STUDY ON THE IMPACT OF AI-GENERATED CLONE VOICES ON STUDENT PERFORMANCE IN LISTENING EXAMINATIONS

Arthur Nguyen
Kanda University of International Studies

ABSTRACT

The rise of artificial intelligence has opened new doors for language instruction and allows for new possibilities that were unthinkable just a few years ago. In particular, this has shown promise in the area of standardized language testing and assessment. While popular standardized tests exist, like the TOEIC and TOEFL, they can be cost-prohibitive for some institutions and/or students. Additionally, these exams may not necessarily fit the needs of the stakeholders, and thus may necessitate the need for tailor-made standardized assessments. However, doing so may prove cost prohibitive for many schools, and this is where AI can help bridge that gap. This study looks at whether AI synthesized cloned voices can be a viable, and cost cutting, alternative to using live human voiced actors for a standardized listening exam. The findings of the study show that there is no significant statistical difference between the two types of recordings, and thus AI synthesized cloned voices can be a legitimate replacement for live human voice actors.

INTRODUCTION

In the realm of language education, the emergence of Artificial Intelligence (AI) technologies has revolutionized various aspects of teaching and learning (Wylie & Roll, 2016). One significant application of AI in this domain is the generation of synthetic voices for listening comprehension exercises (Bione & Cardoso, 2016). This study aims to investigate the impact of voice type—AI-generated clone versus human—on student performance in a listening examination.

Listening exams play a critical role in the English as a Foreign Language (EFL) classroom as they assess students' ability to comprehend spoken language, an essential skill for effective communication (Kapanadze, 2019).

Nguyen, A. (2024). Study on the impact of AI-generated clone voices on student performance in listening examinations. *Literacies and Language Education: Research and Practice, Autumn 2024, 66-77*. English Language Institute, KUIS.

In the field of English language proficiency testing, a range of exams are available to assess individuals' ability to communicate effectively in English (Green, 2006). One of the most widely recognized exams is the Test of English for International Communication (TOEIC) (IM & Cheng, 2019). Additionally, the Test of English for International Communication (TOEFL) is also designed to assess the English language skills of non-native speakers (Kang, 2014).

Unfortunately, the cost of taking these exams can be a significant financial burden for test takers (Powers, Yu, & Yan, 2013). The cost of taking language proficiency exams such as the TOEIC and TOEFL has been a topic of concern for many individuals and institutions alike (Owen et al., 2021). The costs associated with these exams may impact students' motivation, preparation, and overall performance in the test (Netta & Trisnawati, 2019). Furthermore, Netta & Trisnawat (2019) emphasized the need for universities and language centers to consider students' financial constraints when administering these exams. A possible remedy to this could be by creating in-house exams which may reduce costs (Lueg & Engelland, 2007) and can be designed to better suit the needs of the students at that particular school or institution (Zhang, Li, & He, 2022). However, creating such exams obviously come with costs of their own. The school or institution must hire or pay staff and/or instructors to not only create the listening test items, but also devote resources to pay for the voice actors for the dialogue and monologues as well. This can be particularly expensive depending on how long the exam is, and how often fresh and new recordings are required. However, by utilizing AI voice cloning, the costs associated with recording a listening exam can be reduced dramatically. The only question then becomes if there is a significant difference between the two in the students' overall listening test performance.

LITERATURE REVIEW

In recent years, there has been a growing interest in integrating artificial intelligence technology into EFL and English as a Second Language (ESL) education (Nguyen, 2020). Addressing this interest, mounting research has contributed to computer-assisted language learning to enhance EFL learners' content and linguistic knowledge needed for writing tasks using multimedia documents (Liu & Tsai, 2012). However, limited research has been conducted on the application of artificial intelligence voice cloning technology specifically for EFL/ESL listening exams. The current literature review aims to explore the potential benefits and challenges of using AI voice cloning for

EFL/ESL listening exams, as well as propose future research directions in this area (Luong & Yamagishi, 2021).

AI has shown early promise in the development and implementation of listening exams for EFL and ESL learners (Vaerenbergh & Pérez-Suay, 2022). Several studies have explored the potential benefits of using AI in listening exams for such learners. AI voice cloning, also known as text-to-speech synthesis, has gained significant attention and implementation in various domains, including education (Luong & Yamagishi, 2021). Several studies have explored the potential of AI voice cloning in improving the listening skills of students during exams (Srinivasan & Murthy, 2021). Thus, voice cloning technology has emerged as a potential solution for generating natural-sounding speech in various applications, and as a means to create speech that closely resembles the natural speech of a target speaker, including the creation of listening exam dialogues (Luong et al., 2020). This groundbreaking technology may offer a solution to help mitigate the costs of creating a standardized in-house listening exam.

METHOD

This study aims to find out to what extent exists any difference between the test scores of students who were given listening exams voiced by human actors versus those voiced by AI clones of the same actors. This research was conducted on 43 students at Kanda University of International Studies (KUIS), a four-year university located in Chiba, Japan. The participants were all currently enrolled in the school's Global Liberal Arts (GLA) and International Communications (IC) departments. The GLA students were all current freshmen, while the IC students were current juniors and seniors at the time the study was being conducted. All participants were volunteers and were reimbursed for their time in the form of an Amazon gift card.

The human voice acting was performed by one female and one male lecturer at KUIS. Their voices were then cloned and replicated using software created by the software company Eleven Labs. This software was chosen due to a combination of its reputation online, as well as its affordability. The voice actors recorded four short dialogues, as well as two short monologues, as referenced in appendix A. Each dialogue was followed up by one multiple choice question, and each monologue followed up by two.

Prior to the test, the students were separated into two different test groups, Group A and Group B. The students were then given a practice listening pre-test that closely resembled the live exam in scope, structure, and

relative difficulty. This allowed the participants to become more familiar with the makeup of the actual exam. The multiple-choice exam was paper-based and were all identical in terms of the number of questions and multiple-choice items. The students were allowed to view the questions and possible multiple-choice answers prior to listening to the audio recordings. All students listened to four separate short conversations between a male and female speaker, and were asked to choose the best multiple choice answer after each dialogue. Additionally, they listened to two separate short monologues, and had to answer two questions after each dialogue. The scripts were exactly the same for all students taking the exam, however, Group A listened to a recording performed by human voice actors for monologue one, dialogue one, and dialogue two, with monologue two, dialogue three, and dialogue four performed by their AI cloned counterparts. The second recording had the roles reversed, with Group B listening to monologue one, dialogue one, and dialogue two performed by the AI clones, with monologue two, dialogue three, and dialogue four performed by their AI cloned counterparts. The students were not told about the actual nature of the study until after the test finished, and were under the impression that this was simply a listening exam. The exam answers were then transcribed onto an Excel spreadsheet, and then analyzed for statistical significance through SPSS. The final outcome will determine whether or not AI generated voices can replace those voiced by human ones.

RESULTS

To determine the significance of the relationship between student scores, and the type of voice recording they received, an independent t-test was conducted. The students that listened to the AI cloned voiced recordings (M = 60, SD = .693) performed almost entirely similarly to the ones performed by the two human actors (M = 59, SD = .687), $t(168) = -.2222$, $p = .963$. An alpha level of .05 was used for this test, and the differences between the two recordings were not statistically significant. This is further supported by looking at the magnitude of the differences in the means (mean difference = -.024, 95% CI: -.232 to .185), which were also not significant.

DISCUSSION

The findings of this study suggest that the type of voice, whether AI-generated or human, does not significantly impact student performance in listening comprehension examinations. This result contradicts some previous studies that have suggested potential advantages of human voices in language learning contexts.

One possible explanation for these results could be the quality of the AI-generated voice used in the study. The advancements in AI technology have led to increasingly realistic synthetic voices. The participants in this study may have not even noticed that there was any difference in acoustic quality between the two different voice recordings. It is also possible that even if participants still perceived subtle differences between the AI-generated clone and human voices, they were familiar enough with AI technology that it led to comparable performance outcomes. In an era where AI applications are becoming more pervasive, students may be increasingly accustomed to interacting with AI systems, potentially diminishing any perceived difficulty or anxiety associated with AI-generated voices. Furthermore, the nature of the listening examination itself may have influenced the results. The content, difficulty level, and format of the examination could have influenced how students process and comprehend the dialogue and monologues.

It is essential to acknowledge some limitations of this study. Firstly, the sample size may have constrained the statistical power to detect small differences between the two groups. Additionally, the generalizability of the findings may be limited to the specific context and population studied, as this was conducted solely on university students.

Future research could explore other factors that may influence the effectiveness of AI-generated voices in language learning contexts, such as voice characteristics (e.g., accent, intonation), task types, and individual learner differences. Additionally, investigating longitudinal effects and conducting comparative studies across different languages and proficiency levels could provide further insights into the role of AI in language education.

CONCLUSION

In conclusion, the current study found no significant difference in student performance regarding use of AI-generated and human voices in a listening comprehension examination. This suggests that AI-synthesized voices can serve as a viable alternative to human voices in standardized listening exams. The quality of AI-generated voices may have reached a point where they are indistinguishable from human voices to interlocutors, potentially diminishing perceived differences in performance outcomes. Thus, it is reasonable to conclude that, all conditions being equal, listening exams can freely utilize the more cost-effective option.

Nguyen, A. (2024). Study on the impact of AI-generated clone voices on student performance in listening examinations. *Literacies and Language Education: Research and Practice, Autumn 2024, 66-77*. English Language Institute, KUIS.

REFERENCES

Bione, T., & Cardoso, W. (2020). Synthetic voices in the foreign language context. *Language Learning and Technology*, 24(1), 169–186. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/54f66cfe-2240-41a9-8c61-54a2396388fa/content

Dong, Y. (2023). Revolutionizing academic English writing through AI-powered pedagogy: Practical exploration of teaching process and assessment. *Journal of Higher Education Research*, 4(2), 52–52. https://doi.org/10.32629/jher.v4i2.1188

Green, A. (2006). Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing*, 11(2), 113–134. https://doi.org/10.1016/j.asw.2006.07.002

Im, G. H., & Cheng, L. (2019). The test of English for international communication (TOEIC®). Language Testing, 36(2), 315–324.

Kang, C.C. (2014). Pedagogical implications of score distribution pattern and learner satisfaction in an intensive TOEIC course. *Canadian Center of Science and Education*, 7(11), 64–78. https://doi.org/10.5539/elt.v7n11p64

Kapanadze, D.U. (2019). An effective method to develop watching/listening comprehension skills In Turkish teaching . *International Journal of Progressive Education*, 15(6), 66–82. https://eric.ed.gov/?id=EJ1237230

Liu, P E., & Tsai, M. (2012). Using augmented-reality-based mobile learning material in EFL English composition: An exploratory case study. *British Journal of Educational Technology*, 44(1), E1–E4. https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/j.1467-8535.2012.01302.x

Lueg, J. E., & Engelland, B. T. (2007). The development and administration of an in-house knowledge examination for academic program assessment. *Marketing Education Review*, 17(2), 15–24. https://www.tandfonline.com/doi/abs/10.1080/10528008.2007.11489000

Luong, H.-T., & Yamagishi, J. (2021) Preliminary study on using vector quantization latent spaces for TTS/VC systems with consistent performance.

Proc. 11th ISCA Speech Synthesis Workshop (SSW 11), 136–141. https://doi.org/10.21437/ssw.2021-24

Luong, H., Yamagishi, J., Wen, Z., & Zhong, R. (2020). Latent linguistic embedding for cross-lingual text-to-speech and voice conversion. https://doi.org/10.21437/vcc_bc.2020-22

Mahmoud Ghoneim, N. M., & Abdelsalam Elghotmy, H. E. (2021). Using an artificial intelligence based program to enhance primary stage pupils' EFL listening skills. *Journal of Education-Sohag University*, 83. https://edusohag.journals.ekb.eg/article_140694.html

Min, S., Zhang, J., Li, Y., & He, L. (2022). Bridging local needs and national standards: Use of standards-based individualized feedback of an in-house EFL listening test in China. *Language Testing*, 39(3), 425–452. https://journals.sagepub.com/doi/abs/10.1177/02655322211070990

Netta, A., & Trisnawati, I K. (2019). Achenese undergraduate students' strategies in preparing for TOEFL prediction: A preliminary study. *Journal of Language, Education, and Humanities*, 7(1), 41–52. https://doi.org/10.22373/ej.v7i1.5779

Nguyen, M. T. (2020). Understanding listening comprehension processing and challenges encountered: Research perspectives. *International Journal of English Language and Literature Studies*, 9(2), 63–75. https://doi.org/10.18488/journal.23.2020.92.63.75

Owen, N., Shrestha, P., & Hultgren, A. K. (2021). Researching academic reading in two contrasting English as a medium of instruction contexts at a university level. *ETS Research Report Series: Volume 2021*, 1, 1–28. https://doi.org/10.1002/ets2.12317

Powers, D. E., Yu, F., & Yan, F. (2013). The TOEIC® listening, reading, speaking, and writing tests: Evaluating their unique contribution to assessing English-language proficiency. *The research foundation for the TOEIC tests: A compendium of studies*, 2, 3–1. https://www.ets.org/Media/Research/pdf/TC2-03.pdf

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26, 582–599. https://link.springer.com/article/10.1007/s40593-016-0110-3

Srinivasan, V., & Murthy, H N. (2021). Improving reading and comprehension in K-12: Evidence from a large-scale AI technology intervention in India. *Computers and Education: Artificial Intelligence*, 2, 100019–100019. https://doi.org/10.1016/j.caeai.2021.100019

Stock, I., & Eik-Nes, N L. (2016). Voice features in academic texts: A review of empirical studies. *Journal of English for Academic Purposes*, 24, 89–99. https://doi.org/10.1016/j.jeap.2015.12.006

Vaerenbergh, S V., & Pérez-Suay, A. (2022). A Classification of Artificial Intelligence Systems for Mathematics Education. *Springer Nature (Netherlands)*, 89–106. https://doi.org/10.1007/978-3-030-86909-0_5

APPENDIX

Transcript of Listening Exam

Monologue 1

In 1891, an important moment in history occurred during a Physical Education class at YMCA College in Massachusetts, U.S.A. James Naismith, wanted to keep his P.E. students active on a rainy day. Right before the class, Naismith invented and wrote down the rules for a new game on a single piece of paper, which involved two peach baskets. The baskets stood three meters high and were attached to walls at opposite ends of the gymnasium. Naismith's students quickly embraced this imaginative game. It involved engaging in friendly competition by passing a ball to teammates and jumping and shooting it into the baskets.

Little did they know, this would become the first official game of basketball ever in history. Improvements to the game would be gradually added over some time, like removing the bottoms of the baskets for faster gameplay. Additionally, dribbling, or the bouncing of the ball on the floor, was included as part of this game just a few months later. Naismith's game of basketball has changed a lot over the years and has become one of the most beloved sports in the world today. (186 words)

1) Which of the following is correct about James Naismith?

a) He wrote down the rules of basketball in a notebook.

b) His students were not interested in his new game.

c) He was a high school physical education (P.E.) teacher.

d) He invented basketball on the same day it was played. *

2) Which of the following was NOT part of the first game of basketball?

a) passing

b) dribbling *

c) jumping

d) shooting

Monologue 2

You might think that all car and truck fuels are the same, but they aren't. Automobiles can run on one of three types of fuel: gasoline, diesel, and biodiesel. All these kinds of fuel are burned inside of the engine, which creates the heat and energy that is used to run and power the automobile. However, there are important differences between these types of fuel.

Gasoline and diesel are more common than biodiesel, but each burns differently. Diesel fuel is heavier and less burnable than gasoline, so it must be compressed first before it can burn. Gasoline may be lighter than diesel, but both fuels are made from crude oil. On the other hand, biodiesel fuel is made from vegetables. Both biodiesel and diesel fuels must only be used in diesel engines. If gasoline is pumped into a diesel engine, the automobile will not run. Even worse, it could even cause severe damage to the engine itself.

All these kinds of fuel look very similar at the gas station, so remembering the important differences between them is critical in protecting your automobile. (182 words)

1) Which of the following is correct?

a) Biodiesel is more common than diesel and gasoline.

b) Gasoline must be compressed first before it can burn.

c) Diesel and biodiesel fuel can be used in diesel engines. *

d) Gasoline is lighter and more burnable than diesel fuel.

2) Which of the following is correct about gasoline, diesel, and biodiesel fuel?

a) They all look very different at gas stations.

b) They all produce heat and energy in engines. *

c) They are all made mainly from crude oil.

d) They are safe to use on any kind of automobile.

<u>Dialogue 1</u>

M: Oh, no! I think I forgot to bring the concert tickets.

W: Seriously? Aren't the tickets on your smartphone? I thought bought them with your app.

M: I did, but my phone broke this morning, so I printed out the QR code with my PC.

W: Well, I guess we have to run home now and get them. Luckily, we have plenty of time before the show starts.

1) Which of the following is correct?

a) The man bought concert tickets with his smartphone. *

b) The woman will go home alone to get the concert tickets.

c) The woman thinks that she will be late to the concert.

d) The woman forgot to bring the concert tickets.

Nguyen, A. (2024). Study on the impact of AI-generated clone voices on student performance in listening examinations. *Literacies and Language Education: Research and Practice, Autumn 2024, 66-77*. English Language Institute, KUIS.

Dialogue 2

M: I can’t think of any good ideas for this English essay.

W: When I can’t think of anything, I sometimes do something else while listening to some soft music.

M: That sounds like a good idea. I think I’ll just lie down on the couch and listen to some tunes.

W: Well, don’t get too relaxed. Remember, our assignment is due tomorrow morning.

Which one of the statements is correct about the man?

A) He is thinking of taking a nap on the couch.

B) He gives some advice to the woman.

C) He is taking the same class as the woman. *

D) He is almost finished with his English essay.

Dialogue 3

M: Hey, I didn’t see you in history class today.

W: Yeah, I skipped class to finish the chemistry homework that’s due later today.

M: Didn’t you get the e-mail? Dr. Smith gave us two more days to finish it.

W: Really? I checked my inbox like five minutes ago and didn’t get anything from her.

Which of the following is correct?

a) The woman read an e-mail today from Dr. Smith

Nguyen, A. (2024). Study on the impact of AI-generated clone voices on student performance in listening examinations. *Literacies and Language Education: Research and Practice, Autumn 2024, 66-77*. English Language Institute, KUIS.

b) The woman’s chemistry homework is due today.

c) The man attended the history class today. *

d) The woman skipped the chemistry class today.

Dialogue 4

M: Hey, I’m driving to the store. Do you need anything?

W: Sure. Can you pick up some eggplant, onions, tomatoes, ground beef, eggs…

M: Hang on, hang on. Let me write that all down.

W: Actually, why don’t I just come with you?

M: Sure, but can you drive instead of me?

Which of the following is correct?

a) The woman wants the man to drive her to the store.

b) The woman asks the man to buy only food at the store. *

c) The man asks the woman to go to the store with him.

d) The man wants the woman to cook him some dinner.