORGO LM (Table 2) for the DNN AM on the isolated word task, the LibriSpeech unigram LM showed improvements across speakers with moderate and severe dysarthria. This might be because it reduces a large number of insertion errors, resulting from the slow speaking rate, by constraining the output to be a single word. However,

[1] We use the lowest WER instead of the 'knee' of each WER line for results comparison.

LibriSpeech unigram LMs; isolated word task						
	Severe		Moderate		Mild	
Measurements	GMM	DNN	GMM	DNN	GMM	DNN
The lowest WER (%)	84.5	**80.2**	66.4	**64.5**	34.5	**27.0**
Optimal vocab size	5k	15k	30k	30k	50k	50k

LibriSpeech trigram LMs; sentences task						
	Severe		Moderate		Mild	
Measurements	GMM	DNN	GMM	DNN	GMM	DNN
The lowest WER (%)	92.3	**86.4**	67.3	**65.6**	**36.4**	38.4
Optimal vocab size	100k	20k	50k	200k	150k	150k

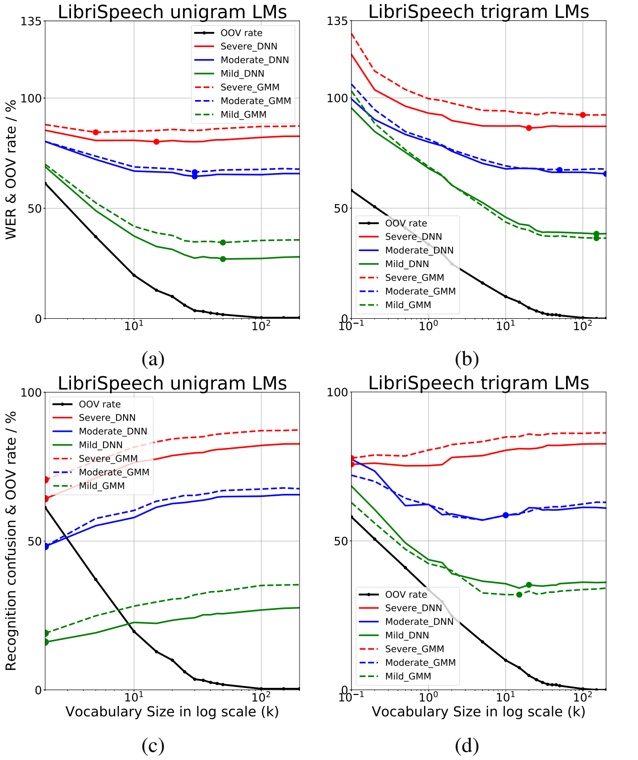
Table 4: ASR performance [WER] using different AMs and the task-specific LibriSpeech LMs for isolated words (LibriSpeech unigram LM) and sentences (LibriSpeech trigram LM) tasks and averaged for speakers with different dysarthria severity.

for mildly impaired speakers, since their speaking rate is similar to the typical speakers, although the LibriSpeech unigram LM constrains the output to make the task easier, it still degrades the performance due to the reduced complexity. Comparing the sentence performances in Table 4 (86.4, 65.6 and 38.4% WER) and those with previous TORGO LM (the last row of Table 2 (48.3, 22.7 and 12.2% WER)), the WER obtained by LibriSpeech trigram LMs are on average 40.5% worse for moderate and severe speakers and even 26.2% for mild speakers. In contrast to the unrealistically small WERs of the TORGO LM, these results present a fairer evaluation.

It is seen that in general speakers with different severity of dysarthria require the LibriSpeech LMs with different vocabulary sizes: the greater the severity of the dysarthria, the smaller the optimal vocabulary size. To explain the possible reasons, we plot the recognition confusion rate across speakers with different degrees of dysarthria in Figure 2c and 2d. We found that at more severe levels there is more confusability in the speech, therefore reducing the vocabulary size reduces the chance of poorly pronounced common words being mistaken for low frequency words that might be better acoustic matches. Typically, for the word recognition task, as the vocabulary size increases, the confusion sees a monotonic increase across all the speakers. While in the sentence recognition task, the confusion rates reach the minimum point with 0.1k, 10k and 20k vocabulary sizes individually for speakers with severe, moderate and mild dysarthria. This might be because that the continually reducing OOV rate, and the increasing number of utterances available, offsets the extra confusions (i.e., some of the extended words are in a recognizable range to reduce some substitution errors caused by OOV words). The greater the severity of dysarthria, the less compensate is made by the decreasing OOV rate. Comparing different AMs, when further increasing the vocabulary size after the optimal vocabulary sizes required by the LibriSpeech LMs, the recognition confusion of the GMM systems will increase more than that of the DNN models.

5. CONCLUSIONS

Very few datasets exists that allows researchers to develop speaker-specific continuous speech recognition systems for people with dysarthria, and they are mostly not designed for ASR, meaning great care has to be taken to choose an appropriate experimental setup. Working on TORGO, this paper presented an in-depth analysis comparing LMs trained on TORGO text prompts with LMs trained on varying vocabulary-sized subsets of LibriSpeech. We found that TORGO LMs (used widely in literature) give a hugely overestimated performance of dysarthric ASR because of prompt overlap between training and test parts. In comparison, the LibriSpeech models gives a lower — but we believe — fairer performance which will better allow for a more reasonable decoding space (in terms of WER). We also found that reporting results on individual tasks (isolated word vs sentence) enabled a more nuanced view of AM performance.



(a)

(b)

(c)

(d)

Fig. 2: WER, recognition confusion and OOV rate for LibriSpeech LMs for speakers with different dysarthria severity.

Exploring different vocabulary sizes for the LibriSpeech LMs, we found that for the most severe cases, performance levels off at about 1000 words. However, in general, the lowest WERs are achieved with the largest vocabulary size, i.e., the continually reducing OOV rate, and the increasing number of utterances available, offsets the extra confusions. In real applications, speaker-specific LMs may be appropriate as, depending on severity and *when not asked to read prompts*, speakers would choose to use different language constructs and words to counteract specific speech impairments.

We believe the results here represent a fair benchmark for the current state-of-the-art for dysarthric read speech ASR. Our results are fully reproducible and Kaldi recipes are available at `https://github.com/zhengjunyue/CADSR-LM`. TORGO remains the best database for exploring continuous dysarthric ASR. Future work will investigate state-of-the-art AMs and end-to-end approaches within this new framework, as well as the free-text recognition task, also available in TORGO.

6. REFERENCES

[1] WR Gowers, "Clinical speech syndromes of the motor systems," *Webb WG, Adler RK, Love RL. Neurology for the Speech-Language Pathologist. Fifth edition. Philadelphia: Butter worth_ Heinemenn*, pp. 196–203, 2001.

[2] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[3] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell, "The nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 3, pp. 1962–1965.

[4] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[5] Kinfe Tadesse Mengistu and Frank Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4924–4927.

[6] Frank Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[7] Cristina Espana-Bonet and José AR Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 97–107.

[8] Neethu Mariam Joy and S Umesh, "Improving acoustic models in torgo dysarthric speech database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 637–645, 2018.

[9] Mauro Nicolao, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain, "A framework for collecting realistic recordings of dysarthric speech-the homeservice corpus," in *Proceedings of LREC 2016*. European Language Resources Association, 2016.

[10] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn, "Scalable modified kneser-ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, vol. 2, pp. 690–696.

[11] Mumtaz Begum Mustafa, Siti Salwah Salim, Noraini Mohamed, Bassam Al-Qatab, and Chng Eng Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PloS one*, vol. 9, no. 1, pp. e86285, 2014.

[12] Myungjong Kim, Beiming Cao, Kwanghoon An, and Jun Wang, "Dysarthric speech recognition using convolutional lstm neural network," *Proc. Interspeech 2018*, pp. 2948–2952, 2018.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] Siddharth Sehgal, Stuart Cunningham, and Phil Green, "Phase-based feature representations for improving recognition of dysarthric speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 13–20.

[15] Jun'ichi Tsujii, "Computational linguistics and natural language processing," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2011, pp. 52–67.

[16] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," Tech. Rep., IEEE Signal Processing Society, 2011.

[17] Andreas Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[18] Heidi Christensen, MB Aniol, Peter Bell, Phil D Green, Thomas Hain, Simon King, and Pawel Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech.," in *INTERSPEECH*, 2013, pp. 3642–3645.

[19] Stanley F Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.