International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 01 | Jan 2024 www.irjet.net p-ISSN: 2395-0072

Usability of input methods within mobile devices: A comparative study between Voice-to-text vs soft keyboard

Rudraksh Gupta

BCA, Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar

---

Abstract - *This study, rooted in Human-Computer Interaction (HCI), meticulously compares mobile input methods— Voice-to-text and Soft Keyboard. The research was conducted by Guru Nanak Dev University student and scrutinizes their advantages and limitations through a 2 × 2 within-subjects design with 10 participants. Contrary to expectations within the HCI framework, Soft Keyboard consistently outperforms Voice-to-text in input speed, indoors and outdoors. This challenges assumptions and calls for future HCI research to embrace a comprehensive evaluation framework, considering factors like user preferences and accuracy, ensuring mobile interfaces align seamlessly with user needs.*

Key Words: *Voice-to-text, Soft Keyboard, Usability*

1.INTRODUCTION

In our rapidly changing, technology-driven world, the devices we use daily have a large impact on how we interact with the digital world. Human-Computer Interaction (HCI) looks to study this essential connection while focusing on enhancing the user experience. Within HCI, the choice of input method is a crucial decision, greatly affecting how users engage with systems. This research delves into two distinct input methods: voice-to-text and standard keyboard input, aiming to shed light on their relative advantages and disadvantages.

Recent years have seen impressive advancements in voice recognition technology. These improvements have created a new era where users can speak to their devices, reducing the need for typical keyboard input. The evolution of this technology is groundbreaking as it completely transforms the way we communicate with digital devices.

The appeal of voice-to-text input lies in its speed and accessibility for users with impairments. Users can talk to their devices and see their words appear as text rather than typing each word out using their fingers. This technology has made voice-to-text input a very intriguing alternative to traditional keyboard input.

However, beneath these advantages, voice-to-text input also produces its own set of challenges. Users often have concerns about its accuracy and overall usability. There are questions surrounding potential trade-offs between the speed of voice-to-text and the precision of typing on a keyboard.

This research seeks to explore HCI by analyzing the comparative advantages and disadvantages of two distinct input methods – voice-to-text and standard keyboard input. Through an in-depth investigation, this study aims to provide insights into the optimal input method for different scenarios, thus contributing to the design of user interfaces.

RELATED WORK

1.1 A Comparison of Speech and Typed Input

The research conducted by Carnegie Mellon, titled "A Comparison of Speech and Typed Input," (Hauptmann, 1990) provides a thorough exploration of the dynamics between voice and keyboard input within the context of digit entry tasks. The primary aim of the study is to conduct a detailed comparison between these two input modalities, shedding light on their respective advantages and limitations.

In terms of experimental setup, the research encompassed two experiments: one involving the presentation of digit strings on a screen and the other requiring subjects to read from a paper (Hauptmann, 1990). Three distinct data entry modes were considered—voice only, voice with keyboard correction, and keyboard only. Each experiment involved subjects entering three lists of 66-digit strings, facilitating a comprehensive examination of various input conditions (Hauptmann, 1990).

Crucially, the study's findings regarding input duration time highlight the efficiency of speech input. The average difference between pronouncing a digit string and typing one was reported to be less than 2 seconds in both experiments (Hauptmann, 2022). Real-time response and accurate speech recognition emerged as pivotal factors influencing speech as a preferable communication mode (Hauptmann, 1990).

Regarding accuracy, the research revealed high typing accuracy for both paper and on-screen presentations. However, recognition word accuracy

was notably higher for on-screen presentations compared to paper presentations, emphasizing the influence of presentation mode on the effectiveness of the input modality (Hauptmann, 1990).

The key findings and solutions proposed by the research suggest that the advantages of each input modality are contingent upon factors such as string length and system response characteristics. The study posits that for more complex tasks requiring a greater number of keystrokes per syllable, speech input could demonstrate superiority (Hauptmann, 1990). Furthermore, in tasks demanding visual monitoring, speech emerges as a preferable input channel, alleviating the cognitive load associated with dividing attention between the keyboard, screen, and paper (Hauptmann, 1990).

Concerning our research on voice-to-text input versus standard keyboard input, the insights gained from the Carnegie Mellon study provide valuable context. The emphasis on real-time response, accuracy, and the impact of visual effort aligns with the core challenges in developing effective voice-to-text interfaces. By leveraging the findings of this related work, we can draw parallels and distinctions, enriching our understanding of the dynamics at play and guiding the design and evaluation of voice-to-text systems.

1.2 Comparing Voice Chat and Text Chat in a communication Tool for Interactive Television

Interactive Television (iTV) involves the use of voice chat and text chat communication modes (Geerts, 2006). This Noteworthy research focused on the pros and cons of these communication mechanisms during TV watching. This paper has been of great help as it offered some insights into how communication modes interact with their audiences while watching television.

One critical aspect of ITV is supporting the social uses of television. Communication has been aided during and after TV watching has widely been accepted. This includes the possibility of communicating with relatives, friends, or neighbors. Such systems enable users, located in different spaces through voice chat or text chat.

The use of CMC has been widespread in instant messaging applications, voice communication, and other areas (Geerts, 2006). Indeed, this mode of communication is being employed in various scenarios such as work, school, and leisure. A lot of these settings involve multitasking to examine the convenience and distractedness of voiced instructions versus text commands.

According to Wikipedia, backchannel communication refers to an online discussion happening simultaneously with live verbal commentary (Geerts, 2006). Previous studies have examined the use of backchannels in academic conferences; however, this one applies the “backchannel” concept for watching TV. The method that is used to carry out this research is quite new but it serves as a basis for understanding how the problem of a backchannel (voice chat or text) works when watching TV or primary material.

Fascinating results were derived from a comparative study that centered around voice chat and text chat in interactive television. It was seen that voice chat was more natural and direct, hence, easier to pay attention to the television program. However, young people and experienced computer-based communicative users preferred instead to text chat (Geerts, 2006). These insights are important in helping to appreciate how user-friendly these communication channels are and the extent they can shape viewers’ experience.

Summarily, this previous research brings into perspective the practical utility, preferences of users, as well as possible distractions that may result from the integration of voice chat with text chats when using it together with television viewing. Using this fundamental perception, this study explores the comparative analysis of voice and text communication within different contexts.

1.3 A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations

Conversational evaluation tools in educational research offer great potential to improve data quality and end-user experience. This study will discuss two separate modalities (text-based vs voice-based) for evaluating courses, building on prior works in the area (Theimo, Naim and Matthias 2022).

Previous studies have shown how conversational interactive surveys can be employed for educational purposes (Theimo, Naim and Matthias 2022). The results demonstrate that these tools are developed to foster dialogic interactions in which the results are more qualitative and deeper than others. This aligns with this research study that considers a user-centred methodology for adaptive conversational interface development for educational purposes. Moreover, the other study has revealed positive effects on customer experience especially social presence and interactive enjoyability (Theimo, Naim and Matthias 2022).

It is becoming increasingly interesting whether text-centered or voice-centered modalities are appropriate within conversational interfaces. Detailed and informative response is a sure way of improving response quality when using text-based conversational tools. However, voice-based interactive learning systems are still in their infancy and have their own advantages. It is a useful point in this study's course evaluations with which it can compare these two modes, noting their advantages. Research on the measurement of response quality has been a common area of study in the literature, focusing on lexical, syntactical, and semantic responses. Consistent with this strategy, this research acknowledges that high-level measures are required to thoroughly appraise response quality, emphasizing attributes such as specificity and pertinence in determining response (Theimo, Naim and Matthias 2022).

For years, researchers have been concerned with improving users' experience in educational settings, including course evaluations. Studies conducted before, show that conversational interface affects self dis-closing and interactional enjoyment and perceived social presence. These trends reflect the findings that show how conversational tools can lift up the user experience for educational purposes.

Research areas for future development are equally numerous and exciting. Voiced education, contexts for conversational interface and applying voice analytics for survey enrichment. Hence, researchers should explore how cultural and contextual matters affect the efficiency of conversational instruments across educational surroundings

1.4 A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations Transcription

Transcription In the domain of text entry tasks, Carnegie Mellon's research paper titled "A Comparison of Speech and Typed Input" (Margaret, Géry and Daniel 2020) emerges as a seminal work, delivering an exhaustive exploration of the interplay between voice and keyboard input, specifically within the context of digit entry tasks. The primary aim of this study is to meticulously compare these two input modalities, shedding light on the associated advantages and limitations (Margaret, Géry and Daniel 2020).

The study conducted two experiments, one presenting digit strings on a screen and the other requiring subjects to read from a paper. Three distinct data entry modes were considered: voice only, voice with keyboard correction, and keyboard only. Significantly, each experiment comprised subjects entering three lists of 66-digit strings, ensuring a comprehensive examination of various input conditions (Margaret, Géry and Daniel 2020).

A critical revelation of the research lies in the significant difference in input duration times between speech and traditional typing. Speech input demonstrated faster input durations compared to traditional typing, with an average difference of less than 2 seconds in both experiments (Margaret, Géry and Daniel 2020). Real-time response and accurate speech recognition were identified as pivotal factors favoring speech as a preferable communication mode.

Concerning accuracy, the study unveiled high typing accuracy for both paper and on-screen presentations. However, recognition word accuracy was notably higher for on-screen presentations than paper presentations, emphasizing the influence of presentation mode on the effectiveness of the input modality (Margaret, Géry and Daniel 2020).

The research's key findings posit that the advantages of each input modality hinge upon factors like string length
and system response characteristics. It suggests that for more complex tasks requiring a greater number of keystrokes per syllable, speech input could exhibit superiority (Margaret, Géry and Daniel 2020). Furthermore, in tasks demanding visual monitoring, speech emerges as a preferable input channel, alleviating the cognitive load associated with dividing attention between the keyboard, screen, and paper (Margaret, Géry and Daniel 2020).

This related work bears significant relevance to this study's exploration of voice-to-text input versus standard keyboard input. The emphasis on real-time response, accuracy, and the impact of visual effort aligns with the core challenges in developing effective voice-to-text interfaces (Margaret, Géry and Daniel 2020).

By leveraging the findings of this related work, parallels and distinctions can be drawn, enhancing the understanding of the dynamics in the comparison between voice and keyboard input. The detailed comparison of input duration times provides a quantitative benchmark for the exploration of similar aspects in the context of voice-to-text interfaces. Additionally, the proposed solutions in the related research form a foundation for addressing challenges in this study's investigation, guiding the design and

evaluation of voice-to-text systems for optimal user experience and efficiency.

2. METHOD

For the purpose of comparing the inputs of both voice-to-text and typing through a soft keyboard, experimental research was conducted. The experiment consisted of a number of people and each participant tested both variables, the voice-to-text and typing using a soft keyboard. Recording each result, a comparison of performance and usability was made.

2.1 Hypothesis

The hypotheses of the user study is split into 2 categories, the performance hypotheses and the usability hypotheses.

For the performance hypotheses:

- H0: No clear difference in terms of input speed between the two methods. ($\alpha = 0.05$)

- Ha : There will be a significant difference in input speed between the two methods.

As for the usability hypotheses:

- H0 : Using either input method will have no difference with respect to errors made. ($\alpha = 0.05$)
- Ha : Using either input method will have a significant difference in respect to errors made.

Given that the nature of this user study is quantitative and repetitive, a paired sample T-test and a F-test was employed to observe any significant differences in the means. This test was appropriate since the same participants tested both input methods.

2.2 Participants

Number of Participants: The study involved a total of 10 participants, split into two groups.

Population from which they were drawn:

Participants were drawn from a diverse population of smartphone users spanning various genders, ages (16-70s), educational backgrounds, and technology experience. This inclusive approach included high school students, Baby Boomers, and millennials, with representation from males, females, and non-binary individuals. In terms of prior experience with technology, participants exhibited varying levels, from older individuals with limited exposure to technology to university students with an abundance of experience.

A diverse group of 10 participants were selected for the study using a combination of recruitment methods tailored to each group. These methods included convenience sampling, giving information on university bulletin boards, corporate channels, social media, community centers, support groups, and online communities.

2.3 Apparatus

For the experiment's apparatus, a standard smartphone can be used to conduct both the voice-to-text and manual typing features. Since most voice assistants are similar to one another there wouldn't be a significant difference in performance when using various devices. However, for consistency's sake, throughout the research, this experiment was conducted using an Apple 13 Pro Max. The built-in voice assistant (Siri) was used to conduct the voice-to-text section of the experiment and the iMessage app was used for the typing section of the experiment.

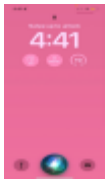

Figure 1. Screenshot of Apple's Siri Voice Assistant on the bottom center

This experiment is easily reproducible, as the voice assistant feature as well as a default messaging app exists in all smartphones.

2.4 Procedure

Participants were greeted and given an overview of the study's objectives. Informed consent was obtained, and participants were assured of the confidentiality of their data. Participants were divided into two equal groups to ensure balanced representation. The order of the trials was counterbalanced to eliminate any order effects.

Participants were introduced to the Apple 13 Pro Max smartphone and the iMessage app. Basic instructions on how to use the voice-to-text feature (Siri) and the soft keyboard were provided. Participants were given a clear description of the experiment task. They were informed that the task involves sending a specific phrase through the iMessage app, either using voice-to-text or manual typing. Participants were informed about the two environmental conditions: sitting indoors and walking in public. They were instructed on how to transition between these conditions.

Participants were allowed a brief practice phase of 5 minutes to familiarize themselves with the voice-to-text feature and the soft keyboard. For the main experiment phase, participants were divided into two equal groups. The experiment was performed in 2 environments i.e., sitting inside and walking in a public setting. For group 1, For each trial, they received the specific phrase: "The

brown fox jumps over the lazy dog". They were instructed to send the phrase via. iMessage using the voice-to-text feature (as input 1) while they were sitting indoors first, and then by soft keyboard (as input 2). Afterwards, they were instructed to send the same phrase while walking outdoors in public using the voice assistant first, then through the soft keyboard.

For group 2, For each trial, they received the same phrase "The brown fox jumps over the lazy dog". They were instructed to send the given phrase via. iMessage using the soft keyboard (as input1) first while walking outside. Then, they were instructed to use the voice input method (as input 2) next. Afterwards, the participants were taken inside, where they sent the same phrase using a soft keyboard first while sitting, then used the voice-to-text feature to send the phrase while remaining seated.

Each participant's time to complete the task and the number of errors made were recorded for every trial. The experiment took approximately 10-20 minutes for each participant, including the practice phase and the main experiment phase

Figure 2. A participant performs the text-to-speech task feature for input [1]

Figure 3. A participant using the soft keyboard as an input method [2]

2.5 Design

The user study applied a 2 × 2 within-subjects design. The participants were each tested using the smartphone's voice-to-text feature or the soft keyboard. Within the user study, each participant was also asked to utilize the input methods in two varying environment conditions. Each input method was tested while sitting indoors and walking in public. Each trial was timed to compare the overall performance results of the two input methods. The independent variables in this study consist of:

- Input method: Voice-to-text or soft keyboard
- Environment conditions: The participants were either sitting indoors or walking outside in public.

The dependent variables in this study consist of:

- Time: The amount of time it takes the participant to complete the task using the voice assistant or the on-screen keyboard.
- Accuracy: The number of errors the voice assistant or the participant may make during the tests.

The participants were divided into two equal groups and the order of the trials will be counterbalanced to reduce the effects of earlier trials on later trials. Each of the 10 participants were told to send the given phrase through the messaging app using both voice-to-text and the on-screen keyboard. Each input method was tested while sitting or walking. Therefore, the total number of trials was 10 participants × 2 input methods × 2 participant/surrounding conditions = 40 total trials

3. RESULTS

After conducting and collecting the resulting data from the user study, the data was then averaged as a mean for performance and usability separately with further separation considering the environmental settings for the trials.

As for performance, the means of the data collected were organized into charts to better visualize the comparison. Overall, when calculating the variances of the input speeds using ANOVA (F-test), the F-value resulted to be 5.55 with a P-value of 0.003. Given that the P-value < 0.05, there is a strong indication of a significant difference in the input speeds so the performance null hypothesis was rejected.

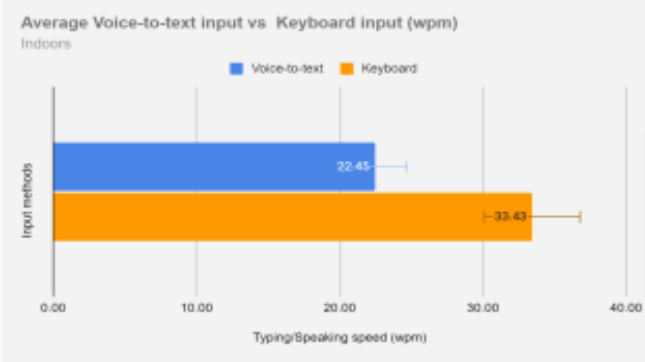

Figure 4. Chart depicting the Average input speed (wpm) for Voice-to-text vs Keyboard while participants were inside.

Figure 4, illustrates the average input speeds for the two input methods compared when the participants were performing their trials while sitting indoors. The average input speed in words per minute (wpm) for voice-to-text input was 22.45, while for soft keyboard was 33.43. The soft keyboard proved to be 39% faster for participants

compared to voice input when the percentage difference was calculated. The difference was statistically significant where (T-statistic = -3.787, $p < .05$)

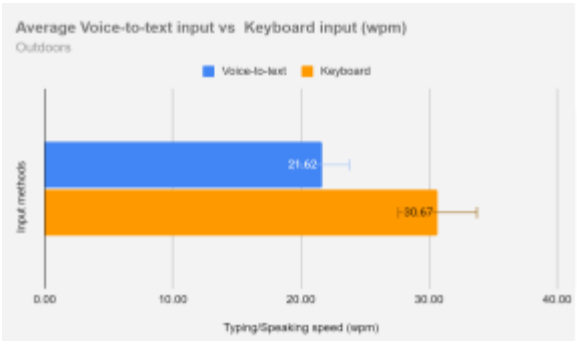

Figure 5. Chart depicting the Average input speed (wpm) for Voice-to-text vs Keyboard while participants were outside.

When the participants were using the two input methods while walking outdoors in public, the mean for the input speed resulted in 21.62 wpm for voice-to-text. On the other hand, the mean for soft keyboard was 30.67 wpm. The percentage difference between the two means is calculated to be 35%. The difference was statistically significant (T-statistic = -2.21, $p < .05$).

The soft keyboard was faster than the voice-to-text input method in terms of input speed for both environmental scenarios. Another finding to mention for input speed was that for both input methods respectively, participants had an overall faster input speed when they were experimenting indoors rather than outdoors.

With regards to usability, Using the ANOVA (F-test), the variance of the number of errors was calculated. The results were, F-value at 2.38 and a P-value at 0.0856. This suggested that there is not a significant difference and not enough evidence to reject the usability null hypothesis at a 0.05 significance level.

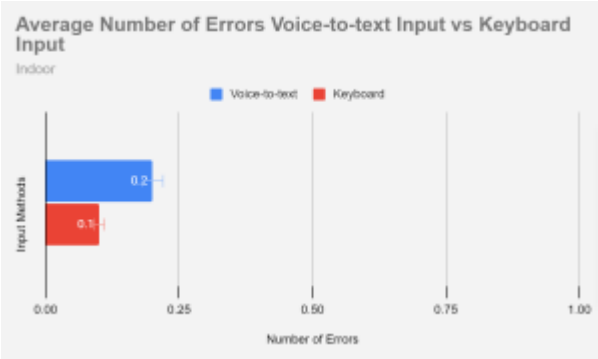

Figure 6. Chart depicting the Average number of errors for Voice-to-text vs Keyboard while participants were inside.

Figure 6 shows the average number of errors for both voice-to-text and keyboard inputs while sitting indoors. Voice-to-text had an average of 0.2 errors while the keyboard input had an average of 0.1 errors. The T-test indicated there is no clear evidence of a significant difference between the two input methods (T-statistic = 0.59, $p > 0.05$).

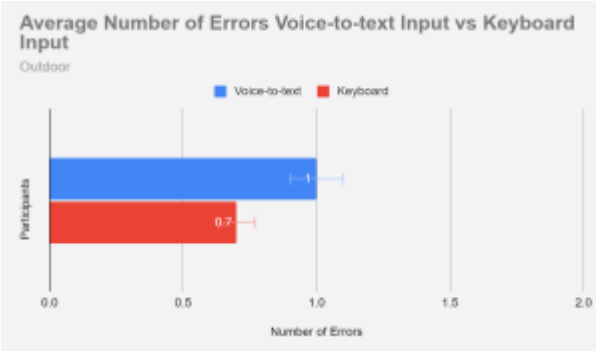

Figure 7. Chart depicting the Average number of errors for Voice-to-text vs Keyboard while participants were outside.

Figure 7 depicts the average number of errors when participants moved outside. Voice-to-text had an average of 1 error while typing with the keyboard had an average of 0.7 errors. Again, the T-test indicated there is no clear evidence of a significant difference between the two input methods (T-statistic = 0.81, $p > 0.05$).

4. DISCUSSION

The results of the user study provide valuable insights into the performance and usability of voice-to-text and soft keyboard input methods in different environmental settings. The findings are presented separately for performance and usability aspects.

In terms of input speed, the soft keyboard consistently outperformed the voice-to-text input method in both indoor and outdoor scenarios. Participants exhibited a 39% and 35% faster input speed with the soft keyboard compared to voice-to-text when inside and outside, respectively. The statistical significance of the differences, as indicated by T-statistics, further supports the conclusion that the soft keyboard is significantly faster for input speed in both settings.

It is noteworthy that participants demonstrated a generally faster input speed when experimenting indoors rather than outdoors, irrespective of the input method. This observation suggests that environmental factors play a role in influencing input speed, with indoor settings potentially providing a more conducive environment for efficient input.

The rejection of the performance null hypothesis emphasizes the importance of considering input speed as a critical factor when comparing voice-to-text and soft keyboard technologies. The practical implications of these findings may influence the design and selection of input methods in various applications, particularly those requiring rapid data entry.

Contrary to the performance results, the usability aspect, as measured by the average number of errors, did not show a significant difference between voice-to-text and soft keyboard input methods. The T-tests indicated no clear evidence of a difference in error rates between the two methods, both indoors and outdoors. This suggests that, at least in terms of error rates, users can expect a comparable experience with either input method.

The absence of a significant difference in error rates may indicate that users can achieve similar levels of accuracy with voice-to-text and soft keyboard inputs. This finding is essential for applications where accuracy is a critical factor, as it suggests that users can adapt to either input method without compromising on error rates.

While the study provides valuable insights, it is essential to acknowledge certain limitations. The study focused on input speed and error rates but did not explore other aspects of user experience, such as user preference, comfort, or fatigue. Future research could delve deeper into these factors to provide a more comprehensive understanding of the overall user experience with voice-to-text and soft keyboard technologies.

In summary, the results of this user study contribute valuable insights into the performance and usability of voice-to-text and soft keyboard input methods. The findings highlight the importance of considering specific use cases and environmental factors when choosing between these two input methods, providing a foundation for future research and design considerations in human-computer interaction.

5. CONCLUSION

In conclusion, this study marks a significant advancement in our understanding of mobile input methods within the realm of Human-Computer Interaction. The unexpected and consistent superiority of the Soft Keyboard in terms of input speed prompts a paradigm shift, challenging prevailing assumptions about the perceived efficiency of Voice-to-text. This newfound insight, critical in the context of HCI, catalyzes future research endeavours. The call is for HCI studies to broaden their scope beyond input speed and delve into a myriad of factors, encompassing user preferences, accuracy, and fatigue. By adopting this comprehensive approach, future HCI research can not only refine the design of mobile interfaces but also provide actionable insights that enhance overall usability and satisfaction for users navigating the dynamic landscape of mobile devices.

6. REFERENCES

1. A participant perform text to speech task feature for input. https://www.wired.com/images_blogs/gadgetlab/2013/03/13 0318_siri_0078.jpg

2. A participant using soft keyboard as a input method. (n.d.) https://heresthethingblog.com/wp-content/uploads/2012/11/7-essential-iPhone-typing-tips.jpg

3. Margaret Foley, Géry Casiez, Daniel Vogel. Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription. CHI 2020 - ACM Conference on Human Factors in Computing Systems, Apr 2020, Honolulu, United States. pp.1-11, ff10.1145/3313831.3376861ff. ffhal02919659f

4. Hauptmann, A. G., & Rudnicky, A. I. (1990). A Comparison of Speech and Typed Input. Proceedings of the ICASSP, 45-48. Retrieved from https://aclanthology.org/H90-1045.pdf

5. Geerts, D. (2006, October). Comparing voice chat and text chat in a communication tool for interactive television. In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles (pp. 461-464).

6. Wambsganss, T., Zierau, N., Söllner, M., Käser, T., Koedinger, K. R., & Leimeister, J. M. (2022). Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1-27.