A Robust End to End Spoken Grammar Assessment System

Sunil Kumar Kopparapu*, Chitralekha Bhat*, Ashish Panda*
* TCS Research, Tata Consultancy Services, Mumbai, India
E-mail:{sunilkumar.kopparapu,ashish.panda}@tcs.com

***Abstract*—Spoken language assessment (SLA) systems restrict themselves to evaluating the pronunciation and oral fluency of a speaker by analysing the read and spontaneous spoken utterances respectively. The assessment of language grammar or vocabulary is relegated to written language assessment (WLA) systems. In this paper, we propose a practical, scalable and robust end-to-end SLA system to assess language grammar from spoken utterances. The use of a large language model (LLM) to bring in variations in the test makes the system largely unteachable thereby making it practical and scalable. The use of a hybrid automatic speech recognition (ASR) with a custom-built language model, enabling processing of read and not spontaneous speech by design, and the novel grammar scoring mechanism that is robust to mis-recognitions of ASR makes the overall spoken grammar assessment (SGA) system robust.**

I. Introduction

The demand for second language (L2) learners to study foreign languages, especially English, leads to the imminent need for the development of a language proficiency assessment system [1], [2]. Although there are several English language assessment tools available, these assessments tend to be lengthy because they evaluate different aspects of language proficiency separately, one after another. The spoken language proficiency assessment is often restricted to assessing the speech articulation of a speaker in terms of pronunciation [3]–[5] and speech delivery in terms of oral fluency [6], [7], which includes speaking rate [8], [9], recognition of pauses, filler words, and analysis of intonation [10] etc. The other aspects of language like grammar or vocabulary are assessed separately through a written language proficiency assessment system. Spoken language assessment (SLA) and written language assessment (WLA) complement each other, providing a comprehensive evaluation of overall language proficiency. Separate SLA and WLA assessments not only make the assessment lengthy but may also encourage learners to not concentrate on language grammar during SLA. Additionally, in most practical settings like call centers and virtual interviews, spoken language communication is important. This motivates the need for a comprehensive SLA system that assesses all aspects of language proficiency.

Grammatical errors in spoken speech can result from either incorrect transcription by ASR or as grammatical errors by the speaker [11]. Language grammar assessment is often delegated to WLA systems because of the performance limitation of automatic speech recognition (ASR), a crucial component, if SLA systems were to be used for grammar assessment. It is well known that the performance of hybrid ASR system is better for read speech compared to spontaneous speech [12] but more importantly, the state-of-the-art end-to-end ASR systems are unable to reproduce the exact spoken word sequence, especially when they are grammatically incorrect because only grammatically correct training data is used to train the ASR. More recently, [13] proposed an end-to-end spoken grammatical error correction architecture by employing an end-to-end ASR foundation model, `whisper` to remove disfluencies from spontaneous speech. However, they do not address the inability of `whisper` to transcribe grammatically incorrect spoken utterances especially because `whisper` has been trained on huge amounts of grammatically correct data which make it biased towards grammatically correct text. A deep learning-based grammatical error detection system, by fine tuning the system designed for written text using ASR transcriptions of spoken data, was shown to improve performance on spoken L2 English. However, challenges of the ASR performance on grammatically incorrect utterances by the learner and disfluency detection limited its performance [14]. Using read speech, [15], explored adopting commercially available ASR to detect spoken grammar errors by utilizing confidence scores or likelihoods obtained from ASR systems at word level. In another study, the impact of grammatical errors stemming from incorrect transcriptions of ASR was studied using a deep learning-based system in [11], however they do not address the grammatical errors by the speaker.

We introduce an end-to-end SLA system to enable assessment of language grammar from spoken speech. The proposed system, by design of the grammar scoring module, is robust to grammatical errors introduced by ASR transcription. In addition the system is able to transcribe accurately any grammatical error introduced by the learner through the design of a custom language model assisting the ASR. Further, the use of a large language model (LLM) makes sure that no two assessment instances are the same; ensuring that the student cannot be coached for the assessment. This makes the SLA system practically usable and scalable. The main contribution of the paper is (a) designing a SLA system that can robustly evaluate language grammar from spoken speech, thereby enabling language proficiency assessment in an unified way [16] without employing any WLA tools leading to reduced assessment time, (b) by design, minimizing disfluencies in the

979-8-3315-7206-8/25/$31.00

learners speech by orienting the learner towards read speech to enable robust SGA, (c) designing a custom-built LM on top of a readily available hybrid ASR system for detecting grammatical error made by learner, (d) proposing a grammar scoring module that is robust to errors in ASR transcription, and (e) employing LLM to enable variations in the assessment to make the SLA system largely unteachable thus making it scalable and practical. The rest of the paper is organized as follows, we describe the spoken language grammar assessment (SGA) system in detail in Section II. We conduct experiments in Section III to show the process of automatic generation of paragraphs that can be used in grammar evaluation and show the need for a custom-built LM for speech transcription and we conclude in Section IV.

II. Spoken Language Grammar Assessment

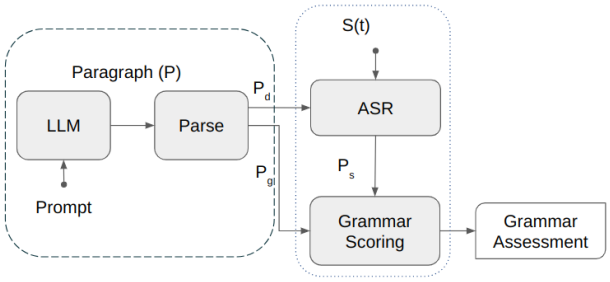
The block diagram of our end-to-end SGA system is shown in Figure 1a. It has two parts, the first part, allows for the generation of a paragraph P (example Figure 4a) by prompting an LLM, and the second part takes an audio utterance $S(t)$, spoken by the learner, corresponding to P_d (example, Figure 4b) and assesses for language grammar using P_g (Figure 4c) and the transcription produced by the ASR. Unlike traditional SLA systems which take an audio input $S(t)$ and use the output P_s of a standard ASR to automatically compute the pronunciation or oral fluency [17], [18] only, in this paper, we enable grammar assessment on spoken speech. The grammar scoring module uses the output transcription of the ASR, namely, P_s and the gold truth P_g to compute the grammar assessment score.

Figure 1b shows a web implementation of the proposed end-to-end grammar assessment system. The interface displays a paragraph P_d generated by an LLM using 1-shot prompt engineering, allows the learner to speak the paragraph and record the utterance ($S(t)$). The ASR transcribes $S(t)$ and uses the grammar scoring module to asses spoken language grammar.

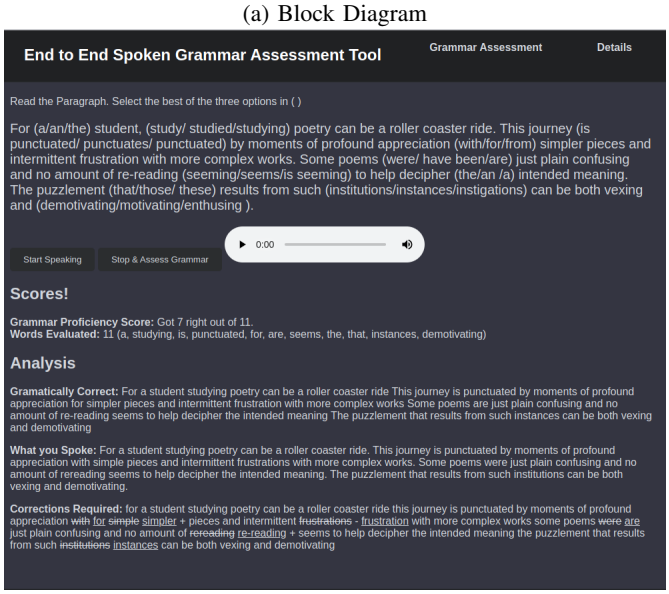
The proposed SGA system is practical, scalable and while being able to detect grammatical errors of the leaner, is robust to grammatical errors introduced by ASR mis-recognition. We describe in detail.

A. Generation of Paragraph

We first describe how to generate a unique assessment paragraph P for each student using ChatGPT. This ensures that the students cannot be coached for the assessment. This ability to generate a *new* paragraph for each assessment test make the SGA system practical from large scale deployment angle. We adopt 1-shot learning prompting style for generating new paragraphs ($P_1, P_2, \cdots$) as described in Figure 2. A wide variety of new paragraphs, namely, P_ns can be generated using the prompt *"Generate just the paragraph. With subject <subject>."* This allows for the generation of a completely new paragraph for each assessment, in the desired format (see Figures 3a, 3b).



(a) Block Diagram



(b) Functional System (Web Application hosted on intranet).

Fig. 1: End to End System for SGA.

#1 User: """ P """ {Sample P in Fig. 4a.} Generate paragraphs like P. One <cor></cor>tag within <gram></gram>tags. Each <gram>tag has three options separated by "/".
#1 ChatGPT: Thank you for providing the specific format and instructions. The grammar choices are marked within <gram>, with the correct option indicated using <cor>.
#2 User: Generate a paragraph similar to the example shown.
#2 ChatGPT: P_1 {Generated paragraph (Fig. 3a)}
#3 User: Generate use subject "learning physics is easy".
#3 ChatGPT: P_2 {Generated paragraph shown in Fig. 3b}

Fig. 2: 1-shot learning prompting to generate new P.

A sample P generated by prompting a LLM[19] is shown in Figure 4a. The tags "<gram></gram>" correspond to the words or phrases that are to be evaluated for grammar. The tag "<cor></cor>" shows the correct choice. For example, "<gram>study/studied/<cor>studying</cor></gram>" indicates that the correct choice of grammar usage is **studying** corresponding to **study/studied/studying** displayed to the learner. In practice, both P_d (Figure 4b) and P_g (Figure 4c) can be obtained by a simple text parser applied on P (Figure 4a).

In <gram><cor>a</cor>/an/the</gram>bustling city, <gram>exploring/ explored/ <cor>exploration</cor></gram>can be an exciting adventure. **<snip>** The challenge <gram><cor>that</cor>/ those/these</gram>comes from such <gram>adventures/<cor>explorations</cor>/explorers </gram>can be both thrilling and <gram>eye-opening/<cor>exhausting</cor>/ insightful</gram>.

(a) A paragraph generated by prompting ChatGPT (P_1).

For <gram><cor>a</cor>/an/the</gram>physics enthusiast, <gram>studying/ studied/ <cor>studying</cor></gram>physics can be a fascinating journey. **<snip>** The understanding <gram><cor>that</cor>/ those/ these</gram>comes from such <gram>endeavors/<cor>pursuits</cor>/ explorations</gram>can be both empowering and <gram><cor>rewarding</cor>/ challenging/exciting</gram>.

(b) A paragraph generated by prompting ChatGPT (P_2).

Fig. 3: Paragraph's generated by prompting ChatGPT.

For <gram><cor>a</cor>/an/the</gram>student, <gram>study/ studied/<cor>studying</cor></gram> poetry can be a roller coaster ride. **<snip>** can be both vexing and <gram><cor>demotivating</cor>/motivating/enthusing </gram>.

(a) A paragraph generated by prompting a LLM (P).

For **(a/an/the)** student, **(study/studied/studying)** poetry can be a roller coaster ride. This journey **(is punctuated/punctuates/punctuated)** by moments of profound appreciation **(with/for/from)**simpler pieces and intermittent frustration with more complex works. Some poems **(were/have been/are)** just plain confusing and no amount of re-reading **(seeming/seems/is seeming)** to help decipher **(the/an/a)** intended meaning. The puzzlement **(that/those/these)** results from such **(institutions/instances/instigations)** can be both vexing and **(demotivating/motivating/enthusing)**.

(b) Paragraph displayed to the student (P_d).

For **a** student, **studying** poetry can be a roller coaster ride. This journey **is punctuated** by moments of profound appreciation **for** simpler pieces and intermittent frustration with more complex works. Some poems **are** just plain confusing and no amount of re-reading **seems** to help decipher **the** intended meaning. The puzzlement **that** results from such **instances** can be both vexing and **demotivating**.

(c) The grammatically correct paragraph (P_g).

Fig. 4: A sample P generated using an LLM. P_d is displayed for the learner to speak and P_g (parsed from P) is used for grammar assessment.

*B. Speech to Text (*ASR*)*

The most crucial block in the SGA is the ASR, which converts the spoken paragraph $S(t)$ into text P_s (see Figure 1a) because ASR transcripts can be erroneous [20] leading to an error in grammar assessment. We hypothesize that the construction of custom language model (CLM) which is *tightly coupled* with the assessment paragraph P performs better than even the state-of-the-art ASR, namely, `whisper`[21]. This is based on the observation that (a) a language model (LM) plays a significant role in the transcript produced by an ASR engine and (b) while `whisper` is trained on extremely large and varied sets of text data, the training data is likely to *lack* grammatically *incorrect* sentences.

As an illustration (see Figure 5) there are three possible options for both the preposition (a/an/the) and the verb (study/studied/studying). Hence, the total number of possible sentences using all options is nine. Most of these (eight of the nine) sentences will rarely occur, in any text databases since they are grammatically incorrect. Hence, text corpora used for training `whisper` will not include these sentences which can result in `whisper` transcripts being biased towards grammatically correct sentences [15]. Shallow fusion is the most popular approach to combine pre-trained ASR model and LM [22]. Shallow fusion can be expressed mathematically as:

$$\text{score}(P_s|S(t)) = \log\left(p(P_s|S(t))\right) + \gamma \cdot \log(p(P_s)) \quad (1)$$

where P_s is the spoken paragraph, $p(P_s|S(t))$ is acoustic score, γ is a scaling factor and $p(P_s)$ is LM score. If P_s is not present in the training text, then $p(P_s) = 0$, which will make $\text{score}(P_s|S(t))$ very small. This results in the ASR choosing the *grammatically correct sentence instead of the spoken wrong sentence*. However, a CLM [23] can, easily, be trained to include all possible variations (including the wrong ones) of the sentence to mitigate this. This is the reason for our belief that an ASR with a custom-built LM (ASR-CLM) can be far more accurate than any state-of-the-art ASR, like `whisper`, with a general-purpose LM.

For (a/an/the) student, (study/studied/studying) poetry can be a roller coaster ride.

(a) Sample sentence displayed to the student

1) For **a** student, **study** poetry can be a roller coaster ride.
2) For **an** student, **study** poetry can be a roller coaster ride.
3) For **the** student, **study** poetry can be a roller coaster ride.
4) For **a** student, **studied** poetry can be a roller coaster ride.
5) For **an** student, **studied** poetry can be a roller coaster ride.
6) For **the** student, **studied** poetry can be a roller coaster ride.
7) *For **a** student, **studying** poetry can be a roller coaster ride.*
8) For **an** student, **studying** poetry can be a roller coaster ride.
9) For **the** student, **studying** poetry can be a roller coaster ride.

(b) Sentences (correct in italics) expected from the student.

Fig. 5: Sample sentence (a) and expected variations (b). Item 7 is the only grammatically correct variant.

We validate our hypothesis that the use of an ASR engine equipped with a custom-built LM based on the generated

paragraph, namely, ASR-CLM is better equipped than even the SOTA ASR like whisper. We used the sota whisper speech recognition engine and a Kaldi-based ASR with a custom-built LM (ASR-CLM) for comparison. The acoustic model of the Kaldi ASR was trained on 960 hours of speech data from Librispeech database [24]. The custom LM was trained on the text comprising all variations, both grammatically correct and incorrect instances of the given sentences (example Figure 5b).

To compare the performance of whisper and ASR-CLM, we recorded speech corresponding to all variations of the below sentence, "It (**was**/is/am) a late afternoon probably (**on**/in/of) the 15th of February, 2019. (I and my friend/**my friend and I**) (was/**were**/will be) walking on the footpath (**in**/inside/into) central Bangalore." namely, $3(\text{was/is/am}) \times 3(\text{on/in/of}) \times 2(\text{I and my friend/my friend and I}) \times 3(\text{was/were/will be}) \times 3(\text{in/inside/into}) = 162$ utterances. We found that ASR-CLM was able to exactly transcribe the utterance (even when there was an error in grammar introduced by the learner) while whisper *corrected* some of the grammatical error of the learner. Table I shows two representative examples. Note that ASR-CLM accurately recognizes the spoken words, regardless of grammatical correctness of the sentence, while whisper falls short. In the first example (Table I) the article "a" was replaced by "the" by whisper while in example two, the article "a" was not recognized by whisper. Overall, the ability of ASR-CLM to recognize what was spoken (even when grammatically incorrect) is 84.7% while that of whisper was 46%. The performance was computed on 137 utterances; 25 of the 162 utterances were discarded because of noise. The poor accuracy of the whisper highlights the need for a CLM-ASR for the purpose of SGA. Note that the CLM needs to be computed afresh for each paragraph during the assessment.

$S(t)$	It was a late afternoon probably on the 15th of February 2019 my friend and I were walking on the footpath in central Bangalore
whisper ASR-CLM	It was the late afternoon probably on the 15th of February 2019 my friend and I were walking on the footpath in central Bangalore "It was a late afternoon probably on the 15th of February 2019 my friend and I were walking on the footpath in central Bangalore".
whisper ASR-CLM	It am ~~a~~ early after noon probably on 15th February 2019 my friend and I was walking on the footpath in central Bangalore It am a late afternoon probably on the 15th of February 2019 my friend and I was walking on the footpath into central Bangalore

TABLE I: Sample $S(t)$. ASR errors, marked in red.

C. Grammar Scoring Module

We propose a grammar scoring module which is robust to ASR transcription errors. The grammar scoring module takes $S(t)$, P_d, and G_w as input and produces a score S_g^s. Namely, $S_g^s = \text{G-SCORE}(P_s, P_d, G_w)$ where, P_d is the displayed paragraph which is used to build ASR-CLM, $P_s = \text{ASR-CLM}(S(t))$, and G_w ($\in P_d, \ll P_d$) a small subset of words (inside <cor></cor>) that are used to assess the grammar of the learner. We compute S_g^s as follows:

1) While maintaining the sequence of the words in P_d and P_s, we create a set $p_1 = \{w \in P_d \mid w \notin P_s\}$ of words that are in P_d but not in P_s.
2) Create $p_2 = \{w \in G_w \mid w \notin p_1\}$, a set of words in G_w but not in p_1.
3) The grammar score, $S_g^s = |p_2|$ is the cardinality of the set p_2. Note that p_2 is a set of all the correctly spoken grammar words by the student.

As an example, using Figure 4 as reference (P_g = {For (*a*/an/the) student, (study/studied/*studying*) poetry can be a roller coaster ride <snip>The puzzlement (*that*/those/these) results from such (institutions/*instances*/instigations) can be both vexing and (*demotivating*/motivating/enthusing)}; correct grammar option in *blue* italics), we show how S_g^s is computed. Let a sample output of ASR-CLM (errors marked in red) be P_s = {For *a* student study poetry can be a roller coaster wide. <snip>The puzzlement *that* results from such instigations can be both vexing and *demotivating*}. We know, G_w = {a, studying, is punctuated, for, are, seems, the, that, instances, demotivating}. We find the difference p_1 = {studying, instigations} (see Item 1) and p_2 = {a, is punctuated, for, are, seems, the, that, instances, demotivating} (see Item 2). The grammar score S_g^s is $|p_2| = 9$, namely, 9 of the 10 grammar words scored right. Note that although the word "ride" was mis-recognized as "wide" by our ASR-CLM, it does not affect S_g^s since "ride" is not a part of the grammar assessment set G_w.

III. Experimental Analysis

To the best of our knowledge and as also reported in [13], [15], a standard speech dataset for SGA with manual annotations of grammatical errors in conversational or read speech is currently unavailable. To evaluate our SGA system, we used an in-house dataset. The student was shown a paragraph P_d on a web interface (Figure 1b) and was asked to familiarize the displayed paragraph. Once ready, while looking at the paragraph and making their word choices, the student could record their utterance. We collected audio recordings from 17 students speaking a generated paragraph. Each spoken paragraph was manually assessed by a linguist to mark the grammar score (${}^*S_g^s$; Table II). We used both whisper and ASR-CLM to convert the spoken paragraph to text and then compute S_g^s as mentioned in an earlier section. The error in assessment is captured in parenthesis for each student in Table II. Larger grammar assessment errors ($\epsilon_g = 20$) due to whisper are observed compared to $\epsilon_g = 3$ for a custom-built LM ASR (ASR-CLM) demonstrating the need for a custom built LM based ASR even in the presence of the SOTA foundational ASR like whisper.

IV. Conclusions

Language proficiency assessment is a common requirement for L2 (non-native) speakers of English. There exists several

Student	Grammar Assessment		
	whisper	ASR-CLM	${}^{*}S_g^s$
	$S_g^s(\epsilon_g)$	$S_g^s(\epsilon_g)$	
#1	14 (1)	15 (0)	15
#2	11 (1)	11 (1)	10
#3	11 (2)	9 (0)	9
#4	12 (1)	13 (0)	13
#5	12 (1)	12 (1)	13
#6	10 (2)	12 (0)	12
#7	6 (2)	8 (0)	8
#8	15 (3)	12 (0)	12
#10	15 (1)	16 (0)	16
#11	3 (0)	3 (0)	3
#12	6 (2)	8 (0)	8
#13	10 (2)	12 (0)	12
#14	15 (1)	15 (1)	16
#15	14 (1)	15 (0)	15
#16	14 (0)	14 (0)	14
#17	13 (0)	13 (0)	13
Total	(20)	(3)	-

TABLE II: Performance evaluation of whisper and ASR-CLM for SGA.

SLA tools to assess learner pronunciation and oral fluency but very few venture into assessing spoken grammar. We designed and implemented a practical, scalable and robust SLA system to assess spoken grammar automatically in an end-to-end architecture. The use of LLM enables the generation of paragraphs that are largely non-repetitive thereby making the SGA system hard to be memorized by students. This aspect makes the SGA system both practical and scalable, thus making it deployable. The robustness of the SGA system is directly dependent on the performance of ASR and the grammar scoring measure. The intervention introduced by design to display the *paragraph with options*, results in *read* speech (devoid of filler words and word repetitions and hence ASR transcripts are less erroneous) by the student, and construction of a custom LM associated with the paragraph being used for assessment enables our ASR-CLM to transcribe grammatically erroneous spoken utterances unlike the SOTA whispermakes SGA robust. Additionally, we can observe that the grammar scoring module, by design, is not affected by ASR mis-recognition of non G_w words making the proposed SGA robust to errors in ASR transcription.

REFERENCES

[1] L. Jin and H. Zhu, "Developing standardized speech and language assessment tools in Mandarin Chinese: A context for improving reading and writing," *Journal of Chinese Writing Systems*, vol. 7, no. 3, pp. 150–160, 2023.

[2] H. Franco, H. Bratt, R. Rossier, *et al.*, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[3] K. Sheoran, A. Bajgoti, R. Gupta, *et al.*, "Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping," *IEEE Access*, vol. 11, pp. 15 485–15 495, 2023. DOI: 10.1109/ACCESS.2023.3244393.

[4] H. Pei, H. Fang, X. Luo, and X. Xu, "Gradformer: A Framework for Multi-Aspect Multi-Granularity Pronunciation Assessment," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 554–563, 2024. DOI: 10.1109/TASLP.2023.3335807.

[5] B. Lin and L. Wang, "Exploiting Information From Native Data for Non-Native Automatic Pronunciation Assessment," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 708–714. DOI: 10.1109/SLT54892.2023.10022486.

[6] A. Preciado-Grijalva and R. F. Brena, "Speaker fluency level classification using machine learning techniques," *arXiv preprint arXiv:1808.10556*, 2018.

[7] S. P. Dubagunta, E. Moneta, E. Theocharopoulos, and M. Magimai Doss, "Towards Automatic Prediction of Non-Expert Perceived Speech Fluency Ratings," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 7–11.

[8] A. Imran, M. Pandharipande, and S. K. Kopparapu, "Speakrite: Monitoring speaking rate in real time on a mobile phone," *International Journal of Mobile Human Computer Interaction (IJMHCI)*, vol. 5, no. 1, pp. 62–69, 2013.

[9] S. K. Kopparapu, *Non-linguistic analysis of call center conversations*. Springer, 2015.

[10] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, no. 3, pp. 254–267, 2010, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2009.11.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639309001708.

[11] Y. Lu, M. J. F. Gales, K. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of ASR Performance on Spoken Grammatical Error Detection," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201126147.

[12] K. Mukherji, M. Pandharipande, and S. K. Kopparapu, "Improved Language Models for ASR using Written Language Text," in *2022 National Conference on Communications (NCC)*, 2022, pp. 362–366. DOI: 10.1109/NCC55593.2022.9806803.

[13] S. Bannò, R. Ma, M. Qian, K. M. Knill, and M. J. F. Gales, "Towards End-to-End Spoken Grammatical Error Correction," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 791–10 795. DOI: 10.1109/ICASSP48485.2024.10446782.

[14] K. Knill, M. 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)*, 2019, pp. 8127–8131. DOI: 10.1109/ICASSP.2019.8683080.

[15] C. Venkata Thirumala Kumar, M. Sirigiraju, R. Vaideeswaran, P. K. Ghosh, and C. Yarra, "Can the decoded text from automatic speech recognition effectively detect spoken grammar errors?" In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 41–45. DOI: 10.21437/SLaTE.2023-9.

[16] S. K. Kopparapu and A. Panda, "Unified spoken language proficiency assessment system," in *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2024, pp. 1–6. DOI: 10.1109/O-COCOSDA64382.2024.10800105.

[17] A. Panda, R. Acharya, and S. K. Kopparapu, "Oral Fluency Classification for Speech Assessment," in *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*, IEEE, 2023, pp. 231–235. DOI: 10.23919/EUSIPCO58844.2023.10289791.

[18] L. Fontan, M. L. Coz, and S. Detey, "Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with Japanese learners of French," in *INTERSPEECH*, 2018.

[19] OpenAI, *GPT-3.5: OpenAI's Generative Pre-trained Transformer 3.5*, https://platform.openai.com, Accessed: 2023-06-26, 2023.

[20] C. Anantaram, S. K. Kopparapu, C. Patel, and A. Mittal, "Repairing General-Purpose ASR Output to Improve Accuracy of Spoken Sentences in Specific Domains Using Artificial Development Approach," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16, New York, New York, USA: AAAI Press, 2016, pp. 4234–4235, ISBN: 9781577357704.

[21] OpenAI, *Whisper: OpenAI's Automatic Speech Recognition (ASR) System*, https://github.com/openai/whisper, Model: Whisper Tiny.en, Accessed: 2023-06-26, 2022.

[22] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-End Speech Recognition: A Survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 325–351, 2024. DOI: 10.1109/TASLP.2023.3328283.

[23] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom: Association for Computational Linguistics, 2011, pp. 187–197. [Online]. Available: https://github.com/kpu/kenlm.

[24] 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)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.