UNIFIED SPOKEN LANGUAGE PROFICIENCY ASSESSMENT SYSTEM

Sunil Kumar Kopparapu[†], *Ashish Panda*

TCS Research, Tata Consultancy Services Limited, Mumbai, India.

ABSTRACT

Language proficiency assessment is a common requirement for L2 speaker of English. There exists separate tools for spoken language assessment (SLA) to assess articulation in terms of pronunciation or oral fluency and for written language assessment (WLA) to assess language grammar and vocabulary. In this paper, we present a SLA system, called unified spoken language proficiency assessment (USLPA) system which is capable of assessing all aspects of language under one hood. From practical usage perspective, the language proficiency test should be easy to administer, should not be repetitive, should be automatic, should be short in duration, and should be reliable and robust in its assessment. The USLPA system has added benefits like removal of inherent human bias in the assessment process, allowing for a greater flexibility of assessment, and dynamic assessment instead of the same assessment tests for all candidates. In this paper, we propose an automatic end-to-end USLPA system to assess language proficiency.

Index Terms— Speech Analysis, Spoken Language Assessment, Spoken Language Proficiency

1. INTRODUCTION

Automatic speech quality and language assessment is crucial not only for services oriented businesses where there is a need for their human agents to voice interact with their customers to resolve their problems but also as a language proficiency assessment tool that is commonly required to assess L2 speaker of English [1, 2]. Spoken language assessment has two aspects (a) quality of articulation of speech in terms of pronunciation [3–5], (b) 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], and (c) language grammar, vocabulary. Today, the speech quality assessment and language grammar and vocabulary assessment are administered as two separate tests using spoken language assessment (SLA) and written language assessment (WLA) tools respectively; together they assess the complete language proficiency. The use of SLA and WLA separately not only makes the language proficiency test lengthier but also allows the learner to be slack in terms of language grammar, for example, during a SLA test.

Speech quality assessment is more of a soft-skill assessment and does not depend on any academic qualification unlike language assessment. A typical SLA would involve either a face-to-face or a telephonic interaction between the candidate and the expert; lasting for about 7-10 minutes. The metrics used for assessment and the decision are mostly business driven and is prone to human expert biases. However, language assessment, generally administered through WLA tools, are fairly automated because of the robust natural language text processing tools available to evaluate language grammar from written text. One of the reason for a wide gap between SLA and WLA tools is because of the differences between spoken and written language text [11]. The main reason SLA systems do not cater to language grammar and vocabulary assessment is most often to do with the tools available for speech analysis which are not as robust as tools available for text analysis. For grammar assessment, as an example, the speech recognition engine needs to be able to determine the precise sequence of words spoken spontaneously. However, most automatic speech recognition (ASR) engines are less than perfect when transcribing spontaneous speech; this leads to erroneous language assessment scores.

In practical scenarios like voice based call centers agent interactions or during online or virtual employment interviews, it is, primarily, the spoken language communication that is dominant. This practical requirement motivates the need to assess all aspects of language proficiency, namely speech quality and language grammar and vocabulary through a single SLA system. Some of the requirements of such a system are (a) it should not be repetitive, meaning the test administered should have a large variety of administered tests, (b) should be as short as possible, meaning a single test to assess language proficiency, (c) should be economical, meaning fully automatic with no human involvement, and (d) be devoid of human biases. In this paper, we propose and describe an end-to-end automatic speech quality and spoken language assessment system called USLPA.

The main contribution of the paper is designing and implementing an USLPA system that can robustly evaluate all aspects of language proficiency, without employing additional WLA tools, thereby significantly reducing the time taken to take the test. We make use of our earlier work of (a) proposing

[†]Corresponding Author (SunilKumar.Kopparapu@TCS.Com)

a mechanism to incorporate language grammar assessment by exploiting the superior performance of available speech analysis tools on *read speech*, (b) automatic grammar and vocabulary assessment using a custom built language model on top of a readily available hybrid ASR system [12], (c) proposing grammar scoring module that is robust to errors in ASR, (d) employing LLM to to bring in variations [12] in the test to make the SLA system largely unteachable thus making it scalable and practical. The rest of the paper is organized as follows, we describe an end-to-end automatic unified spoken language proficiency assessment system in detail in Section 2 and describe in brief its implementation in Section 3. We elaborate, in Section 4, an assessment system of the future keeping GenZ in mind and conclude in Section 5.

2. USLPA SYSTEM

The complete end-to-end automatic unified spoken language proficiency assessment system is shown in Fig. 1. Our system consists of three parts, (B1) dynamic text generation using a large language model (LLM) which generates a wide variety of non-repetitive paragraphs P which makes it possible to administer unique assessment test [12], (B2) a spoken language assessment module that assess language grammar and vocabulary from spoken speech $S(t)$ [12], and (B3) a speech quality assessment module.

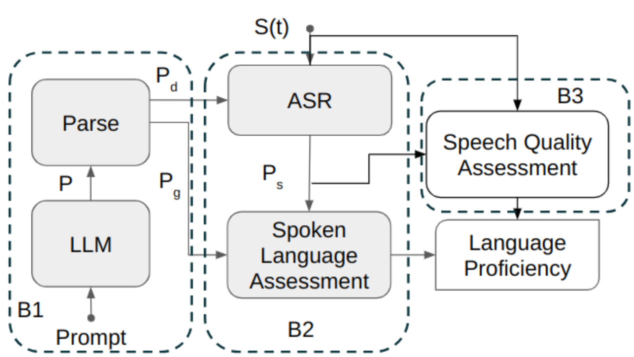


Fig. 1: USLPA System.

2.1. Generating large variety of P using ChatGPT

We adopt 1-shot learning prompting style for generating new paragraphs $(P_1, P_2, \cdots)$ as described in Fig. 2 [12]. We provide a sample paragraph which has both correct and incorrect grammar and vocabulary options to ChatGPT. As can be seen in Fig. 3, the tags "<grammar> </grammar>" and "<vocabulary> </vocabulary>" corresponds to the words or phrases that are to be evaluated for grammar and vocabulary respectively. The tag "<correct> </correct>" shows the correct choice. In practice, both P_d (Fig. 3b) and P_g (Fig. 3c) can be extracted by applying simple text parser on P (Fig. 3a) . A wide variety of P's can be generated using the prompt *"Generate paragraph. With subject <subject>."* This allows for the generation of a completely new paragraph in the desired format for every prompt; sample generated P is shown in Fig. 4a, and 4b. Note that the use of large language model (LLM) makes USLPA system scalable and practical because no two assessment instances are the same; making it very hard for the student to be coached for the assessment.

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

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

A student, who is to be assessed, is shown P_d and given a finite amount of time to familiarize herself with the displayed paragraph on her computer screen. Post that she reads into the microphone the displayed text (Fig. 3b) by making an appropriate choice from among the three possibilities enclosed in "(" and ")". This *read* speech $S(t)$ (see Fig. 1) is used for both speech quality and language assessment.

2.2. Language Assessment

The student is shown a paragraph P_d on a web interface as shown in Fig. 5a. As mentioned earlier, she has to choose one of the words from among the three shown in parenthesis. The correct choice of words determine the language grammar and vocabulary proficiency of the student.

The language assessment system (B2) first uses the paragraph P that is displayed to the student to build a customized language model (CLM) to be used along with a hybrid automatic speech recognition (ASR) engine. The performance of ASR-CLM for assessment of language grammar and vocabulary is far superior than the state-of-the-art end-to-end ASR system like `whisper` [13] as reported in [12].

As shown in Fig. 5b, using $P_s = \text{ASR-CLM}(S(t))$ the transcript of $S(t)$, the language (grammar and vocabulary) score S_l is computed by extracting the set of words that are in P_d but not in the transcript P_s, say $p_1 = \{w \in P_d \mid w \notin P_s\}$ and then extracting $p_2 = \{w \in G_w \mid w \notin p_1\}$; the language score, $S_l = |p_2|$ is the cardinality of the set p_2 (see [12] for complete details). Here G_w is the list of words in P that are

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

(a) Sample paragraph for prompting 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. 3: A sample P usage.

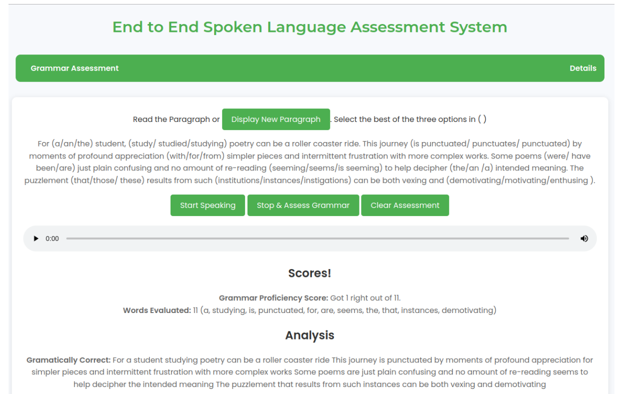
In <grammar><correct>an</correct>/a/the</grammar> bustling city, <grammar>exploring/ explored/ <correct>exploration</correct> </grammar> can be an exciting adventure. **<snip>** The challenge <grammar><correct>that</correct>/ those/these</grammar> comes from such <grammar>adventures/<correct>explorations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

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

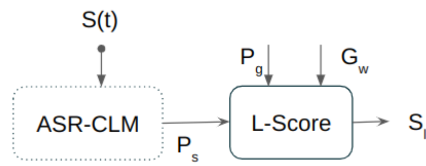
For <grammar><correct>an</correct>/a/the</grammar> physics enthusiast, <grammar>studying/ studied/ <correct>studying</correct> </grammar> physics can be a fascinating journey. **<snip>** The understanding <grammar><correct>that</correct>/ those/ these</grammar> comes from such <grammar>endeavors/<correct>pursuits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

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

Fig. 4: Paragraph's generated by prompting ChatGPT.



(a) Interface.



(b) Language Assessment using custom language ASR.

Fig. 5: Spoken Language Assessment.

being evaluated for language, namely, all the words between the "<comment>" tags. Note that both P_d and G_w can be obtained by text parsing P. Overall, the language assessment (see Fig. 5b) takes $S(t)$, P_d, and G_w as input and produces a score S_l. Namely, $S_l = \text{L-SCORE}(P_s, P_d, G_w)$.

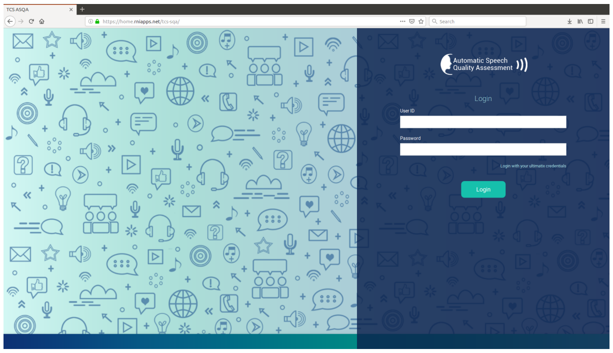
2.3. Speech Quality Assessment

The *read* speech $S(t)$ that was used for language assessment is used for speech quality assessment as well. Unlike the spoken language grammar assessment, there are several SLA systems like [14, 15] that assess speech quality in terms of pronunciation, speaking rate, stress, oral fluency, and emotion. The speech quality assessment interface, developed in-house, is shown in Fig 6a and the parameters that are measured, namely, pronunciation (Pr) [16,17], speaking rate (Sr) [16, 18, 19], stress [20] (Ln), oral fluency [21] (Of), emotion [22] (Ta) which have been developed in-house based on published literature are shown in Fig 6b.

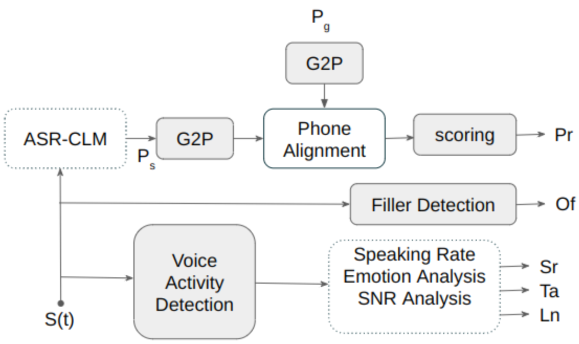
3. IMPLEMENTATION

The USLPA is implemented as a web application. The backend architecture of USLPA is a base Flask Framework; and is built using Flask[1], a lightweight web framework for Python. We used `kaldi` based ASR [23] with a custom built language model (ASR-CLM). The acoustic model of the ASR-CLM was trained on 960 hours of speech data from Librispeech

[1]https://flask.palletsprojects.com/en/3.0.x/



(a) Interface.



(b) Parameters Assessed.

Fig. 6: Speech Quality Assessment.

database [24]. The CLM is trained on the text encompassing all possible variations of the given paragraph P displayed to the student on the fly. For example if a sentence P contained "*For (a/an/the) student, (study/studied/studying) poetry can be a roller coaster ride.*" then all nine possible combination, namely, "$\cdots$ *a* $\cdots$ *study* $\cdots$", "$\cdots$ *an* $\cdots$ *study* $\cdots$", $\cdots$, "$\cdots$ *the* $\cdots$ *studying* $\cdots$" are used to train the language model. A CLM [25] trained to include all possible variations (including the grammatically wrong ones) of the sentence to allows the ASR-CLM to correctly recognize grammatically *incorrect* sentences when spoken by the student. On the other hand `whisper` with a general purpose language model "corrects" grammatically incorrect sentences because *only* grammatically correct sentences are part of `whisper`'s training set [12].

The front-end is based on HTML, CSS, and JavaScript and the web-app allows users to record audio via the laptop microphone using JavaScript, which is saved and transcribed and the HTML Templates uses Flask's Jinja2 templating engine to render dynamic content, including the paragraph (P) and language (grammar and vocabulary) and speech quality assessment (pronunciation score etc) results.

As this time, the spoken grammar assessment (Fig. 5) and speech quality assessment (Fig. 6) are two different interfaces, however they work on the same paragraph P and use the same audio recording of the student, namely, $S(t)$ for assessing the grammar score, goodness of pronunciation, speaking rate, tonality, loudness, etc. We are in the process of integrating these two interfaces to enable a single interface for unified language assessment (USLPA).

4. LOOKING FORWARD: USLPA++

While several tools like [14, 15] for speech quality assessment exist there are none, to the best of our knowledge, for spoken language assessment. However, almost all tools are web interfaces, display a paragraph of text, record the student's speech using a microphone, and analyze the recorded speech. These systems while being usable, are not very attractive especially considering the fact that most of the GenZ students are not only quite deft at using hand held mobile devices but also have a very short attention span. Which means they are not equipped to take lengthy assessment test in one sitting.

We hypothesize that going forward, an USLPA would be a mobile application (along the lines of [26]) as shown in Fig. 7 called USLPA++. The interface would provide several *picture icons* for the student to select. Each *picture icon* could result in either the generation of a new text paragraph (P) or an image of a word-cloud as shown in Fig. 8; the displayed paragraph or word-cloud would assess a very specific attribute of the language without the student being aware of what aspect of the language is being assessed. For example, a selected *picture icon* could lead to evaluating language grammar, while another *picture icon* could lead to evaluating of oral fluency, and a third *picture icon* could lead to evaluating pronunciation, and so on.

As an example, to keep up with the diminished attention span of the student, on selection a *picture icon*, USLPA++ could show the student Fig. 8 and ask the student to form a correct sentence using *only* the words seen in the word-cloud. Evaluation for language is along the lines described earlier, using the correct sentence *"I went to the movies having purchased the tickets online"* to assess the language. This process of introducing *variety*, would help retain the interest of the digitally savvy students while administering the language assessment test.

Transforming the assessment onto a mobile application can have some additional benefits, for example some important metrics like *"How much time did the student take to assimilate the information that was displayed to him and construct the statement?"* can also be measured as part of the assessment process, if required. While this might not be directly related to language proficiency, it can assess an aspect of the candidate, that is not measured explicitly today, but useful for an enterprise to identify candidates that demonstrate an *"ability to think on their feet"*. The essential idea of USLPA++ is to introduce novel mechanisms for language assessment using the available infrastructure (mobile | tablets) and structuring assessment modules that are not repetitive and more fun to

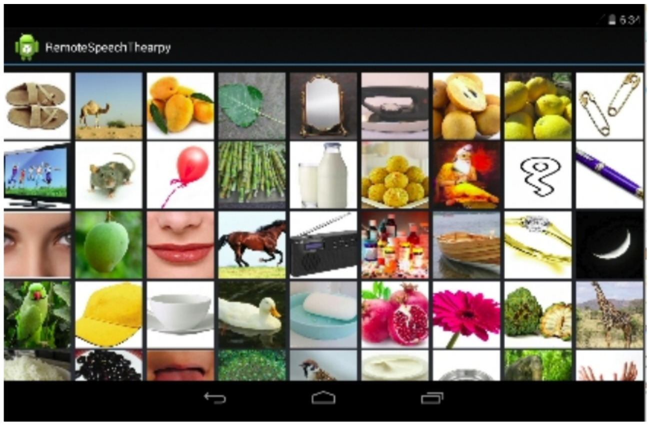


Fig. 7: USLPA++: A mobile application to allow the student choose a picture icon. Each icon could result in assessing different aspect of the language.

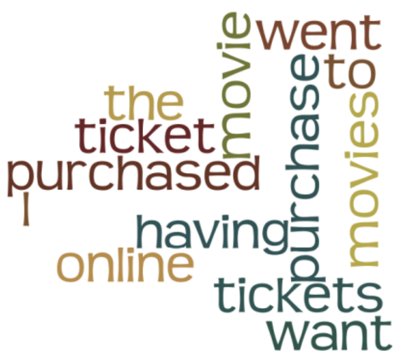


Fig. 8: Word Cloud instead of a text paragraph.

use to cater to a cohort that is shot of attention span.

5. CONCLUSIONS

Language proficiency assessment is a common requirement for L2 speaker of English. There exists several SLA assessment tools to analyze speech (pronunciation, oral fluency, etc) but none of these assessment tests venture into assessing language grammar; instead they depend on WLA systems. In this paper, we designed and implemented a practical, scalable and a robust USLPA system that is capable of assessing both speech quality (pronunciation, oral fluency) and spoken language (grammar, vocabulary) assessment. Thereby unifying the assessment of language into a single system. Most of the language assessment tools suffer from repeating the text used in assessment, often choosing from a large bank of text paragraphs, for evoking spoken utterance (speech) from a student. The proposed system overcomes this by making use of the capability of Generative AI (LLM's) to enable generation of text paragraphs that are largely non-repetitive thereby making the proposed system hard to be memorized by students. We further propose a language proficiency assessment system for the new digital generation keeping their short attention span and their ability to interact with digital devices like smart phones in mind. The idea is to make the assessment process fun for the student without sacrificing the strictness in the assessment process.

6. ACKNOWLEDGEMENTS

Several former and current members of TCS Research have contributed to the development of speech analysis tools over a long period of time. Several of these find place in the USLPA system.

7. REFERENCES

[1] Lixian Jin and Hua 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] Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda, "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] Kavita Sheoran, Arpit Bajgoti, Rishik Gupta, Nishtha Jatana, Geetika Dhand, Charu Gupta, Pankaj Dadheech, Umar Yahya, and Nagender Aneja, "Pronunciation scoring with goodness of pronunciation and dynamic time warping," *IEEE Access*, vol. 11, pp. 15485–15495, 2023.

[4] Hao-Chen Pei, Hao Fang, Xin Luo, and Xin-Shun Xu, "Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 554–563, 2024.

[5] Binghuai Lin and Liyuan Wang, "Exploiting information from native data for non-native automatic pronunciation assessment," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 708–714.

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

[7] S Pavankumar Dubagunta, Edoardo Moneta, Eleni Theocharopoulos, and Mathew 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] Ahmed Imran, Meghna Pandharipande, and Sunil Kumar 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] Sunil Kumar Kopparapu, *Non-linguistic analysis of call center conversations*, Springer, 2015.

[10] Juan Pablo Arias, Nestor Becerra Yoma, and Hiram Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[11] Kaustuv Mukherji, Meghna Pandharipande, and Sunil Kumar Kopparapu, "Improved language models for asr using written language text," in *2022 National Conference on Communications (NCC)*, 2022, pp. 362–366.

[12] Sunil Kumar Kopparapu, Chitralekha Bhat, and Ashish Panda, "Spoken grammar assessment using LLM," `https://arxiv.org/abs/2410.01579`, 2024.

[13] OpenAI, "Whisper: Openai's automatic speech recognition (asr) system," `https://github.com/openai/whisper`, 2022, Model: Whisper Tiny.en, Accessed: 2023-06-26.

[14] Microsoft, "Pronunciation assessment tool," `https://speech.microsoft.com/portal/pronunciationassessmenttool`, 2024, Accessed: 2024-06-26.

[15] SpeechAce, "Speechace pronunciation assessment tool," 2023, Accessed: 2024-06-26.

[16] Imran Ahmed, Meghna Pandharipande, and Sunil Kumar Kopparapu, "A suite of mobile applications to assist speaking at right speed," in *Proc. Speech and Language Technology in Education (SLaTE 2013)*, 2013, pp. 106–108.

[17] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kopparapu, "Automatic assessment of articulation errors in hindi speech at phone level," in *TENCON 2015 - 2015 IEEE Region 10 Conference*, 2015, pp. 1–4.

[18] Sunil Kumar Kopparapu and Meghna Abhishek Pandharipande, "System and method for monitoring a voice in real time," Oct. 14 2014, US Patent 8,861,708.

[19] Imran A. Sheikh, Meghna Pandharipande, and Sunil Kumar Kopparapu, "Speakrite: Monitoring speaking rate in real time on a mobile phone," *Int. J. Mob. Hum. Comput. Interact.*, vol. 5, pp. 62–69, 2013.

[20] Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu, *Analyzing Emotion in Spontaneous Speech*, Springer, 2017.

[21] Ashish Panda, Rajul Acharya, and Sunil Kumar Kopparapu, "Oral fluency classification for speech assessment," in *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*. 2023, pp. 231–235, IEEE.

[22] Upasana Tiwari, Meet Soni, Rupayan Chakraborty, Ashish Panda, and Sunil Kumar Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7194–7198.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011, IEEE Catalog No.: CFP11SRW-USB.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[25] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, 2011, pp. 187–197, Association for Computational Linguistics.

[26] Chitralekha Bhat, Anjali Kant, Bhavik B. Vachhani, Sarita Rautara, Ashok Kumar Sinha, and Sunil Kumar Kopparapu, "Moparest - mobile phone assisted remote speech therapy platform," in *Interspeech*, 2017.