Impact of ASR Performance on Spoken Grammatical Error Detection

Y. Lu, M.J.F. Gales, K.M. Knill, P. Manakul, L. Wang, Y. Wang

ALTA Institute / Engineering Department
Cambridge University, UK
{ytl28,mjfg,kate.knill,pm574,lw519,yw396}@eng.cam.ac.uk

Abstract

Computer assisted language learning (CALL) systems aid learners to monitor their progress by providing scoring and feedback on language assessment tasks. Free speaking tests allow assessment of what a learner has said, as well as how they said it. For these tasks, Automatic Speech Recognition (ASR) is required to generate transcriptions of a candidate's responses, the quality of these transcriptions is crucial to provide reliable feedback in downstream processes. This paper considers the impact of ASR performance on Grammatical Error Detection (GED) for free speaking tasks, as an example of providing feedback on a learner's use of English. The performance of an advanced deep-learning based GED system, initially trained on written corpora, is used to evaluate the influence of ASR errors. One consequence of these errors is that grammatical errors can result from incorrect transcriptions as well as learner errors, this may yield confusing feedback. To mitigate the effect of these errors, and reduce erroneous feedback, ASR confidence scores are incorporated into the GED system. By additionally adapting the written text GED system to the speech domain, using ASR transcriptions, significant gains in performance can be achieved. Analysis of the GED performance for different grammatical error types and across grade is also presented.

Index Terms: speech recognition, grammatical error detection

1. Introduction

Over 1.5 billion people worldwide are expected to be using and learning English as an additional language by 2020 [1]. Computer assisted language learning (CALL) systems are essential to support this level of interest, allowing learners to check their progress and identify areas they need to improve. Free speaking tasks, where a candidate speaks for 15-60 seconds in response to a series of prompts, are preferred to assess speaking ability. A key spoken communication skill is the use of English, putting words together using "grammar" that is consistent with a native speaker. Even native speakers don't tend to follow all grammatical rules when free speaking. There are, however, phrases and word sequences that a native speaker is highly unlikely to say but a learner might in error. Detection of and feedback of these "grammatical errors" can therefore help learning.

Attempts have been made to provide feedback on spoken learner English, each with coverage constraints. In [2] a rule-based grammatical error collection system was proposed which is tailored to a subset of error types. For non-spontaneous speech, [3] proposed a grammatical error detection (GED) system comparing a candidate's answer to a matching reference focused on a question-answering style test. By contrast, in recent years significant developments have been achieved in detecting the full range of grammatical errors in written text through deep learning methods [4]. It is therefore interesting to adapt these text GED systems and apply advanced deep learning based techniques to developing spoken GED on free speech.

In a CALL system, the candidate response to free speaking tasks is not known to the system. The spoken GED, therefore, has to run on the output of an automatic speech recognition (ASR) system. It was shown in [5] and [6] that the word error rate (WER) of words corresponding to grammatical errors is higher than fluent, grammatical speech for speakers of the same proficiency level. This poses a greater challenge of providing reliable feedback to spoken learner English as it would be confusing to the learner to provide feedback on an error that resulted from incorrect transcriptions rather than learner errors.

To reduce confusion, it is important to reduce false positives arising from errors in ASR transcriptions. One way of mitigating this problem is to reduce errors in transcriptions through improving ASR systems. Reducing the WER of an ASR system has been shown to help improve language assessment [7, 8, 5]. Another possible approach is to make use of a richer set of features derived from the ASR output; in particular, ASR confidence scores can be incorporated into the GED system as an additional condition to help determine whether or not to give feedback to candidates. This paper compares three ASR systems with decreasing WER and analyses the impact of ASR performance on the spoken GED system. The effect of incorporating confidence scores into GED system is evaluated. Analysis of the GED performance for different grammatical error types and across proficiency levels is also presented.

2. Grammatical Error Detection

In this work, a bidirectional LSTM based model[1] [9] was used for the grammatical error detection (GED) system. GED is modelled as a sequence labelling task [6]. For an input word sequence $\mathbf{w} = \{w_1, \ldots, w_N\}$, a reference label y_i (1:incorrect; 0:correct) is given to each word w_i. The probability distribution of $\hat{y}_i$ over the two labels is the target to be predicted. The training objective function is the log-likelihood of $\hat{y}_i$ summing over all L sentences and all $N^{(r)}$ words in each sentence:

$$\sum_r \sum_i \log(P(y_i^{(r)}|\mathbf{w}^{(r)})); \quad r = 1, ...L; i = 1, ...N^{(r)} \quad (1)$$

Written text trained GED system can be extended to handle ASR transcriptions by making use of ASR confidence scores. When the ASR system generates an incorrect transcription, it would be confusing for GED to give feedback on that word as the true candidate response is lost in the transcription. Words with low confidence scores are more likely to be ASR errors, therefore for the GED system to give feedback to the candidate indicating a grammatical error, two criteria are to be met: the

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for support and access to the BULATS data.

[1] https://github.com/marekrei/sequence-labeler

ASR confidence score passes a threshold α and the GED system predicts that the probability of a grammatical error occurring is over β:

$$\hat{y}_i^* = \begin{cases} 1; \ P(w_i^*|\mathbf{x}^*) > \alpha \text{ and } P(\hat{y}_i^* = 1|\mathbf{w}^*) > \beta \\ \emptyset; \text{ otherwise} \end{cases}$$

where $P(w_i^*|\mathbf{x}^*)$ is the ASR confidence that word w_i^* was correctly recognised given input audio sequence $\mathbf{x}^*$. Based on this feedback rules, the objective function is modified by reducing the words considered for utterrance r in the cross entropy loss to:

$$i : P(w_i^{(r)}|\mathbf{x}^{(r)}) > \alpha$$

i.e. when the ASR confidence is below α, the GED system is not penalised regardless of the prediction made on this word.

3. ASR

Three automatic speech recognition (ASR) systems are considered, all having the same vocabulary. Each is a hybrid deep learning-HMM graphemic system. The acoustic models (AM) are trained on non-native learner English speech recorded on business English tests (BULATS [10]). The first system, ASR1, is a speaker-independent joint decoding of a stacked hybrid DNN and LSTM system that was used in [5, 6, 11]. Both DNN and LSTM are trained on 39-dimensional bottleneck (BN) features. The output targets of the neural networks are global state-position tri-grapheme targets generated by a set of graphemic PLP GMM-HMMs trained on the same data. The models are trained on combined crowd-sourced transcriptions [12]. An in-domain trigram language model (LM) is trained on the same transcriptions and then interpolated with a general pre-trained Broadcast News English [13] LM. The transcriptions from ASR1 are decoded using this interpolated trigram.

The second system, ASR2, is a speaker adaptive sequence teacher-student (TS) trained lattice-free maximum mutual information (LF-MMI) factorised time-delay neural network (TDNN-F) system [14, 15, 16, 11]. Three TDNN-F models are trained with different random initialisations to form a random ensemble. They are trained on 40-dimensional filterbank features with 100-dimensional i-vectors. Lightly-supervised transcriptions produced by ASR1 are used for training [11]. A single student TDNN-F model with the same input is trained by minimising the Kullback-Leibler (KL) divergence between the sequence-level posteriors produced by the student TDNN-F model and the combined posteriors from the teacher ensemble. This TS trained model is decoded with a trigram LM to produce the ASR2 transcriptions. The third system, ASR3, is obtained by rescoring the lattices generated by ASR2 using a succeeding word recurrent neural network LM (su-RNNLM) [17]. Table 1 summarises the 3 systems with the amount of training data used. Note that the su-RNNLM is trained on a semi-supervised set by augmenting the supervised training set with a unsupervised set that is decoded by ASR1.

Table 1: *ASR systems compared.*

System	AM		BULATS LM	
	Model	Tng (hrs)	Model	Tng (wds)
ASR1	DNN+LSTM	330	Trigram	1.8M
ASR2	TDNN-F TS	500	Trigram	2.6M
ASR3			su-RNNLM	25.6M

4. Experimental results

4.1. Data and setup

Due to the lack of annotated spoken learner corpora, GED training was initially conducted on the written Cambridge Learner Corpus (CLC) following previous work in [6, 9] then fine-tuned to various speech domain corpora using both manual and ASR transcriptions. Grammatical errors are carefully annotated in the CLC data, which consists of text responses to written examinations targeting candidates at different proficiency levels. The same train/dev/test split defined in [6] was used. To adapt the written corpora to be closer to speech transcriptions, spelling mistakes, punctuation and capitalisation were removed. The FCE-public dataset [18] is a subset of the CLC corpus, thus falling under the same domain as CLC. It was used for evaluation in the fine-tuning experiments as a sanity check.

Following [6], one public and one proprietary spoken corpora were used to evaluate spoken GED performance. The publicly available NICT Japanese Learner English (JLE) Corpus [19] provides manual transcriptions of a English speaking test involving candidates at A1-B2 levels on the CEFR scale [20]. JLE is labeled with meta-data and grammatical errors [19], but no audio information is available. The other test corpus came from the free speaking parts of the spoken BULATS test [21]. Manual transcriptions are annotated with meta-data, speech units and grammatical errors [22]. ASR transcriptions were generated using the ASR systems discussed above.

Speech transcriptions are fundamentally different from written text in two main aspects. Firstly, disfluencies, such as false starts and repetitions, only exist in spoken language. They disrupt the flow of the transcriptions, yet by definition they cannot be categorised as a grammatical error. Secondly, sentence breaks are not automatically predicted in ASR output, which might lead to overly long sequences. These discrepancies pose great challenges in adapting written GED to spoken corpora, and they remain active research topics e.g. [23, 24]. The focus of this work is to investigate the impact of ASR on spoken GED, therefore several pre-processings were applied to the test sets to bridge the gap between text and spoken corpora, such that reasonably good baseline GED results can be achieved: disfluencies marked in the meta-data were manually removed for JLE and BULATS; both manual and ASR transcriptions of the BULATS test set were segmented into short sentence units using reference speech units; no segmentation was applied to JLE as its conversational turns are sufficiently short.

In GED training, word embeddings were initialised with Google's 3 million word word2vec [25]. ASR confidence scores used in error detection were returned by the ASR engines, followed by piece-wise linear mapping [26]. Following [6], the fine-tuning of the GED models used for domain adaptation adopted a 10-fold cross-validation approach. Unlike [6], abbreviations with apostrophes were tokenised using the RASP convention [27], e.g.: $it's \rightarrow it +' s$.

4.2. Text GED performance on ASR transcriptions

To assess GED performance on ASR output, the manual and ASR transcriptions were aligned using a modified Damerau-Levenshtein algorithm [6]. The reference GE labels were mapped to the aligned ASR word. When an ASR error is seen, there are two choices of mapping. A grammar-based approach maps GE labels "as is" regardless of ASR errors. This approach does not penalise the system for its ASR performance, thus it often scores GED performance higher than the alternative,

feedback-based, approach. The feedback-based scheme labels GEs as correct where an ASR error occurs. The rationale is that giving feedback on words identified as grammatical errors due to an ASR error will cause confusion to the learner. Feedback-based scoring better depicts the system performance from the user perspective, thus was used for comparison across various corpora as well as manual transcriptions. Grammar-based scoring was considered to help analyse the nature of the grammatical errors as well as the GED system. For example:

MAN	she	**say**	what	i	**made**	do	...
ref	c	i	c	c	i	c	
ASR	she	**way**	what	i	**made**	do	...
grammar	c	i	c	c	i	c	
feedback	c	c	c	c	i	c	

Table 2: *Effect of ASR system on GED* $F_{0.5}$.

System	WER	grammar		feedback	
		—	+conf	—	+conf
ASR1	25.5	30.5	30.5	25.7	27.2
ASR2	21.3	32.5	32.6	27.8	29.9
ASR3	19.5	33.6	33.8	29.4	31.2

Table 2 gives GED $F_{0.5}$ scores for three ASR transcriptions using the system trained on the CLC written corpora. It is unsurprising that as the WER decreases the GED performance improves for both grammar-based and feedback-based scoring. There is a gap with GED on manual transcriptions, with $F_{0.5}$ of 42.5, 29.4, respectively for manual and ASR3 feedback GED.

To reduce erroneous feedback caused by incorrect transcriptions, ASR confidence scores were incorporated into the GED system. When the confidence threshold is set to be 0, all transcribed words receive a GE label (i.e. they may be labelled as grammatically incorrect). When the confidence threshold increases, ASR words with confidence scores lower than the threshold are rejected, labelled as grammatically correct.

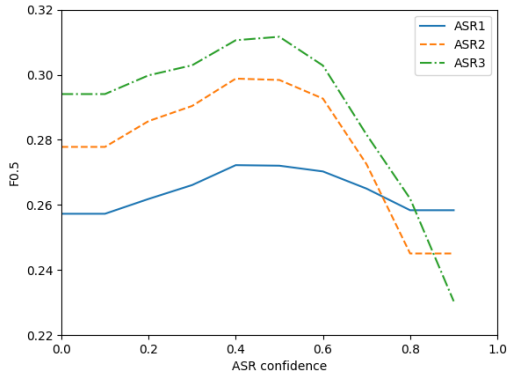

Figure 1: *Feedback-based* $F_{0.5}$*-confidence curve.*

Figure 1 shows the feedback-based $F_{0.5}$ scores of the CLC-trained GED system evaluated on BULATS against the ASR confidence thresholds. In the low confidence region, GED performance gradually rises as the confidence threshold increases. Low confidence scores often imply ASR errors, thus rejecting words with low confidence helps to rule out some of the ASR errors, which helps to reduce the number of false positives under feedback-based scoring. After reaching a peak at around a confidence of 0.4, the GED performance drops dramatically as the confidence threshold goes up. High confidence thresholds force more words to be marked as grammatically correct even if they are correctly identified and recognised, which significantly increases the number of false negatives thus reducing $F_{0.5}$. Grammar-based scoring plateaued in the low confidence region followed by the same decreasing trend in the high confidence region. When words with low confidence scores are rejected, the reduced false positives complements the increased false negatives. As the threshold is raised, the recall rate drops dramatically and dominates the $F_{0.5}$ score. Table 2 quotes the highest $F_{0.5}$ scores at the a confidence score threshold of 0.4. The GED performance can be seen to have gained approximately 2 points over the feedback-based scoring; whereas the grammar-based performance is unchanged.

4.3. GED fine-tuning

One of the challenges facing spoken GED is the lack of annotated spoken learner corpora. One option is to adapt text GED to spoken English. Here the trained written CLC-trained GED model was fine-tuned to the speech data in the 10-fold cross-validation fashion used in [6]. Table 3 contrasts the GED performance before and after fine-tuning for various test sets.

Table 3: *Precision (P), Recall (R) and* $F_{0.5}$ *scores with a CLC trained GED system and fine-tuned using cross-validation on the test data*[2]*. Feedback-based ASR3 used for BULATS-asr.*

	Test	fine-tune	P	R	$F_{0.5}$
Written	FCE	✗	74.0	29.6	56.9
		✓	72.1	30.8	56.9
Spoken	NICT-JLE	✗	60.7	27.5	48.9
		✓	69.6	31.8	56.3
	BULATS-man	✗	45.7	33.3	42.5
		✓	64.8	35.9	55.8
	BULATS-asr	✗	29.1	30.6	29.4
		✓	46.8	25.1	39.9
	+conf	✗	33.2	24.9	31.2
		✓	46.8	26.6	40.6

As expected the written FCE corpus model does not benefit from fine-tuning as the CLC and FCE both operate in the same domain. GED performance on JLE and BULATS manual transcriptions both improved significantly through domain adaptation, reaching similar performance. This shows that GED is an extremely domain sensitive task, and it is essential to tune the GED model to the domain of interest before further evaluation. The best performing ASR3 system was used to supply ASR transcriptions. Feedback-based scoring was adopted in order to analyse GED performance from a candidate's perspective. Fine-tuning on ASR transcriptions boosted the performance by 10.5 to 39.9 $F_{0.5}$. This is a slightly smaller gain than for manual, which may be due to the feedback-based scoring penalising the system for ASR mistakes, as observed in Table 2.

To further reduce false positives caused by ASR errors, confidence scores were taken into account during fine-tuning, using a threshold of 0.4 based on Figure 1. Here the confidence-based ASR fine-tuning does not consider words with confidence scores lower than the threshold of 0.4, consistent with those rejected from GED feedback. The performance was 0.7 higher than the vanilla fine-tuning result, reaching an $F_{0.5}$ of 40.6.

[2]The slight performance drop compared to our previous work [6] is due to change of tokenisation to follow the RASP convention.

Figure 2 compares the GED performance on manual and ASR transcriptions before and after fine-tuning (using vanilla and confidence-based fine-tuning for manual and ASR respectively). It is worth noting that the fine-tuned ASR model outperformed the baseline manual system in the high precision region of most interest for deployment. The PR curves start at a lower precision rate for the ASR transcription than manual. This is due to the feedback-based scoring labeling all ASR errors as grammatically correct. The initial precision gap between manual and ASR transcriptions measures the direct penalty caused by ASR errors on the GED system.

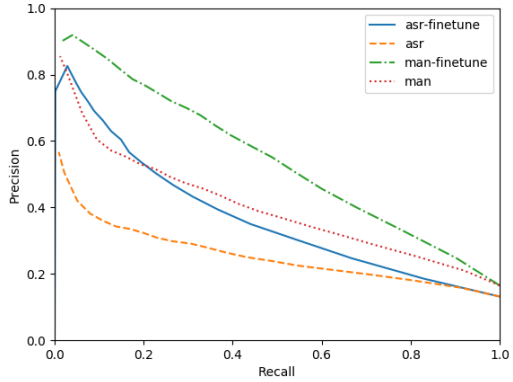

Figure 2: *GED precision-recall curve on BULATS manual and ASR (feedback-based ASR3) transcriptions.*

5. Analysis

Table 4: *Top 5 error tags on BULATS identified as incorrect by manual and ASR confidence fine-tuned systems.*

	manual			asr		
Posn	Tag	No.	% corr	Tag	No.	% corr
1	DV	89	80.9%	MC	412	77.4%
2	AGV	159	77.4%	DV	48	75.0%
3	MC	462	76.0%	AGV	115	71.3%
4	F	11	63.6%	FN	45	60.0%
5	MD	506	58.5%	AGN	161	53.4%

The top 5 error tags correctly detected by the manual and ASR3 confidence fine-tuned GED systems on the BULATS data are shown in Table 4. The systems operate similarly both in terms of which tags they detect well and the level of correct detection achieved. Errors in verb derivation (DV), verb agreement (AGV) and missing connectors (MC) [22] are most likely to be correctly detected. Some form errors (F, noun form (FN)), missing determiner (MD) and noun agreement (AGN) also appear in the top 5 tags. Both the manual and ASR tuned GED perform poorly, with very low % correct, on: replacement (substitution) errors (R$*$) where the word or phrase is valid and has the correct part-of-speech but needs replacing; other missing (insertion) (M$*$) types; and word order (W) errors.

It is interesting to contrast the spoken errors and GED rates achieved with the written FCE GED results. On the latter DV, AGV and MC error tags are correctly detected 61.5%, 56.3% and 5%, respectively, so the spoken system is actually detecting these tags more reliably. The most accurately detected errors on FCE are the 96% countability of noun error (CN), 75-91% incorrect inflections (IN,IJ,IV), and 73% adverb derivation (DY) [28]. By contrast on BULATS countability errors are not marked and only a handful of adverb derivation errors are observed in the data. The spoken system struggles to detect inflection (I$*$) errors, with 29-36% accuracy.

In practice the GED would be set to operate in the high precision/low recall region to minimise incorrect feedback. [5] showed that the WER, including on grammatical errors, decreases with increasing ASR confidence scores. Taking a P/R threshold from the ASR3 fine-tune system in Figure 2 to give 83% P/22% R, and an ASR confidence score of ≥ 0.9, all error tags are detected with $\geq$63% accuracy and the top 12 tags at $\geq$90%. Of the highest scoring tags in Table 4, over 75% missing connectors are detected with 98% accuracy. Agreement and verb derivation errors, however, tend to be excluded due to lower confidence. In contrast, R$*$, I$*$ and F$*$ errors score highly, with 90% accuracy for 43% and 52% of commonly seen replacement verb and preposition errors.

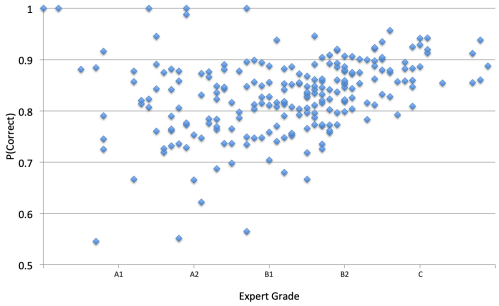

Figure 3: *Probability of grammatical correctness per speaker against expert grade for BULATS.*

Figure 3 shows that a speaker's proficiency level will affect the likelihood that they make grammatical errors. There is a wide variation in the probability of grammatical correctness for the lowest grade speakers. This gradually narrows as the speaker level improves, with the lower bound of probability of correctness increasing with proficiency. Note, candidates at the lower end of the grade scale who make very few errors tend to use a restricted vocabulary and say very little, e.g. "I do not understand". A similar pattern of detection of error tags to Table 4 is observed across grades A1-B2. The verb derivation (DV) errors mostly occur at the A2 and B1 levels. C speakers make very few AGV and F$*$ errors. A caveat to this analysis is that manual disfluency detection and segmentation has been applied. Automatic approaches may degrade GED performance.

6. Conclusions

This paper investigated the impact of ASR performance on grammatical error detection (GED) with spontaneous, second language learner, speech. A deep learning based bidirectional LSTM framework was used for GED, initially trained on text corpora. Three ASR systems with decreasing WER were considered. As expected, GED improves with higher transcription accuracy. Incorporating ASR confidence scores into the GED system reduced the number of false alarms for each ASR output, thus boosting GED from the user perspective. Domain adaptation through fine-tuning proved to be extremely successful in adapting text GED to non-native English free speaking corpora. The final confidence-based fine-tuned spoken GED system is able to detect some error tags with a high precision for feedback. Ongoing research aims to automate the full process including disfluency detection and speech segmentation.

7. References

[1] B. Council, "The English Effect," Aug 2013, Research Report.

[2] K. Lee, S. Ryu, H. Seo, S. Kim, and G. G. Lee, "Grammatical error correction based on learner comprehension model in oral conversation," in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, 2014, pp. 283–287.

[3] B. P. de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice and feedback in an ASR-based CALL system," *Computer Assisted Language Learning*, vol. 28, no. 6, pp. 550–576, 2015.

[4] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, "The CoNLL-2014 Shared Task on Grammatical Error Correction," in *Proc. of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2014 Shared Task)*, 2014.

[5] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines, "Impact of ASR performance on free speaking language assessment," in *Proceedings of INTERSPEECH*, 2018, pp. 1641–1645.

[6] K. Knill, M. Gales, P. Manakul, and A. Caines, "Automatic grammatical error detection of non-native spoken learner English," 2019, to appear in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). [Online]. Available: http://mi.eng.cam.ac.uk/~mjfg/ALTA/publications/Knill_ICASSP2019_AcceptedPaper.pdf

[7] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment," in *Proceedings of INTERSPEECH*, 2016, pp. 3122–3126.

[8] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional LSTM-RNN for Improving Automated Assessment of Non-native Children's Speech," in *Proceedings of INTERSPEECH*, 2017, pp. 1417–1421.

[9] M. Rei, G. K. O. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *Proc. COLING 2016, 26th International Conference on Computational Linguistics*, 2016, pp. 309–318. [Online]. Available: http://aclweb.org/anthology/C/C16/C16-1030.pdf

[10] "BULATS. Business Language Testing Service," Available: http://www.bulats.org/computer-based-tests/online-tests.

[11] Y. Wang, J. H. M. Wong, M. J. F. Gales, K. M. Knill, and A. Ragni, "Sequence teacher-student training of acoustic models for automatic free speaking language assessment," *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, pp. 994–1000, 2018.

[12] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2015.

[13] J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," in *Proceedings of DARPA Speech Recognition Workshop*, 1997.

[14] D. Povey, V. Peddinti1, D. Galvez, P. Ghahrmani1, V. Manohar1, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Proc. Interspeech*, 2016, pp. 2751–2755.

[15] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of INTERSPEECH*, 2018, pp. 3743–3747. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1417

[16] J. H. M. Wong and M. J. F. Gales, "Sequence student-teacher training of deep neural networks," in *Proc. Interspeech*, 2016, pp. 2761–2765.

[17] X. Chen, X. Liu, A. Ragni, Y. Wang, and M. J. F. Gales, "Future word contexts in neural network language models," in *Proc. ASRU*. IEEE, 2017, pp. 97–103.

[18] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 180–189.

[19] E. Izumi, K. Uchimoto, and H. Isahara, "The NICT JLE Corpus Exploiting the language learners' speech database for research and education," *International Journal of The Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, May 2004.

[20] C. of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[21] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[22] A. Caines, D. Nicholls, and P. Buttery, "Annotating errors and disfluencies in transcriptions of speech," University of Cambridge, Computer Laboratory, UK, Tech. Rep. UCAM-CL-TR-915, Dec. 2017. [Online]. Available: https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-915.pdf

[23] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency Detection Using a Bidirectional LSTM," in *Proceedings of INTERSPEECH*, 2016, pp. 2523–2527.

[24] E. Shriberg, A. Stolcke, D. Z. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[25] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop*, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[26] G. Evermann, H. Y. Chan, M. J. F. Gales, B. Jia, D. Mrva, P. C. Woodland, and K. Yu, "Training LVCSR systems on thousands of hours of data," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[27] T. Briscoe, J. Carroll, and R. Watson, "The second release of the RASP system," in *Proc. of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 77–80.

[28] D. Nicholls, "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT," in *Proc. of the Corpus Linguistics 2003 conference; UCREL technical paper number 16.*, 2003. [Online]. Available: http://ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf