Form interaction through Speech Recognition

Carmine D'Angelo
c.dangelo23@studenti.unisa.it
Università degli studi di Salerno
Fisciano, SA, Italia

Emanuele Vitale
e.vitale19@studenti.unisa.it
Università degli studi di Salerno
Fisciano, SA, Italia

Francesco Aurilio
f.aurilio@studenti.unisa.it
Università degli studi di Salerno
Fisciano, SA, Italia

Abstract

In recent years, the devices we interact with every day have changed with the advent of many mobile devices. As a result, the ways in which we interact with these devices have also evolved, especially with regard to voice interaction. The study's goal is to examine different types of textual voice input. Depending on the method of interaction used by the user to input the appropriate values in the fields, several modes are available. Short text fields, long text fields, short select fields, and long select fields are all included in four different formats. The trials shown how the task completion time is greatly influenced by the insertion mode, ranging from an average of 88.45 seconds for interactive mode to 27.41 seconds for mixed mode. The type of field that needed to be filled out text or select was also important for the experiment's outcomes, with the form with just text fields yielding noticeably different outcomes.

Keywords: Form, Form interaction, Human-computer interaction, Speech recognition, Text to speech, Voice, Voice interaction, Web browsing

1 Introduction

Human-computer interaction is changing significantly in the digital age as more and more individuals access the internet through mobile and linked devices. People may communicate with gadgets and obtain information using simple voice commands thanks to voice interfaces, which are a fundamental advance in the user experience.

With innovations in speech recognition technology, natural language processing, and artificial neural networks, speech interface technology has come a long way. The development of speech interfaces in a variety of situations, including mobile devices, smart speakers, and automobiles, has been sparked by the introduction of voice assistants from leading technology companies similar to Siri, Alexa, and Google Assistant. The majority of research on voice interfaces, however, has concentrated on accessibility for illiterates or persons with disabilities, omitting a complete examination of the utility of such interfaces for those without disabilities and acculturate.

By examining how the general public uses voice interfaces and assessing potential advantages over conventional systems, our study attempts to close this research gap. In order to enhance the user experience, we will pay particular attention to analyzing the possibilities and restrictions of speech interfaces. In order to accomplish this, we will run an experiment in which several voice interfaces will be tested in order to determine their efficacy and use, more precisely we will test four interfaces, which will be more detailed later.

We aim to uncover potential obstacles and opportunities that emerge from this particular setting by gathering data, examining user responses, and observing interactions with speech interfaces. We will also examine the benefits and drawbacks of voice interfaces in comparison to other forms of interoperability.

The paper is organized as follows: Section 2 contains related work. Section 3 contains our proposal presenting the various modes. Section 4 specifies the evaluation, which deals with describing how the experiment was designed and carried out. Section 5 shows the results of the experiments by dividing them into general results, select-only form results, and text-only form results. This section also presents the results of the SUS and final feedback questionnaires completed by the participants. Finally, Section 6 presents the conclusions that could be drawn by analyzing the results obtained.

2 Related Work

In this section, we give a summary of the pertinent prior research, with a particular emphasis on voice interfaces for users who are not disabled and who have a high level of literacy. We go over the methodology, approaches, and major conclusions of these studies, emphasizing the contributions they made to the field of voice interfaces.

S. Usharani, P. Manju Bala, and R. Balamurugan's paper, "Voice-Based Form Filling System for Visually Challenged People"[12], which was presented at the IEEE ICSCAN 2020 conference, sought to develop a system to aid visually impaired and illiterate people in filling out forms. The authors created a smartphone app that allowed users to complete forms using voice input. The study demonstrated how the voice-based method effectively increases accessibility and usability for the intended user population.

A new artificial intelligence-based speech recognition system was suggested to assist with form filling in the article "Voice Assisted Form Filling for the Differently Abled"[9] by S.S. Ramasubramanian, S. Koodli, and Pranav S. Nair, which was presented at the IEEE Xplore conference in 2022. The study addressed issues including the lack of program tutorials and users' difficulties understanding when to talk.

The study emphasized how speech recognition technologies could help people with disabilities fill out forms more easily.

The purpose of the paper "SpeechForms: From Web to Speech and Back"[1] by Luciano Barbosa, Diamantino Caseiro, Giuseppe Di Fabbrizio, and Amanda Stent was to assess the effectiveness of two language models. The study showed a tool that lets people make bookings by speaking their answers into forms. This work provided the impetus for our investigation because it proved that voice-based form completion was feasible and could be integrated with web-based applications.

The usage of voice interfaces for information gathering by rural Indian farmers was covered in the paper "A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India"[8] by Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S. Parikh, which was presented at CHI 2009. The outcomes demonstrated the farmers' high level of satisfaction with the technology and highlighted the potential of voice interfaces for information retrieval activities.

The benefit of voice-controlled web browsers in comparison to conventional mouse-based surfing is examined in the paper "A comparison of voice-controlled and mouse-controlled web browsing"[3]. The study conducted a within-subjects experiment with 18 participants (12 men and 6 women) to examine if numbering links improves voice navigation and whether three common forms of hypertext (forms, linear slide shows, grid/tiled maps, and hierarchical menus) are suitable for voice navigation. According to the study, using voice control increases performance times for some jobs by about 50%. According to subjective satisfaction tests, text links are preferred over numbered links for voice surfing.

In addition, a number of current applications can be used as models for voice-based form input, these include well-known voice assistants that can be used to fill out forms. Applications and voice assistants are listed below:

- **Alexa**: users of the Amazon Alexa app can dictate notes and have them written down. For notes, there is a cap of 500 characters. It is crucial to keep in mind that Alexa's dictation could not always be correct, which could lead to transcription errors when taking notes;
- **Siri**: users of Apple's Siri can utilize the input to take notes, create emails, blog posts, documents, or conduct Google searches. You must issue a number of voice instructions in order to begin these processes. Saying "Hey Siri, take notes" will activate Siri; for example, you are able to utilize the voice to type the document. Additionally, you can use specific words that Siri understands as commands. For example, the commands **"New line"** go to the next line, **"New paragraph"** creates a new paragraph, and **"Cap"** capitalizes the following word, respectively;
- **Google Gboard**[7]: on mobile devices, Google Gboard can be installed as a keyboard. The "Talk to Write" option of Gboard lets you use your voice to enter text into any program that supports text input, including Gmail and Keep. Simply open a text entry-enabled app, tap a text entry field, tap and hold the microphone button at the top of the keyboard, and when "Talk Now" displays, speak the text you wish to enter;
- **Google Docs**[5]: the "Type with your voice" feature of the Google Docs application allows users to create and format text using voice commands. For example, say "Select [word]" or "bold" to make a word bold or "Increase font size" to make the text larger;
- **Speech2Forms**[11]: is a smartphone application that enables table creation through voice or keyboard. When the software launches, a screen appears where you may add a new table by clicking the "+" button, choosing the table name, and entering a name. After that, you may add table fields by selecting the "+" button and typing the field's name. After filling out the forms, you can go back to the home page, where the table will be presented automatically if there is just one, or you can choose which table to use. You can type data into the various table fields by clicking the "+" icon, or you can use the entries to fill them up. There is a button with a microphone icon that, when pressed, changes the color of the button to yellow and enables voice input. Clicking the button again returns the button's color to green and disables voice input.

3 Proposal

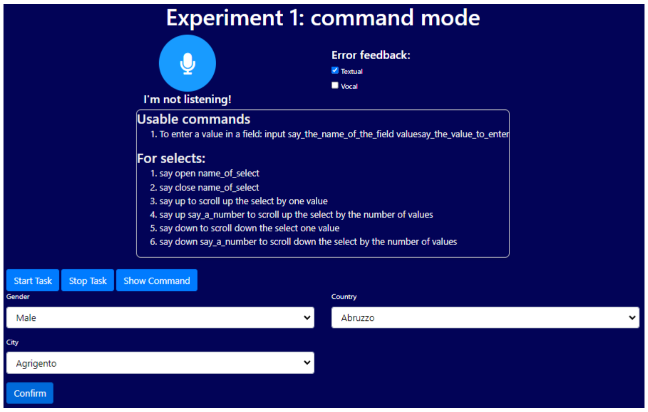
Our proposal was presented using four forms on web pages that the user can fill out using different modes. The Web Speech API[6] was utilized for speech recognition, while the Artyom[4] library was utilized for speech synthesis.

The four proposed modalities are as follows:

3.1 Command mode

Without using the mouse or keyboard, the user may complete forms that have both text fields and select fields by speaking the data. The user can communicate with the system using the following commands:

- Input "field_name" value "value_to_enter" to enter a value in a field;
- Open "Select_name" to open a select;
- Close "Select_name" to close a select;
- Up to scroll up the select by one value;
- Down to scroll down the select by one value;
- Up "number" to scroll up the select by the number of values;
- Down "number" to scroll down the select by the number of values


Experiment 1: command mode

Usable commands

1. To enter a value in a field: input say_the_name_of_the_field valuesay_the_value_to_enter

For selects:

1. say open name_of_select
2. say close name_of_select
3. say up to scroll up the select by one value
4. say up say_a_number to scroll up the select by the number of values
5. say down to scroll down the select one value
6. say down say_a_number to scroll down the select by the number of values

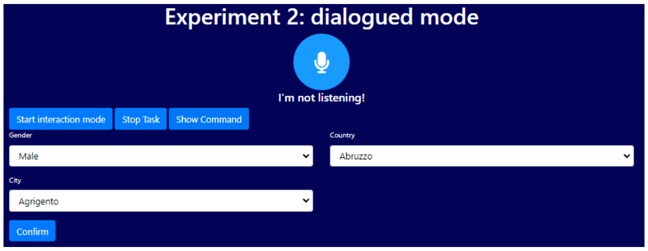
Gender: Male
Country: Abruzzo
City: Agrigento

Figure 1 Command mode interface

3.2 Dialogued mode

In this mode, the voice assistant interacts with the user, asking him to input a value for each form field. By clicking the "Start interaction mode" button, the user starts the mode. You may use the following commands to move in the select in this scenario:

- Up to scroll up the select by one value;
- Down to scroll down the select by one value;
- Up “number” to scroll up the select by the number of values;
- Down “number” to scroll down the select by the number of values


Experiment 2: dialogued mode

Gender: Male
Country: Abruzzo
City: Agrigento

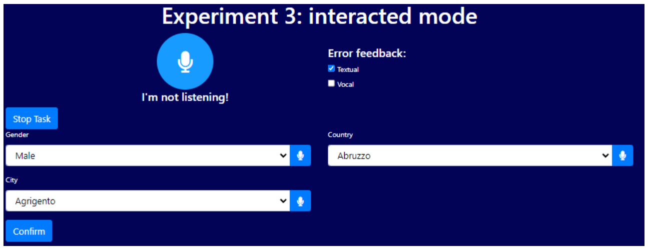
Figure 2 Dialogued mode interface

3.3 Interacted mode

In this mode, the user can click a button near the fields. In the case of the select field, the voice assistant asks the participant for each option if he wants to choose it; the option is chosen after the user has said yes. In this mode, the user does not need additional commands to interact.

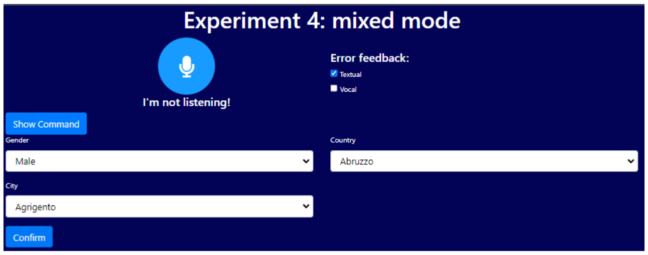
3.4 Mixed mode

In this mode, the user utilizes a mouse to navigate between the different fields, selects a field, and then uses speech recognition to enter data into the field. The commands that are helpful to the user in this mode are the same as those in dialog mode (subsection 3.2).


Experiment 3: interacted mode

Gender: Male
Country: Abruzzo
City: Agrigento

Figure 3 Interacted mode interface


Experiment 4: mixed mode

Gender: Male
Country: Abruzzo
City: Agrigento

Figure 4 Mixed mode interface

4 Evaluation

This section will describe the methodologies and apparatus used, the participants who supported the experiment, and the design of the experiment.

4.1 Participants

Sixteen volunteer participants who have had more or less experience with filling out online forms in the past participated in the experiment. Participants of both sexes (68.75% men and 31.25% women) ranged in age from 22 to 30 years and had a medium to high level of education. It’s important to note that 75% of participants had prior experience with voice assistants like Siri, Google Home, and Alexa. This information was gathered by examining at how users responded to the initial questionnaire’s inquiry about their prior experience with voice assistants.

4.2 Apparatus

Experiments were performed using an Acer Aspire 5 A515-52G-559E using the built-in microphone. The laptop has the following technical data:

- Intel core i5-8265u;
- NVIDIA GeForce MX130;
- 8GB DDR4 Memory;
- 256GB PCIe NVMe SSD.

4.3 Procedure

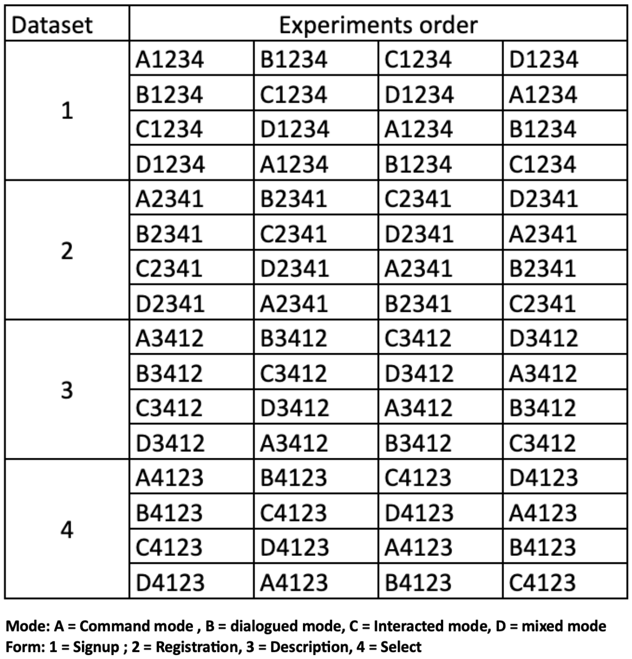
The procedure followed to carry out the experiment consists of 4 main steps:

Figure 5 Acer Aspire 5 A515-52G-559E

1. Preparation Phase: Providing participants with a cognitive questionnaire to fill out, aimed at gathering information about their experience with voice commands and filling out online forms.
2. Test Form: Before each mode, participants will be guided through a test form to fill out using that specific mode, which will consist of an exercise to familiarize them with the voice command system used. They will be encouraged to perform several voice commands to fill out the form.
3. Completion of Main Forms: Participants will have to fill out four main forms, each with all the data entry modes. If participants notice that they have made any mistakes during data entry, they can proceed to correct them. There is no break between modes except for the period when the participant can try the test form of the next mode. Participants will have to enter predetermined and defined values from four separate datasets. At the end of each mode, the user will have to complete the SUS questionnaire.
4. Final Questionnaires: Participants will be required to complete two final questionnaires:
   - Final Feedback: In this questionnaire, they will be asked to rate their overall experience in using voice commands to fill out forms and provide suggestions for improvements.
   - System Usability Scale (SUS)[2]: Participants, at the end of each mode, will complete the SUS questionnaire to measure the overall usability of the voice command system.

4.4 Design

The experiment was a 4 x 4 within-subject, and the experiments were balanced due to the following latin matrix (figure 6):



Dataset	Experiments order			
1	A1234	B1234	C1234	D1234
	B1234	C1234	D1234	A1234
	C1234	D1234	A1234	B1234
	D1234	A1234	B1234	C1234
2	A2341	B2341	C2341	D2341
	B2341	C2341	D2341	A2341
	C2341	D2341	A2341	B2341
	D2341	A2341	B2341	C2341
3	A3412	B3412	C3412	D3412
	B3412	C3412	D3412	A3412
	C3412	D3412	A3412	B3412
	D3412	A3412	B3412	C3412
4	A4123	B4123	C4123	D4123
	B4123	C4123	D4123	A4123
	C4123	D4123	A4123	B4123
	D4123	A4123	B4123	C4123

Mode: A = Command mode , B = dialogued mode, C = Interacted mode, D = mixed mode
Form: 1 = Signup ; 2 = Registration, 3 = Description, 4 = Select

Figure 6 Latin Matrix

By dataset we mean the data that was provided to the user to enter within the forms through the various modes described above.

5 Results and discussion

We made the decision to examine the task completion time from three angles in order to properly respond to the study questions:

- the total time spent looking through all the data gathered (Figure 7);
- the time spent filling out forms with just select fields (Figure 8);
- the time spent filling out forms with only text fields (Figure 9).

The mixed mode finished the overall work in the quickest amount of time (on average, 27.41 seconds), followed by the dialogue mode (on average, 35.65 seconds), and then the command mode (on average, 35.82 seconds). The interactive mode takes an average of 88.45 seconds, which is finally and clearly discernible. This is mostly because more time is needed for the compilation of the choose fields.The above-mentioned differences were found to be statistically significant ($F_{3,45} = 193.15$, $p < .001$).

When we analyzed the data pertaining to the times of the forms with only a select fields, we found that the mixed mode was still the fastest with an average time of 15.18 seconds, followed by the dialogue mode with a time average of 20.79 seconds, the command mode very closely behind with a time average of 21.6 seconds, and the interactive mode

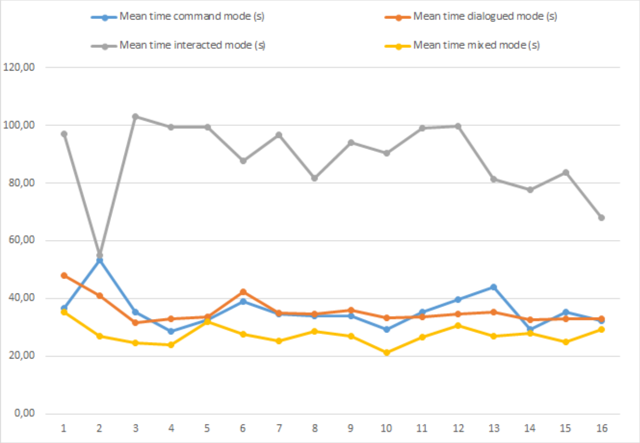


Figure 7 Mean time for every partecipant

with a time average of 142.68 seconds. This analysis makes us understand even better how the select are temporally expensive in the dialogue mode. The difference was also statistically significant ($F_{3,45}$ = 129,156, p < .0001).

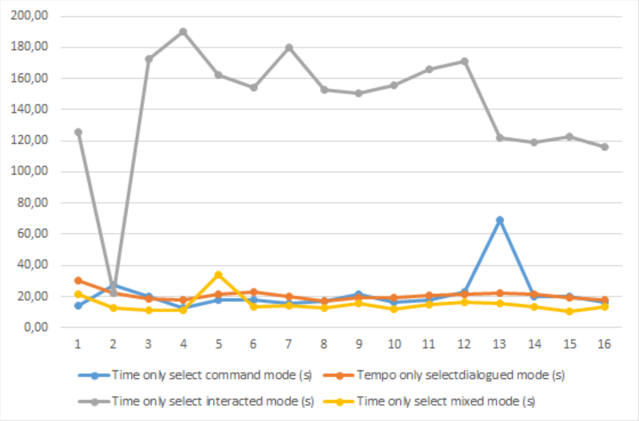


Figure 8 Time for every partecipant only selec

According to data gathered from an observation form with only text fields, the mixed mode is the fastest with an average time of 26.33 seconds, followed by the interactive mode with an average time of 26.74 seconds, the command mode with an average time of 31.57 seconds, and the dialogue mode with an average time of 33.95 seconds. The difference was found to be statistically significant ($F_{3,45}$ = 6,761, p < .001).

5.1 Evaluation of Usability through System Usability Scale (SUS)

The four suggested modes' usability was evaluated using SUS. After interacting with each of the four modalities, participants were encouraged to complete the SUS questionnaire. Ten statements made up the questionnaire, to which participants responded with a score ranging from 1 (strongly disagreed) to 5 (strongly agreed). The mixed mode has a score of 93.91, followed by the dialogue mode with 85.47, the interactive mode with 71.72, and the command mode with 66.1, according to our analysis of the information gathered from the questionnaires created by the participants.

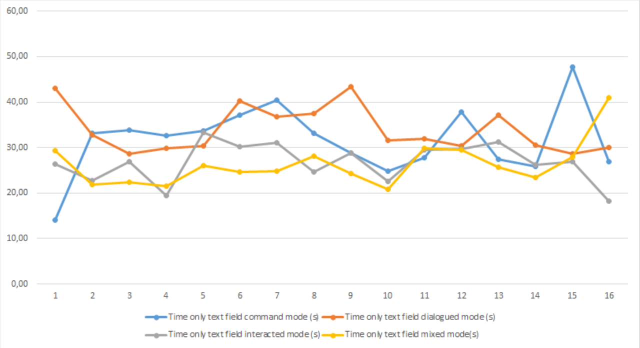


Figure 9 Time for every partecipant only text field

5.2 User Feedback and Response Analysis

Users were invited to score the modes according to their preferences in the final feedback form, with the most liked option at the top and the least favored at the bottom (Table 1). The data was then analyzed by tallying the number of times each mode received a vote in each slot. Finally, we used a weighted method to create a ranking by multiplying the number of times a mode received the first vote by 3, the number of times it received the second vote by 2, the number of times it received the third vote by 1, and the number of times it received the fourth vote by 0, then adding the results together. The results of these calculations led to the following final categorization of the modes:

- Mixed mode: 39 points
- Dialogue mode: 33 points
- Command mode: 17 points
- Interactive mode: 7 points

The mixed mode is preferred by users, getting the highest score. It follows the dialogue mode in second place, followed by the command mode in third place, and finally the interactive mode in fourth place.

Mode	1st	2nd	3th	4th	Total
Command	2	1	9	4	17
Dialogued	5	7	4	0	33
Interacted	0	2	3	11	7
Mixed	9	6	0	1	39

Table 1 Number of preferences for each mode

Users got the chance to give their thoughts on the many modes our system offers during the final feedback questionnaire. To better understand their preferences and the factors that influence their decisions, we need this qualitative input. Following are some user opinions:

- One participant said, "The voice system seems well organized, easy to understand, and practical to use; there is no need for much training to use the various modes. Perhaps I would narrow down the fields of consecutive selection (e.g., regions with provinces), otherwise, it seems like a very good program";
- Another user shared the following comment: "Delete interactive version selections.";
- A third participant said, "I would often use this system, especially in dialogue and mixed mode; in fact, they seem to me the most intuitive and easy to use, even by older people.";
- Finally, one user said, "Command mode is not very intuitive".

Analyzing the opinions of users reveals several significant pieces of information, such as the inconvenience of using interactive mode in very long selects. Or how the command mode for some users has few intuitive commands. Finally, some comments mention the ease of use of the system, especially in dialogue and mixed modes.

6 Conclusion

There are four suggested methods for inputting text into the fields using the entry. Sixteen participants in an experiment completed four forms across all formats. Data were gathered on participant preferences indicated through the surveys as well as the speed at which each manner of compilation was completed. At the conclusion of this investigation, we can state that the mixed mode was the fastest and most fascinating option, while the interactive mode was the least interesting and slowest, likely due to the waiting time in the compilation of select fields.

References

[1] Luciano Barbosa, Diamantino Caseiro, Giuseppe Di Fabbrizio, and Amanda Stent. 2011. SpeechForms: From web to speech and back. *Twelfth Annual Conference of the International Speech Communication Association* (2011).

[2] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[3] Kevin Christian, Bill Kules, Ben Shneiderman, and Adel Youssef. 2000. A Comparison of Voice Controlled and Mouse Controlled Web Browsing. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies* (Arlington, Virginia, USA) *(Assets '00)*. Association for Computing Machinery, New York, NY, USA, 72–79. https://doi.org/10.1145/354324.354345

[4] Carlos Delgado. [n. d.]. Artyom. https://sdkcarlos.github.io/sites/artyom.html

[5] Google docs editor. [n. d.]. Type with your voice. https://support.google.com/docs/answer/4492226?hl=en#zippy=%2Cselect-text%2Cformat-your-document

[6] Google. [n. d.]. Web Speech API. https://developer.chrome.com/blog/voice-driven-web-apps-introduction-to-the-web-speech-api/

[7] Google LLC. 2013. Gboard: la tastiera Google. Google play store. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=it&gl=US Versione 12.9.20.521739039-release-arm64-v8a.

[8] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S. Parikh. 2009. A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India. (2009), 51–54. https://doi.org/10.1145/1518701.1518709

[9] Shreya Sri Ramasubramanian, Sunit Koodli, Pranav S Nair, Mahah Sadique, and Hr Mamatha. 2022. Voice Assisted Form Filling for the Differently Abled. *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)* (2022), 40–44.

[10] Jahanzeb Sherwani, Sooraj Palijo, Sarwat Mirza, Tanveer Ahmed, Nosheen Ali, and Roni Rosenfeld. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. *2009 International Conference on Information and Communication Technologies and Development (ICTD)* (2009), 447–457. https://doi.org/10.1109/ICTD.2009.5426682

[11] Simple Seo Solutions. 2020. Speech2Forms - voice task list. Google play store. https://play.google.com/store/apps/details?id=com.speech2forms.android&hl=it&gl=US Versione 1.21.

[12] S. Usharani, P. Manju Bala, and R. Balamurugan. 2020. Voice Based Form Filling System For Visually Challenged People. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (2020), 1–5. https://doi.org/10.1109/ICSCAN49426.2020.9262431