electronics

MDPI

Article

The Synergy between a Humanoid Robot and Whisper: Bridging a Gap in Education

Akshara Pande and Deepti Mishra *

Educational Technology Laboratory, Department of Computer Science (IDI), Norwegian University of Science and Technology, 2815 Gjøvik, Norway; akshara.pande@ntnu.no
* Correspondence: deepti.mishra@ntnu.no

Abstract: Students may encounter problems concentrating during a lecture due to various reasons, which can be related to the educator's accent or the student's auditory difficulties. This may lead to reduced participation and poor performance in the class. In this paper, we explored whether the incorporation of the humanoid robot Pepper can help in improving the learning experience. Pepper can capture the audio of a person; however, there is no guarantee of accuracy of the recorded audio due to various factors. Therefore, we investigated the limitations of Pepper's speech recognition system with the aim of observing the effect of distance, age, gender, and the complexity of statements. We conducted an experiment with eight persons including five females and three males who spoke provided statements at different distances. These statements were classified using different statistical scores. Pepper does not have the functionality to transcribe speeches into text. To overcome this problem, we integrated Pepper with a speech-to-text recognition tool, Whisper, which transcribes speech into text that can be displayed on Pepper's screen using its service. The purpose of the study is to develop a system where the humanoid robot Pepper and the speech-to-text recognition tool Whisper act in synergy to bridge the gap between verbal and visual communication in education. This system could be beneficial for students as they will better understand the content through the visual representation of the teacher's spoken words regardless of any hearing impairments and accent problems. The methodology involves recording the participant's speech, followed by its transcription to text by Whisper, and then evaluation of the generated text using various statistical scores. We anticipate that the proposed system will be able to increase the student's learning experience, engagement, and immersion in a classroom environment.

Keywords: Pepper; speech-to-text; whisper; social robot

Citation: Pande, A.; Mishra, D. The Synergy between a Humanoid Robot and Whisper: Bridging a Gap in Education. *Electronics* **2023**, *12*, 3995. https://doi.org/10.3390/electronics12193995

Academic Editors: Rania Hodhod and Mohammad Jafari

Received: 18 August 2023
Revised: 19 September 2023
Accepted: 20 September 2023
Published: 22 September 2023

1. Introduction

Social robots have the potential to bring a progressive change in our way of living and can have a tremendous impact on numerous sectors, including education. They can intelligently interact with humans [1]. They can offer interesting and engaging learning experiences to individuals [2,3]. By incorporating social robots into classroom settings, we can enhance educators' capacity to deliver lectures more efficiently [4,5]. Social robots can play various roles in the education field, such as teaching assistant, playmate, and entertainer. Alemi et al. [6] observed that students' comprehension improved significantly when learning with the assistance of robots. Furthermore, Kennedy et al. [7] found that children who interacted with robots displayed better nonverbal behavior and showed significant improvements in their learning. Mishra et al. [8] found that students' enjoyment of learning programming increased when it involved robots rather than computer-based coding. Pandey et al. [9] also showcased how social robots connect with primary-level students in quiz-based interactions, emphasizing their interactive capability in educational settings. Previous studies suggest that social robots can significantly benefit and improve learning outcomes and achievements [9–11]. However, robot-assisted learning is not

effective for all subjects, as found by Rani et al. [12], who showed that the utilization of social robots is more effective for programming than for mathematics. With the progression of robotics technology, there is scope for social robots to provide a wide range of benefits, making them applicable to various fields and subjects.

Students may often encounter difficulties in understanding the accent of a teacher, particularly if the teacher is not a native speaker of the language [13]. A potential solution could be the use of assistive technologies such as the use of speech recognition systems and automated subtitling of spoken words. In this study, we used the humanoid robot Pepper, as it has a speech recognition system and a screen where the text can be displayed corresponding to spoken words. However, there is a need to find the effectiveness of a robot's speech recognition system. In order to evaluate the accuracy of the speech recognition system, experiments were performed. The objective of the experiment was to examine the impacts of various factors such as age, gender, distance, and complexity of statements on the speech recognition system of a robot. Further, a better understanding of spoken words can be received through visual aids such as screens and tablets. With this idea, we propose a system that converts the person's speech into text and displays that on the humanoid robot Pepper's screen. Figure 1 presents the overview of the proposed system.

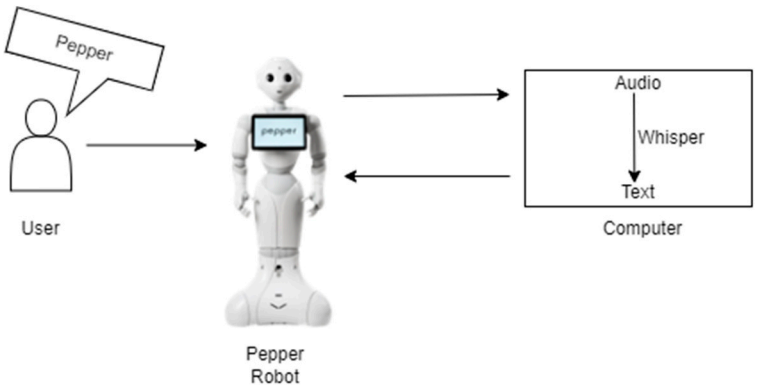

Figure 1. Display of text on Pepper's tablet using audio input by a user.

The Pepper robot [14] was developed by SoftBank Robotics, and it is a humanoid robot with the potential to interact and perceive its environment through sensors, cameras, and microphones. Pepper has four microphones on its head, through which it can capture sound from different directions. Additionally, it has a speech recognition system through which it can establish communication with humans and record their speech statements. However, this speech recognition system does not have the facility to transcribe the recorded audio to text. This limitation inhibits the robot from providing written spoken language representations on its screen. Along with this incapability, there are many challenges associated with this system as well. Speech recognition is the most significant challenge, as Pepper struggles to understand different accents. Furthermore, due to its limited vocabulary, it is difficult for Pepper to comprehend complex words. Meanwhile, background noise poses another problem for Pepper to accurately perceive spoken utterances. Consequently, these issues collectively highlight the insufficiency of Pepper as an effective communicator in educational settings. To overcome the inherent limitations of Pepper's speech recognition system, we propose a novel solution of establishing synergy between Pepper and an external speech-to-text recognition tool so that the text subtitling of spoken statements can be displayed on Pepper's screen. Our research aims to increase the ability of the Pepper robot by seizing its existing restrictions in speech processing, eventually helping both teachers and students.

Speech-to-text recognition is a crucial component of human–robot interaction, as it enables people to interact with robots naturally and conveniently. Additionally, it promotes greater accessibility for individuals who have hearing impairments [15]. It is widely used in various domains, including hospitals and education. Shadiev et al. [16] reviewed speech-

to-text technology in education, including its capabilities. Speech-to-text recognition can help to enhance learning [16]. Debnath et al. [17] showed that audio-visual information in combination with automatic speech recognition can effectively provide education to people with disabilities. Gross et al. [18] conducted a survey, and their findings showed speech recognition technology is generally perceived as useful for clinical documentation, though they recognized some challenges associated with its use.

Currently, many open-source speech-to-text conversion tools are available. Pande et al. [19] compared various speech recognition tools, such as Google speech recognition, Vosk, CMUSphinx, DeepSpeech, and Whisper speech recognition tools on the basis of evaluation metrics and found that Whisper has the best accuracy among them. OpenAI developed Whisper [20], is an open-source speech recognition tool. It recognizes speech in multiple languages and performs various tasks efficiently. Whisper is an innovative development and is currently being utilized in various research studies to leverage its capabilities. The findings of Correa et al. [21] suggested that using synthetic data to adapt Whisper-based Automatic Speech Recognition (ASRs) for specific domains leads to performance improvements. Machacek et al. utilized Whisper for real-time speech transcription [22]. Spiller et al. [23] suggested that the use of Whisper for audio transcription in mental health research can greatly simplify the transcription process and streamline data analysis. Considering all of these research findings, we decided to integrate the Whisper speech-to-text recognition tool with Pepper's speech recognition system for the present study.

In the past, few studies incorporated speech recognition systems with robots. Fujii et al. [24] demonstrated the dialogue established between humans and robots about food and recipes by integrating open-source speech recognition systems along with different components of a robot system. Deuerlein et al. [25] integrated a cloud-based speech recognition system along with human–robot interaction for different purposes. Grasse et al. [26] showed that the use of robots, which could be controlled by speech, could safely deliver items in healthcare settings. However, it has been found that none of these studies explored the integration of speech-to-text conversion with humanoid robots in educational settings. Although existing studies have highlighted the potential benefits of using social robots to increase student engagement and enjoyment in educational settings, they often focus on programming and language learning. To the best of our knowledge, the utilization of social robots' speech recognition systems integrated with speech-to-text recognition tools for better understanding and conveying the teacher's speech statements is an unexplored area.

The goal of this study is to assess the speech recognition capability of Pepper based on four factors: age, gender, distance, and sentence complexity. The other objective of the study is to implement a system that combines a humanoid robot, 'Pepper', with the speech-to-text recognition OpenAI tool 'Whisper', which can be applied to classroom settings. We aim at creating an environment that will be more engaging, immersive, and interactive for students. Furthermore, this system can help students to increase their understanding of the material being taught, as well as foster their critical thinking and communication skills. Additionally, this system can provide support in accessibility for students who may have difficulty with auditory processing. The proposed system will have the potential to help people with special needs [27,28]. The structure of the paper is as follows. The methods utilized in this study are outlined in Section 2. The results and related discussions are presented in Section 3. Finally, the conclusions and future work are summarized in Section 4.

2. Materials and Methods

In the present paper, an experiment was performed to explore the functionality of the speech recognition system of Pepper. The experiments were conducted in the Educational Technology Laboratory of NTNU Gjøvik, Norway. The laptop utilized in this experiment had an 11th Gen Intel(R) Core (TM) i5-1145G7 @ 2.60GHz 1.50 GHz processor,

16 GB RAM, and was running the Windows 10 operating system. Two services, namely ALAudioRecorder [29] and ALTabletService [30], were utilized by Pepper. To accomplish this task, we used Python version 2.7.16 and Python version 3.9.13. The overall working of proposed system is shown in Figure 1.

2.1. Experimental Setup

Eight persons were recruited for this experiment including five females and three males. Their age was in the range of 15 to 55. The sample included participants with diverse educational backgrounds, including bachelor's, master's, and Ph.D. degrees. None of the participants speak English as their first language. Six randomly selected statements were provided to participants to say, and they were recorded at distances 1 m, 3 m, and 5 m from Pepper. These statements were classified into three categories based on their complexity; they were simple, moderate, and complex sentences. The experiment duration was around 30 min for each participant.

2.2. Complexity of Statements

The nltk library [31] was used to import word_tokenize and cmudict packages. Tokenizers help in dividing the sentence into smaller substrings and form a list of these substrings. The CMU pronouncing dictionary was used for syllable counting based on pronunciation. A textstat [32] library was also used to get the statistical information from the sentences. Different scores were obtained using textstat such as flesch_reading_ease, flesch_kincaid_grade, gunning_fog, coleman_liau_index, smog_index, automated_readability_index, lix, and dale_chall_readability_score. The overall complexity of statements was obtained by adding all of these values. Further, these statements were categorized into simple, moderate, and complex sentences by imposing a threshold on overall complexity. The definitions of these scores are as follows:

1. Flesh reading ease: This refers to easiness of reading the text [33];
2. Flesh Kincaid grade: This is mostly used in the education field. It indicates the necessary level of education to comprehend a text [34];
3. Gunning fog: This is a test utilized for readability in English writing [35];
4. SMOG: This is a readability index that determines the required year of education to comprehend the text [36];
5. Automated Readability Index: This is used to determine whether or not a text is comprehended [34];
6. LIX: This denotes the challenges related to text reading [37]:
7. Dale–Chall Readability: This is related to understanding problems that the user faces during text reading [38].

2.3. Pepper Robot Function

This paper utilized the Pepper robot for two functions (Figure 2). Firstly, it captures the human voice and stores it within its system. This was made possible using the ALAudioRecorder service, which allowed the robot to record audio using its microphones. The generated audio format is '.wav'. Additionally, the ALTabletService enabled the Pepper robot to display text on its screen. To create instances of these two services, Python 2.7.16 was employed. The instance of the ALAudioRecorder module was created by using the ALProxy ("ALAudioRecorder", pIp, PORT) command, where "pIp" represents the IP address of Pepper and "PORT" represents the port number. Although there are four channels available for recording on Pepper, we only used the front channel. The "startMicrophoneRecording" and "stopMicrophoneRecording" methods were used to start and stop recording, respectively. The audio files were saved with the ".wav" extension. The ALTabletService was used to create tablet applications that could be displayed on Pepper's screen. The method "showWebview" was used to display the text on the tablet, and we utilized a text file generated by an audio file as the input to this method.

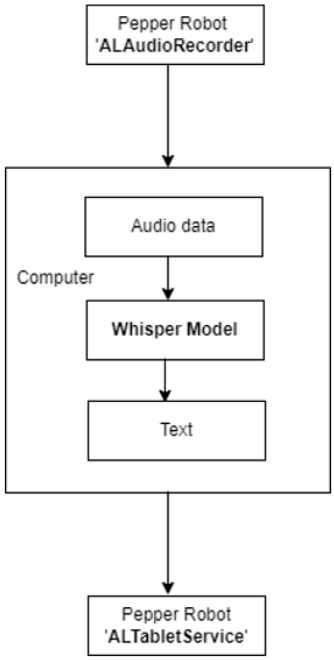

Figure 2. Integration of Pepper's services with speech-to-text recognition tool.

2.4. Paramiko Function

Paramiko [39] is a Python library that enables secure communication with remote servers using the SSH protocol. Paramiko can be installed in a computer system using the pip command. Paramiko allows establishing an encrypted connection to a remote server by authenticating either with a private key or password and facilitates secure file transfer between the computer and Pepper robot. A connection set up with Pepper required the creation of a transport object using 'paramiko.Transport(pIp, PORT)'. The 'pIp' parameter demonstrates the IP address of the Pepper robot, while the 'PORT' parameter indicates the port. After the creation of the transport object, authentication is required with a username and password. For setting up the SSH connection, there is a need for an SFTP client object from the transport object ('t') which can be made using the 'paramiko.SFTPClient.from_transport(t)' method.

2.5. Whisper Function

Whisper [40] is an open-source speech recognition model with a wide scope of applications. A comprehensive collection of audio data is used to train it. Whisper can be used to perform multiple tasks, such as speech recognition in multiple languages, speech recognition, and translation. The audio recordings from Pepper can be transferred to the computer system, which can be then used as input for Whisper. With the help of Whisper, the corresponding text file of recorded audio can be generated. Further, Pepper can be programmed to display this generated text on its screen.

We used Python 3.9.13 to convert audio into text. The Whisper library can be installed in and imported into Python code. Next, a pre-trained model called 'base' was loaded. The pre-trained model was already trained on a large quantity of audio data and can convert speech to text efficiently. This model was applied to convert the recorded audio data (with the '.wav' extension) into a text format using the transcribe () method.

3. Results

We were able to establish the connection between the computer system and Pepper. However, we encountered problems while retrieving audio files from Pepper directly. It was guessed that the challenges could be due to errors in Pepper's file system or other technical issues. In order to ensure that Pepper was able to record a person's speech, all audio files were transferred using the Paramiko library to the computer system. All audio files were successfully downloaded and saved in a designated path using this approach.

3.1. Integration of Pepper with Speech-to-Text Recognition Tool Whisper

The original statements that were given to participants were saved in the computer as reference files in text format. The recorded audio files were provided as input to Whisper to transcribe them in text format. A new text file was generated for each audio file by Whisper. There is the possibility of the misspelling of some words by the user, which can result in differences between the Whisper-generated text file and the reference file.

The audio-generated text was displayed on Pepper's tablet. For this purpose, audio-generated text from the computer's notepad was taken as input to Pepper's ALTabletService. The display for the corresponding text of the recorded audio file is shown in Figure 3. Pepper displays the text corresponding to the recorded audio. If the speaker misspells a word, the incorrect spelling of that word is displayed on Pepper's screen. A student can read the sentence on screen and either infer the meaning by himself or can request that a teacher repeat the sentence. This helps to make sure that any important material is not skipped during the lecture.

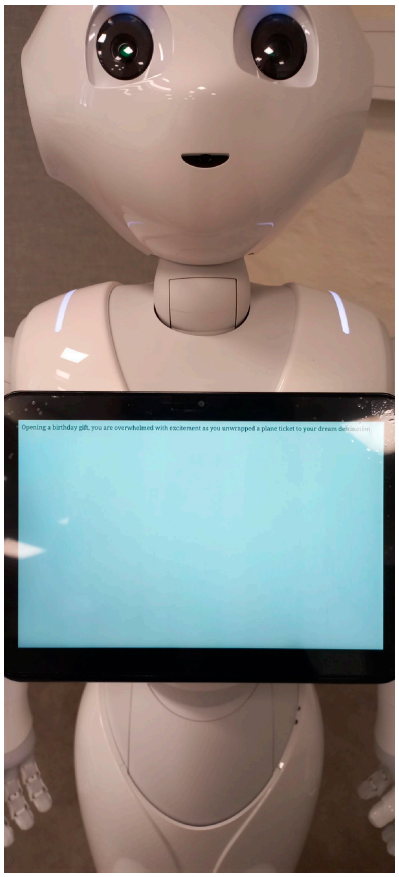

Figure 3. Display of generated text from recorded audio on Pepper's screen.

3.2. Exploring the Effectiveness of the Speech Recognition System based on Various Factors

For the first experiment, a total of 144 audio recordings were made, which include six sentences each stated by eight persons at three different distances from Pepper. Six statements were categorized into three categories based on their complexity. These categories were named simple, moderate, and complex sentences and denoted by 1, 2, and 3, respectively. To evaluate the efficacy of Pepper's speech recognition system, we converted the speech to text by using Whisper. We used the original sentence as a reference text and compared them with Whisper-generated text using evaluation measures such as the Word Error Rate (WER) [41], Match Error Rate (MER) [42], Word Information Lost (WIL) [43], and Character Error Rate (CER) [44].

Figure 4 illustrates the characteristics used in this study, such as the demographics of the participants, statement numbers, distance from Pepper, and different evaluation measures present in the dataframe. This dataframe contains 144 records, out of which 8 records are displayed in Figure 4. Since some recordings contained no speech, we received no values for WER, MER, WIL, and CER and assumed them to be missing values. Missing values were imputed using the mean value for that particular column. Furthermore, the analysis of evaluation measures was done with the help of box plots. Figure 5 suggests that there were two outliers for WER and CER, which corresponds to one of the statements. That statement was dropped for further analysis.

Participant	Age	Gender	Statements	Distance	WER	MER	WIL	CER
Participant 1	39	Female	1	1	0.07	0.07	0.14	0.02
Participant 2	16	Female	1	1	0.00	0.00	0.00	0.00
Participant 3	53	Male	1	1	0.14	0.14	0.21	0.16
Participant 4	30	Male	1	1	0.21	0.19	0.25	0.14
Participant 5	51	Female	1	1	0.07	0.07	0.14	0.01
Participant 6	28	Male	1	1	1.00	1.00	1.00	0.86
Participant 7	35	Female	1	1	0.14	0.14	0.21	0.07
Participant 8	29	Female	1	1	0.07	0.07	0.14	0.02

Figure 4. Top 8 rows (out of 144 rows) containing the details of participants, statements, distance, and evaluation measures.

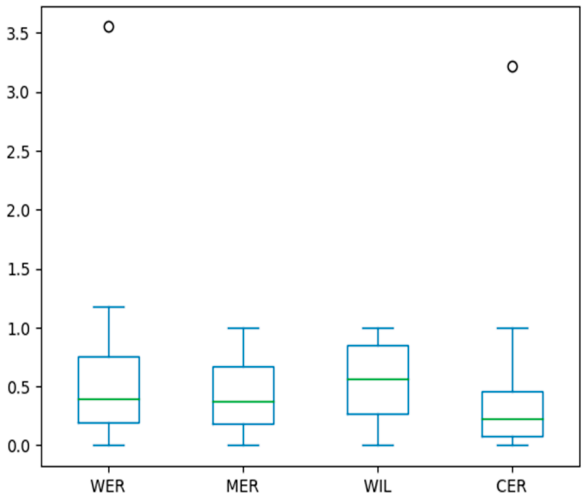

Figure 5. Boxplot for evaluation measures with outliers shown in circle.

Figure 6 contains the overall visualization pairplot, which is used to analyze the relationships and distributions among the key variables: age, distance, and complexity of statements. The data points are distinguished based on color provided by the WER's varying values. Each graph in the matrix shows information about the interactions between these variables. By using the pairplot, the possible patterns and correlations within the dataset can be identified, which can help in gaining insights into the data and making informed decisions. The scatter plots present in Figure 6 will be examined in detail later on in the present section.

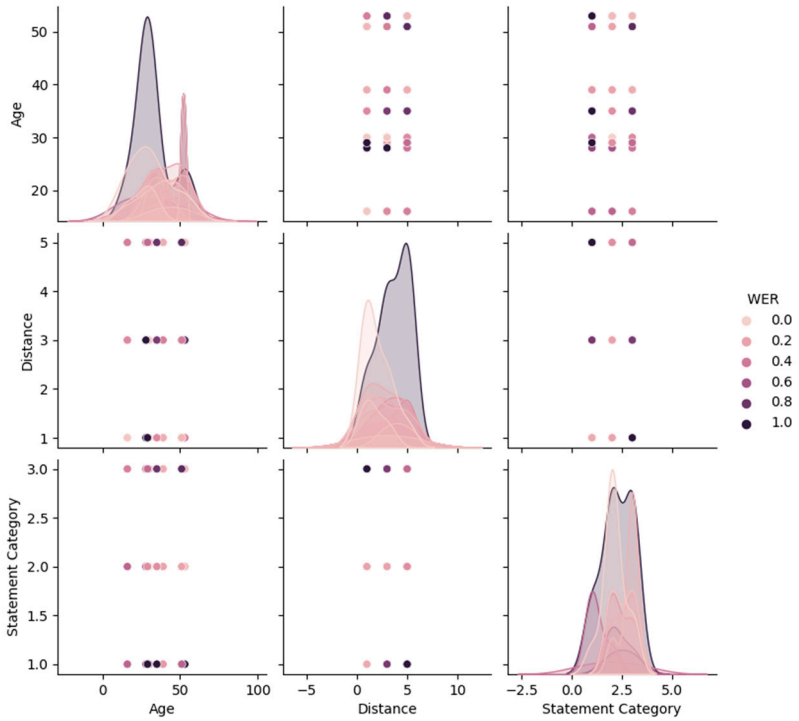

Figure 6. Pairplot of variables age, distance, and complexity of statements with WER as hue.

The boxplot in Figure 7 shows how the variables age, distance, WER, and complexity of statements are distributed. Each variable's interquartile range (IQR) is represented by a box with a median line inside the box. This visualization provides valuable insights into each variable's spread and central tendency, helping to understand how they are distributed.

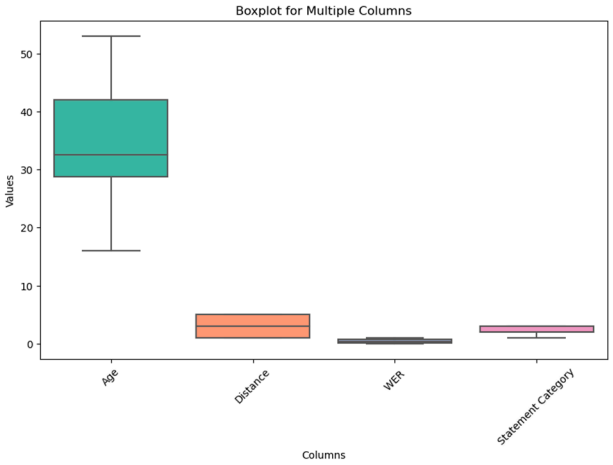

Figure 7. Boxplot of variables age, distance, WER, and complexity of statements.

In Figure 8, the correlations among the variables age, distance, WER, and the complexity of statements are depicted. The correlation between age and distance reveals an extremely weak and practically negligible relationship. Similarly, the correlation between age and WER demonstrates a very weak negative relationship. Age's correlation with the complexity of statements also indicates an extremely weak and almost negligible relationship. On the other hand, the correlation between distance and the WER of statements shows a positive but weak relationship. In contrast, the correlation between distance and the complexity of statements is extremely weak and practically negligible. Lastly, the correlation between the complexity of statements and WER exhibits a very weak negative relationship. Thus, the correlation matrix does not give any insight into how one variable affects another. To further explore, it is important to focus on variables depending on varying ranges of WER.

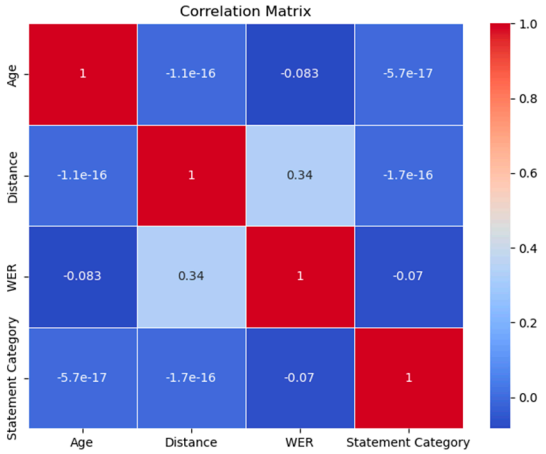

Figure 8. Correlation matrix of variables age, distance, WER, and statement category.

In Figure 9, age, distance, and complexity of statements were focused on. The scale of the WER showed the range of WER values. A lower WER value indicates higher speech recognition accuracy [45]. The blue bubbles in Figure 9 indicated the WER values less than 0.5. The darker color of blue represented high accuracy, while the red bubbles denoted the WER values greater than 0.5. The dark red color was indicative of less accuracy. Some WER values overlapped, making it difficult to interpret any pattern. Thus, further exploration is required to visualize this more efficiently. Hence, 2-dimensional values of these characteristics were considered.

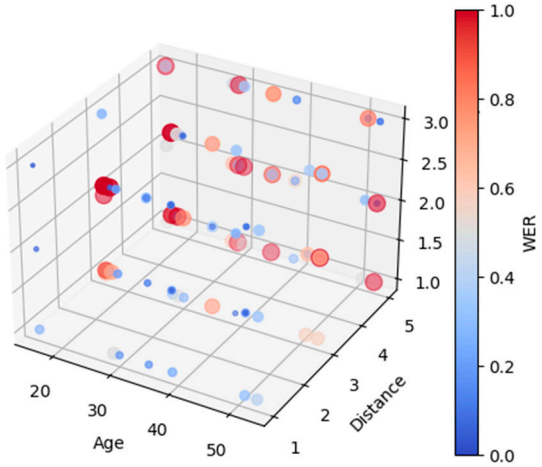

Figure 9. Demonstration of WER values for age, distance, and complexity of statements.

The relationships between age and distance were observed for simple, moderate, and complex statements with varying ranges of WER values. It is evident from Figure 10a that when simple statements were considered, irrespective of the age of the persons, the values of WER were high when the person stood at a maximum distance (5 m) from Pepper. These WER values suggest that accuracies were highest when the person was standing near Pepper, i.e., a 1 m distance away.

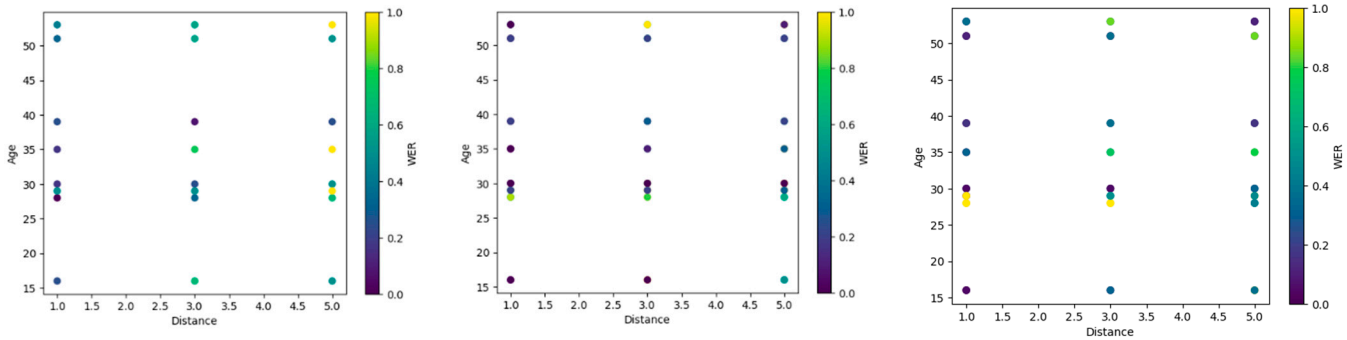

Figure 10. Illustration of impact of age and distance from Pepper on WER calculation for (**a**) simple statements, (**b**) moderate statements, (**c**) complex statements.

In the case of moderate statements, Figure 10b shows that the value of the WER was high when a person was placed at a 3 m distance away from Pepper. But again, the lowest values of WER could be found at a 1 m distance from Pepper. Figure 10c illustrates that for complex speeches, the WER values were high for even a 1 m distance. From this, it could be inferred that Pepper faces difficulty in understanding complex statements. With the increased distance from Pepper, the accuracy of the speech recognition system decreases.

Next, statements spoken at different distances were considered to further explore the accuracy of Pepper's speech recognition system based on different age groups. Four age groups were considered here, ages below 20, ages from 21 to 30, ages from 31 to 40, and ages above 50. Figure 11 demonstrates the WER values based on the distance from Pepper and the spoken statements' complexity. It should be noted here that many WER values overlapped, due to which it was difficult to interpret anything. As we were interested in knowing the limitations of Pepper's speech recognition system, only the maximum WER values in different age groups were taken into account (Figure 12).

Figure 12a illustrates that with a below-age-20 person, Pepper's speech recognition system did not have a problem recognizing statements of various complexity spoken by a person standing at a 1 m distance. However, higher WER values were reported for 3 m and 5 m distances with moderate and complex statements, respectively. This analysis suggests that the best selection of distance from Pepper is 1 m for the age group below 20.

Figure 12b shows that Pepper's speech recognition system got mostly high WER values for the age group of 21 to 30. At a distance of 1 m, the maximum WER value was for complex statements. For the 3 m distance, there were two maximum WER values for moderate and complex statements. Meanwhile, for a 5 m distance, irrespective of sentence complexity, maximal WER values were obtained every time. This suggests that Pepper faces difficulty in understanding statements in this age group.

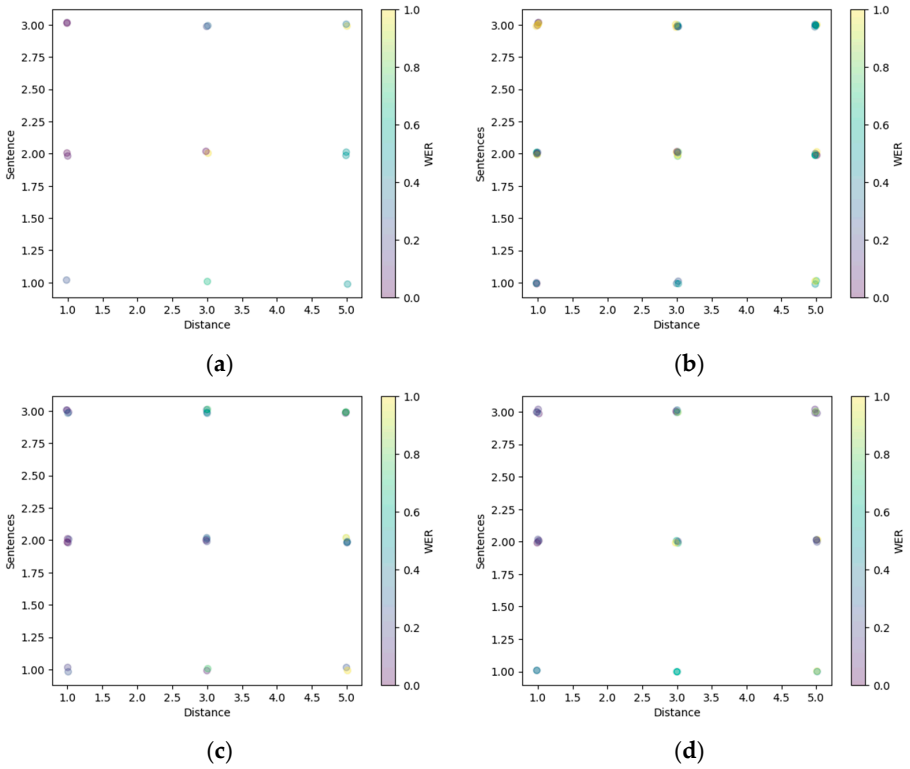

Figure 11. Overlapped WER values for different age groups, (**a**) ages below 20, (**b**) ages from 21 to 30, (**c**) ages from 31 to 40, (**d**) above age 50.

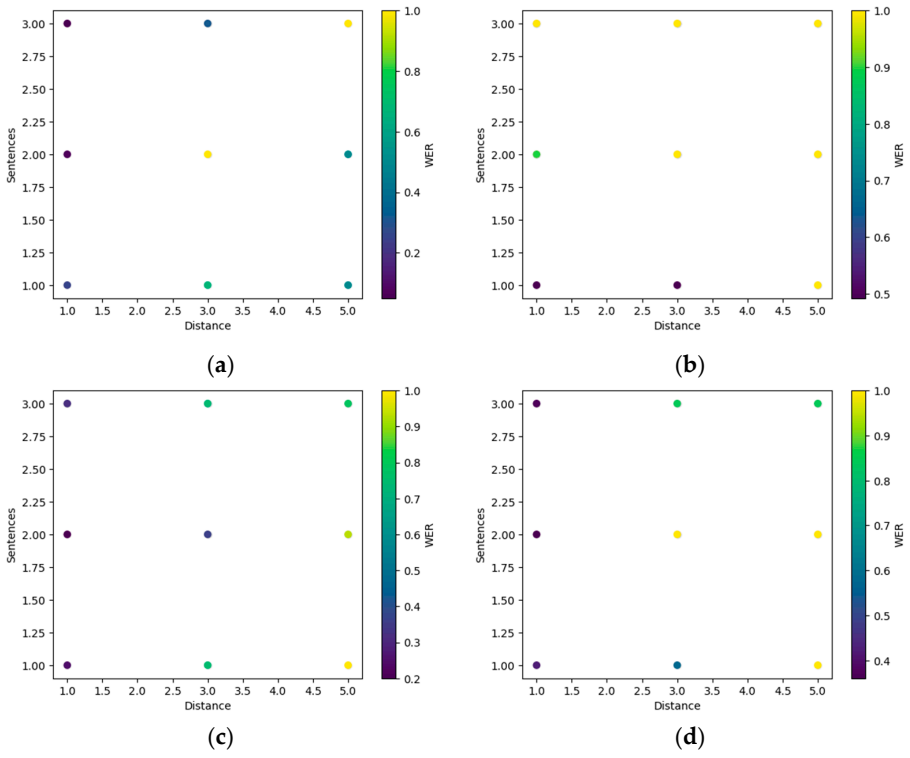

Figure 12. Maximum WER values for various statements and distances for different age groups, (**a**) ages below 20, (**b**) ages from 21 to 30, (**c**) ages from 31 to 40, (**d**) above age 50.

Figure 12c explains that Pepper's speech recognition works properly with a 1 m distance for the age group ranging from 31 to 40. However, the WER was highest for a 5 m distance even for a simple statement. For other distances and complexity of statements,

WER values were higher, but not maximum. This analysis indicates that 1 m is the best distance at which to place a person from Pepper.

Figure 12d describes that for the age group above 50, a 1 m distance is the best selection irrespective of the complexity of statements. The WER values were maximal at a 3 m distance in the case of moderate statements, while at a 5 m distance, WER values were found to be high for both simple and moderate statements.

4. Discussions

Research has shown that speech recognition systems can be significantly impacted by the age and complexity of statements spoken. Often, there may be some variations in children's pronunciation when they state a word. Li et al. [46] demonstrated that children's voices are more difficult for automatic speech recognition systems to process accurately compared to adults. Kennedy et al. [47] highlighted the importance of improving speech recognition for children, as it poses challenges for robots to comprehend. Speech features commonly observed in older people, including speech rate, pauses, redundancies, and diminished articulation and pronunciation, have been recognized by previous studies for increasing the Word Error Rate in automated speech recognition systems [48–51]. Differences in speech patterns and articulation between younger and older individuals can pose a challenge for automatic recognition systems. Due to above-mentioned challenges, the present study included participants from other age groups and identified that the WER is not affected by the age of the participant in the included age groups. This finding is also supported by Aman et al. [52], who explored elderly persons' voice effects on ASR and showed a significant 34% increase in the average Word Error Rate between elderly and non-elderly individuals; however, they also showed that WER is not associated with age.

Research on speech recognition has explored gender-based differences in speech, including variations in vocal tract length [53] and pitch [54] between males and females that can impact recognition systems. Decker et al. [55] demonstrated that average speech recognition was better for females than males for English and French speech. Garnerin et al. showed [56] that there was a large difference in the performance of ASR when compared on the basis of the WER of men and women. In contrast to these studies, the findings of the present study did not identify any differences between male and female utterances in terms of their impact on the speech recognition system.

The distance between the speaker and the speech recognition system can impact audio quality and accuracy. Rodrigues et al. [57] also demonstrated that the accuracy of speech recognition was decreased with the increasing distance between the user and the microphone. Nematollahi et al. [58] suggested that the lesser distance provides better accuracy outcomes for speech recognition. The present study confirms the same.

From the overall analysis of present study results, it could be inferred that the distance between the person and Pepper should be small. For most of the age groups, it was found that the speech recognition system understood every statement properly from a 1 m distance, except for the age group of 21–30. There could be many possible reasons which might have impacted these results. One such possibility is different speech patterns. Speech patterns may vary depending on people's accents and pronunciations [59]. This might cause Pepper's speech recognition system to understand the same statements differently. The other reason could be the energy levels of a person during the experiment, as it may impact the person's speech. Background noise also influences the speech recognition capability of Pepper to properly understand the statements. Attawibulkul et al. [60] suggested that environmental noise can be one of the major challenges for the speech recognition system of robots. Gnanamanickam et al. [61] showed that noise signals can be suppressed using different methods, such as iterative signal enhancement, subspace-based speech enhancement, and nonlinear spectral subtraction.

5. Conclusions, Limitations, Implications, and Future Work

Traditional classroom teaching and learning can present challenges for students and educators both. Students may struggle to understand lectures for a variety of reasons. In some cases, educators may also feel frustrated by the lack of performance of students. Multimodal learning, which allows students to learn through both spoken words and text displayed on a humanoid robot's screen, may offer a potential solution to these issues. The integration of a humanoid robot with the speech-to-text recognition tool Whisper will assist in conveying communication. Further, from the results, it can be observed that the gender and age of a person did not have an impact to a similar extent. By integrating speech recognition tools with a humanoid robot, this proposed system could enhance students' understanding of lectures and improve accessibility, eventually motivating students to engage and participate more enthusiastically in classrooms.

The Pepper robot was selected for this study since it is one of the robots with a screen capable of displaying text subtitles, and due to its availability to the research team. However, there is a challenge with the speech recognition system of Pepper in that it cannot transcribe speech to text. Through a series of experiments, the other limitations of Pepper's speech recognition system were noticed. The findings suggest that one of the limitations is related to variability in accent, pronunciation, and speech patterns. It was identified through the analysis of evaluation measures that Pepper understood the same stated sentence differently. The limited vocabulary of Pepper also imposes another issue in understanding the complex words present in the statements. Background noise is another hindrance that inhibits the proper recognition of utterances.

Integrating a social robot with Whisper, a speech-to-text conversion tool, has significant implications across various domains, including education. This synergistic integration between social robotics and advanced speech recognition technology has the potential to bring forth a range of benefits and applications. Healthcare is one of the settings where this integration can help, by transcribing patient–doctor interactions. Health practitioners and patients both will benefit; patients with communication challenges can express their problems more conveniently, and patients can better understand their doctor's advice, by reading it on the robot's screen. The other domain is the service industry, such as retail, hotels, and tourism. In the service industry, customer service can be provided with the help of a social robot, which can assist customers in displaying particular information, such as product information, menus, and travel information on its screen. Additionally, a social robot can provide multilingual support to customers.

To validate the potential advantage of the proposed system, we plan to experiment in a real-world educational setting in the future. In this context, the present study is an important step, since it is crucial to be aware of and address the limitations of the proposed system in laboratory settings before employing it in a real classroom setting. The findings of this study suggest that distance from Pepper and speech complexity affect Pepper's speech recognition capability. Based on our research, it is recommended that teachers should be positioned within a minimum distance from the robot. As part of our ongoing research efforts, we have yet to investigate whether the positioning of the microphone device on participants relative to Pepper can impact Pepper's speech recognition. Additionally, we will aim to include the robot's responses alongside the text on its screen, which, we believe, will increase human–robot interaction and engagement. In order to improve the learning experience, we plan to synchronize the robot's screen with a projector in the future. This will enable students to view the text displayed on the projector screen that corresponds with what the teacher is saying. This will not only lead to better understanding but also ensure that all students have access to the same information at the same time. The proposed system will serve as the foundation for enhanced speech understanding through which the teaching–learning process will be strengthened.

In order to better understand user preferences, the gathering of qualitative data is important. In the future, in the classroom setting, interviews and surveys can also be conducted among students and teachers to get their feedback about the system. The

interviews will focus on collecting insights from user experience. This experience will include the person's perceptions regarding the correctness of recorded speech. Furthermore, we will also identify any challenges they faced and collect their suggestions for further improvement. A more comprehensive analysis can be done by incorporating qualitative insights along with quantitative results, which can provide a complete understanding of the performance of the speech recognition system and enhance the findings of this study. This will help us to improve the system further.

Author Contributions: A.P. and D.M. designed the project. A.P. carried out the study and analyzed the results. A.P. and D.M. wrote the overall manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data can be made available on request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Dautenhahn, K. Socially intelligent robots: Dimensions of human–robot interaction. *Philos. Trans. R. Soc. B Biol. Sci.* **2007**, *362*, 679–704. [CrossRef]
2. Engwall, O.; Lopes, J. Interaction and collaboration in robot-assisted language learning for adults. *Comput. Assist. Lang. Learn.* **2022**, *35*, 1273–1309. [CrossRef]
3. Lytridis, C.; Bazinas, C.; Papakostas, G.A.; Kaburlasos, V. On measuring engagement level during child-robot interaction in education. In *Robotics in Education: Current Research and Innovations 10*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–13.
4. Christodoulou, P.; Reid, A.A.M.; Pnevmatikos, D.; del Rio, C.R.; Fachantidis, N. Students participate and evaluate the design and development of a social robot. In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 31 August–4 September 2020.
5. Wang, K.; Sang, G.-Y.; Huang, L.-Z.; Li, S.-H.; Guo, J.-W. The Effectiveness of Educational Robots in Improving Learning Outcomes: A Meta-Analysis. *Sustainability* **2023**, *15*, 4637. [CrossRef]
6. Alemi, M.; Meghdari, A.; Ghazisaedy, M. Employing humanoid robots for teaching English language in Iranian junior high-schools. *Int. J. Humanoid Robot.* **2014**, *11*, 1450022. [CrossRef]
7. Kennedy, J.; Baxter, P.; Senft, E.; Belpaeme, T. Higher Nonverbal Immediacy Leads to Greater Learning Gains in Child-Robot Tutoring Interactions. In *Social Robotics*; Springer International Publishing: Cham, Switzerland, 2015.
8. Mishra, D.; Inal, Y.; Parish, K.; Romero, G.A.; Rajbhandari, R. Exploring Active and Critical Engagement in Human-Robot Interaction to Develop Programming Skills: A Pilot Study. In *Design, User Experience, and Usability*; Springer Nature: Cham, Switzerland, 2023.
9. Alemi, M.; Meghdari, A.; Ghazisaedy, M. The effect of employing humanoid robots for teaching English on students' anxiety and attitude. In Proceedings of the 2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, 15–17 October 2014.
10. Belpaeme, T.; Kennedy, J.; Ramachandran, A.; Scassellati, B.; Tanaka, F. Social robots for education: A review. *Sci. Robot.* **2018**, *3*, eaat5954. [CrossRef] [PubMed]
11. Movellan, J.; Eckhardt, M.; Virnes, M.; Rodriguez, A. Sociable robot improves toddler vocabulary skills. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, La Jolla, CA, USA, 9–13 March 2009.
12. Rani, A.; Pande, A.; Parish, K.; Mishra, D. Teachers' Perspective on Robots Inclusion in Education–A Case Study in Norway. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023.
13. Tsang, A. Why English accents and pronunciation 'still'matter for teachers nowadays: A mixed-methods study on learners' perceptions. *J. Multiling. Multicult. Dev.* **2020**, *41*, 140–156. [CrossRef]
14. Pepper Robot Homepage. Available online: https://www.aldebaran.com/en/pepper (accessed on 7 May 2023).
15. Shezi, M.; Ade-Ibijola, A. Deaf chat: A speech-to-text communication aid for hearing deficiency. *Adv. Sci. Technol. Eng. Syst. J.* **2020**, *5*, 826–833. [CrossRef]
16. Shadiev, R.; Hwang, W.-Y.; Chen, N.-S.; Huang, Y.-M. Review of speech-to-text recognition technology for enhancing learning. *J. Educ. Technol. Soc.* **2014**, *17*, 65–84.
17. Debnath, S.; Roy, P.; Namasudra, S.; Crespo, R.G. Audio-Visual Automatic Speech Recognition Towards Education for Disabilities. *J. Autism Dev. Disord.* **2022**, *53*, 3581–3594. [CrossRef]
18. Goss, F.R.; Blackley, S.V.; Ortega, C.A.; Kowalski, L.T.; Landman, A.B.; Lin, C.-T.; Meteer, M.; Bakes, S.; Gradwohl, S.C.; Bates, D.W. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *Int. J. Med. Inform.* **2019**, *130*, 103938. [CrossRef]

19. Pande, A.; Shrestha, B.; Rani, A.; Mishra, D. A Comparative Analysis of Real Time Open-Source Speech Recognition Tools for Social Robots. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023.
20. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
21. Vásquez-Correa, J.C.; Arzelus, H.; Martin-Doñas, J.M.; Arellano, J.; Gonzalez-Docasal, A.; Álvarez, A. When Whisper Meets TTS: Domain Adaptation Using only Synthetic Speech Data. In Proceedings of the International Conference on Text, Speech, and Dialogue, Pilsen, Czech Republic, 4–6 September 2023.
22. Macháček, D.; Dabre, R.; Bojar, O. Turning Whisper into Real-Time Transcription System. *arXiv* **2023**, arXiv:2307.14743.
23. Spiller, T.R.; Ben-Zion, Z.; Korem, N.; Harpaz-Rotem, I.; Duek, O. Efficient and Accurate Transcription in Mental Health Research—A Tutorial on Using Whisper AI for Sound File Transcription. 2023. Available online: https://osf.io/9fue8/ (accessed on 7 May 2023). [CrossRef]
24. Fujii, A.; Kristiina, J. Open source system integration towards natural interaction with robots. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Sapporo, Japan, 7–10 March 2022.
25. Deuerlein, C.; Langer, M.; Sessner, J.; Heß, P.; Franke, J. Human-robot-interaction using cloud-based speech recognition systems. *Procedia CIRP* **2021**, *97*, 130–135. [CrossRef]
26. Grasse, L.; Boutros, S.J.; Tata, M.S. Speech interaction to control a hands-free delivery robot for high-risk health care scenarios. *Front. Robot. AI* **2021**, *8*, 612750. [CrossRef]
27. Mitsea, E.; Drigas, A. A journey into metacognitive learning strategies. *Int. J. Online Biomed. Eng.* **2019**, *15*, 4–22. [CrossRef]
28. Mitsea, E.; Drigas, A.; Skianis, C. Metacognition in Autism Spectrum Disorder: Digital Technologies in Metacognitive Skills Training. *Tech. Soc. Sci. J.* **2022**, *31*, 153. [CrossRef]
29. Naoqi API Documentation-ALAudioRecorder. Available online: https://doc.aldebaran.com/2-5/naoqi/audio/alaudiorecorder.html (accessed on 7 May 2023).
30. Naoqi API Documentation ALTabletService. Available online: https://doc.aldebaran.com/2-5/naoqi/core/altabletservice.html (accessed on 7 May 2023).
31. Natural Language Toolkit. Available online: https://www.nltk.org/ (accessed on 19 June 2023).
32. Textstat. Available online: https://pypi.org/project/textstat/ (accessed on 19 June 2023).
33. Farr, J.N.; Jenkins, J.J.; Paterson, D.G. Simplification of Flesch reading ease formula. *J. Appl. Psychol.* **1951**, *35*, 333. [CrossRef]
34. Kincaid, J.P.; Fishburne, R.P., Jr.; Rogers, R.L.; Chissom, B.S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*; Institute for Simulation and Training, University of Central Florida: Orlando, FL, USA, 1975.
35. Gunning, R. *Technique of Clear Writing*; McGraw-Hill: New York, NY, USA, 1952.
36. McLaughlin, G. SMOG grading—A new readability formula. *J. Read.* **1969**, *12*, 639–646.
37. Björnsson, C.-H. Readability of newspapers in 11 languages. *Read. Res. Q.* **1983**, *18*, 480–497. [CrossRef]
38. Dale, E.; Chall, J.S. A formula for predicting readability: Instructions. *Educ. Res. Bull.* **1948**, *27*, 37–54.
39. Paramiko Documentation. Available online: https://www.paramiko.org/ (accessed on 7 May 2023).
40. OpenAI Whisper. Available online: https://openai.com/research/whisper (accessed on 7 May 2023).
41. Wikipedia Page Word Error Rate. Available online: https://en.wikipedia.org/wiki/Word_error_rate (accessed on 7 May 2023).
42. Match Error Rate Documentation. Available online: https://torchmetrics.readthedocs.io/en/stable/text/match_error_rate.html (accessed on 7 May 2023).
43. Word Information Lost Documentation. Available online: https://torchmetrics.readthedocs.io/en/stable/text/word_info_lost.html (accessed on 7 May 2023).
44. Character Error Rate Documentation. Available online: https://torchmetrics.readthedocs.io/en/stable/text/char_error_rate.html#:~:text=character%20error%20rate%20is%20a,0%20being%20a%20perfect%20score (accessed on 7 May 2023).
45. Lhoussain, A.S.; Hicham, G.; Abdellah, Y. Adaptating the levenshtein distance to contextual spelling correction. *Int. J. Comput. Sci. Appl.* **2015**, *12*, 127–133.
46. Li, Q.; Russell, M.J. An analysis of the causes of increased error rates in children's speech recognition. In Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002.
47. Kennedy, J.; Lemaignan, S.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F.; Senft, E.; Belpaeme, T. Child speech recognition in human-robot interaction: Evaluations and recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017.
48. Diaz-Asper, C.; Chandler, C.; Turner, R.S.; Reynolds, B.; Elvevåg, B. Acceptability of collecting speech samples from the elderly via the telephone. *Digit. Health* **2021**, *7*, 20552076211002103. [CrossRef] [PubMed]
49. Errattahi, R.; El Hannani, A.; Ouahmane, H. Automatic speech recognition errors detection and correction: A review. *Procedia Comput. Sci.* **2018**, *128*, 32–37. [CrossRef]
50. Horton, W.S.; Spieler, D.H.; Shriberg, E. A corpus analysis of patterns of age-related change in conversational speech. *Psychol. Aging* **2010**, *25*, 708. [CrossRef] [PubMed]
51. Young, V.; Mihailidis, A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assist. Technol.* **2010**, *22*, 99–112. [CrossRef] [PubMed]

52. Aman, F.; Vacher, M.; Rossato, S.; Portet, F. Analysing the performance of automatic speech recognition for ageing voice: Does it correlate with dependency level? In Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France, 21–22 August 2013.
53. Pépiot, E. Voice, Speech and Gender: Male-female acoustic differences and cross-language variation in english and french speakers. In Proceedings of the XVèmes Rencontres Jeunes Chercheurs de l'ED 268, Paris, France, June 2012; (à paraître), ffhalshs-00764811f. Available online: https://shs.hal.science/halshs-00764811/document (accessed on 7 May 2023).
54. Tsantani, M.S.; Belin, P.; Paterson, H.M.; McAleer, P. Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception* **2016**, *45*, 946–963. [CrossRef] [PubMed]
55. Adda-Decker, M.; Lamel, L. Do speech recognizers prefer female speakers? In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
56. Garnerin, M.; Rossato, S.; Besacier, L. Gender representation in French broadcast corpora and its impact on ASR performance. In Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, Nice, France, 21 October 2019.
57. Rodrigues, A.; Santos, R.; Abreu, J.; Beça, P.; Almeida, P.; Fernandes, S. Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. In Proceedings of the XX International Conference on Human Computer Interaction, Donostia, Spain, 25–28 June 2019.
58. Nematollahi, M.A.; Al-Haddad, S.A.R. Distant speaker recognition: An overview. *Int. J. Humanoid Robot.* **2016**, *13*, 1550032. [CrossRef]
59. Braber, N.; Smith, H.; Wright, D.; Hardy, A.; Robson, J. Assessing the Specificity and Accuracy of Accent Judgments by Lay Listeners. *Lang. Speech* **2023**, *66*, 267–290. [CrossRef]
60. Attawibulkul, S.; Kaewkamnerdpong, B.; Miyanaga, Y. Noisy speech training in MFCC-based speech recognition with noise suppression toward robot assisted autism therapy. In Proceedings of the 2017 10th Biomedical Engineering International Conference (BMEiCON), Hokkaido, Japan, 31 August–2 September 2017.
61. Gnanamanickam, J.; Natarajan, Y.; KR, S.P. A Hybrid Speech Enhancement Algorithm for Voice Assistance Application. *Sensors* **2021**, *21*, 7025. [CrossRef]