L2 Pronunciation Intelligibility in Google Voice-to-text Applications

Michael Olayinka GBADEGESIN, PhD[1]
Deborah Adejumoke ADEJOBI[2]
[1, 2] Department of Languages and Literature, Lead City University, Ibadan
[1]gbadegesinmike@gmail.com, +2347084478356
[2]adejobi.deborah@lcu.edu.ng, +2348033788621

Abstract
There is an increase in the application of new technologies in all spheres of human endeavour. Industrial 4.0 is one of the recently birthed industrial revolutions through which machines understand human speech, thinks and comprehends human intentions. It structures critical components for intelligent vehicles, intelligent offices, intelligent service robots, intelligent industries, and so on. This furthers the structure of the intelligent ecology of the Internet of Things. At the centre of all these is human speech which is used to give order to and ask questions from Artificial Intelligence (AI) and robots. Previous studies on AI and linguistics have discussed the use of AI in language classroom using AI modelling pronunciation to enhance pronunciation performance of the second language learners. This study examines Automatic Speech Recognition (ASR) using Word Error Rate (WER) to measure the level of intelligibility of the pronunciation of man to machine. It investigates if there will be communication breakdown if the pronunciation is not intelligible. To achieve this, 30 L2 speakers of English were selected from Igbo, Hausa and Yoruba to read 135 crafted words into 5sentences using Google ASR application as primary data. The secondary data was drawn from journal articles, textbooks and the Internet. The result showed that the pronunciation model used to develop the application has made provision for several L2 speakers of English in reality of the new world Englishes.

Keywords: Artificial Intelligence, Human Voice-to-text, Communication Breakdown, Intelligible Pronunciation, L2

Introdu ction

Speech interaction in a globalised setting has gone beyond conversation among non-native speakers to conversation between native and non-native speakers of English as well as human and language modelled Artificial Intelligence (AI). In this new communicative setting, pronunciation plays a key role in achieving mutual intelligibility and enhancing understanding that sustains collaboration. Research has seen a growing recognition of the crucial role pronunciation plays in human-machine interaction. For any effective relationship to be established between two parties (human-human, human-animal, animal-animal, human-machine, machine-machine), there is need for a kind of communication that will be intelligible to both parties. The level of intelligible communication between two parties determines the level of interactional success. The new era of industrial revolution has given birth to the integration of AI into almost all areas of human endeavour. There are two levels of communication with AI -the creator or inventors that use special language or special coding to communicate with AI and end users that employ the use of human language to interact with AI.

Literature Review

AI technology is changing the world very fast, traditional methods of doing things is fading out for AI technology. Getchell et al. (2022) submits that AI technology is gaining popularity in the world due to its numerous advantages. For instance, the adoption of AI technology in communication enhances the speed, accuracy, and efficiency of communication. Fountaine et al. (2019) also observes that AI helps companies to analyze and interpret large volumes of data, provides valuable insights into communication patterns, employee engagement, and organizational performance. Google voice to text in the mobile devices is designed for a large vocabulary task that uses a language model designed for mixture of search queries and dictation. In other words, voice instructions can be given to the device to perform an action, dictation can be done while AI writes the words dictated. The appropriateness of the action that will be performed is based on the intelligibility of pronunciation of the speaker to the model designed.

Speech recognition is essential for interacting with devices. Tseng, (2021) notes that Google has the capability and databases to translate spoken language to text exerting the processing power of its own servers. AI technology is designed to collect, annotate, and analyze large speech data in many languages. In YouTube speech recognition task, the goal is to transcribe YouTube data while mobile device voice input does not have strong language model to constrain the interpretation of the acoustic information. Good discrimination requires an accurate acoustic model. Smith (1992), observes that understanding interactions between speaker and listener occur through intelligibility, comprehensibility, and interpretability is important. He defined intelligibility as referring to the listener's recognition at the word/utterance level; comprehensibility as referring to the listener's ability to understand word/utterance meaning and interpretability as referring to the listener's ability to infer meanings behind the word/utterance.

Global communication is affected by the realities of L2 pronunciation now that the COVID-19 pandemic has changed the way people live, work, and learn overnight. In the face of recent pandemic, classrooms were transformed overnight from the physical to digital space; the aftermath is that human-machine interaction is longer a luxury but a great necessity. In the education sector, for instance, AI becomes a double edged-sword, the instructor is facilitating the classroom interaction, as he speaks AI is transcribing what he is saying into text that could be given to the learners after the class. However, when the AI device cannot meaning of what the instructor is saying because of unintelligible pronunciation, the level of accuracy will be very low with high rate of deletion of unintelligible words, substitution of words due to wrong pronunciation, and insertion of words to fill the empty space all because of wrong pronunciation. Taking about developing capacity for intelligible pronunciation among second and foreign language learners, Low (2021) observes that teachers need to take advantage of technological affordances and to connect with students using digital platforms to keep learning going at any time. According to him, a deliberate effort can be put in place to avail the learners opportunities to be exposed to varieties of standard and intelligible pronunciation through video compilation and documentaries.

Intelligible Pronunciation is not the same thing as native-like proficiency. Low (2021) argues that in communicative contexts, a minimum number of standard pronunciation elements should be available to ensure mutual intelligibility, but achieving native-like proficiency for L2 learners should be recognised as being unrealistic. Intelligibility should take priority in both L2 communication and L2 pronunciation teaching and learning. Since attaining native-like proficiency in pronunciation is unrealistic for L2, an appropriate model of pronunciation is ideally one that has some core phonological features to ensure

international intelligibility and one that also incorporates some local pronunciation features (Low, 2021). While Jenkins' (2000) was arguing in favour of pronunciation unintelligibility among non-native speakers of English as against native-like proficiency, he proposes the Lingua Franca Core model for teaching and learning pronunciation where she emphasised segmental features above suprasegmental features.

Low (2015:143) states that the issue of intelligibility becomes a complicated matter in a context where English is spoken by people from many different countries, speaking many different background languages, and belonging to many different cultures. It is no longer what constitutes components of intelligibility, rather, who one is intelligible to, and for what purpose.

The importance and relevance of digital communication in human endeavour cannot be denied, many organisations are turning to AI technology to optimize their internal communication procedures. Russell and Norvig, (2016) describes AI as the development of computer programs that can perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation. Davenport (2019) explains further that AI involves using of algorithms and statistical models (a special language or code) to enable computers learn from data, recognize patterns, and make decisions based on the input received. AI is different from other traditional software that follows pre-programmed instructions because AI has the capacity to analyze and modify its processes based on new data. In other words, AI is flexible and versatile.

Matsuda and Matsuda (2018) observes that while preparing teachers for teaching in L2 situation, they must be equipped to teach pronunciation with understanding of how to provide their students with exposure to different varieties of English and how such exposure can be incorporated and implemented in the curriculum and how it may be. Automatic Speech Recognition (ASR) systems is an AI technology that converts a speech signal into a sequence of words either for text-based communication purposes or for device controlling. ASR, speech-to-text technology, is a technology that converts spoken language into written text. It is a branch of AI and Natural Language Processing (NLP) that focuses on the recognition and transcription of spoken words and phrases. ASR systems take audio input in the form of spoken language, process it and produce an output in form of text. The input come from various sources, including microphones, phone calls, or recorded audio. The systems incorporate linguistic knowledge in the form of phonetic and language models. Phonetic models help recognize individual speech sounds (phonemes), while language models consider the likelihood of word sequences based on the language's grammar and vocabulary.

Scholars have reiterated the benefits of using ASR for dictation such as vocabulary acquisition, customised pronunciation training, pronunciation improvement and ease of use (McCrocklin, 2016; Liakin et al., 2017; McCrocklin, 2019 Golshan, et al. 2021). Another benefit of AI in teaching and learning of intelligible pronunciation is the 'development of a digital assessment diagnostic tool in which AI can be used to automatically tag learners' conversations to identify instances of communication breakdown caused by pronunciation issues' (Low 2021:31). Also, technology can be used to identify speakers with pronunciation difficulties and propose independent learning solutions to enable them learn at their pace without a teacher. Low (2012) concludes that supporting the learners to be digitally literate, open to exploring, synthesising, and creating new ways of learning in a virtual space, with ethical consideration is the way forward in keeping teaching and learning going in the post-pandemic era. The adoption of AI in communication does not come without its challenges. George et al. (2023) identifies some of these challenges as concerns for data privacy and

security; risk of compromising sensitive information and potential reputation damage for the company.

L2 Pronunciation Intelligibility in Google Voice-to-text Applications

English as the most widely spoken language is spreading around the globe as it becomes more globalized. There are currently more non-native speakers of the language than native speakers worldwide. Sonako (2022) submits that international intelligibility is the capacity to make yourself understood in your target language (L2) when communicating with people from different first language (L1) backgrounds. For example, many English language learners do not need or want to communicate with native English speakers. The speakers who are most likely to be successful in global communication are not necessarily those who have a native-speaker accent; rather, international intelligibility is the product of competence in the pronunciation features that are common to many accents (Sonako, 2022). As a result, effective pronunciation teaching should be focused on preparing learners to deal with multiple native and non-native speaker accents.

The Common European Framework of Reference for Languages (CEFR) observes that in language teaching, the phonological control of an idealised native speaker has traditionally been seen as the target, with accent being seen as a marker of poor phonological control. The focus on accent and on accuracy instead of on intelligibility has been detrimental to the development of the teaching of pronunciation.

Teaching international intelligibility clearly requires a shift in priority for teaching pronunciation skills. AI plays several significant roles in language teaching and learning. These roles leverage AI's ability to process large amounts of data, provide personalized experiences, and offer real-time feedback. AI analyzes learners' pronunciation and provides real-time feedback, helping them improve their speaking skills. Some AI tools even offer visual representations of pronunciation, making it easier to understand and correct errors.

Cao et al. (2023) investigates how automatic speech recognition (ASR) errors influence discourse models of small group collaboration in noisy real-world classrooms. Their dataset consisted of 30 students recorded by consumer off-the-shelf microphones (YetiBlue) while engaging in dyadic- and triadic- collaborative learningin a multi-day STEM curriculum unit. They found that two state-ofthe-art ASR systems (Google Speech and OpenAI Whisper) yielded very high word error rates, although, Google speech to text was adjudged better than some others. It was established that errors in human -machine communication affect the output of the machine because there is a communication breakdown.Overall, their results provide insights into how different types of NLP-based tasks might be tolerant of ASR errors under extremely noisy conditions and provide suggestions for how to improve accuracy in small group modelling settings for a more equitable, engaging, and adaptive collaborative learning environment.

Maje and Hermann (2007) also worked on Word Error Rates using corpus data from Spanish and English from European Parliamentary Plenary Session. They found that WER gives a better overview of the nature of translation errors in human -machine interaction. The study presents a framework for extraction of linguistic details from standard word error.

FutureBee (2023) identifies some factors that affect ASR performance such as speaker variability which is the acoustic model obtained from speech data of a speakers at a given time and situation. Spoken language variability deals with the spontaneous and accented speech and the high degree of pronunciation variation due to dialects, and co-articulation. In another study, Rahhal, El Hannania, and Ouahmanea (2018) identify other factors that determine the effective performance of ASR as poor articulation, speaking rate, noise, side-

speech, accents, unintelligible pronunciation, hesitation, repetition, interruptions. Almost all the factors identified in the two studies are human related factors. This establishes the fact that the effective performance or otherwise of ASR is, most of the time, human dependent.

Word Error Rate (WER) is a method of analysis used to measure speech-to-text accuracy. WER counts the number of incorrect words identified during recognition and then divides it by the total number of words provided in the correct transcript, which is often created by human labelling. WER empowers end-to-end education solutions for computer-assisted language learning. Pronunciation assessment involves multiple criteria to assess learners' performance at multiple levels of detail, with perceptions similar to human judges. ASR systems are generally evaluated by WER. This entails aligning the ASR transcript to a human transcribed transcript, then counting the number of words missed (deletion), altered (substitution) or inserted relative to the original text (Cao et al, 2023). Somnath Roy, (2021) opines that WER is domain-agnostic, in the sense that the quality of transcription is assessed purely by straightforward word-level matching between ground truth and hypothesis. Scholars believe that WER reported alongside published models may give overly optimistic assessments of real-world performance (Terenzini et al: 2001; Goldwater et al: 2010; Cao et al, 2023). This study, therefore, identified the errors, analysed the errors, calculated the WER and discussed it in relation to the place of pronunciation intelligibility in Google human voice to text ASR.

Methodology

The study used read aloud test and observation to gather data from 30 L2 speakers of English in Nigeria. 10 Igbo, Hausa and Yoruba speakers of English were selected to read 135 words crafted into five sentences using Google ASR system. Most of the respondents recorded the text in the presence of the researchers which allowed for on the spot observation. Also, all the factors that could affect the output of ASR were guided against. A mini studio (language laboratory was used for the recording. The study draws insight from L2 Pronunciation models and Words Error Rate (WER) for theoretical insight and analytical guide. The choice of Google ASR is informed by the fact that Google is more conservative because of its low tolerance for unintelligible pronunciation (Cao et al, 2023).

Results and Analysis

Table 1 showing NoE, WER and Error Types

Key: *NoE =Number of Errors WER = Word Error Rate Type of Errors 1=Deletion 2 =Insertion 3 =Substitution 4 =Wrong Word Merge 5= Wrong Word Break*

S/N	**NoE**	**WER**	**%**	**Types of Error**				
				1	**2**	**3**	**4**	**5**
1	4	0.03	3	1		2	1	
2	16	0.12	11.85	1	3	12		
3	15	0.11	11.11	4	4	7		
4	20	0.15	14.81	2	3	11	1	3
5	20	0.15	14.81	3		17		
6	26	0.19	19.26	3	3	18	2	
7	25	0.19	18.52	3	6	13	2	1

8	19	0.14	14.07	4	4	10	1	
9	17	0.13	12.59	1	5	8	1	2
10	11	0.08	8.15	4	4	3		
11.	8	0.06	5.93	2		6		
12.	10	0.07	7.41		2	7	1	
13	14	0.10	10.37	1	3	7	1	2
14	6	0.04	4.44		1	3		2
15	11	0.08	8.15		1	7	1	2
16	5	0.04	3.70		1	3		1
17	20	0.15	14.81	2	4	11	1	2
18	21	0.16	15.56	1	7	8	2	3
19	17	0.13	12.59		3	12	1	1
20	16	0.12	12.59	1		12	2	1
21	9	0.06	11.85	2		5	2	
22	5	0.04	6.67		1	2		2
23	5	0.04	3.70			4	1	
24	18	0.13	3.70	1	5	9	1	2
25	22	0.16	13.33	3	2	17		
26	17	0.13	16.29	3	2	9	1	2
27	19	0.14	14.07	2	4	10	1	2
28	21	0.16	15.56	1	3	13	3	1
29	20	0.15	14.81	2	5	9	1	3
30	24	0.18	17.78	3	3	16	2	
Grand Total	461			50	79	271	29	32
Average WER		**0.11**	**11.38**	**1.66**	**2.6**	**9**	**0.96**	**1.06**

The results presented in the above table show that there are 461 errors in all with 15.3 as the average, this amount to 11.38% while the WER is 0.11

Substitution has the highest occurrence in the list of errors identified with 271 instances representing 58.7% of the total errors. It accounts for more than half of the total recorded errors.

Example:

Original Text: The **professor says** this while answering question on a weekly national television on the growth of the oil region. The oil region, agrarian society and the economic hub should be allowed to develop separately.

ASR Text 1: The professor ***sees*** this ***when*** answering questions on the weekly national television on the growth of the ***all*** region the **all** region agrarian society and economy ***call*** should be allowed to develop separately.

ASR Text 2: The ***professional*** says this ***when*** answering questions on the weekly national television on the growth of the oil ***rejoin*** the oil region agrarian society and the economic ***home*** should be allowed to develop separately.
In ASR text 1, 'says' was substituted for 'sees', 'while' was substituted for 'when', 'oil' was substituted for 'all', 'hub' was substituted for 'call' while in text 2 'professor' was substituted for 'professional', 'while' for 'when', 'region' for 'rejoin' and 'hub' for 'home'.
Insertion followed substitution (though with a wide margin) with 79 instances representing 17.13%. Insertion is another type of error identified which. includes addition of a complete word, a morpheme or a letter to the word pronounced. Examples include: a, an, the, her, secret, and, own, to, -d, -s. The study found that not only words are inserted, plural and past tense morphemes are inserted. The most common insertions are –s, -d, a, an, the, are, and etc.

Example 1
Original Text: There is need for the federal government of Nigeria to develop the different regions base on their peculiarities and allow them to grow at their own pace.
ASR Text: There is (a) need for the federal government of Nigeria to develop (to) the different regions base(d) on the picture realities and allow them to grow their own peace Example1 above shows that 'a' was inserted after the second word and letter 'd' was inserted after 'base'.

Example 2
Original Text: The rate at which churches and mosques are springing up in our society without character transformation is alarming.
ASR Text: The rate at which (it) charges and mosses springing over in our society without (the) character transformation is (an) alarming
There is insertion of 'it', 'the' and 'an' to the original text in this example. This validate the presence of insertion in ASR output.
Deletion: error closely follows insertion with 50 recorded instances representing 10.84%. Some of the element of the original text that were deleted in ASR output include –a, are, boy, says, chieftain, need, different, former, -uate (from undergraduate). Example 1:
Original Text: To be factual, the boy must be a riff raft to have done that to his undergraduate law student girlfriend of university of Lagos.
ASR Output: To be factual (**the boy**) must be (**a**) riffraff to have done that to his undergraduate law student girlfriend of the University of Legos.
The original text has 24 words while the ASR output has 22 words; this implies that **three** words have been deleted. The words in bracket were deleted.
Example 2:
Original Text: The former People's Democratic Party chieftain left the party, he said that many of their programmes are not people oriented.
ASR Output: The (former) people ('s) democratic party chief (tain) left the party (he) said (that) many of their programs are not people oriented.
The original text has 20 words while the ASR output has 17 words from which there are two deletions of morphemes, this implies that there are **five** deletions. The words in bracket were deleted.
The other errors are wrong word break and wrong word merge with 32 representing 6.94% and 29 representing 6.29% occurrences respectively.

a. Examples of wrong word break

	Original Text	**ASR Text**
i.	region	rig in

ii.	chieftain	chief tim
iii.	chieftain	chief thing
iv.	chieftain	chief and
v.	churches	church is
vi.	agrarian	are green
vii.	undergraduate	going to graduate
viii.	former	for my
ix.	peculiarities	picture realities

b. Examples of Wrong Word merge

	Original Text	**ASR Text**
i.	oil region	origin
ii.	oil region	original
iii.	aspirin	are springing
iv.	the approve	their programmes
v.	riff raft	Refracts
vi.	the former	depalma
vii.	the boy	define

The examples in 'a' above show instances where a word is broken into two or three other words. There are cases of none word forming the second word, while in other cases all the words are correct but not in the original text. Systematically, combinations of most of the wrong word break have similar sound with the original word. For example 'chieftain' was broken to 'chief thing', 'agrarian' to 'are green', 'former' to 'for my'. The examples on 'b' also show sound similarities in the original text and ASR text. For instance 'oil region' was merged to 'original' and 'origin', are 'springing' merged to 'aspirin', 'riff raff' to 'rifracts' while 'the former' becomes 'depalmer'.

Discussion

There are two dimensions to consider while measuring accuracy in ASR these are human and machine. However, many of the previous studies focused on machine with little or no attention to human factor or dimension. In the process of recording, it was observed that some of the errors are machine errors such as 'the approve' instead of 'their programmes', 'a river' instead of 'factual' produced by the respondents. This is as a result of its inability to catch up with the speed at which some of the respondents read. It is observed that when a speaker speaks at a speed faster than the ASR model speed it gives room for deletion, insertion, substitution, wrong word break or wrong word merging. This is in line with the submission of Rahhal, El Hannania and Ouahmanea (2018) that speech rate affect the ASR output.

Most of the instances of substitution are occasioned by pronunciation that is not intelligible to ASR, it substitutes such words with the nearest related word. The unintelligible pronunciation of the respondents is evident in the high percentage recorded (58.7%) by substitution in the analysis of errors. Although, the focus of L2 pronunciation training is not to achieve native-like accent but international intelligibility which is the capacity to make oneself understood in the target language (L2) when communicating with people from different first language (L1) backgrounds (Sonako, 2022). The place of intelligible pronunciation in the face of industrial revolution cannot be denied.

Previously, insertion (words in the transcription that don't appear in the source content), substitution (words in the transcription that are incorrectly transcribed from the

source content) and deletion (words that don't appear in the transcription but are in the source content) have been the three errors identified in ASR and subject to WER analysis (FutureBee, 2023). This study also found errors of insertion, substitution and deletion in the ASR text that corroborates the previous studies. However, two other new errors were discovered which are wrong word break (where a word in original text or ground truth is broken into two or more words in ASR text) and wrong word merge (where two words in original text or ground truth are merged into one word in ASR text). The study corroborates the existing literature on the need for AI in the development of our society. However, it states that there is need for human pronunciation that will meet the standard of AI model pronunciation for effective human-machine interaction.

The study also is in line with FutureBee's (2023) assertion that different speaking styles, such as conversational speech or formal speech, can also affect the accuracy of ASR systems. Some speaking styles may be more difficult for the system to recognize accurately. Speech AI models are trained on different speech data to overcome this error like general conversations (Informal, general, between friends, family), contact centre conversations (customer service calls, etc). Different speakers may have different accents, and ASR systems may struggle to recognize speech accurately if they are not trained in those accents. In every 30 to 40 kilometres, accents of speakers of the same languages seem to be different and this is one of the reasons ASR gets different results for different people. This makes accuracy or otherwise of ASR to be based more on speakers' intelligible pronunciation.

Conclusion

In terms of level of accuracy, going by the submission of scholars that when WER is 5-10% it is considered to be of good quality and accurate and when WER is 20% it is acceptable but not accurate, however, when WER is more 20% it is not acceptable, it is therefore correct to say that with WER of 11%, the pronunciation could be described as of good quality and accurate. This implied that base on the result of ASR and the WER judgment, the pronunciation of Nigerian English is intelligible to other speakers of English especially non-native speakers and that the pronunciation model used to develop the Google ASR application has made provision for several L2 speakers of English in reality of the new world Englishes. There is no communication breakdown between L2 speakers of English and ASR machine as a result of the pronunciation. Although, there are tribal flavour and colouration observed (as confirmed by the result from ASR text), the ASR is modelled to accommodate most of the L2 tribal colouration and accent.

References

Cao, J. Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. (2023). A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus. ACM, New York, NY, USA, pp: 250-262. https://doi.org/10.1145/3565472.3595606

Council of Europe (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages.

Davenport, T. H. (2018). The AI advantage: How to put the artificial intelligence revolution to work. MIT Press

Errattahia, R., El Hannania,A., Ouahmanea, H. (2018) Automatic Speech Recognition Errors Detection and Correction: A Review (https://creativecommons.org/licenses/by-nc-nd/4.0/

Fountaine, T., McCharty, B. and Saleh, T. (2019). Building the AI-Powered Organization. Harvard Business Review, July-August issue, pp: 63-73.

FutureBee. (2023, July 29) Breaking Down Word Error. https://www.linkedin.com/pulse/breaking-down-word-error-rate-asr-accuracy-optimization-futurebeeai/

Getchell, K., Carradini, S., Cardon, P. W., Aritz, J., Fleischmann, C., Stapp, J., & Ma, H. (2022). Artificial Intelligence in Business Communication: The changing Landscape of Research and Teaching. Business and Professional Communication Quarterly. https://doi.org/10.1177/23294906221074311

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which Words are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase Speech Recognition Error Rates. *Speech Communication (52)*3, p: 181–200.

Golshan, M., Nejad M. Z., Naeimi A. (2021). The Effect of Synchronous and Asynchronous Computer-Mediated Communication (CMC) on Learners' Pronunciation Achievement. Cogent Psychology, *8*(1), pp: 1-18. https://doi.org/10.1080/23311908.2021.1872908

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, M., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep Neutral Networks for Acoustic Modelling in Speech Recognition. IEEE Signal Processing Magazine. Digital Object Identifier 10.1109/MSP.2012.2205597 1053-5888/12/$31.00©2012IEEE https://doi.org/10.1016/j.system.2015.12.013.

Jenkins, J. (2000) The Phonology of English as an International Language. Oxford: Oxford University Press.

Liakin, D., Cardoso W., Liakina N. (2017). Mobilizing Instruction in a Second-Language Context: Learners' perceptions of two speech technologies. Languages, 2(3), pp: 11–32. https://doi.org/10.3390/languages2030011

Low EL (2015) Pronunciation for English as an International Language. London/NY: Routledge.

Low, E. L. (2021), EIL Pronunciation Research and Practice: Issues, Challenges, and Future Direction. Research Perspectives on Prcatice. RELC Journal, *52*(1), pp: 22–34. DOI: 10.1177/0033688220987318

Matsuda A, Matsuda PK (2018) Teaching English as an International Language: A WE-Informed Paradigm for English language Teaching. In: Low EL and Pakir A (eds) World Englishes: Rethinking Paradigms. London/NY: Routledge, pp. 64–77.

McCrocklin, S. (2016). Pronunciation Learner Autonomy: The Potential of Automatic Speech Recognition. System, 57(1), pp: 25-42.

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. Journal of Second Language Pronunciation, 5(1), pp: 98-118. https://doi.org/10.1075/jslp.16034.mcc.

Patrick, T., Cabrera, A., Colbeck, C., Bjorklund, S., and Parente, J. (2001). "Racial and Ethnic Diversity in the Classroom. Does It Promote Student Learning?" *Journal of Higher Education 72*(5), pp: 509–31.

Smith LE (1992) Spread of English and issues of intelligibility, In Kachru,B. B (ed.) The Other Tongue: English across cultures (2nd edn). Urbana: University of Illinois Press, pp. 75–90.

Somnath Roy. (2021). Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability. arXiv preprint arXiv:2106.02016.

Tseng, J.-L. (2021). Intelligent Augmented Reality System Based on Speech Recognition. *International Journal of Circuits, Systems and Signal Processing*, 15, pp: 178–186. https://doi.org/10.46300/9106.2021.15.20

Yams, N.B., Richardson, V., Shubina, G.E., Albrecht, S., Gillblad, D. (2020). Integrated AI and Innovation Management: The Beginning of a Beautiful Friendship. *Technology Innovation Management Review*, *10*(11): pp: 5-18. http://doi.org/10.22215/timreview/1399