Can the decoded text from automatic speech recognition effectively detect spoken grammar errors?

Chowdam Venkata Thirumala Kumar[1], *Meenakshi Sirigiraju*[2], *Rakesh Vaideeswaran*[1], *Prasanta Kumar Ghosh*[1], *Chiranjeevi Yarra*[2]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bengaluru, 560012, India
[2]Speech Lab, Language Technologies Research Center (LTRC), IIIT, Hyderabad, 500032, India

kumarcvt55@gmail.com, meenakshi.sirigiraju@research.iiit.ac.in, rakeshvaideeswaran.nitt@gmail.com, prasantg@iisc.ac.in, chiranjeevi.yarra@iiit.ac.in

Abstract

Language learning involves the correct acquisition of grammar skills. To facilitate learning with computer-assisted systems, automatic spoken grammatical error detection (SGED) is necessary. This work explores Automatic Speech Recognition (ASR), which decodes text from speech, for SGED. With current advancements in ASR technology, often it can be believed that these systems could capture spoken grammatical errors in the decoded text. However, these systems have an inherent bias from the language model towards the grammatically correct text. We explore the ASR-decoded text from commercially available current state-of-the-art systems considering a text-based GED algorithm and also its word-level confidence score (CS) for SGED. We perform the experiments on the spoken English data collected in-house from 13 subjects speaking 4110 grammatically erroneous and correct sentences. We found the highest relative improvement in SGED with CS is 15.36% compared to that with decoded text plus GED.

Index Terms: spoken grammar error detection, computer-assisted language learning, automatic speech recognition

1. Introduction

In today's world, English has become the lingua franca of education, science, technology, and employment [1]. Thus, it is crucial for individuals to acquire proficiency in the English language for effective communication. Consequently, English language learners (non-native speakers) have grown significantly [2]. Effective language learning involves the correct acquisition of grammar and pronunciation skills. Spoken grammar refers to the rules governing how words are used to construct sentences in spoken English. It plays a crucial role in enhancing effective communication, enabling individuals to express themselves clearly and ensuring that their intended message is accurately understood. However, they often make grammatical errors in sentence construction due to the influence of their native language's grammar structures, limited English language exposure, and the complex nature of English grammar rules.

To support language learners, Computer Assisted Language Learning (CALL) systems are employed for both pronunciation and grammar skills, providing automatic assistance and feedback akin to a teacher's supervision. Typically, these systems utilize ASR technology. The ASR systems are designed to recognize (decode) the text as a sequence of words from speech capturing its properties by considering various factors such as context, grammar, and pronunciation [3]. Thus, it has been hypothesized that ASR systems can recognize grammatically incorrect sentences. Consequently, grammatical errors are assumed to be detected with text-based grammatical error detection methods [4]. Further, the current advancements in ASR technology that involves end-to-end modelling made to believe the hypothesis stronger. However, the correct grammatical structure bias in ASR often limits the utilization of decoded text from ASR for detecting grammatical errors to build CALL systems.

In [4], a deep learning-based Grammatical Error Detection (GED) system originally designed for written text [5, 6] was fine-tuned using ASR transcriptions of spoken data. For each token/word in the transcript, GED assigns a label indicating grammatically correct or incorrect in that context. Similarly, in [7], the GED system was combined with ASR transcripts along with the decision of whether to pass a token from the transcript to the GED system based on the word-level confidence score obtained from ASR and a predefined threshold. Tokens with confidence scores exceeding the threshold were passed to the GED system for error detection, while those below the threshold were labelled as correct, indicating potential ASR errors. These works highlight the limitation of solely relying on ASR transcripts for detecting grammatical errors and emphasize the need for better spoken grammatical error detection systems with improved accuracy and efficacy.

In this work, we perform analysis by utilizing confidence scores or likelihoods obtained from ASR systems to assess their effectiveness in identifying grammatical errors, eliminating the need for additional systems. The likelihoods are obtained corresponding to the five best paths from the ASR decoding process, and the confidence scores are obtained at the word level for the text from the best decoding path. For this, ASR is built with an open-source tool kit considering three different LMs, namely, Librispeech [8], Tedlium [9] and Zamia. We compare the performance of the proposed approach with the GED system considering the decoded text obtained from three state-of-the-art (SOTA) ASR systems, which include a commercially available ASR. For evaluation, we collected speech data from 13 speakers reading both grammatically correct and incorrect sentences, serving as a test set. The spoken grammar error detection accuracy obtained with the confidence score-based analysis has the highest relative improvement of 15.36% compared to that with the GED on ASR-decoded transcripts.

2. Background

Figure 1 shows the block diagrams of statistical [10] and end-to-end ASR systems [11, 12, 13]. Both systems have the following three common components: AM, LM and Decoding. Besides these, the statistical ASR has a pronunciation model that maps the phonemes to words. Each of the three components is explained as follows.

Acoustic Model (AM): The AM captures the relationship between the acoustic features extracted from the input

10.21437/SLaTE.2023-9

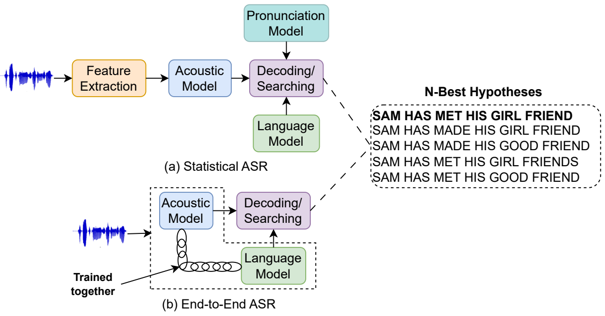

Figure 1: *Block diagrams of Statistical and end-to-end ASRs*

speech and the corresponding phoneme/sub-word/character sequence. These acoustic features generally are Mel Frequency Cepstral Coefficients (MFCCs) [14] in the case of statistical ASR. Whereas for E2E ASR, these features are learnt during the training with convolutional neural networks [15]. Thus, the AM provides likelihood/probability for each unit, given the input acoustics.

Langauge Model (LM): The LM captures the grammatical structure of the spoken language. The LM takes a sequence of words as input and predicts the probability of each word being the most likely choice to follow the given grammatical context. Thus, the LM is trained by considering correct text sentences. In statistical ASR, LM is trained separately from AM. However, in E2E ASR, the LM is trained together with AM in an end-to-end fashion.

Decoding: The beam search [16] algorithm is typically used for decoding. This algorithm efficiently explores the space of possible word or unit sequences to find the most likely transcription. It maintains a beam, a set of the most promising hypotheses, and at each step, it expands the beam by considering multiple candidate paths. These candidates are scored using a combination of probabilities from the AM and LM, allowing the algorithm to prioritize sequences that align well with both the acoustic and linguistic aspects of the input.

N-best: The N-best hypotheses are generated by selecting the top N candidates from the final beam. These candidates represent different possible transcriptions or interpretations of the input speech, ranked by their likelihood. An example case is shown in Figure 1 for the 5-best hypotheses.

3. Data collection

3.1. Text data collection

The text data for the recording is chosen such that it contains both grammatically correct and incorrect sentences. The dataset comprises sentences from a set of Multiple Choice Questions with a blank in each question with one correct and varying number of incorrect options. Sentences were prepared by taking each option at a time in the blank position. The sentences with the correct option will be correct sentences and those with the incorrect option will be incorrect. Questions were chosen from 37 topics that include a wide range of grammatical aspects including the use of prepositions, irregular verbs, tenses, pronouns, etc. These questions were designed by English teachers having 5 years of teaching experience, out of 37 topics 5 topics have 2 options, 1 topic has 3 options and 31 topics have 4 options for each question and each topic consists of 30 questions resulting in 1110 questions with 1110 correct and 3000 incorrect sentences. All the sentences were in the range of 3 to 33 words long with mean and standard deviation being 8.92 and 3.13 respectively.

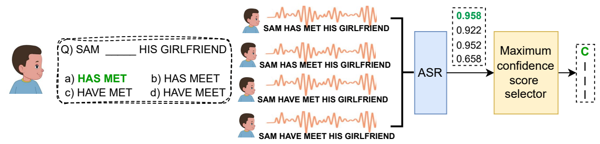
Q) SAM _____ HIS GIRLFRIEND
a) HAS MET
b) HAS MEET
c) HAVE MET
d) HAVE MEET
SAM HAS MET HIS GIRLFRIEND
SAM HAS MEET HIS GIRLFRIEND
SAM HAVE MET HIS GIRLFRIEND
SAM HAVE MEET HIS GIRLFRIEND
ASR
0.958
0.922
0.952
0.658
Maximum confidence score selector
C
I
I
I

Figure 2: *Block diagram of spoken grammar error detection (C - Correct and I - Incorrect).*

3.2. Speech data collection

The speech data were collected from 13 speakers reading the above text data. The subjects span across 3 states in India, 10 speakers from Andhra Pradesh, 2 speakers from Gujarat and 1 speaker from Karnataka, all of whom were college students. The data was collected using a web application written using ReactJS and Firebase backend. Each sentence will be displayed to the user with the option to record, listen to his/her recording and submit the recording. After the completion of recording 4110 sentences, the recordings were verified manually to check whether the recordings are matching with the text or not, and the recordings which are not matching with the text were flagged and the flagged sentences were sent to the speaker for re-recording. All the audios were recorded at 48kHz and then downsampled to 16kHz. The audio clips were in the range of 1-12 seconds long, with most of the audio clips being 3-4 seconds long. The audio clips received were single-channeled, in wav format.

4. Spoken grammar error detection

We analyse the detection of grammatically correct and incorrect sentences considering the following three measures: 1) LM likelihoods 2) Average ASR word-level confidence score (CS) 3) ASR decoding likelihoods of N best hypotheses. We compare the performance in terms of classification accuracy and average F1 score of the target class being correct and incorrect obtained in a question-specific (QS) manner. For each question, the sentence with the highest likelihood/confidence score is classified as correct and the rest as incorrect. A schematic diagram of SGED is shown in Figure 2. Each of the analyses is explained along with its motivation as follows:

4.1. LM likelihood based analysis

The analysis is performed by considering the LM loglikelihoods which explores the effectiveness of the text data alone as ASR-decoded text would have errors that would propagate when speech-based analysis is performed. Also, this analysis provides an upper bound for the speech-based analysis. Generally, LM is trained with a large number of correct sentences, so it has N-Gram likelihoods only for the correct sequence of words. If we pass the incorrect sentence then the N-Grams which are incorrect will be considered either unknown or the corresponding discounting will be applied which will result in lowering the overall sentence likelihood. One such example case is shown in Table 1 where the correct sentence got high log-likelihood than all incorrect sentences.

Experimental setup: We consider the most widely used N-Gram language models in statistical ASR systems i.e. Librispeech, Tedlium and Zamia. We consider 3-Gram small, 3-Gram medium and 3-Gram Large LMs trained on Librispeech text corpus using SRILM toolkit [17], 4-Gram small and 4-

Gram large LMs trained on Tedlium text corpus using PocoLM toolkit and 4-Gram and 5-Gram LMs trained on Zamia text corpus using KenLM toolkit [18] resulting in a total of 7 LMs covering different N-Grams from different text corpora. Librispeech LMs taken from Open Speech and Language Resources website [1], Tedlium LMs taken from the Kaldi website[2] and Zamia LMs were taken from Zamia Speech's official website [3]. We calculate LM likelihoods using the SRILM toolkit.

Baseline: We consider the widely used LSTM [19] based GED [6] algorithm trained with FCE Grammatical Error Detection corpus [20]. Since the GED is a sequence labeler, it labels every word as correct or incorrect in a sentence. We classify the sentence as incorrect if at least one incorrect label is predicted in the sentence, otherwise correct. An NVIDIA GeForce RTX 2080 GPU is used for training and testing the GED model with the default parameter setup mentioned in the paper.

4.2. Confidence score based analysis

Confidence score-based analysis is performed by considering the average word-level likelihoods of the first best hypothesis for the classification. The hypothesis in considering confidence scores as a direct measure of classification is that when an incorrect sentence is decoded, the confidence score for words at the incorrect position would be penalised because of LM's inherent bias towards correct grammar. Thus, it would reflect in the average score. One such example case is shown in Table 2. For the correct sentence, all words have the highest CS, in incorrect sentences the words in incorrect positions got very less CS than other words, resulting in a low average score shown in the last column.

Experimental setup: We consider two types of ASRs, the first type of ASR is statistical which is a pre-trained Librispeech ASR model available on the Kaldi website[4] and decoded using different LMs mentioned in the previous section, the second type of ASR is Google ASR, which is end-to-end based. Along with the most widely used LMs, we have trained LMs with only correct sentences from the test set and used them for decoding with statistical ASR. The intention behind using these LMs is to check whether the reduced searching space of the words while decoding improves the score-based classification. The use case for this type of LMs is when we know the text that we are expecting from the learner while learning through a CALL system, but giving answers to the practice questions through speech by inserting the option that the learner thinks is the correct answer for a fill in the blank question with multiple options. The same is illustrated in Figure 2. In this scenario, the word space will be limited to the words in the test set and will be useful in evaluating the answer given by the user. The confidence scores from Librispeech ASR are calculated using the Kaldi ASR toolkit [21] and confidence scores from Google ASR are obtained through paid speech-to-text API.

Baseline: We consider the commonly used cascaded type of systems with ASR as the first block and text-based GED as the second block. We consider 3 end-to-end ASRs such as Google ASR, wav2vec2.0 [5] [22] ASR and whisper [23] large v2 multilingual ASR with WERs 21.26%, 19.70% and 11.89% respectively.

Table 1: *Example LM likelihoods for 4 options case*

Correct (C)	sam	has	met	his	girlfriend	-16.70
Incorrect1 (I1)	sam	has	**meet**	his	girlfriend	-18.9
Incorrect2 (I2)	sam	**have**	met	his	girlfriend	-16.78
Incorrect3 (I3)	sam	**have**	**meet**	his	girlfriend	-18.9

Table 2: *Example confidence scores for 4 options question*

	Ground Truth	sam	has	met	his	girlfriend	
C	Decoded	sam	has	met	his	girlfriend	
	CS	1.00	1.00	0.93	1.00	0.86	0.958
	Ground Truth	sam	has	meet	his	girlfriend	
I1	Decoded	sam	has	**made**	his	girlfriend	
	CS	1.00	1.00	**0.62**	0.99	1.00	0.922
	Ground Truth	sam	have	met	his	girlfriend	
I2	Decoded	sam	have	met	his	girlfriend	
	CS	1.00	**0.84**	0.92	1.00	1.00	0.952
	Ground Truth	sam	have	meet	his	girlfriend	
I3	Decoded	sam	have	meet	his	girlfriend	
	CS	1.00	**0.40**	**0.38**	1.00	0.51	0.658

4.3. N-best scoring based analysis

The N-best scoring analysis is performed by considering the decoding likelihood of each of the N-best at a time for classification. The reason behind choosing N-best is that not always the first best is the best transcription. An example case is shown in Table 3 where the bolded one is actual text but is the 4th best hypothesis. So we analyse 1 to 5 best hypothesis likelihoods for classification.

Experimental setup: From the statistical ASR i.e. Librispeech ASR, we have calculated decoding likelihoods for the 5 best hypotheses for every audio using different LMs using the Kaldi ASR toolkit. We perform this analysis to check if there is any pattern with change in the best hypothesis as the decoded audio consists of grammatically incorrect text. We use the same **baseline** mentioned in Section 4.2 for comparison.

Table 3: *Example N-best hypotheses with decoding likelihoods (correct transcription **bolded**)*

1st best	some	of	the	beats	are	missing	3.0892
2nd best	some	of	the	beets	are	missing	3.0718
3rd best	some	of	the	beats	on	missing	3.0616
4th best	**some**	**of**	**the**	**beads**	**are**	**missing**	**3.0608**
5th best	some	of	the	beasts	are	missing	3.0602

5. Results and discussion

5.1. LM likelihood based analysis

Table 4 shows option-wise text-based classification accuracies with average F1 scores. It is clearly seen that the classification accuracies through LM likelihoods with the QS method are higher than the baseline state-of-the-art GED. In particular, Tedlium 4-Gram large, Zamia 5-Gram and Tedlium 4-Gram large show better accuracy than other LMs in the 2, 3 and 4 options cases respectively. But the performance difference between the 1st and 2nd highest is not much. This tells us that N-Gram LMs have the capability in detecting grammatical error occurrences by penalising the likelihood scores for incorrect sentences as the occurrence of words sequence in incorrect sentences is not seen in the training and is treated as unknown words. This can boost the score-based classification with statistical ASR as the LM's contribution is more while decoding. Among all, Zamia 5-gram LM shows the highest average accuracy & average F1 score. With this LM, the highest relative per-

[1] https://www.openslr.org/11/

[2] https://kaldi-asr.org/models/m5

[3] https://goofy.zamia.org/zamia-speech/lm/

[4] https://kaldi-asr.org/models/m13

[5] https://huggingface.co/speechbrain/asr-wav2vec2-librispeech

Table 4: *LM likelihood based analysis (accuracies in percentage (%) with average F1 scores in brackets)*

Options (#)	**Librispeech**			**Tedlium**		**Zamia**		**Baseline GED**
	3 (Small)	**3 (Medium)**	**3 (Large)**	**4 (Small)**	**4 (Large)**	**4**	**5**	
2	67.33 (0.67)	68.67 (0.69)	74.00 (0.74)	71.33 (0.71)	**76.00 (0.76)**	68.00 (0.68)	72.00 (0.72)	56.00 (0.56)
3	57.78 (0.53)	55.56 (0.50)	62.22 (0.58)	51.11 (0.45)	60.00 (0.55)	84.44 (0.83)	**88.89 (0.88)**	67.78 (0.63)
4	76.94 (0.69)	75.59 (0.67)	80.11 (0.73)	78.71 (0.72)	**80.75 (0.74)**	77.96 (0.71)	79.62 (0.73)	57.39 (0.52)
Average	67.35 (0.63)	66.61 (0.62)	72.11 (0.68)	67.05 (0.63)	72.25 (0.68)	76.8 (0.74)	**80.17 (0.78)**	60.39 (0.57)

Table 5: *Confidence score based analysis (accuracies in percentage (%) with average F1 scores in brackets)*

#	**Average ASR word level confidence score based classification**										**Baselines(ASR+GED)**		
	Librispeech pretrained ASR									**Google ASR**	**Google ASR**	**wav2vec2 ASR**	**whisper ASR**
	Librispeech LMS			**Tedlium LMs**		**Zami LMs**		**Test set LMs**					
	3 (Small)	**3 (Medium)**	**3 (Large)**	**4 (Small)**	**4 (Large)**	**4**	**5**	**3**	**4**				
2	50.51(0.51)	50.77(0.51)	52.10(0.52)	53.08(0.53)	53.95(0.54)	50.26(0.50)	52.05(0.52)	**55.44(0.55)**	55.28(0.55)	54.10(0.54)	52.70(0.52)	**55.10(0.54)**	53.71(0.54)
3	57.95(0.53)	**60.00(0.55)**	57.61(0.52)	57.61(0.52)	59.32(0.54)	57.78(0.53)	57.61(0.52)	59.15(0.54)	59.32(0.54)	57.01(0.52)	51.94(0.46)	**55.64(0.46)**	50.77(0.46)
4	63.43(0.51)	63.73(0.52)	64.33(0.52)	64.10(0.52)	64.05(0.52)	63.99(0.52)	64.17(0.52)	66.46(0.55)	**66.47(0.55)**	63.85(0.52)	57.35(0.50)	**63.18(0.53)**	52.50(0.48)
Avg	57.30(0.52)	58.17(0.53)	58.01(0.52)	58.26(0.52)	59.11(0.53)	57.34(0.52)	57.94(0.52)	60.35(0.55)	**60.36(0.55)**	58.32(0.53)	54.00(0.49)	**57.97(0.51)**	52.32(0.49)

Table 6: *N best scoring based analysis*

	N-Gram	**Options**	**1st**	**2nd**	**3rd**	**4th**	**5th**
Librispeech	**3 (Small)**	**2**	49.44	48.92	48.92	48.77	48.36
		3	52.99	52.99	52.82	52.99	52.99
		4	61.83	61.81	61.77	61.76	61.72
	3 (Medium)	**2**	49.33	48.97	49.13	48.72	48.51
		3	53.16	53.16	53.50	53.50	53.50
		4	61.99	61.86	61.81	61.83	61.79
	3 (Large)	**2**	49.59	49.33	48.46	48.26	48.26
		3	53.68	53.50	53.50	52.99	52.82
		4	62.20	62.10	61.96	61.88	61.82
Tedlium	**4 (Small)**	**2**	49.33	49.49	49.28	49.08	48.77
		3	54.19	54.19	54.02	54.02	54.02
		4	62.01	61.91	61.82	61.85	61.77
	4 (Large)	**2**	49.69	49.38	49.28	49.23	48.72
		3	54.36	54.19	54.19	53.68	53.50
		4	62.09	62.05	61.92	61.83	61.73
Zamia	**4**	**2**	49.28	49.08	48.97	48.51	48.15
		3	53.68	53.68	53.33	53.50	53.33
		4	62.08	61.97	61.83	61.79	61.73
	5	**2**	49.74	49.23	48.87	48.56	48.05
		3	54.02	53.50	53.68	53.50	53.50
		4	62.17	62.02	61.90	61.85	61.80
Test set	**3**	**2**	50.82	49.59	48.72	48.10	47.59
		3	55.73	55.21	54.87	54.70	54.36
		4	62.96	62.65	62.49	62.33	62.23
	4	**2**	**50.87**	49.54	48.62	48.05	47.69
		3	**55.90**	55.38	55.04	54.70	54.53
		4	**62.96**	62.65	62.48	62.35	62.24

formance observed is 32.75% with the GED baseline in terms of accuracy and 36.84% in terms of average F1 score.

5.2. Confidence score based analysis

Table 5 shows the option-wise CS based classification comparison with baselines. Initially, if we compare the classification accuracies of baselines with E2E Google ASR, score-based accuracies with the QS classification method are higher than the baseline in all cases except in 2 options case with the highest being wav2vec2 with a very less margin of 1%. This concludes the classification with score-based classification is better than the cascaded type of system in the detection of spoken grammar errors as the errors will be directly reflected in CS which helps in better classification to overcome the main disadvantage of the cascaded type of system i.e. error propagation. Further, if we compare the performance of E2E Google ASR with the statistical ASR without considering test set LMs, in 2 options case Googe ASR performs slightly better than the statistical ASR but the performance gap is very less i.e. 0.15% with Tedlium 4-Gram large LM. And in cases of 3 options and 4 options, Librispeech 3-Gram medium LM and Librispeech 3-Gram large LM show better performance than Google ASR. Again the performance gap in the 4 options case is very less. If we consider the test set LMs for comparison, then the accuracies with test set LMs are higher than all other cases except in the 3 options case where Librispeech 3-Gram medium is higher. From these results, it is clearly shown that the score-based classification is better than the widely used cascaded system and in particular statistical ASR shows a better advantage because of LM, further reducing the search space with LM trained with only test set text will be added advantage. Among all, Librispeech pretrained ASR with test set 4-gram LM has the highest average accuracy & average F1 score. With this, the highest relative performance in terms of accuracy observed is 15.36% with whisper ASR baseline. And 11.80% and 4.12% for google ASR and wav2vec ASR baselines respectively. In terms of average F1 score, the relative performance observed is 12.24% with google & whisper ASR and 7.84% with wav2vec ASR.

5.3. N-best scoring based analysis

Further, we show the performance of decoding likelihood-based classification using only statistical ASR in Table 6 with 1 to 5 best hypotheses to compare the performance variation. The demonstrated results indicate there is no performance variation with 1 to 5 best hypotheses and very negligible variation among different LMs in terms of classification accuracy. Even the best cases are not better than CS based classification but are almost equal to the baseline except in the 2 options case. A similar trend is observed in average F1 scores. This shows that the decoding likelihood score-based approach is almost equal to the baseline but not better than the CS based classification and there is no improvement with 5-best hypotheses.

6. Conclusion

In this work, we analyse the widely used cascaded type of spoken grammar error detection with SOTA ASR and a score-based classification. Experimental results showed better performance with the CS based method than the cascaded type of system and even further statistical ASR shows better performance than the end-to-end. Reduced search space while decoding in the case of statistical ASR is an added advantage. Decoding likelihood-based classification from statistical ASR is almost equal to the baseline cascaded system but not better than the CS based method. No effect of changing LM and different N-best hypotheses is observed with decoding-based likelihood classification with statistical ASR. Achieving the problem in a generic case from question specific and detecting the error locations will be the future scope of this work.

Acknowledgment - We thank Department of Science and Technology (DST), Govt. of India for supporting this work.

7. References

[1] P. S. Rao, "The role of english as a global language," *Research Journal of English*, vol. 4, no. 1, pp. 65–79, 2019.

[2] J. Prior, "English language statistics: How many people learn english?" *DoTEFL Report*, 2023. [Online]. Available: https://www.dotefl.com/english-language-statistics/

[3] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[4] K. M. Knill, M. J. Gales, P. Manakul, and A. Caines, "Automatic grammatical error detection of non-native spoken learner english," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8127–8131.

[5] M. Rei, G. K. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," *arXiv preprint arXiv:1611.04361*, 2016.

[6] M. Rei and H. Yannakoudakis, "Compositional sequence labeling models for error detection in learner writing," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1181–1191. [Online]. Available: https://aclanthology.org/P16-1112

[7] Y. Lu, M. Gales, K. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of asr performance on spoken grammatical error detection." ISCA, 2019.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[9] A. Rousseau, P. Deléglise, and Y. Esteve, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[11] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[12] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[16] X. L. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 89–114, 2002.

[17] A. Stolcke, "Srilm—the sri language modeling toolkit," 1999.

[18] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 180–189. [Online]. Available: https://aclanthology.org/P11-1019

[21] 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. CONF. IEEE Signal Processing Society, 2011.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356