WIP: Automated Speech Proficiency Assessment for Conversations on Technical Subjects

Akhila Yaragoppa
Human Technology Interaction Lab
Plaksha University
Mohali, India
akhila.yaragoppa@plaksha.edu.in

Utkarsh Agarwal
Human Technology Interaction Lab
Plaksha University
Mohali, India
utkarsh.agarwal@plaksha.edu.in

Arnav Rustagi
Human Technology Interaction Lab
Plaksha University
Mohali, India
arnav.rustagi@plaksha.edu.in

Anushka Desai
Human Technology Interaction Lab
Plaksha University
Mohali, India
anushka.desai@plaksha.edu.in

Siddharth
Human Technology Interaction Lab
Plaksha University
Mohali, India
siddharth.s@plaksha.edu.in

Brainerd Prince
Centre for Thinking, Language and Communication
Plaksha University
Mohali, India
brainerd.prince@plaksha.edu.in

Abstract*—This research-to-practice WIP paper presents a multimodal artificial intelligence (AI) system that assesses speech proficiency especially tuned for conversations on technical subjects particularly on the theme of AI. Critical thinking and clear communication skills are as important as technical career-specific skills. Automated speech proficiency tools only test for English speaking skills. However, the assessment of domain-specific technical communication skills is of utmost importance. The objective of this paper is to introduce an AI system for automated assessment of technical communication competency by providing an assessment for anyone desiring to converse on the theme of AI. The following three steps outline the approach followed in this paper. First, a multi-modal AI system has been designed by bringing together three pre-trained machine learning models each specializing in one domain related to speech proficiency estimation. Second, the multi-modal AI system is used to assess competence in proficiency of language communication in the context of technical conversations which particularly include: 1) content relevance 2) grammatical correctness, and 3) fluency of speech. Third, insights drawn from the multimodal system are compared to the speaker's self-perception of their proficiency. The uniqueness of our AI system is that it is able to combine the insights from these three models together towards a holistic speech proficiency analysis, especially for the technological domain. We validate the performance of our AI system under a controlled setting at a technological university. Our experiments led us to make the following discoveries about the use of our AI system to assess technical speech proficiency: 1) our AI system correctly analyzes that participants exhibit higher speech proficiency on topics in their domain of expertise, 2) grammatical correctness of speech remains unaffected when testing participants on topics with varied familiarity, and 3) participants' self-perception of their knowledge of various topics aligns with their ability to speak on the topic as measured by our AI system. We propose that our AI system would provide anyone interested in conversing proficiently on technological themes, especially AI, with the necessary tools to self-assess and track their progress.

***Keywords*— AI, technical communication, automated speech assessment**

I. Introduction

Assessments constitute an essential part of student learning. It helps students understand whether they can achieve their learning goals and keep track of their ongoing performance towards their learning goals. Various forms of assessments have been used in the past to evaluate student performance across disciplines such as written assignments, viva, and debates [1]. Verbal communication forms an important portion of all learning assessments used. In this paper, we focus on speech assessments.

Providing manual assessment in educational contexts, especially for descriptive assignments, is both challenging and time-consuming, thus not scalable. This also limits the number of assessments a student can take for self-evaluation purposes. Automated assessments can mitigate these issues.

Automated speech proficiency assessment is becoming a point of growing importance and interest in language learning because of the increasing numbers of English as a Second Language (ESL) learners worldwide. Early works on assessing speech proficiency used automatic speech recognition (ASR) outputs and prosodic analyses to calculate scores [2]. In recent years, fast-growing AI-inspired deep learning (DL) models have been applied to the task of speech proficiency assessment.

Automated assessment tools have been developed for speech proficiency assessments of English and other languages. Automated assessment tools have also been developed for various science domains [3], particularly for programming assessments in Computer Science [4] and written assessments in Biology [5]. However, to the best of our knowledge, no automated assessment tools have been built for assessing speech

proficiency for technical conversations such as, for a discourse on AI.

To address this gap, we introduce an automated tool for the assessment of speech proficiency, particularly for technical conversations. While in this paper we test our tool on discourses on AI themes, it can be adapted for discourses in any other technical field of study as well. We choose three unique AI speech proficiency models for a holistic analysis of technical speech proficiency. We begin by using a model that analyses content relevance by classifying speech into relevant or non-relevant categories corresponding to a topic. Furthermore, we employ two other models that can assess fluency and grammatical accuracy.

Our contributions in this research are the following:

- We show how existing AI models trained on the English language for automated speech proficiency assessment can be adopted to evaluate domain-specific technical communication skills for AI themes.
- We show that taking a multimodal approach to assess 1) content relevance, 2) grammatical correctness, and 3) fluency helps in providing holistic feedback to participants to improve their conversational skills on technical subjects.
- We show a positive correlation between participants' self-perception of proficiency and the multimodal AI system's scores affirming the AI system's efficacy.

II. Related Works

Automated assessments of spoken proficiency take a speech sample as input from the learner and predict a level of holistic proficiency or a specific feature of proficiency [6], [7]. Most traditional approaches design hand-crafted features from the audio input and feed these into classification models [8], [9], [10], [11]. Similar approaches are used to combine two or more such hand-crafted features to predict the proficiency level [12], [13], [14], [15]. The effectiveness of these methods heavily relies on the underlying assumptions made for the constituent handcrafted features and their effect on the proficiency level. For a more holistic grading of speech, in recent years, rapidly advancing deep learning models have been applied to the task of automated speech assessment. AI models either predict an overall proficiency level or a constituent feature determining proficiency [16], [17], [18], [19].

Sentence similarity matching methods have been used to measure the relevance of a text in comparison to a topic [20], [21], [22]. Apart from models trained for a specific task, in the recent past, large language models such as ChatGPT, and Llama have been trained to perform a wide variety of tasks based on a specific prompt, giving state-of-the-art performance on several language tasks. These large language models can also be prompt-engineered to predict content relevance.

Grammatical evaluation for speech is done in two steps. First, an Automatic Speech Recognition (ASR) system [25], [26], [27] is used to convert speech to text. Next, Grammar Correction tools such as Grammarly are used to correct the grammar of textual data.

Dysfluencies in speech such as sound repetitions, word repetitions, and blocks are common among everyone and are especially prevalent in people who stutter. Datasets such as SEP-28k [23] and UCLASS [24] have collected and annotated speech samples for disfluency analysis. Previous state-of-the-art works in disfluency detection [28], [29], [30] predict the presence of various types of disfluencies in a speech sample - such as sound repetition and prolongation. Fluency of speech can be understood as varying inversely with respect to disfluency.

III. Methodology

A. Data Collection

Participant information: The participants for this study were members of staff at Plaksha, a technological university, with diverse technical and non-technical backgrounds. The participants have degrees in engineering, management, liberal arts, or human resources.

Sample size: We collected data from N = 17 participants, who consented for their speech samples to be used for this study. 8 of the participants were female and 9 were male.

Speech recording: We used a smartphone audio recording application with an audio sampling rate of 48 kHz to record participant audio responses. Each speech sample was 1-2 minutes long.

Data collected: We asked three questions to each participant and recorded their response to each question. We also asked three survey questions to understand the participant's perception of their knowledge corresponding to each question. We chose each question based on the expected degree of familiarity our participants would have with the questions. The questions ranged from most familiar, somewhat familiar, to least familiar.

The three questions asked were:

Q1: Tell us what you know about AI/ML.

Q2: How do you think AI/ML is going to transform your job role?

Q3: How do you think ChatGPT works in the background?

Participant Survey: Participants were asked to rate themselves on a scale of 1-10 on each question. The three survey questions asked for self-rating were:

R1: Rate your knowledge of AI/ML.

R2: Rate your familiarity and knowledge of the working of AI/ML tools that can be used in your present job role to boost productivity.

R3: Rate your knowledge of the working of neural networks behind ChatGPT.

B. Automated Proficiency Assessment Models

In this section, we outline the three AI models used for analyzing speech input on three diverse aspects of speech proficiency. We used one speech-based feature, one meaning-based feature and one sentence-structure-based feature to build a holistic understanding of the spoken responses. We choose state-of-the-art models for each task, as described in the

"related works" section. Two of the tasks – grammatical correctness, and disfluency – are performed using Deep Learning models trained for each specific task respectively. The content relevance task prompts a large language model (LLM) to perform the task.

- **Content Relevance**: We used the GPT-4o model to analyse the content relevance. We first transcribed the speech samples to text using an ASR model [25]. We prompt engineered the GPT-4o model which can be accessed at https://openai.com/index/hello-gpt-4o/, to classify each response as "relevant" or "non-relevant" with respect to the question asked. To evaluate the performance of the model, the responses from the LLM were compared with human ratings on the same task, for a subset of 30 speech samples, and an accuracy of 0.9 was observed on this subset.

- **Grammatical Correctness**: We first converted the speech sample to text - using an ASR model as in [25]. We use the " base" model from [25]. We then passed this text to an open-source grammar correction model called LanguageTool which can be accessed at https://languagetool.org/. The final output from the model gave the corrected grammar along with the number of grammatical errors. We then calculated percentage grammatical correctness as 100 * (total number of words – number of errors) / (total number of words). The ASR model used reports a low word error rate while also offering fast computation. The grammar correction model LanguageTool, while does not report performance metrics, is a popular open-source tool.

- **Disfluency**: We implemented the disfluency model suggested in [30] and trained our model on the SEP-28k dataset [23]. Before inputting data, the data was split into audio samples of 3s each. For each sample, the model predicted the presence of five types of disfluencies. We combined three of the stutter types – block, prolongation, and interjection to obtain a new stutter variable, indicating the presence of stutter in that 3s snippet. We computed the stutter ratio of the model as a ratio of the number 3s positive stutter predictions, and the total number of 3s samples. [30] reports an overall accuracy of 50.79 for their disfluency model.

IV. Results

In this section, we present the results obtained from our multimodal AI system on our test set of 17 participants. We first show the efficacy of our content relevance model. "Fig. 1" plots the percentage of relevant responses given by participants for each question. We observe that a higher percentage of participants give relevant answers when asked about their field, as compared to when asked questions about a field chosen at random. We therefore infer that most participants are aware of their own field and the influence of AI on it. However, not many participants are familiar with the general field of AI and lack in-depth knowledge of the machine learning models behind it.

Next, we look at predictions from the disfluency model, considering the relevance of answers to the question asked. "Fig. 2" shows that participants tend to stutter a lot more when their answers are irrelevant, compared to when they give relevant answers. This can be explained by their uneasiness in speaking on a topic they are unfamiliar with and our AI system's disfluency model was able to detect this trend. One outlier is observed in the graph for a participant answering Q1, where the answer is detected to be non-relevant. This is a case of incorrect model prediction, due to inaccuracies in the model. This shows that for AI speech, fluency is positively correlated to overall proficiency, aligning with the literature on general speech proficiency.

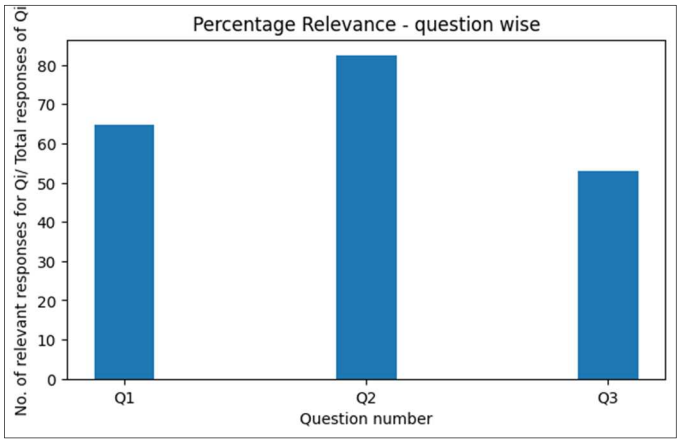

Fig. 1. Bar graph showing the question-wise percentage of relevant responses.

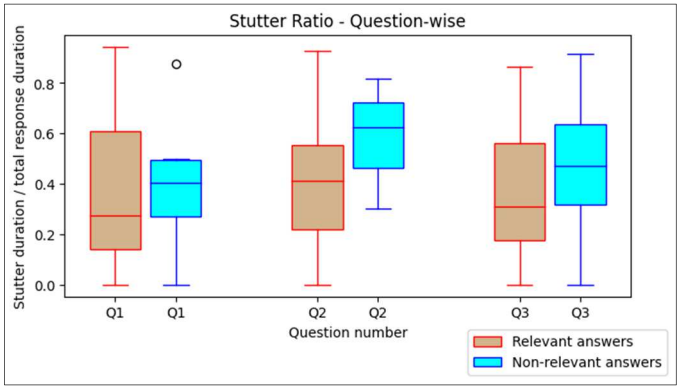

Fig. 2. Boxplot of stutter ratio distribution for relevant and non-relevant answers.

We then look at prediction from the grammar correction model, considering the relevance of answers. "Fig. 3" shows no pattern for the variation of the grammatical scores between questions or relevance of answers. The variations observed in the grammar scores can be attributed to the varying grammatical proficiencies of the speakers and to the AI model's accuracy. This points to the observation that grammatical correctness does not have any effect on a person's ability to converse on unfamiliar topics. Looking further into the 3 outliers in this figure we observe the following. The datapoint with the low grammar score & non-relevant answer is not far from a relevant

answer, where such a grammar score would not be considered an outlier. The datapoint with the high grammar score and non-relevant answer shows that the participant does not attempt to answer the question and says they do not know the answer. The final datapoint for Q1 with a relevant answer is an outlier because the participant seems to partially know the answer but possibly doubts themself before answering, leading them to make grammatical mistakes, repeat words, and answer in one very long sentence. Although this is what we initially expected would happen to the larger population when they are not confident about the answer, however, as observed through the data, most people were able to answer comfortably to whatever degree they were aware of the answer. The inferences made here for grammar score indicate that for technical speech proficiency, grammar is not an important indicator. This observation is not in alignment with the previous literature on English speaking proficiency.

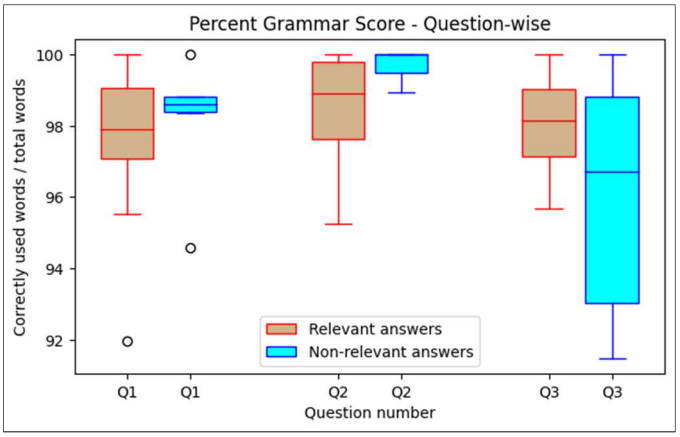

Fig. 3. Boxplot of grammar score distribution for relevant and non-relevant answers.

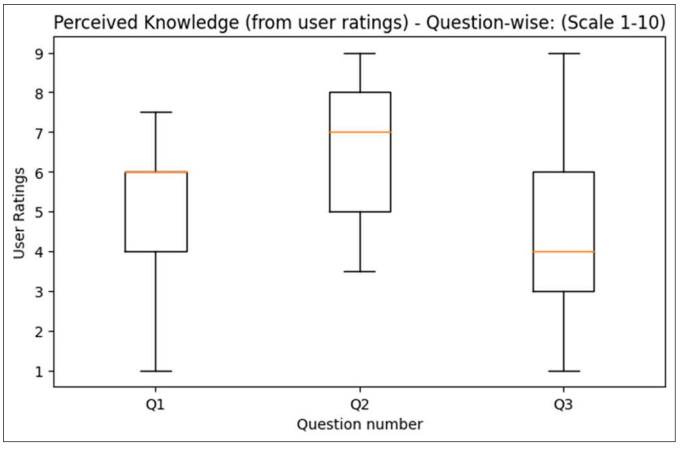

Fig. 4. Boxplot of question-wise perceived knowledge distribution.

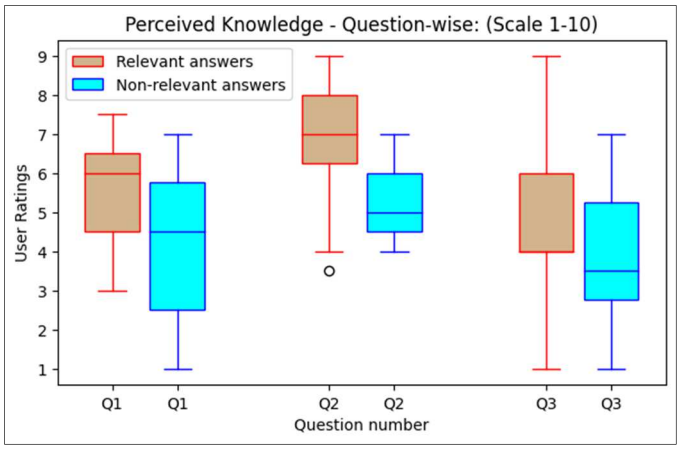

Fig. 5. Boxplot of perceived knowledge distribution for relevant and non-relevant answers.

Finally, we look at the perceived self-proficiency of participants on the three questions they were asked. “Fig. 4” shows the distribution of the self-ratings. We observe that the distribution of the user’s perceived self-knowledge of each question shows a positive correlation with the relevance scores obtained from the model as in “Fig. 1”. This shows that the relevance model can competently detect relevance in answers, as attested by the users themselves. In “Fig. 5” we see that the perceived knowledge of the participants giving relevant answers is on average higher than for participants giving non-relevant answers. This demonstrates that people are generally self-aware of their knowledge of different AI topics. This result also indicates that relevance is a good indicator of topic familiarity, as supported by English proficiency literature.

V. Conclusion and Future Work

The ability to think clearly and communicate proficiently in your domain of study is of utmost importance. Automated tools for evaluating domain-specific speech proficiency do not exist. We introduce a multimodal AI system based on existing speech proficiency models, which can be used to test domain-specific speech proficiency, particularly for AI discourse. We show that participants exhibit higher speech proficiency on topics in their domain of expertise. We also observe that grammatical correctness of speech remains unaffected by changing familiarity of topics. We finally show that participants’ self-perception of their knowledge of various topics aligns with their ability to speak on the topic.

This work-in-progress pilot study showed us that content relevance and fluency are highly indicative of technical speech proficiency, whereas grammatical correctness adds little knowledge to speech proficiency analysis. In the future, we plan to exclude grammatical accuracy when analyzing technical speech proficiency. We also plan to further improve the AI models used in this study, by finetuning them on more technological speech data. Furthermore, the interpretability and usefulness of the AI system can be improved by including more AI models that can predict speech parameters such as

intonation, pacing, speech rate, and content parameters such as clarity of thought, ease of understanding.

REFERENCES

[1] D. Hounsell et al., “An analytical review of the literature,” 2007.

[2] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, “Automatic scoring of non-native spontaneous speech in tests of spoken English,” Speech Communication, vol. 51, no. 10, pp. 883–895, Oct. 2009, doi: 10.1016/j.specom.2009.04.009.

[3] O. L. Liu, J. A. Rios, M. Heilman, L. Gerard, and M. C. Linn, “Validation of automated scoring of science assessments,” Journal of Research in Science Teaching, vol. 53, no. 2, pp. 215–233, 2016, doi: 10.1002/tea.21299.

[4] J. C. Paiva, J. P. Leal, and Á. Figueira, “Automated Assessment in Computer Science Education: A State-of-the-Art Review,” ACM Trans. Comput. Educ., vol. 22, no. 3, p. 34:1-34:40, Jun. 2022, doi: 10.1145/3513140.

[5] R. H. Nehm, M. Ha, and E. Mayfield, “Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations,” J Sci Educ Technol, vol. 21, no. 1, pp. 183–196, Feb. 2012, doi: 10.1007/s10956-011-9300-9.

[6] N. Iwashita, A. Brown, T. McNamara, and S. O’Hagan, “Assessed Levels of Second Language Speaking Proficiency: How Distinct?,” Applied Linguistics, vol. 29, no. 1, pp. 24–49, Mar. 2008, doi: 10.1093/applin/amm017.

[7] N. H. de Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, and J. H. Hulstijn, “FACETS OF SPEAKING PROFICIENCY,” Studies in Second Language Acquisition, vol. 34, no. 1, pp. 5–34, Mar. 2012, doi: 10.1017/S0272263111000489

[8] L. Chen, K. Evanini, and X. Sun, “Assessment of non-native speech using vowel space characteristics,” in 2010 IEEE Spoken Language Technology Workshop, Dec. 2010, pp. 139–144. doi: 10.1109/SLT.2010.5700836

[9] E. Coutinh et al., "Assessing the prosody of non-native speakers of English: Measures and feature sets", Proceedings of 10th International Conference on Language Resources and Evaluation (LREC'16), 2016.

[10] S. Bhat and S.-Y. Yoon, “Automatic assessment of syntactic complexity for spontaneous speech scoring,” Speech Communication, vol. 67, pp. 42–57, Mar. 2015, doi: 10.1016/j.specom.2014.09.005.

[11] H. Strik and C. Cucchiarini, "Automatic assessment of second language learners' fluency", Proceedings of the International Congress of Phonetic Sciences (ICPhS), 1999.

[12] P. Müller, F. D. Wet, C. V. D. Walt, and T. Niesler, “Automatically assessing the oral proficiency of proficient L2 speakers,” in Speech and Language Technology in Education (SLaTE 2009), ISCA, Sep. 2009, pp. 29–32. doi: 10.21437/SLaTE.2009-8.

[13] Y. Wang et al., “Towards automatic assessment of spontaneous spoken English,” Speech Communication, vol. 104, pp. 47–56, Nov. 2018, doi: 10.1016/j.specom.2018.09.002.

[14] Z. Liu et al., “Dolphin: A Spoken Language Proficiency Assessment System for Elementary Education,” in Proceedings of The Web Conference 2020, in WWW ’20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 2641–2647. doi: 10.1145/3366423.3380018.

[15] S. Crossley and D. McNamara, “Applications of Text Analysis Tools for Spoken Response Grading,” Language Learning & Technology, vol. 17, no. 2, pp. 171–192, Jun. 2013.

[16] S. Bannò and M. Matassoni, “Proficiency Assessment of L2 Spoken English Using Wav2Vec 2.0,” in 2022 IEEE Spoken Language Technology Workshop (SLT), Jan. 2023, pp. 1088–1095. doi: 10.1109/SLT54892.2023.10023019.

[17] K. Takai, P. Heracleous, K. Yasuda, and A. Yoneyama, “Deep Learning-Based Automatic Pronunciation Assessment for Second Language Learners,” in HCI International 2020 - Posters, C. Stephanidis and M. Antona, Eds., Cham: Springer International Publishing, 2020, pp. 338–342. doi: 10.1007/978-3-030-50729-9_48.

[18] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, “ASR-Free Pronunciation Assessment.” arXiv, May 24, 2020. doi: 10.48550/arXiv.2005.11902.

[19] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, “End-to-End Neural Network Based Automated Speech Scoring,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 6234–6238. doi: 10.1109/ICASSP.2018.8462562.

[20] J. Mueller and A. Thyagarajan, “Siamese Recurrent Architectures for Learning Sentence Similarity,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, Art. no. 1, Mar. 2016, doi: 10.1609/aaai.v30i1.10350.

[21] B. Hu, Z. Lu, H. Li, and Q. Chen, “Convolutional Neural Network Architectures for Matching Natural Language Sentences,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2014.

[22] S.-Y. Yoon and C. M. Lee, “Content Modeling for Automated Oral Proficiency Scoring System,” in Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, and T. Zesch, Eds., Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 394–401. doi: 10.18653/v1/W19-4441.

[23] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, “SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, pp. 6798–6802. doi: 10.1109/ICASSP39728.2021.9413520.

[24] P. Howell, S. Davis, and J. Bartrip, “The University College London Archive of Stuttered Speech (UCLASS),” Journal of Speech, Language, and Hearing Research, vol. 52, no. 2, pp. 556–569, Apr. 2009, doi: 10.1044/1092-4388(2009/07-0129).

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in Proceedings of the 40th International Conference on Machine Learning, PMLR, Jul. 2023, pp. 28492–28518.

[26] A. Hannun et al., “Deep Speech: Scaling up end-to-end speech recognition.” arXiv, Dec. 19, 2014. doi: 10.48550/arXiv.1412.5567.

[27] C. Wang et al., “fairseq S2T: Fast Speech-to-Text Modeling with fairseq.” arXiv, Jun. 14, 2022. doi: 10.48550/arXiv.2010.05171.

[28] M. Jouaiti and K. Dautenhahn, “Dysfluency Classification in Stuttered Speech Using Deep Learning for Real-Time Applications,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 6482–6486. doi: 10.1109/ICASSP43922.2022.9746638.

[29] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, “Introducing ECAPA-TDNN and Wav2Vec2.0 Embeddings to Stuttering Detection.” arXiv, Apr. 04, 2022. doi: 10.48550/arXiv.2204.01564.

[30] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, “StutterNet: Stuttering Detection Using Time Delay Neural Network,” in 2021 29th European Signal Processing Conference (EUSIPCO), Aug. 2021, pp. 426–430. doi: 10.23919/EUSIPCO54536.2021.9616063.