Available online at IJEE (Indonesian Journal of English Education) Website:

http://journal.uinjkt.ac.id/index.php/ijee

AUTOMATIC CAPTION FEATURES ON GOOGLE MEET AS A PRONUNCIATION ASSESSMENT TOOL

Boris Ramadhika*, Rolisda Yosintha, Sukma Shinta Yunianti
Universitas Tidar, Indonesia, Indonesia
(ggramadhika@untidar.ac.id)

Received: 13th September 2021; Revised: 28h Juni 2022; Accepted: 27th December 2022

ABSTRACT

This study aims to find out the use Google Meet Automatic Caption feature to assist teachers of non-native English to assess their students' English pronunciation. We used a mixed-method with the explanatory-sequential approach following it. This research study was done at Tidar University with 12 participants, further reduced to 4 in the participant selection step. As the data were both quantitative and qualitative, we used the Word Error Rate (WER) formula quantitatively and Qualitative Content Analysis qualitatively. The findings show that Artificial Intelligence (AI) has a very sensitive system in transforming sounds into written forms. It also has an auto-correction system that sometimes can substitute a word with a meaningless one if a speaker pronounces the word unclearly or to the nearest word if a speaker mispronounces it. Even though it does not accurately process the punctuation and there is no sufficient correction on grammar, we believe the AI can help teachers in a pronunciation assessment.

Key Words: assessment; automatic caption; google meet; pronunciation

ABSTRAK

Penelitian ini bertujuan untuk mengetahui penggunaan fitur Teks Otomatis pada Google Meet untuk membantu para guru Bahasa Inggris bukan penutur asli saat menilai pengucapan bahasa Inggris siswa mereka. Kami menggunakan metode penelitian campuran dengan pendekatan eksplanatori-sekuensial. Penelitian ini dilakukan di Universitas Tidar dengan jumlah peserta sebanyak 12 orang yang dikurangi menjadi 4 orang pada tahap seleksi peserta. Data penelitian bersifat kuantitatif dan kualitatif. Untuk menganalisa data secara kuantitatif digunakan rumus Word Error Rate (WER). Sedangkan, secara kualitatif mengunakan Analisis Isi Kualitatif. Hasil penelitian menunjukkan bahwa AI (Artificial Intelligence/ Kecerdasan buatan) memiliki sistem yang sangat sensitif dalam mengubah suara menjadi bentuk tulisan. Fitur ini juga memiliki sistem koreksi otomatis yang terkadang dapat menggantikan kata yang tidak berarti jika pembicara mengucapkan kata dengan tidak jelas atau diubah ke kata terdekat jika pembicara salah mengucapkan kata tersebut. Meskipun tidak memproses tanda baca secara akurat dan tidak ada koreksi yang memadai pada tata bahasa, kami yakin AI dapat membantu para guru dalam penilaian pengucapan.

Kata Kunci: *penilaian; teks otomatis; google meet; pengucapan*

How to Cite: Ramadhika, B., Yosintha, R., Yunianti, S. S. (2022). Automatic Caption Features on Google Meet as a Pronunciation Assessment Tool. *IJEE (Indonesian Journal of English Education), 9(*2), 396-410. doi:10.15408/ijee.v9i2.22482

* Corresponding author

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://doi.org/10.15408/ijee.v9i2.22482

INTRODUCTION

In English as a Foreign Language (EFL) contexts, teachers often neglect or avoid pronunciation. Insufficient time allocation, methodological uncertainties, pedagogical priorities, inadequate materials, and inadequate pronunciation teaching training are often factors causing this situation (Dixon, 2018; Isabelle, 2018). In the Indonesian context, lack of confidence caused by non-nativeness only adds to the teaching and assessment process challenges of students' pronunciation. Meanwhile, pronunciation is essential and bonded with all language skills (Hunt-Gómez & Navarro-Pablo, 2020), and thus teaching pronunciation should be prioritized. This pedagogical gap should be addressed to help improve the English teaching and learning process results. To do so, utilizing technology could be an alternative for teachers to develop students' pronunciation skills.

The integration of technology with the teaching and assessment process of pronunciation has been extensively implemented. Numerous Mobile Assisted Language Learning (MALL) media have been developed and used to teach and learn pronunciation. MALL is defined as a learning mode in which students can manage their own learning with the help of mobile devices such as mobile phones, tablets, etc. (Cohen & Ezra, 2018; Hoi & Mu, 2021). Kim and Kwon (2012) list four types of mobile application services for MALL, namely Mobile Social Networking, Mobile Podcasting, Course Management Service, and Automatic Speech Recognition.

The use of MALL in teaching pronunciation is highly beneficial for both the teachers and students. The exposure to native pronunciation provided by MALL makes it possible for students to develop their skills properly. For non-native teachers, MALL lessens their burden in delivering the lesson. In the assessment process, however, teachers might still encounter some problems. By nature, the testing of productive skills, including pronunciation, is not as simple as that of receptive skills. Various MALL media for testing reading and listening are vastly available since the answers to the questions are mostly closed-ended. Those for testing writing, speaking, and pronunciation skills are relatively limited since the responses could be varied and open-ended. With no fixed answers to the questions, human raters are needed to review the answers manually. Consequently, the time required for this process is considerably long, and the cost of developing such media for testing the productive skills is highly-priced.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

To help teachers in the assessment process of pronunciation, some MALL alternatives have been extensively explored and developed. Rayshata and Ciptaningrum (2020) analyzed the development of fifteen Indonesian EFL students' pronunciation using Automatic Speech Recognition (ASR) in Google's Voice Search Application (GVSA). Their study showed that GVSA could help students improve their pronunciation skills in consonant sounds and vowels, and diphthongs. In another study, Evers and Chen (2020) explored the use of ASR in Speechnotes to examine the difference in Taiwanese adults' pronunciation performance with peer feedback and individual practice. They found that the two groups' pronunciation results were significantly different in that ASR in Speechnotes was better used in ASR-based pronunciation activities with peer feedback. In other words, ASR in Speechnotes could not be successfully used for individual or independent learning.

Cheng, Lau, Lam, Zhan, and Chan (2020) developed a Phonics Learning Voice Chatbot (PLVC) by combining ASR and Triple Neural Networks (TNN). Using this MALL media, they compared users' performance in phonic learning when assessed by PLVC and professional English teachers. With a correlation coefficient of 0.71, the results showed that the assessment result from PLVC was reliable. In a more recent study, Spring and Tabuchi (2021) investigated the practicality of an ASR tool called NatTos in improving Japanese EFL students' pronunciation in an online learning setting. After a series of implementations, they concluded that NatTos could objectively improve students' pronunciation, particularly in their mastery of consonant and vowel sounds, even though the lesson was delivered online.

Despite the potential advantages of ASR-based MALL media for assessing pronunciation, some areas could be improved. Most studies about ASR-based pronunciation MALL media focused on the teaching and learning of pronunciation, not on the assessment process (Rayshata & Ciptaningrum, 2020; Ryan & Ryuji, 2021). In addition, some of the ASR tools being explored were not freely and readily available for public use (e.g., Phonics Learning Voice Chatbot and NatTos). Therefore, other ASR-based MALL media should be further explored that could help teachers assess students' pronunciation skills. Google Meet, a video-communication service developed by Google in 2017, has an ASR feature and thus could be developed for this purpose. ASR in Google Meet converts speech to text to provide live captions

to assist participants with hearing problems or those who simply cannot follow spoken language well. Until December 2020, this feature has been expanded to five languages: English, French, German, Portuguese (Brazil), and Spanish (Nelson, 2020).

Considering the potential of ASR used in Google Meet, the researchers assumed that automatic caption features on Google Meet could be further explored in testing pronunciation. Given this motive, the researchers aimed to evaluate the feasibility of automatic caption features on Google Meet as a pronunciation assessment tool. The following research questions guided this study: Can the Google Meet Auto Caption feature help teachers assess students' pronunciation? The findings of this study are expected to contribute to the development of students' pronunciation through the utilization of effective and efficient MALL media. Furthermore, teachers are expected to make use of this study as a reference and basis for consideration when they are about to assess their students' pronunciation with ease and accuracy.

METHOD

The method used in this research study was the mixed method because we had quantitative and qualitative data. We also adopted the explanatory-sequential approach as we were interested in following up the quantitative results with qualitative data. This approach can be seen figure 1.

Figure 1. The Explanatory-Sequential Approach

The words 'quan' and 'qual' are written and used according to their own purposes. The word 'quan' means that our quantitative data were taken as secondary to qualitative data, while the word "qual" means the study was driven qualitatively (Edmonds & Kennedy, 2020).

Research design

We used the participant-selection design as our research design which a involves two-phase process (Edmonds & Kennedy, 2020). The first phase is the participant selection which was done by collecting and analyzing quantitative data; while the second phase is collecting and analyzing qualitative data, which was done to the selected participants and was used to interpret the quantitative data at the final step. The research design can be seen in the following figure 2.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

Figure 2. The Participant-Selection Design

Research site and participants

The total participants were twelve people: eleven university students and one native speaker of English. These students were randomly selected. For the native speaker, the authors asked one person from America to analyze the feature of Automatic Caption in Google Meet.

Data collection and analysis

Based on the research design that the authors used, the data were gathered synchronously using the Google Meet application. In the first phase, to collect the "quan" data, we first asked each participant to read a passage (426 words) taken from a TOEFL test question in silence for one minute. Then, they were asked to read it aloud while activating the Automatic Caption from Google Meet. The authors then recorded each participant's reading using XBOX Game Bar, which is available in Windows 10, by pressing Windows logo + G. After that, the authors used Word Error Rate (WER) (Klakow & Peters, 2002) to analyze the data. We used WER because we gathered the data based on the system performance (Prabhavalkar et al., 2018). WER contains 1) insertion: words that are not spoken but are detected, 2) deletions: words that are spoken but are not written, and 3) substitutions: words that are spoken and written differently in which the formula is as seen figure 3.

$$\text{Word Error Rate} = \frac{\text{Insertions (I)} + \text{Deletions(D)} + \text{Substitusions (S)}}{\text{Number of Words (N)}}$$

Figure 3. The Word Error Rate Formula

For the insertions and deletions, we were looking at the caption while, at the same time, also hearing what the participants said. Some words written in the caption but were not said by the participants were considered insertions. On the other hand, words that were spoken but not written in the caption were considered deletions. For substitutions, we focused on the caption and matched it with the text that the participants read. After we had the data, we put it into the formula and had the results for the first phase. Based on the quantitative data analysis result using WER, we then partially selected some participants for the next phase. We categorized the participants into high, mid, and low in WER. At this stage, it was the end of the first phase, as the objective in the first phase was to select participants for the second phase (Edmonds & Kennedy, 2020).

In the second phase, we started to collect the "QUAL" data by

interviewing the selected participants. The interviews lasted about 10 to 15 minutes and were digitally recorded and saved on a secure laptop. After that, the data were analyzed qualitatively using Qualitative Content Analysis (Mayring, 2019). We would like to have the data related to the educational background of the selected participants, the device used, and the experience they had. Next, the categorization was made based on the interview results such as English and Non-English students, high, mid, and low devices, and years of learning English.

Furthermore, we also used Wondershare Filmora 9 software to see the audio pulse track, especially in the error parts. This was done to determine whether there was a difference between correct and incorrect pronunciation. Finally, we interpreted the quantitative data to qualitative data in the final stage in describing the findings qualitatively. First, we looked at the WER results and matched them with the qualitative data. Second, we analyzed the errors made by the participants. For example, whether the low-quality device would affect the caption (the AI performance) or not. Third, we tried to look at the audio pulse whether the errors had the same pulse or not. Lastly, we interpreted the findings qualitatively.

FINDINGS AND DISCUSSION

Findings

Phase 1

In phase 1, stage 1, we asked the participants to read aloud a passage taken from a TOEFL test question. We then recorded it and analyzed the results quantitatively using WER. The table below shows the results of phase 1 in quantitative results in table 1.

Table 1. The Results of WER

Initial Names	Insertions	Deletions	Substitutions	Number of Errors	Number of Words	Words Error Rate	Percentage
NS	0	0	4	4	436	0.0092	0.917%
ILLE 1	4	7	37	48	446	0.1076	10.762%
EE 1	4	6	25	35	422	0.0829	8.294%
EE 2	3	5	8	12	430	0.0372	3.721%
EE 3	13	7	65	85	436	0.1950	19.495%
OTU 1	21	4	56	81	456	0.1776	17.763%
OTU 2	10	5	38	53	439	0.1207	12.073%

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

Initial Names	Insertions	Deletions	Substitutions	Number of Errors	Number of Words	Words Error Rate	Percentage
OTU 3	8	4	13	25	430	0.0581	5.814%
ILLE 2	18	15	66	99	455	0.2176	21.758%
SA 1	5	4	23	32	431	0.0742	7.425%
SA 2	6	6	16	28	434	0.0645	6.452%
OU 4	10	9	38	57	437	0.1304	13.043%

The terms that are used in the table are as follows:
NS : Native Speaker
ILLE : Indonesian Literature and Language Education students
EE : English Education students
SA : State Administration students
OTU : Outside Tidar University students

As shown in Table 1, the 'quan' data show that the native speaker had the lowest percentage (0.917%) in making errors. Some of the students also made notable efforts since they could make less than 10% errors. Although we cannot say that the feature is perfect, by looking at the results, we could also say that the Google Meet Automatic Caption feature did work to translate spoken forms to written forms. However, we might think many things could be analyzed further. Therefore, we wanted to go deeper, assuring that the Google Meet Automatic Caption feature can be a supportive learning tool to assess speaking skills.

The next step that we would do after getting the 'quan' data was to select some participants for phase 1, step 2. We then selected NS, EE 3, OTU 3, and ILLE 2. First, we asked NS because she is a native American; we wanted to know why the automatic caption feature still showed some substitutions. Second, we chose EE 3 because this student is actually an English Education student. In fact, the automatic caption feature noticed many errors in her pronunciation; we were curious whether it was her pronunciation or the machine translation that caused the errors. Third, we were interested in OTU 3's results because he was not from the English Education study program. However, he still could make it, reaching less than 10% errors. Lastly, we selected ILLE 2 because this student had the most errors; we might find out whether ILLE 2 made the mistakes.

Phase 2

After we selected some participants, in phase 2, step 3, we asked them one by one using

WhatsApp messages. They were asked open-ended questions. We tried to figure out their educational background, the device used, and their experience when they read the passage. After that, we used IPA to find out mainly about the learning experience that they had. The results would be the "QUAL" data used for further analysis in interpreting 'quan' data to qualitative data. Here are qualitative data results for each participant that are also being the analysis in step 4:

NS

Ns is a native speaker from the USA. However, she has been in Indonesia for more than five years. Although she is a native, she still made some errors according to the machine even though it is the lowest one. Her errors were only four substitutions even though she used a laptop without an external microphone. The errors are presented in the following table2.

Table 2. NS Error Words

Error Words	Error Types	Correct Words
In Visage	Substitution	Envisaged
Our	Substitution	Are
They're	Substitution	Their
Vectors	Substitution	Factors

She admitted that she found some words that were unusual for her. She added that some terms were very English for an American like her. She had to look up first for idioms that did not make sense to her, such as "come a cropper." Moreover, she argued that some sentences were too long to read and did not have enough punctuation, which made her hard to take a break from reading the passage. Therefore, at some points, she had to make pauses that made the errors such as 'their', which was, as a result, substituted with 'they're'. In that case, pauses have affected the AI processing of sound into written forms.

EE 3

EE is a fifth-semester student of English Education. She used a laptop and a headset to read the passage. Ironically, she had so many errors with more than 19% of WER. We tried to find out the reasons by interviewing her. First, she argued that she was so nervous and made reading very fast. This might cause the machine to catch the sound in error. Second, some vocabularies were unfamiliar to her; she did not know how to pronounce them. She also stated that even when she thought she mispronounced it, she still believed that she made it correctly. She was very confident with her skill that she rated herself 8.9 out of 10, and,

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

from what we heard from our ears, we believed that she spoke too fast that the AI from the machine missed the sound. It was because AI needed a little time to transform sound into written forms. It was like when she had already pronounced three words; the AI only managed to process one word, causing deletion errors. Apart from it, she mispronounced some words, causing the AI to process them in errors.

OTU 3

OTU 3 is a fourth-semester student of the Agricultural Industry Technology study program. She used a laptop without a headset to read the passage. Based on our interview with her, we found out that she thought her English skill was only 5 out of 10, and it was not in line with the result of her WER, which is only at around 5%. She also stated that she was not too sure of her pronunciation as there were some words that, in her opinion, were hard to pronounce. As a result, she finished reading the passage by pronouncing the words carefully and slowly.

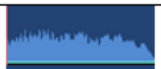
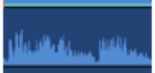
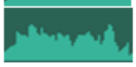
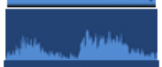
ILLE 2

As for the last, we decided to look at the most errors produced by ILLE 2. She was a student of Indonesian language and Literature Education. It might be logical for a non-English Education student to make such errors. However, we were still interested in this because we would like to compare the AI to recognize errors. She used her smartphone with a headset on it when she read the passage. As for her, English was not something that she would jump into, which was why she chose Indonesian instead as her major. She also did not hesitate when she said she was between 3 or 4 out of 10 for her English level. She was so nervous and read the passage uncalmly as she found some difficulties pronouncing certain words. She also had some pauses, causing the AI to process her sound into random words.

After we look at the errors caused by the selected participants, in step 5, an analysis to relate and describe the "quan" to "qual" data was carried out. EE 3, OTU 3, and ILLE 2 had common substitution error words. What we did after was to find out the audio pulse track both for the error and correct

words using Wondershare Filmora 9 software to track down whether the error words were audio-visually the same as the correct ones or not. We took the audio pulse track of the correct words from NS, who pronounced those words correctly. The results can be seen in the following table 3.

Table 3. Audio Pulse Track in Common Substitution Errors

Error Words	Audio Pulse Track	Correct Words	Audio Pulse Track
Screen		Green	
So don't		Seldom	
This		These	
Over		Offer	
Led		Lead	

The table 3 shows that the pulse tracks on the error and correct words are different, and it means that the speaker needs to pronounce the words correctly and carefully. For example, many participants pronounced the word "screen" instead of "green" because they sounded the phrase "promotesgreen" (words 's' and 'g' were like in the one word) which resulted in "promote screen" (not "promotes green") by the AI; they did not give a clear space to pronounce the words "promotes" and "green." Meanwhile, the word "seldom" to "so don't do", happened because the participants did not pronounce the 'l' in "seldom." This made the AI recognize the word "sedom"-which the 'l' was not clear enough- to the nearest pronunciation "so don't".

As for the words "this"-"these", "over"-"offer", and "led"-"lead", they were purely mispronounced by the participants. For example, in the pulse track for the word "this", we can see that the participants only produced /I/ (short 'i') instead of /i:/ (long 'i') to produce "these". The same thing happens to words "over"-"offer". The participants pronounced /v/ instead of /f/. The last words, "led"-"lead", were clear. It can be heard that the participants pronounced /led/ rather than /li:d/.

Discussion

From this point, we have found several findings from the analysis results. First, it is interesting that as long as the speaker pronounces words correctly, even at a fast tempo, the AI will still function precisely in transforming sounds into written forms. The AI is very sensitive to speakers' sounds that a word has to be pronounced correctly to be captioned

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

accurately. As long as the speaker pronounces the words correctly, the AI will not substitute the words for other nearest-pronounced words.

Second, Google Meet Auto Caption's AI has an auto-correction system. However, the system is only limited to words. The AI will transform the sounds into written words that have the nearest pronunciation produced by the speaker. On some occasions, it can arbitrarily form meaningless words. For example, when a speaker wants to say 'architectural' but he/she does not pronounce it clear enough, it can become 'RC tectoral". We know that 'tectoral' is meaningless.

Third, the other thing found from the AI's auto-correction system is that it does not care about grammar. It can be said that the auto-correction system does not correct the syntax for its captions. The AI will be forming captions according to the sounds produced by the speaker ignoring the structures of the sentences. For example, if the speaker mispronounces the word 'these' to 'this' in a phrase 'these documents', the AI will stand in 'this documents' for the caption.

Fourth, if the speaker speaks too fast, it can cause deletion errors. This is because the AI needs time to process sounds to words. On average, when a speaker already pronounces three words, the AI will only have processed one word. As a result, there will be some delays in processing the captions. If the speaker corrects the words or restates the same words during the 'delaying time', it can cause the AI to make errors in captioning.

Fifth, we also noticed that the AI is not really accurate in processing punctuations because of some factors. However, it recognizes punctuations; it still arbitrarily transforms pauses to either a comma or a full stop for a pause between two sentences. Sometimes, the next word does not start with a capital word, even after a full stop. While in some random occasions, a word with a capital letter is written in the middle of a sentence. Other punctuations such as apostrophes, colons, and quotation marks are greatly ignored.

The AI in Google Meet has proven useful in captioning speech to text based on the findings. This is supported by what Soni, Sheikh, and Kopprapu (2019) have found that Google Speech Command is a good tool even without an enhancement. Even so, at some times, the AI has failed to contain accurate captions, leading to confusion. This is similar to the AI on Youtube. Lee and Cha (2020) found that the AI on Youtube also occasionally fails to caption the speech accurately,

especially when spoken by non-native people. The difference is that the AI in Google Meet will caption the speech the same as what the speaker pronounces, even resulting arbitrarily and meaningless. In contrast, the AI on Youtube will caption the speech with the nearest pronounced word (Malik et al., 2021).

However, we have identified and acknowledged our study's limitations. Our data only can prove that the AI in Google Meet can sensitively detect one's English pronunciation. We still do not know, for instance, if there is an English preference, such as British or American English, or not in the Google Meet Automatic Caption feature. Examination of key issues, impacts, and effectiveness of this feature to be applied in an English as a Foreign Language (EFL) class setting has not been explored yet due to our limited data. As a result, we only focus on whether the AI in Google Meet can help non-native English teachers in assessing the pronunciation skill of non-native English students or not.

We believe that our research findings will be a good starting point to explore the potential of the Google Meet Automatic Caption feature. For other researchers, they can identify whether the Automatic Caption feature in Google is better than the others, such as the AI on Youtube, or not. Teachers can also start to use Google Meet for a meeting and, especially, teaching and assessing the pronunciation skill. Even students can use it as a self-improvement tool for their pronunciation skills as they can notice the caption if they mispronounce a word. Thus, further works related to the AI in Google Meet should be related more in practicality, especially in an EFL class setting. Moreover, there is a possibility that this tool can be used to teach and assess other English skills.

CONCLUSIONS AND SUGGESTION

To conclude, Automatic Speech Recognition technologies have been particularly developed. Based on the findings, Google Meet as a pronunciation assessment tool will be a handy tool to assess students' pronunciation. It can, especially for the non-natives of English, reduce the teachers' subjective perspective in assessing students' pronunciation since the AI is very sensitive to word pronunciation. In that case, it will help non-native teachers to assess students' English pronunciation more comprehensively. For students, we suggest that they use the Google Meet Auto Caption feature to help them correct their English pronunciation.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

Moreover, our study raises a number of opportunities for future research that may be done to find other possible findings which we cannot reach due to our data limitations. First, we believe that Google Meet can be used as a pronunciation assessment. However, its practicality is still unknown in terms of its effectiveness and efficiency in being used in an EFL class setting. Second, this study can also be extended in comparative ways. There may be other ASR tools that can be used to assess pronunciation, such as on Youtube. Comparing these two ASR AI will need further works to do. Finally, it is necessary to do similar research studies to examine whether the Google Automatic Caption feature can be used to assess other skills.

Acknowledgments

The authors acknowledge the support received from the Institute of Research, Community Service, and Education Quality Assurance of Universitas Tidar. In addition, the authors want to thank the research participants for their contribution to this research.

REFERENCES

Cheng, V. C. W., Lau, V. K. T., Lam, R. W. K., Zhan, T. J., & Chan, P. K. (2020). Improving english phoneme pronunciation with automatic speech recognition using voice chatbot. *Communications in Computer and Information Science* (Vol. 1302, pp. 88–99). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-33-4594-2_8

Cohen, A., & Ezra, O. (2018). Development of a contextualized MALL research framework based on L2 Chinese empirical study. *Computer Assisted Language Learning*, *31*(7), 764–789. https://doi.org/10.1080/09588221.2018.1449756

Dixon, D. H. (2018). Use of Technology in Teaching Pronunciation Skills. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0692

Edmonds, W. A., & Kennedy, T. D. (2020). An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods. In *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781071802779

Evers, K., & Chen, S. (2020). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*.

https://doi.org/10.1080/09588221.2020.1839504

Hoi, V. N., & Mu, G. M. (2021). Perceived teacher support and students' acceptance of mobile-assisted language learning: Evidence from Vietnamese higher education context. *British Journal of Educational Technology*, *52*(2), 879–898. https://doi.org/10.1111/bjet.13044

Hunt-Gómez, C. I., & Navarro-Pablo, M. (2020). ANALYSIS OF PRE-SERVICE FOREIGN LANGUAGE TEACHERS' INCORRECT ARTICULATIONS: FREQUENCY, INFLUENCE ON COMMUNICATION, AND A SPECIFIC CORRECTIVE STRATEGY. *Problems of Education in the 21st Century*, *78*(6), 933–947. https://doi.org/10.33225/pec/20.78.933

Isabelle, D. (2018). Powerful and effective pronunciation instruction: how can we achieve It? *The Catesol Journal*. *30.1*, 13-45.

KimHeyoung, & Kwon, Y. (2012). Exploring Smartphone Applications for Effective Mobile-Assisted Language Learning. *Multimedia-Assisted Language Learning*. *15(1)*, 31-57. https://doi.org/10.15702/mall.2012.15.1.31

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, *38*(1–2), 19–28. https://doi.org/10.1016/S0167-6393(01)00041-3

Lee, J. H., & Cha, K. W. (2020). An analysis of the errors in the auto-generated captions of university commencement speeches on youtube. *Journal of Asia TEFL*, *17*(1), 143–159. https://doi.org/10.18823/asiatefl.2020.17.1.9.143

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

Mayring, P. (2019). Qualitative content analysis: Demarcation, varieties, developments. *Forum Qualitative Sozialforschung*, *20*(3). https://doi.org/10.17169/fqs-20.3.3343

Nelson, P. (2020, December 6). *Google Meet live caption support now available in French, German, Portuguese, and Spanish | Google Cloud Blog*. https://cloud.google.com/blog/products/google-meet/live-captioning-in-google-meet-expanding-to-4-new-languages

Prabhavalkar, R., Sainath, T. N., Wu, Y., Nguyen, P., Chen, Z., Chiu, C. C., & Kannan, A. (2018). Minimum Word Error Rate Training for Attention-Based Sequence-to-Sequence Models. *ICASSP, IEEE*

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v9i2.22482
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

International Conference on Acoustics, Speech and Signal Processing - Proceedings (Vol. 2018-April, pp. 4839-4843). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICASSP.2018.8461809

Rayshata, C. E., & Ciptaningrum, D. S. (2020). Automatic speech recognition to enhance EFL students' pronunciation through Google's Voice Search Application. In *English Linguistics, Literature, and Language Teaching in a Changing Era* (pp. 115–122). Routledge. https://doi.org/10.1201/9780429021039-15

Ryan, S., & Ryuji, T. (2021). Assessing the Practicality of Using an Automatic Speech Recognition Tool to Teach English Pronunciation Online. *STEM Journal*, *22*(2), 93–104. https://doi.org/10.16875/stem.2021.22.2.93

Soni, M., Sheikh, I., & Kopparapu, S. K. (2019). Label-Driven Time-Frequency Masking for Robust Speech Command Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11697 LNAI, pp. 341–351). Springer Verlag. https://doi.org/10.1007/978-3-030-27947-9_29