Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SGED-Probe: Probing E2E ASR decoder and aligner for spoken grammar error detection under three speaking practice conditions

Chowdam Venkata Thirumala Kumar, Chiranjeevi Yarra

Speech Lab, Language Technologies Research Center (LTRC), IIIT, Hyderabad, 500032, India
chowdam.v@research.iiit.ac.in, chiranjeevi.yarra@iiit.ac.in

Abstract

Grammar error detection in one's speech, referred to as spoken grammar error detection (SGED), is a critical component for building computer-assisted language learning systems. Traditionally, ASR followed by text-based GED has been used, but ASR errors can limit SGED performance. In this work, we analyse SGED under three speaking conditions: spontaneous, semi-spontaneous, and memorized, using five state-of-the-art ASRs. We examine the decoded text and ASR cues such as confidence scores and aligner probabilities, out of which, for the latter one, the ground-truth grammatically correct (GGC) text is used. Experiments revealed that autoregressive ASRs show bias toward GGC text, leading to suboptimal performance. It is observed that there is an increase in performance in response to the decreased freedom of speech, with semi-spontaneous speech outperforming memorized speech. Aligner probabilities outperform confidence scores despite the alignment and overconfidence issues.

Index Terms: spoken grammar error detection, speaking practice, computer assisted language learning, E2E ASR

1. Introduction

Second language learning is becoming popular, especially for English [1], which has been used as the global language for education, science, technology, and employment [2]. Among the different learning aspects, practising grammar while speaking is also an important component. It has been emphasized that the practice should be done under three speaking style conditions based on the freedom of the spoken content, starting from memorized, semi-spontaneous and spontaneous speaking [3]. Thus, automatic grammar error detection in learners' speech, referred to as spoken grammar error detection (SGED) across all three conditions, would benefit developing computer-assisted language learning (CALL) systems. This work analyses SGED approaches under these three conditions, considering end-to-end (E2E) automatic speech recognition (ASR) systems.

Almost all the existing works have considered a cascaded pipeline of ASR, which decodes text from speech, followed by grammar error detection/correction (GED/C), which detects or corrects grammatical errors in the decoded text. Several studies, such as [4–7], have aimed to enhance GED/C by fine-tuning pre-trained written text-based models with manually transcribed spoken text along with correct and incorrect tags. Moreover, these works often focus only on GED/C, while leaving the ASR component largely unchanged. Further, all of these works have addressed SGED tasks under spontaneous conditions only. As the spontaneous speech contains disfluencies and improper chunks in its decoded text, majority of the works dealt with disfluency detection [8] and obtaining meaningful chunks [9] along with GED/C. However, the other two speaking practice conditions are free from disfluencies and improper chunking issues. Moreover, for these two conditions, it is important to analyse ASRs, which are typically trained on grammatically correct data and often struggle to decode grammatically incorrect speech due to bias from language models.

In this work, we perform a comprehensive analysis using five state-of-the-art (SOTA) E2E ASRs, representing both autoregressive and non-autoregressive categories across all three types of speaking practices for SGED. Initially, we compare the performance of ASR models based on the decoded text for SGED. Further, we perform a detailed analysis of the decoder and aligner, considering cues such as confidence scores (CS) [10, 11] and aligner probabilities (AP), respectively, obtained from token-level posterior probabilities. Experiments reveal that non-autoregressive ASRs are proven better suited for spontaneous and semi-spontaneous speaking, while autoregressive models perform better for memorized speaking. It is noticed that the conditional dependency nature of the decoder in autoregressive models leads to decoding the majority of grammatically incorrect utterances as correct, making them less suitable for the SGED task. Our analysis shows that CS is less reliable than AP in distinguishing between correctly and incorrectly decoded speech. However, AP is observed to suffer from alignment and overconfidence problems.

2. Motivation

Generally, proficiency in spoken grammar can be achieved through regular speaking practice. However, unlike grammar feedback on written content, the spoken grammar feedback needs to be provided immediately. But, in traditional classroom settings, it is often challenging to receive personalized feedback due to the large student-teacher ratio. As a result, there is a growing need for automated tools to support speaking practice in CALL systems that can assess spoken grammar. Generally, in the existing spoken English practices, the spoken grammar has been practised under three conditions: 1) spontaneous, 2) semi-spontaneous, and 3) memorized [3]. These three practice conditions were considered in the order of decreasing the amount of spontaneity (freedom) or increasing the amount of constraints (controls) imposed on the spoken content.

Spontaneous: Speaking practices that have the highest freedom of expression fall into this category. These kinds of speaking practices are essential for developing fluency and confidence in language use, particularly for advanced learners. In these exercises, learners are encouraged to construct sentences spontaneously, ensuring intended meaning with simple to complex sentences by adhering to grammatical rules. This practice often involves the ability to handle unexpected questions or topics, requiring quick thinking and adaptability and prepares the

10.21437/Interspeech.2025-2500

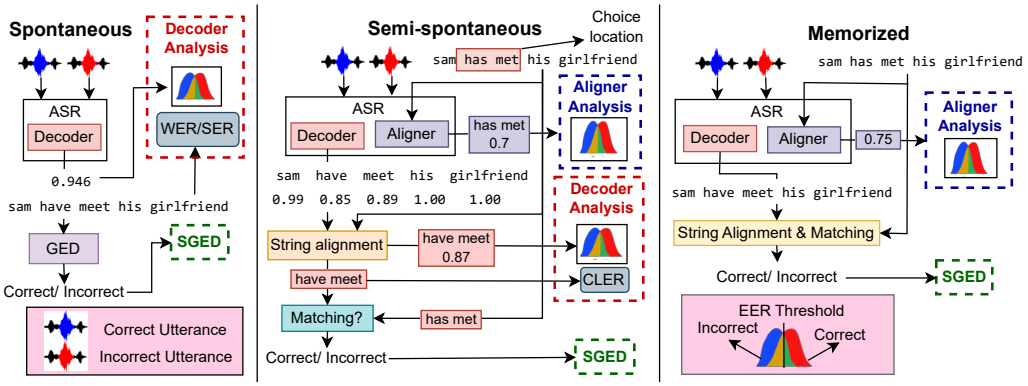

Figure 1: *Block diagram of proposed analysis (Numbers shown along with decoded text are CS and from aligner are AP)*

Table 1: *Details of the ASRs considered for the analysis (AED: attention-based encoder-decoder, CTC: connectionist temporal classification [19], SSL: self-supervised learning, CD: Conditional Dependency, AR: Auto Regressive)*

ASR	Type	Training	ASR Data (in Hours)	CD/AR
Whisper [20]	AED	Supervised	680 K	Yes
Wav2vec2 [21]	CTC	SSL+Supervised	960	No
Hubert [22]	CTC	SSL+Supervised	960	No
Nvidia-C [23]	CTC	Supervised	24.5 K	No
Vakyansh-C [24]	CTC	SSL+Supervised	700	No

learners for real-world communication. CALL systems such as GenieTutor [12] and SPELL [13] offer this type of speaking practice through dialogue-based conversations, where the learners are prompted to describe a scene or picture.

Semi-spontaneous: Speaking practices with controlled freedom of expression fall into this category and are beneficial for intermediate learners. These speaking practices are designed to enhance the repertoire of lexical, morphological, and syntactical elements, and to teach learners how to recombine them effectively. These exercises offer a balance between freedom and structure, where certain control points are introduced by the instructor to guide the learning process. CALL systems such as [14–18] provide exercises which require learners to construct sentences on the fly, either by ordering the words or by choosing the right word from multiple choices for a specified location and then speaking the sentence.

Memorized: Speaking practices focused on memorization with no freedom of expression fall into this category. These practices are designed to help learners internalise sentence structures by mimicking predefined sentences with a goal of learning and remembering how sentences are constructed and used. Such exercises are suited for beginner learners, who are just starting to learn the language, as they provide a structured way to build foundational skills.

In this work, we analyse the SGED under all three practice conditions, considering cues from five SOTA ASRs.

3. Background

E2E ASR systems are primarily classified into two categories: (1) autoregressive models and (2) non-autoregressive models. These models are composed of two key components: (1) an encoder that processes raw speech or features and converts them into acoustic representations; and (2) a decoder that transforms these acoustic representations into text units (tokens). The key distinction between these categories lies in their decoder in terms of handling the conditional dependence of the next token prediction on the current token. This dependency acts as an internal language model (LM) for maintaining syntax, such as grammar, in ASR predictions. This dependency exists mainly in autoregressive models. In this work, five ASR models are considered as indicated in Table 1.

4. Dataset

We consider the speech corpus reported in the work [10] for the analysis. The data contains recordings from 13 Indian English speakers (8 females and 5 males) from three different Indian states, reading grammatically correct and incorrect prompts. These prompts were prepared by English teachers with 5 years of experience, covering 37 grammar topics, each consisting of 30 sets of prompts. These 30 sets were equally distributed among three difficulty levels based on the grammatical complexity: easy, medium, and hard. For a given topic, there were 2 to 4 prompts in each set, out of which one is grammatically correct and the remaining are incorrect. The grammatically incorrect prompts were deduced from the correct prompt based on the grammatical topic. One such exemplary set for the topic of "present perfect" is: (1) sam *has met* his girlfriend - ground truth grammatically correct (GGC), (2) sam *has meet* his girlfriend, (3) sam *have met* his girlfriend, (4) sam *have meet* his girlfriend. Overall, the data contains a total of 4110 prompts, of which 1110 are grammatically correct and 3000 are incorrect. These 4110 sentences were read by 13 speakers, resulting in a total of 53,430 recordings. All recordings were manually verified, and no reading errors exist.

The same dataset is used for three speaking conditions: (1) for *Spontaneous*: all utterances are considered without their GGC text, as the the text is not known in advance in this case, (2) for *Semi-spontaneous*: all utterances are considered along with their corresponding GGC text and the right word in the choice location, since the freedom is at choice location only and GGC is known in advance in this case and (3) for *Memorized*: all utterances are considered along with their corresponding GGC text, since the GGC is known in advance in this case.

5. Experimental methodology

We analyse SGED performance and the role of decoder confidence scores and aligner probabilities in SGED under all three practice conditions, considering all five ASRs as illustrated in Figure 1. We consider large version 2 of Whisper and large versions of Wav2vec2 and Hubert models downloaded from HuggingFace's Transformers [25] library. Nvidia Nemo toolkit [26] is used to obtain NVIDIA-conformer (Nvidia-C) and Vakyansh-conformer (Vakyansh-C) models.

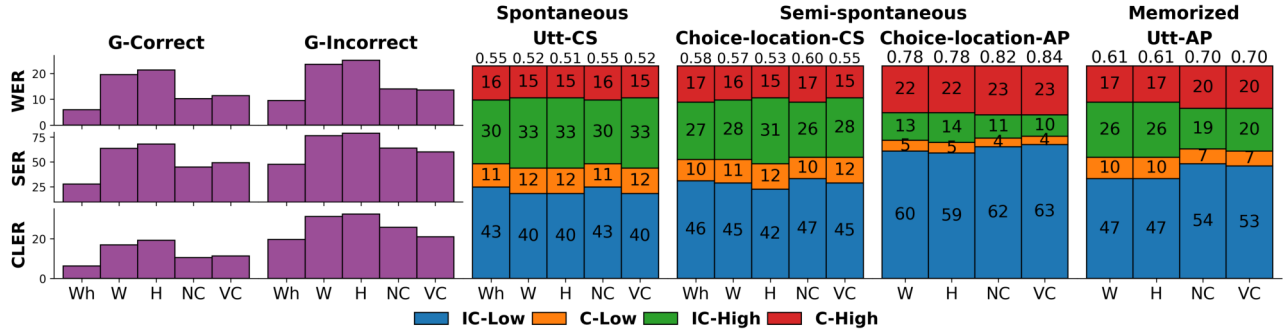

Figure 2: *Decoder and aligner based analysis of five ASRs (Wh:Whisper, W:Wav2Vec2, H:Hubert, NC:Nvidia-C, VC:Vakyansh-C), each stacked bar represents the percentage of CS/AP that are lower (Low) and higher (High) than the EER threshold under grammatically correct (C) and incorrect (IC) set of utterances*

Table 2: *Decoded text-based SGED performances in F1 scores separately for grammatically correct (G-C) & incorrect (G-IC) utterances and their average (Avg)*

ASR	Spontaneous			Semi-spontaneous			Memorized		
	G-IC	G-C	Avg	G-IC	G-C	Avg	G-IC	G-C	Avg
Whisper	0.59	0.52	0.56	0.91	0.81	0.86	0.91	0.74	0.82
Wav2vec2	0.67	0.47	0.57	0.91	0.77	0.84	0.88	0.50	0.69
Hubert	0.69	0.46	0.58	0.92	0.77	0.85	0.88	0.46	0.67
Nvidia-C	0.65	0.51	0.58	0.91	0.78	0.84	0.89	0.64	0.76
Vakyansh-C	0.70	0.52	0.61	0.93	0.82	0.87	0.90	0.63	0.77

5.1. Spontaneous speaking

SGED: We perform the SGED in the traditional ASR-GED pipeline framework as shown in the Figure 1. For GED, we use existing GEC model [27]. If decoded text is modified by GEC, we classify the decoded text as grammatically incorrect and vice versa. We compare SGED performance with F1 score.

Decoder based analysis: As SGED depends on ASR quality, WER (Word Error Rate) and SER (Sentence Error Rate) computed on the decoded text are analysed for ground-truth grammatically correct (GGC) and incorrect (GGIC) utterances. Also, utterance-level CS (Utt-CS) values from the decoder are analysed. The Utt-CS values are obtained by averaging the token-level CS values extracted from the decoder across the entire utterance. Typically, as the ASRs trained on GGC data, when an equal error rate (EER) criterion [28] is applied on the CS values as shown in Figure 1, it is expected that the percentage of GGC utterances whose CS values above EER threshold is more than that of GGIC. Based on this, we further classify the utterances based on the EER threshold criterion, i.e, utterances with CS value above threshold as grammatically correct and vice versa.

5.2. Semi-spontaneous speaking

SGED: As shown in the Figure 1, SGED is performed by aligning the decoded text of an utterance from ASR with the respective GGC text, and when the aligned words at the choice location are matched then the utterance is classified as grammatically correct else incorrect.

Decoder based analysis: Since the SGED performance in this condition depends on the error at the word choice location, we analyse the word errors at the choice location as indicated in the Figure 1 referred to as choice location error rate (CLER). Further, as similar to the spontaneous case, we also perform EER-based analysis using choice location word-level (CLWL) CS computed by averaging the token-level CS within the respective choice location word segment.

Aligner based analysis: Although the correct word at the choice location is known, the computation of decoder CS values does not consider it. In contrast, the aligner provides CLWL probabilities indicating the match between the spoken word segment and the respective right word transcription at the choice location. Thus, it is expected that for GGC utterance, the CLWL aligner probabilities are higher compared to GGIC as the spoken segment matches with the right word. EER-based analysis is performed on the CLWL aligner probabilities obtained from all five ASRs except Whisper, as it does not contain a CTC network and hence is not feasible to extract aligned probabilities.

5.3. Memorized speaking

SGED: In this, the GGC text is known, hence, we compute the string alignment distance between the decoded text and the GGC text. We classify the utterance as grammatically correct when the distance is zero and vice versa.

Aligner based analysis: As the AP indicate matching between the utterance and its respective transcription, we perform EER-based analysis on the probabilities obtained by aligning GGC and GGIC utterances with only GGC text.

6. Results and analysis

6.1. Spontaneous speaking

SGED performance: Table 2 shows SGED performance in terms of F1 scores for all three speaking conditions. Under *spontaneous speaking*, it is observed that, among all five ASRs, the Whisper model excelled in detecting GGC utterances along with Vakyansh-C. While the Vakyansh-C alone performed better in detecting GGIC utterances. On average, Vakyansh-C achieved a higher F1-score considering both GGC and GGIC utterances, indicating superior performance. The lower performance of Whisper is due to its bias towards decoding the grammatically correct text. It is observed that Whisper decodes 89% of GGC and 59% of GGIC utterances as grammatically correct. However, Vakyansh-C decodes a lower percentage, i.e., 67% of GGC and 44% of GGIC utterances as correct. The bias in Whisper may stem from the conditional dependency within Whisper's autoregressive model, where the decoder functions as an internal LM. As a result, autoregressive models like Whisper struggle to accurately decode grammatically incorrect utterances. Hence, a non-autoregressive model could be better for the SGED task since there is no conditional dependency.

Decoder based analysis: Figure 2 shows the WER and SER obtained from all five ASRs. The autoregressive model Whisper results in the least WER and SER for both GGC and GGIC utterances. However, Table 2 shows that Vakyansh-C achieves the highest F1-score. A plausible hypothesis is that the conditional dependency in Whisper causes it to mostly decode both

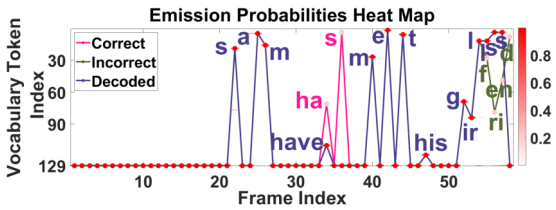

Figure 3: *Illustrative CTC paths (with heat map) across all the tokens for all the frames belonging to a grammatically incorrect utterance, whose prompt is "Sam have met his girlfriend", force-aligned separately with the text from following three cases 1) (Incorrect) same GGIC "Sam have met his girlfriend" 2) (Correct) its respective GGC prompt "Sam has met his girlfriend" and 3) (Decoded) decoded text from ASR "Sam have met his girls"*

GGC and GGIC utterances as grammatically correct. Whereas Vakyansh-C might decode most of the GGIC as another grammatically incorrect sentence. As a result, Whisper would classify a higher percentage of GGIC utterances as correct compared to Vakyansh-C. This observation suggests that the performance in WER and SER can't correlate with SGED performance. Also, the decoding patterns from non-autoregressive ASRs, which are conditionally independent, could be useful for SGED task.

From Figure 2, observing the stack of bars under the spontaneous case, it is noticed that all ASRs show a similar pattern in utterance-level CS despite the change in conditional dependency. A high percentage of GGIC utterances having CS scores above the EER threshold reveals that CS scores are not reliable measures for detecting grammatical errors. Moreover, the highest average F1 score achieved based on the EER criterion (shown at the top of each bar) is 0.55, which is 0.06 less than the best average F1 score achieved with SGED in Table 2.

6.2. Semi-spontaneous speaking

SGED performance: From Table 2, under *semi-spontaneous speaking*, SGED performance is higher across all five ASRs compared to the spontaneous condition. Vakyansh-C outperformed all other models, followed by Whisper. However, the performance differences between the ASRs are relatively small. This could be due to the result of the stringent constraint imposed on the speaking condition. This suggests that imposing constraints would be helpful in SGED.

Decoder based analysis: From Figure 2, observing CLER, it is noticed that Whisper is performing better than other models, with the second best being Vakyansh-C in GGC and GGIC utterances. In contrast, Vakyansh-C showed better F1-scores than Whisper in Table 2. This suggests that CLER doesn't correlate with SGED performance as similar to the conclusion from the spontaneous condition.

From Figure 2, observing a stack of bars under choice-location-CS of the semi-spontaneous case, a similar pattern is observed as in the spontaneous case across all ASRs, with the higher percentage of GGIC having CS scores above the EER threshold. Nvidia-C showed the best classification average F1 score, i.e. 0.60, but it is 0.27 lower than the average F1 score achieved using SGED in Table 2. It can be concluded that CS scores are not reliable for detecting grammatical errors in the semi-spontaneous condition.

Aligner based analysis: From Figure 2, observing the stack of bars under choice-location-AP of the semi-spontaneous case, the percentages of GGC utterances below (orange colour) EER threshold and GGIC above (green colour) EER threshold are lower, indicating better possible discrimination between GGC and GGIC utterances. Further, comparing average F1-scores (on top of each bar) obtained from EER-based classification and those from Table 2, AP-based F1-scores are lower. This indicates that the decoded text alone can be reliably used without the need for an additional alignment process.

6.3. Memorized speaking

SGED performance: From Table 2, observing F1-scores under *memorized speaking*, the SGED performance falls between spontaneous and semi-spontaneous condition. Further, Whisper outperforms all other ASRs. This highlights the benefit of autoregressive model for the SGED task when the text is known.

Aligner based analysis: Observing the F1-scores on top of bars stack under memorized case in Figure 2 and those from Table 2, the highest average F1-score with EER based classification on AP is lower than the highest decoded text based classification by 0.07 (i.e. Vakyansh-C). This indicates that the decoded text-based classification suffices across all ASRs. This suggests that aligner probabilities do not improve the performance further. We investigate the reasons considering an illustrative exemplary decoder and aligner path in Figure 3.

Detailed aligner path based analysis: From Figure 3, it is observed that most of the repetitions occur at the blank label, corresponding to the 129th token (the last token index). This indicates that the posterior probability is skewed at each frame, meaning only one token consistently receives a very high probability, indicating the decoder's overconfidence. As expected, force-aligned tokens that match the decoded tokens have higher probabilities, but mismatched tokens receive very low values due to this skewness (ex: has and friend).

It's important to note that the aligner has certain issues, such as 'have' being treated as a single token, whereas 'has' results from two tokens. In this case, repeated blank labels after 'have' compensate for the extra tokens. This misalignment causes subsequent tokens to not align properly, even if they match the ground truth. As a result, the aligned token probabilities decrease for the tokens facing this issue, making it difficult to identify tokens that match the ground truth. In conclusion, CTC-based aligners exhibit alignment challenges along with the high confidence of decoders, which limit the performance of CTC-based ASRs in tasks that rely on the alignment path.

7. Conclusion

In this work, we analyse the performance of five SOTA E2E ASR models, covering both autoregressive and non-autoregressive architectures, across three types of speaking practices: spontaneous, semi-spontaneous, and memorized speaking, for SGED. The experimental results indicate that non-autoregressive ASRs are better suited for SGED tasks due to the conditional independence nature of their decoder in spontaneous and semi-spontaneous conditions. Confidence scores from the decoder struggle to distinguish between grammatically correct and incorrect utterances in all three practice conditions compared to the traditional approach. In the case of semi-spontaneous and memorized practices, aligner probabilities demonstrated better performance at the word level but struggled at the sentence level due to alignment errors and decoder overconfidence. Overall, SGED based on decoded text proved more effective for all three conditions. The future scope of this work will be to analyse SGED at the word-level in spontaneous and memorized speaking conditions, both with and without disfluencies.

8. Acknowledgements

The authors would like to thank iHub Anubhuti IIITD Foundation for supporting this work.

9. References

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

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

[3] A. Chang, "Semi-free conversation: A beachhead leading to liberated spontaneous conversation," *The Journal of the Association of Teachers of Japanese*, vol. 10, no. 2/3, pp. 119–130, 1975. [Online]. Available: http://www.jstor.org/stable/489022

[4] 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.

[5] A. Caines, C. Bentz, K. Knill, M. Rei, and P. Buttery, "Grammatical error detection in transcriptions of spoken english," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2144–2162.

[6] S. Bannò, M. Rais, and M. Matassoni, "Grammatical Error Correction for L2 Speech Using Publicly Available Data," in *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 136–140.

[7] S. Bannò, R. Ma, M. Qian, K. M. Knill, and M. J. Gales, "Towards end-to-end spoken grammatical error correction," *arXiv preprint arXiv:2311.05550*, 2023.

[8] Y. Lu, M. J. Gales, and Y. Wang, "Spoken language'grammatical error correction'." ISCA, 2020.

[9] Y. Lu, M. J. Gales, K. Knill, P. Manakul, and Y. Wang, "Disfluency detection for spoken learner english." in *SLaTE*, 2019, pp. 74–78.

[10] 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 *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 41–45.

[11] Y. Lu, M. J. Gales, K. M. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of asr performance on spoken grammatical error detection." in *INTERSPEECH*, 2019, pp. 1876–1880.

[12] O.-W. Kwon, K. Lee, Y.-K. Kim, and Y. Lee, "Genietutor: a computer assisted second-language learning system based on semantic and grammar correctness evaluations," in *Critical CALL–proceedings of the 2015 EUROCALL conference, Padova, Italy*. Research-publishing. net, 2015, pp. 330–335.

[13] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[14] B. W. P. de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice in call: The effect of corrective feedback and education level in adult l2 learning," *Language Teaching Research*, vol. 24, no. 5, pp. 714–735, 2020. [Online]. Available: https://doi.org/10.1177/1362168818819027

[15] S. B. H. S. Bart Penning de Vries, Catia Cucchiarini 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. [Online]. Available: https://doi.org/10.1080/09588221.2014.889713

[16] H. Wang, C. J. Waple, and T. Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition," *Speech Communication*, vol. 51, no. 10, pp. 995–1005, 2009.

[17] C. Cucchiarini and H. Strik, "Second language learners' spoken discourse: Practice and corrective feedback through automatic speech recognition," in *Smart technologies: Breakthroughs in research and practice*. IGI Global, 2018, pp. 367–389.

[18] C. Cucchiarini, J. Van Doremalen, and H. Strik, "Disco: development and integration of speech technology into courseware for language learning." in *INTERSPEECH*, 2008, pp. 2791–2794.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[21] 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.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[23] "Stt en conformer-ctc large," https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_large.

[24] "Vakyansh open source models," https://github.com/Open-Speech-EkStep/vakyansh-models.

[25] "Huggingface's transformers," https://github.com/huggingface/transformers.

[26] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[27] "Gramformer," https://github.com/PrithivirajDamodaran/Gramformer.

[28] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.