Improving Accessibility and Independence for Blind/Visually Impaired Persons based on Speech Synthesis Technology

Manpreet Kaur Dhaliwal
Department of Computer Science and Applications,
Panjab University, Chandigarh

Rohini Sharma
Department of Computer Science and Applications,
Panjab University, Chandigarh

ABSTRACT
Speech is a crucial communication tool and Text-to-Speech systems are revolutionizing the world by enabling disabled persons to access information and achieve independence. This study investigates the relevance and effects of speech synthesis systems in enhancing the independence and accessibility of people with visual impairments. An overview of voice synthesis technology, followed by categories of speech synthesis systems is given in this study. Studies that increase BVIPs' freedom and accessibility are also considered in the analysis. To evaluate the speech quality of synthesis systems in terms of naturalness and intelligibility, the pilot study is carried out utilizing the gTTS, pyttsx3, SpeechT5, and Bark models. It has been observed that SpeechT5 and pyttsx3 are performing very well in terms of naturalness and intelligibility.

Keywords
Blind/Visually impaired persons, gTTS, Speech Synthesis, SpeechT5, pyttsx3, Mean Opinion Scale.

1. INTRODUCTION
Vision is the most important sense organ in the human body and 80% of information is perceived by vision. Blindness/Vision impairment negatively affects the quality of life and increases the economic burden globally [1]. A study of 2044 participants was undertaken to determine the significance of eye health in the US population. The study indicated that having a strong vision is essential to overall health and losing eyesight is more detrimental to health than hearing, memory, speech, or limb loss [2]. 7.08 million US population is predicted to have visual acuity loss, of these, 1.08 million are blind. Of them, 1.62 million people with reduced vision are under 40 years old [3]. More than 25% of Indians over the age of 50 are visually impaired, according to the National Blindness and Visual Impairment Survey 2015–2019. Blindness is significantly correlated with illiteracy and older age. Cataracts, corneal opacity (CO), cataract surgical complications, etc., were the leading causes of blindness [4][5].

The Blind/Visually impaired persons (BVIPs) face numerous obstacles in their quest for information and independence. The accessibility and usability of traditional methods of accessing information, like braille or audio recordings, also have limitations. Consequently, an efficient and user-friendly solution is required to improve the accessibility and independence of the visually impaired. The assistive technology for BVIPs market is anticipated to reach US $4.2 billion globally in 2023 and expand at a 13.1% compound annual growth rate (CAGR) from 2023 to 2033[6]. Few technological devices are available for BVIPs like screen readers [7], NVDA (Non-Visual Desktop Access) [8], JAWS (Job Access With Speech) [9], OrCam MyEye [10] that convey visual information audibly, The vOICe (visual-to-auditory) or BrainPort sensory substitution devices (SSDs), which convert visual data into audio [11], smart cane [12], smart glasses [13], Tactile graphics [14], Computer Algebra System Aimed at Visually Impaired People (CASVI) [15], etc. The main form of interpersonal communication is speech. For many years, researchers have been working on synthetic creation of speech. Stewart unveiled the first complete electrical synthesis apparatus in 1922. Although many synthesizers have been designed with excellent intelligibility and naturalness, achieving genuine sound quality remains difficult. It has been observed that if a person lacks one of the sense organs, the capability of other senses is enhanced. Danielle Bragg et al. [16] conducted a study on 453 participants to study the human listening rates and the effect of textual intricacy. It has been observed that experienced users prefer robust voices at rapid speeds, whereas inexperienced users prefer speech that sounds human. 120–180 words per minute is the average speaking rate. A few people who use screen readers adjust the speed to 500 words per minute. It was concluded that BVIPs typically listen at a higher rate than sighted persons. Notably, BVIP students at the University have also been surveyed regarding speech-based outputs, listening rates, preferences, and pitch. BVIPs claim a preference for speech-based output, and occasionally become tired of listening to voices and turn to Braille systems. Second, as the years have gone by, their listening rates have accelerated. Thirdly, it has been observed that people prefer the voices of men and women opposite their genders, but the clarity, naturalness, and calming quality of the voice ultimately determine which one to choose. BVIPs don't like synthetic voices and find it annoying.

Speech synthesis systems can improve their ability to access different types of written content such as books, documents, websites, and digital media. BVIPs will be able to lead more autonomous and satisfying lives as a result. Through utilizing technological developments and addressing the shortcomings of current approaches, researchers can work to close the accessibility gap and improve inclusivity for BVIPs. Speech-based technologies that support BVIPs in their daily activities, that promote independence and accessibility are the main focus of this study. For the benefit of the blind community, it is hoped that the knowledge and insights shared in this work will stimulate more investigation, creativity, and cooperation in the field of voice synthesis technology.

This article has the following structure. The overview of speech synthesis systems is discussed in Section 2. Studies on TTS that promote BVIPs' independence and accessibility are covered in Section 3. The studies on natural language processing are considered in Section 4, and experiments for assessing speech quality are performed in Section 5. The conclusion is provided in Section 6.

2. OVERVIEW OF SPEECH SYNTHESIS SYSTEMS

Text-to-speech (TTS), or speech synthesis, is a technology that translates written text into spoken words. To produce speech that sounds human, computer algorithms, deep learning models, transformers, corpus-based systems, and artificial voices are used in this procedure. Speech synthesis is used in many different applications, such as telecommunications, vocally disabled persons, voice assistants, e-governance services, navigation systems, accessibility features for those with vision impairments [17], and more. The basic steps involved in the speech synthesis systems are shown in Fig. 1 [18]:

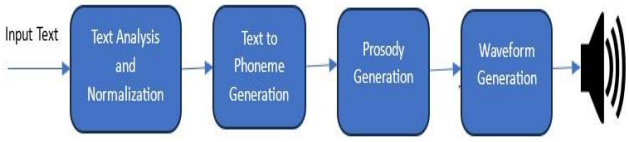

Fig 1: Flow of Text-to-Speech Synthesis

1. Text Analysis and Normalization: The input text is first examined by the system to recognize words, punctuation, and other linguistic components and put the content into a uniform format to improve speech and processing [19]. For example, the input sentence "The quick brown fox jumps over the lazy dog." In terms of text analysis, it can be broken down as follows:

- It consists of 43 characters, including spaces and punctuation.
- There are 9 words in the sentence.
- The sentence contains all 26 letters of the English alphabet.
- It is a simple declarative sentence with a subject ("fox") and a verb ("jumps").

In text normalization, spelling or grammatical errors are corrected, abbreviations are expanded and ensure consistent formatting.

2. Text to Phoneme Generation: The system then generates phonemes, the basic units of sound in a language, based on the text analysis. For example, the alphabet of the International Phonetic Association contains diacritical marks, phoneme symbols, and other symbols related to speech.

3. Prosody Generation: The patterns of stress and intonation in speech are referred to as prosody. To enhance the naturalness of the synthesized speech, the system produces prosodic elements. The pitch, tone, and tempo of the synthetic voice are all modeled to give the speech a more realistic feel. Prosody, or speech rhythm, emphasis, and intonation, is also thought to improve the synthetic voice's naturalness.

4. Waveform Generation: Finally, the system creates a waveform, the real sound signal that symbolizes the synthetic speech by converting the phonemes and prosody data.

2.1 Categories of Speech Synthesis Systems

Speech synthesis systems are broadly divided into three categories as shown in Fig. 2 [20].

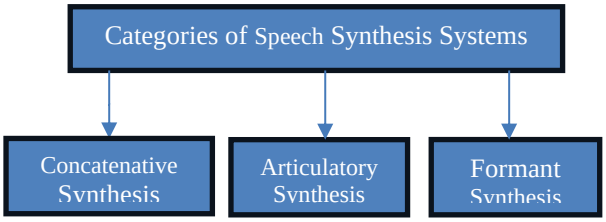

Fig 2: Categories of Speech Synthesis Systems

2.1.1 Concatenative Synthesis: In this method, short, pre-recorded audio clips known as "units" or "grains" are combined to produce new sounds in digital audio processing. These time intervals are usually very short, lasting anywhere from a few milliseconds to a few hundred milliseconds. Complex and human-like sounds that imitate genuine or artificial sources can be produced by choosing and placing these elements in a particular order. To achieve smooth transitions between the units and a coherent and continuous sound output, concatenative synthesis uses algorithms that examine the spectral and temporal properties of the units. Concatenative synthesis is further divided into domain-specific synthesis, phoneme synthesis, and unit synthesis systems [20].

2.1.2 Articulatory Synthesis: This technique simulates the locations and motions of the articulatory organs—such as the tongue, lips, and vocal folds that are used to produce speech. By simulating the physiological mechanisms involved in human voice creation, it seeks to produce speech. These are computational simulations of the articulatory organs' motions and locations [21].

2.1.3 Formant Synthesis: It modifies the formants of the human vocal tract to produce artificial speech. The resonant frequencies that the vocal tract produces during speech production are called formants [21]. Speech that sounds understandable and natural can be produced by adjusting the frequencies and amplitudes of these formats. To extract the formant frequencies and amplitudes from recorded speech, software is needed that can analyze speech. This software aids in recognizing the traits of various phonemes and formants. Software like Wavesurfer, Praat, or specialist formant analysis tools are a few examples of this type. After extracting formant information, software is needed to manipulate formant frequencies and amplitudes to create phonemes and words. Some software includes Festival, espeak, Klatt, and Linear Predictive Coding Synthesizer.

All the synthesis systems have some pros and cons. Concatenative synthesis provides naturalness [22] and intelligibility like real human voice but glitches, high memory requirements, and time consumption are some of the limitations. In concatenative, unit synthesis is a widely used method that is based on corpus[20]. Formant synthesis is based on the source-filter model and produces more sounds compared to concatenative synthesis but audio output is more robotic and unnatural. Formant synthesis is a technique used by many TTS systems to produce speech from text input. Instead of using phonetic representations to control the formant synthesis process, these systems use linguistic rules and algorithms to convert written text into phonetic representations. Some of the text-to-speech systems that use formant synthesis are Apple's Siri, Microsoft Speech Platform, and Google Text-to-Speech. Articulatory synthesis is a very sophisticated speech production system and is difficult to implement. To evaluate speech output quality, Mean Opinion Score (MOS) is a widely used method.

3. TEXT-TO-SPEECH SYSTEMS THAT BOOST BVIPS' INDEPENDENCE AND ACCESSIBILITY

As discussed in Section 1 BVIPs have good listening capabilities and prefer speech-based systems for assistance and learning. Speech synthesis systems can enhance their access to various textual material, including books, mathematical documents, webpages, and digital media, improve surrounding accessibility, and can use various systems like email systems, ChatGPT, etc., as normal beings. Fig. 3 shows the speech-based technologies used by BVIPs. Advancements in assistive

devices and speech-based technologies have the potential to enhance accessibility, independence, and Social Inclusion to mitigate the challenges faced by BVIPs in life [1]. This will enable them to live more independent and fulfilling lives. This section considers the studies that improve accessibility and independence for BVIPs based on speech systems.

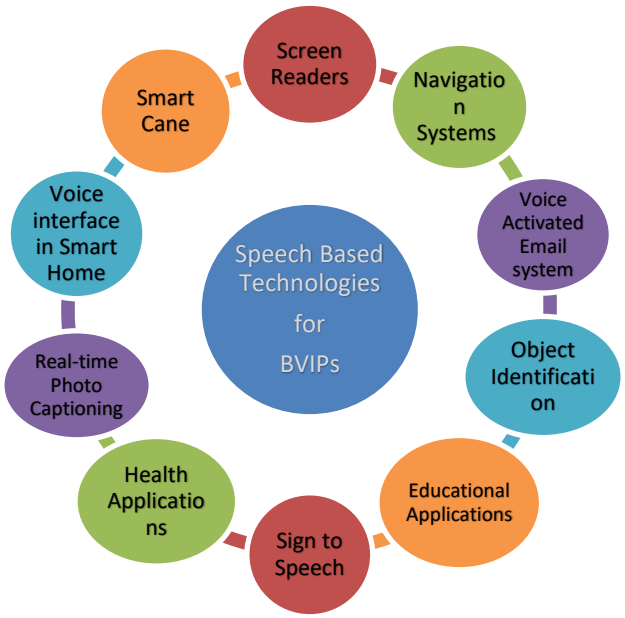

Fig 3: Speech-Based Technologies for BVIPs

Malathi D et al. [23] developed a voice-activated email system that enables BVIPs to send and receive emails independently. Training and testing are done with convolutional neural networks. The system is separated into two parts. A speech recognition system that can identify a voice, and translate it to text, and Text to speech converter that uses the gTTS module in Python to convert text to speech. The system is tested for actual users.

G.Nirosha and Dr. Velmani Ramasamy [24] proposed a sign-to-speech conversion system for older people, mute communities, and physically disabled persons. Flex sensors are used for sign conversion to voice and a 3.5mm audio jack is used for audio output. Raspberry Pi microprocessor is used.

R. Prabha et al. [25] designed a Vivoice device that can identify objects. It consists of a Raspberry Pi 3 module, earphones, a push switch, and a camera. Images are taken with a CMOS camera and processed using Python's Tesseract OCR package. To convert text to speech, eSpeak is utilized. Ashveena A et al.[26] proposed a speech-based object identification and recognition system. The ultrasonic camera is used for capturing the image and a single shot detection algorithm is used for identifying and converting the image to text and it uses the pyttsx3 package for speech synthesizing. Tahani Jaser Alahmadi et al. [27] proposed the YOLOv4_Resnet101 model for object identification, and obstacle detection and provides real-time auditory output about the object to BVIPs. pyttsx3 library is used for audible output. Yi-Chin Huang and Cheng-Hung Tsai [28] suggested a methodical approach to building a voice interface, with automated speech recognition and customized speech synthesis that would allow BVIPs to understand their surroundings. For speech synthesis, an HMM-based Mandarin speech synthesis system is employed.

Valbon Ademi and Lindita Ademi [29] discussed the use of optical character recognition (OCR) and text-to-speech technology for visually impaired people. The paper highlights the development of camera-based products for reading printed text. The paper also discusses concatenative synthesis, norm-based synthesis, and MFM-based synthesis. Researchers also discussed the limitations of types i.e. Database size increased when different phonetic contexts had to be considered. The norm-based synthesis produced mechanical sound. Developing norms for controlling synthesis is difficult. Sneha.C.Madre et al. [30] used an OCR to convert the image to text and then matlab16 TTS was used for speech conversion. Vagdevi Adusumilli et al. [31]proposed a vision-based assistive system that can provide voice-based assistance in Telugu and English language. Tesseract is used for optical character recognition (OCR) and Python TTS is used for audio output. Sameer Agrawal and Neelam Agrawal's [32] study is based on OCR and audio conversion using the gTTS (Google text-to-speech Conversion) package. For character recognition, CNN is used. P. Swetha et al. [33] proposed a TTS and image recognition-based AI assistant for VIPs. Used Raspberry Pi camera, GPS module, and ultrasonic sensor for navigation and eSpeak synthesis for audio output.

Safiya K M and R Pandian [34] proposed a real-time photo captioning system based on audio output. Three models VGGNet-16, ResNet-152, and MobileNet-V3 are trained on the Flickr30k dataset having a total of 31k photos, 8k photos from Flickr8k, and 2k custom dataset with five matching captions each and VGGNet-16 is selected for deployment on raspberry pi and TTS API is used for TTS conversion.

Indrianto et al. [35] proposed a Finite state automaton (FSA) based voice system for a health monitoring system in the Indonesian language. FSA processes the beats per minute, temperature, and spo2 data and converts it into audio. Ervasti et al. [36] proposed a BlindsNFC (Near Field Communication) based medication management system for VIPs. NFC-based PDA reads aloud the dosage and medicine name by touching the medicine packet. This system is tested on 39 elderly people suffering from vision impairment in Spain and Finland. It was found that elderly VIPs learned, used, and showed great satisfaction with the BlindNFC device.

Branko LuIiT et al. [37] focused on the importance and creation of educational applications based on speech systems for BVIPs in the Serbian language. Two applications Lugram and Mastermind are created based on voice commands and results are encouraging after initial mentoring in BVIPs. For playing these games the keyboard is used as a tactile device and for voice commands ASR application is used. Concatenative and hidden Markov models are used for building a TTS system. Evaluation of these applications is challenging because of the constant feedback. Amjad Ali and Shah Khusro [38] proposed a speech-based SA-MEAS (Sympy-based Automated Mathematical Equation Analysis and Solver) software application to enhance accessibility to mathematics and learning among BVIPs. This application is used for solving and analyzing mathematical equations. Pyttsx3 is used for text-to-speech conversion. The main limitation of this application is unclear voice, lack of analysis, and solution in TTS. M. Khan et al.[39] presented a voice-based tool for text correction. The tool can be used as a text editor and utilizes Google text-to-speech and speech-to-text API for text correction. However, this system is not completely speech-based. C. Edirisinghe et al. [40] created interactive picture books with touch, smell, and sound for BVIP children. The multi-sensory interactive picture book was found to be entertaining during the user evaluation

with 10 educators and 25 visually impaired students from a special school in Malaysia. However, the book is quite simple, with only a few pages, and educators prefer lengthier tale books with more varied noises.

D. Vander Wilt et al. [41] audio description of live theater shows (musical and non-musical) is provided to BVIPs using sound processing techniques. For aligning the audio description with live performances, an online time-warping algorithm is utilized. Online time warping is a technique that synchronizes the two sequences of data that are different in the time domain. In this study recorded audio description is used and Apple's Tom's TTS is used for voice. In [42] the value of voice interface in smart homes for the elderly is assessed. The findings suggested that voice interfaces in smart homes can improve safety and independence for the elderly. The voice interface makes it possible for the elderly and blind to call for assistance from anywhere in the house and helps identify distress, falls, and dangerous situations. When it comes to speech interfaces, natural voices are preferred over synthetic voices.

Shivang Sunil Singh et al. [43] proposed an object detection and prevention system for providing better navigation systems for blind persons. This system is based on multiple algorithms and technologies such as crowd, object, and distance detection algorithms. Text-to-speech algorithms have been used for voice messages. For collision detection in real-time, cloud computing and machine learning are used. Lilit Hakobyan et al.[44] examined assistive devices based on information technologies for BVIPs. Researchers focused on tactile representation and haptic feedback methods for navigation. The quality of life is significantly impacted by visual impairment. The use of technology can improve the lives of those who are blind or visually impaired. Raihan Bin Islam et al.[45] proposed an intelligent system for BVIPs to navigate independently in their environment. MobileNetV2 SSDLite technique is used for object detection and surrounding environment description. Google text-to-speech library generates audio feedback for each class label and object detection model files are saved in MP3 format when an object is detected, the corresponding MP3 file is played back to the user for audio feedback, enabling the system to generate audio outputs of the detected objects.

Similarly, Abbineni Charishma et al.[46] proposed OCR and audio-based assistive devices for reading books and newspapers. In [47][48][49][50] text detection and audio feedback are provided. Memoona Mushtaq et al.[51] proposed a smart cane that recognized objects and eSpeak for TTS conversion.

Apart from the speech-based studies, review studies are also considered to gain in-depth insights into the technologies for BVIPs, and drawbacks in existing systems. In this study, Muhsin, Zahra J. et al. [1] examined the assistive technologies that BVIPs use for navigation, sound-based devices, mobile apps, and smartphone-based gadgets. In this study, it is observed that assistive devices are underutilized because of the lack of BVIP involvement in their development, and the low adoption rate in most technologies is caused by their emphasis on functionality rather than experience. The majority of research was done on navigation systems. White cane obstacle detection methods can be enhanced by low-cost and renewable energy sources. Better assistive technology solutions may result from distinguishing the demands of people who are colorblind and partially sighted from those who are completely blind. Effective algorithms and multimodal interfaces are essential for enhancing assistive technologies' usability for BVIP. For assistive aids to be successful and widely accepted, BVIP must be included in their creation and testing. F. Hugo, et al. [52] provided a thorough analysis of the BVIPs' use of navigation and spatial orientation technology. The study discusses many direct sensing-based localization approaches, including computer vision, RFID, GPS, and Wi-Fi, and highlights their benefits and drawbacks in terms of accuracy, dependability, and affordability. M.D Messaoudi et al. [53] examined assistive technology, navigation strategies such as Wi-Fi, Bluetooth, RFID, and ultrasonic sensors, as well as aural feedback solutions for BVIPs. The shortcomings of assistive technology were also mentioned by the researchers, along with suggestions for ways to increase their mobility both inside and outside. While the majority of the techniques make sense in theory, it could be unduly complex or arduous for the user in actual use. B. Kuriakose et al. [54]examined BVIP navigation systems for both indoor and outdoor environments. Moreover, Identified the essential components that navigation systems lack. Due to the size of the devices, the difficulty of carrying them, and the exclusion of BVIPs from design and implementation methods, the majority of the solutions are theoretical and challenging to put into practice. A. Façanha et al. [55] examined the mobility and orientation environments intended for indoor navigation. Additionally, it emphasized the use of joysticks and spatial audio for orientation support.

After reviewing the literature, it has been observed that very little has been done in the field of speech synthesis and recognition in terms of assistive devices. Although very advanced systems like Alexa, Siri, etc. are present. In the scenario of BVIPs major focus is given to object detection, object recognition, screen readers, navigation systems, distance measurement, etc. but most of these devices are yet in theory the practical implementation of these devices is less because of inclusion of BVIPs in designing and implementation is nowhere. Most of the devices are not tested on BVIPs and feedback is not taken from them. It has also been observed that to assist BVIP in navigating their surroundings, technological devices can frequently be expensive and contradictory. It has been found that most of the devices are proposed but not commercially available for BVIPs because of the lack of resources. It has been noted that very little is discussed about speech-based output in the review studies. It is worth mentioning that speech-based output is used in navigation, object detection, screen readers, sign-to-speech conversion, etc. as discussed above but the quality of speech is not assessed by a single study. It was mentioned by researchers in some cases that gTTS or pyttsx3 TTS are used, but experiments are not performed to test the BVIPs acceptance for these TTS system. So, to remove this gap a pilot study is performed in Section 5 on BVIPs and normal human beings to assess the quality of TTS systems. In the next section, natural language processing-based studies are considered to understand how text-to-speech conversion is performed and how speech quality is assessed.

4. NATURAL LANGUAGE PROCESSING SYSTEMS

Tamrat Delessa Chala et al. [56] developed an Afan Oromo language text-to-speech synthesizer based on a unit selection synthesis approach. The research involves the collection of a text corpus of 1000 sentences from newspapers, and books and the recording of sentences in a native speaker's voice. For building the Afan Oromo speech synthesis system, the Festival tool is used. Speech synthesizer performance is tested in terms of naturalness and intelligibility on 3 users and the MOS score for intelligibility is 3.06 and 4.44 for naturalness. Yuxuan Wang et al. [57] proposed an end-to-end generative text-to-speech model based on the attention paradigm and seq2seq model. This model synthesizes voice directly from characters as input. This model is trained on an internal North American

English dataset, which contains about 24.6 hours of speech data spoken by a professional female speaker. In US English, Tacotron received a subjective mean opinion score of 3.82 on a 5-scale. As it uses frame-level autoregressive techniques, it operates more quickly than sample-level autoregressive models. As frame-level autoregressive methods operate at a coarser degree of granularity than sample-level approaches, they usually have a lower computing cost. However, it may sacrifice some minor details in the generated speech.

Monika Podsiadło et al. [58] conducted a pilot study on 90 users of screen readers in three US, Spain, and England countries to understand the preferences of VIP/Blind for the TTS system. The pilot study tested the naturalness, intelligibility, and male/female voice preferences. Male voices are preferred if it is narrative and clear. On the other hand, for conversation, a female voice is preferred. Voices that are calm, pleasant, emotionally neutral, with clear pronunciation, are most preferred in TTS systems by users. MOS and AB comparison tests were performed. Mukta Gahlawat et al. [59] developed a natural-sounding speech synthesizer. For this system, a speech corpus was created that consisted of 849 words and 168 sentences using the neutral, sad, and happy emotions of one female speaker. Speech synthesis is performed using Matlab TTSBOX. The speech system tested on 10 blind persons of age group 14-42 of Akhil Bhartiya Netrahin Sangh, Residential School and Training Center for Blinds Raghubir Nagar, New Delhi using listening test and naturalness, intelligibility, usability, localization awareness, expressions are the parameters considered for analysis. Good MOS was received for naturalness and usability, and fair MOS for intelligibility and localization.

In [60] vowels are synthesized using the cascade formant structure method, and the MOS scores for vowels [a, e, i, o, u] are 4.6, 4, 4.5, 4.1, and 4.5, respectively. In [61] Using the DC + SSRN model, ten participants were tested for intelligence and naturalness on two datasets: LibriSpeech and VCT-K. The results showed a MOS score of 3.47 ± 0.094 for the LibriSpeech dataset and 3.56 ± 0.093 for the VCT-K dataset. Sangramsing Kayte et al. [62] synthesized 30 sentences using a hidden Markov model and unit selection method. Ten sentences are used to compute each subject's MOS score. For instance, the MOS score for phrase 1 in the unit selection is 5, except for subject 8, receives a score of 4. Alakbar Valizada et al.[63] evaluates Tacotron and DC TTS, two speech synthesis systems based on the Azerbaijani language dataset collected over 24 hours from news websites, for both quality and intelligibility. Tacotron performed better for the In-Vocabulary words, as evidenced by the Mean Opinion Score, whereas DC TTS (Deep Convolutional Text-to-Speech) showed superior performance for the Out-Of-Vocabulary word synthesis. With an MOS of 3.49 ± 0.193, Tacotron performs somewhat better than DC TTS, which obtains an MOS of 3.36 ± 0.187.

The Mean Opinion Scale (MOS) technique is the most popular and straightforward way to assess speech quality, it is utilized to evaluate the synthesized text. The TTS is rated for intelligibility and naturalness using the arithmetic mean of the ratings provided by human raters. The corresponding points on the scale were Poor, Fair, Good, Very Good, and Excellent, ranging from 1 to 5 [23].

5. PILOT STUDY

There are various methods used to evaluate text-to-speech systems such as subjective test, objective test, one-word synthesis evaluation etc. In this study, subjective test is performed where listener judges the speech quality using two main metrics: naturalness and intelligibility. Naturalness means the clarity and fluency of the voice whereas intelligibility refers to the accuracy with which the phrases were interpreted.

To assess the speech quality in terms of naturalness and intelligibility of speech synthesis systems a pilot study has been conducted using gTTS, pyttsx3, SpeechT5, and Bark models as discussed below.

1. **SpeechT5:**SpeechT5 is based on transformer architecture that generates high-quality output [64]. SpeechT5 model is trained on large text corpus and unlabeled speech. SpeechT5 achieved a 2.91 MOS score for naturalness and an MOS of 3.65 ± 0.04 on the 460-hour LibriTTS dataset.
2. **Bark:** Bark is a text-to-audio model developed by Suno using transformers. Bark can produce a variety of audio, such as music, background noise, and basic sound effects, in addition to extremely lifelike, multilingual speech [65].
3. **gTTS:** gTTS commonly known as Google Text-to-Speech API converts text-to-speech and it supports several languages like English, French, Hindi, Tamil, etc., and by choosing speaker and rate of speech customization of speech output is possible. There is one limitation that active internet connection is required for gTTS to work because text input is processed by the Google server [66].
4. **pyttsx3:** A Python text-to-speech conversion library is called pyttsx3. It operates in offline mode. Several TTS engines are supported by the library, including espeak, nsss, and sapi5 [67].

Ten university students voluntarily participated in this study. All are of Indian descent, mentally fit, and do not have any hearing impairment. All students are from various departments of Panjab University Chandigarh. Each participant is given detailed information about the experiment and the entire method, and the goal of data collection and utilization is defined. Before starting the test, students were informed about the two-performance metrics naturalness and intelligibility are going to be graded from 1-5. Of these Ten students, 5 are BVIPs and 5 students are without any disability. Fig. 4 shows BVIP subjects listening to the various speeches for grading.

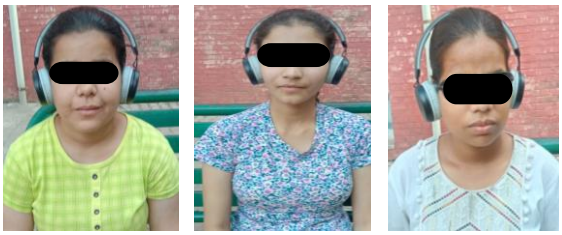

Fig 4: Subjects Listening to the Various Speeches for Grading

A verbal consent is taken from BVIPs before capturing and using their photographs. The following sentence is selected for testing the performance of speech synthesis systems. Results of performance measures in terms of intelligibility and naturalness are shown in Fig. 5 and Fig. 6.

"Do it again. Play it again. Sing it again. Read it again. Write it again. Sketch it again. Rehearse it again. Run it again. Try it again. Because again is practice, and practice is improvement, and improvement only leads to perfection."- Richelle E. Goodrich

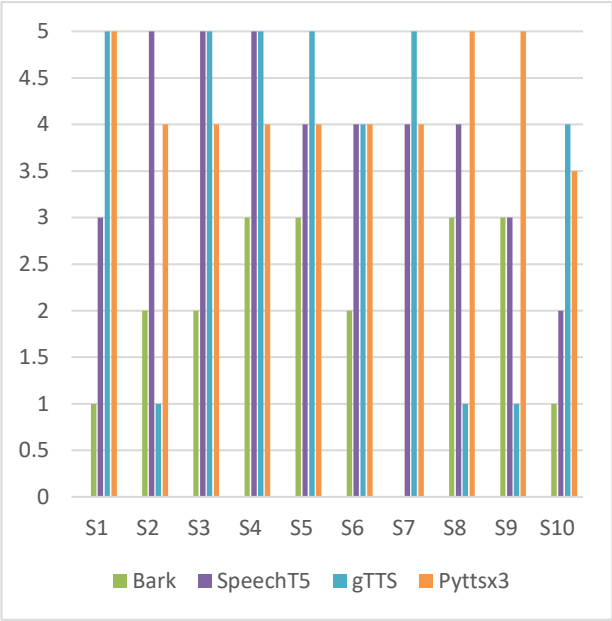

Fig 5: Subject wise performance measure TTS in terms of Intelligibility

It has been observed that pyttsx3 and SpeechT5 are performing similarly in terms of intelligibility with a MOS value of 4.11 and 4.22 respectively which means that speech quality is very good in terms of Intelligibility. As per BVIP students, excellent for SpeechT5, pyttsx3, and gTTS is graded. Similarly, in terms of naturalness, the pyttsx3 score is a little better than the SpeechT5 score with a MOS value

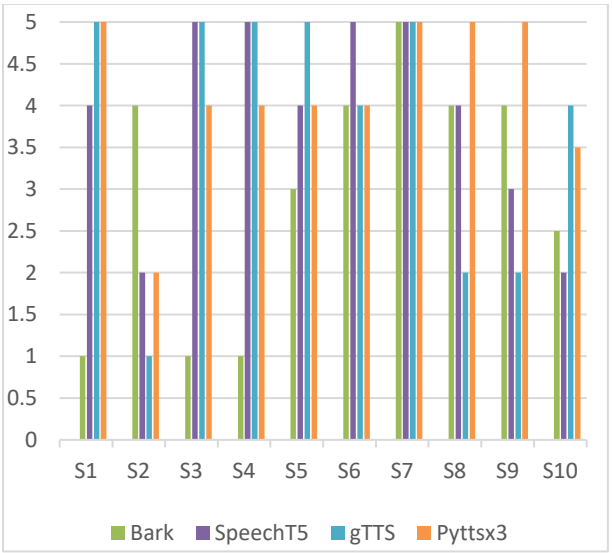

Fig 6: Subject wise performance measure TTS in terms of Naturalness

of 4.22. BVIPs have graded both models as excellent for naturalness.

The study's primary weakness is that each test is only administered on a single sentence. The speech synthesis system needs to be tested on a variety of sentences in order for its quality to be accurately assessed. Second, the study only included 10 pupils. Five of these ten are BVIPs. The test should include more BVIPs.

6. CONCLUSION

Vision is very important for perceiving the information. But for BVIPs speech is of greater importance because, with the help of speech synthesis technology, their lives are progressing toward independence and self-sufficiency in terms of information access and movement. This paper reviews the research that uses speech synthesis technology to increase BVIPs' independence and accessibility. It has been observed that all the focus has been given to object detection, recognition, mathematical tools, screen readers, etc., but the quality of speech-based systems has not been given that much importance. Studies on natural language processing have also been explored and it is concluded that naturalness and intelligibility are very crucial for good speech synthesis systems, especially for BVIPs. Four Speech synthesis systems pyttsx3, gTTS, SpeechT5, and Bark are used to test the quality of speech. It has been observed that SpeechT5 and pyttsx3 are the best systems in terms of naturalness and intelligibility. In the future, the speech quality of models will be tested on more subjects, and multi-sentence testing on the same model will be performed. Other performance metrics will also be explored.

7. ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance of Nisha, Manjot Kumar, Rekha, Anita, and Kishor Kumar from the Department of Hindi, Department of Punjabi, and Institute of Educational Technology & Vocational Education, Panjab University, Chandigarh for their continuous feedback, information sharing about BVIPs.

8. REFERENCES

[1] Z.J. Muhsin, R. Qahwaji, F. Ghanchi, M. Al-Taee, Review of substitutive assistive tools and technologies for people with visual impairments: recent advancements and prospects, J. Multimodal User Interfaces. 18 (2024) 135–156. https://doi.org/10.1007/s12193-023-00427-4.

[2] A.W. Scott, N.M. Bressler, S. Ffolkes, J.S. Wittenborn, J. Jorkasky, Public Attitudes About Eye and Vision Health, JAMA Ophthalmol. 134 (2016) 1111. https://doi.org/10.1001/jamaophthalmol.2016.2627.

[3] A.D. Flaxman, J.S. Wittenborn, T. Robalik, R. Gulia, R.B. Gerzoff, E.A. Lundeen, J. Saaddine, D.B. Rein, K.N. Baldonado, C. Davidson, M.C. Dougherty, M.R. Duenas, D.S. Friedman, K.M. Jackson, C.E. Joslin, B.E.K. Klein, P.A. Lamuda, Y. Liu, F.C. Lum, N.L. Okeke, N.P. Sinha, B.K. Swenor, J.P. Todd, E. Tolbert, Prevalence of Visual Acuity Loss or Blindness in the US, JAMA Ophthalmol. 139 (2021) 717. https://doi.org/10.1001/jamaophthalmol.2021.0527.

[4] P. Vashist, S.S. Senjam, V. Gupta, N. Gupta, B.R. Shamanna, M. Wadhwani, P. Shukla, S. Manna, S. Yadav, A. Bharadwaj, Blindness and visual impairment and their causes in India: Results of a nationally representative survey, PLoS One. 17 (2022) 1–14. https://doi.org/10.1371/journal.pone.0271736.

[5] K.A. Vashist Praveen, National Blindness & Visual Impairment Survey 2015-19: A Summary Report, Dir. Gen. Heal. Serv. (2019) 1–18. https://npcbvi.mohfw.gov.in/writeReadData/mainlinkFile/File341.pdf.

[6] Assistive Technologies for Visually Impaired Market, (n.d.). https://www.factmr.com/report/4635/assistive-technologies-demand-for-visually-impaired-market.

[7] A.F. for the Blind, Screen Readers - Browse by Category - American Foundation for the Blind, (2019). https://www.afb.org/blindness-and-low-vision/using-technology/assistive-technology-products/screen-readers (accessed January 23, 2024).

[8] J.P. Bigham, C.M. Prince, R.E. Ladner, WebAnywhere: A screen reader on-the-go, W4A'08 Proc. 2008 Int. Cross-Disciplinary Conf. Web Access. W4A. (2008) 73–82. https://doi.org/10.1145/1368044.1368060.

[9] S. Sandhya, K.A.S. Devi, Accessibility evaluation of websites using screen reader, in: 2011 7th Int. Conf. Next Gener. Web Serv. Pract., IEEE, 2011: pp. 338–341. https://doi.org/10.1109/NWeSP.2011.6088201.

[10] O. MyEye, OrCam MyEye For People Who Are Blind or Visually Impaired, (n.d.). https://www.orcam.com/en/myeye2/ (accessed January 23, 2024).

[11] C. Jicol, T. Lloyd-Esenkaya, M.J. Proulx, S. Lange-Smith, M. Scheller, E. O'Neill, K. Petrini, Efficiency of Sensory Substitution Devices Alone and in Combination With Self-Motion for Spatial Navigation in Sighted and Visually Impaired, Front. Psychol. 11 (2020). https://doi.org/10.3389/fpsyg.2020.01443.

[12] M.A. Rahman, S. Siddika, M.A. Al-Baky, M.J. Mia, An automated navigation system for blind people, Bull. Electr. Eng. Informatics. 11 (2022) 201–212. https://doi.org/10.11591/eei.v11i1.3452.

[13] H. Ali A., S.U. Rao, S. Ranganath, T.S. Ashwin, G.R.M. Reddy, A Google Glass Based Real-Time Scene Analysis for the Visually Impaired, IEEE Access. 9 (2021) 166351–166369. https://doi.org/10.1109/ACCESS.2021.3135024.

[14] A. Nasser, K. Zhu, P.V.M. Rao, Poster: Colortact: A Finger Wearable Audio-tactile Device Using Customizable Color Tagging, in: UbiComp/ISWC 2018 - Adjun. Proc. 2018 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2018 ACM Int. Symp. Wearable Comput., ACM, New York, NY, USA, 2018: pp. 178–181. https://doi.org/10.1145/3267305.3267583.

[15] P. Mejia, L.C. Martini, F. Grijalva, A.M. Zambrano, CASVI: Computer Algebra System Aimed at Visually Impaired People. Experiments, IEEE Access. 9 (2021) 157021–157034. https://doi.org/10.1109/ACCESS.2021.3129106.

[16] D. Bragg, K. Reinecke, R.E. Ladner, Expanding a Large Inclusive Study of Human Listening Rates, ACM Trans. Access. Comput. 14 (2021). https://doi.org/10.1145/3461700.

[17] Á. Csapó, G. Wersényi, H. Nagy, T. Stockman, A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research, J. Multimodal User Interfaces. 9 (2015) 275–286. https://doi.org/10.1007/s12193-015-0182-7.

[18] D. Sasirekha, E. Chandra, Text To Speech : a Simple Tutorial, Int. J. Soft Comput. Eng. 2 (2012) 275–278. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.682.5362&rep=rep1&type=pdf.

[19] S. Kayte, M. Mundada, J. Gujrathi, Hidden Markov Model based Speech Synthesis: A Review, Int. J. Comput. Appl. 130 (2015) 35–39. https://doi.org/10.5120/ijca2015906965.

[20] R. A., J. S., Concatenative Speech Synthesis: A Review, Int. J. Comput. Appl. 136 (2016) 1–6. https://doi.org/10.5120/ijca2016907992.

[21] L.A. Valbon ADEMI1, NATURAL LANGUAGE PROCESSING AND TEXT-TO-SPEECH TECHNOLOGY Valbon, J. Nat. Sci. Math. (2023) 1–9. https://doi.org/UDC:003.2:[004.51/52:004.934.5.

[22] S. Tiomkin, D. Malah, S. Shechtman, Z. Kons, A hybrid text-to-speech system that combines concatenative and statistical synthesis units, IEEE Trans. Audio, Speech Lang. Process. 19 (2011) 1278–1288. https://doi.org/10.1109/TASL.2010.2089679.

[23] D. Malathi, S. Gopika, D. Awasthi, D. Jayaseeli, Voice Automation Mail System for Visually Impaired, in: 2023 Int. Conf. Netw. Commun., IEEE, 2023: pp. 1–6. https://doi.org/10.1109/ICNWC57852.2023.10127558.

[24] G. Nirosha, R. Dr Velmani, Raspberry Pi based Sign to Speech Conversion System for Mute Community, IOP Conf. Ser. Mater. Sci. Eng. 981 (2020) 042005. https://doi.org/10.1088/1757-899X/981/4/042005.

[25] R. Prabha, M. Razmah, G. Saritha, R. Asha, S.G. A, R. Gayathiri, Vivoice - Reading Assistant for the Blind using OCR and TTS, in: 2022 Int. Conf. Comput. Commun. Informatics, IEEE, 2022: pp. 01–07. https://doi.org/10.1109/ICCCI54379.2022.9740877.

[26] A. Ashveena, J. Bala Deepika, S.P. Mary, D.U. Nandini, Portable Camera based Identification System for Visually Impaired People, 7th Int. Conf. Trends Electron. Informatics, ICOEI 2023 - Proc. (2023) 1444–1450. https://doi.org/10.1109/ICOEI56765.2023.10126008.

[27] T.J. Alahmadi, A.U. Rahman, H.K. Alkahtani, H. Kholidy, Enhancing Object Detection for VIPs Using YOLOv4_Resnet101 and Text-to-Speech Conversion Model, Multimodal Technol. Interact. 7 (2023). https://doi.org/10.3390/mti7080077.

[28] Y.-C. Huang, C.-H. Tsai, Speech-Based Interface for Visually Impaired Users, in: 2018 IEEE 20th Int. Conf. High Perform. Comput. Commun. IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Sci. Syst., IEEE, 2018: pp. 1223–1228. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00206.

[29] V. Ademi, L. Ademi, NATURAL LANGUAGE PROCESSING AND TEXT-TO-SPEECH TECHNOLOGY, J. Nat. Sci. Math. 8 (2023) 299–306. https://doi.org/https://eprints.unite.edu.mk/1528/1/JNSM%202023-299-306.pdf.

[30] S.C. Madre, S.B. Gundre, OCR Based Image Text to Speech Conversion Using MATLAB, in: Proc. 2nd Int. Conf. Intell. Comput. Control Syst. ICICCS 2018, IEEE, 2018: pp. 858–861. https://doi.org/10.1109/ICCONS.2018.8663023.

[31] V. Adusumilli, M.F. Shaik, N. Kolavennu, L.B.M.T. Adepu, A. V. Prabhu, I.R. Raja, Reading Aid and Translator with Raspberry Pi for Blind people, 2023 9th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2023.

1 (2023) 327–331. https://doi.org/10.1109/ICACCS57279.2023.10113042.

[32] S. Agrawal, N. Agrawal, Recogniition and Speech Conversion of Devnagri Script using CNN, in: 2023 2nd Int. Conf. Innov. Technol., IEEE, 2023: pp. 1–4. https://doi.org/10.1109/INOCON57975.2023.10101034.

[33] P. Swetha, AI Based Assistance for Visually Impaired People Using TTS (Text To Speech), Int. J. Innov. Res. Sci. Technol. 01 (2021) 8–014. www.ijirst.com.

[34] S. K M, R. Pandian, Real-Time Photo Captioning for Assisting Blind and Visually Impaired People Using LSTM Framework, IEEE Sensors Lett. 7 (2023) 1–4. https://doi.org/10.1109/LSENS.2023.3327565.

[35] Indrianto, Abdurrasyid, M.N.I. Susanti, A. Ramadhan, Text-to-speech on health monitoring bracelet for the visually impaired, Bull. Electr. Eng. Informatics. 12 (2023) 3826–3836. https://doi.org/10.11591/eei.v12i6.5369.

[36] M. Ervasti, M. Isomursu, I. Idigoras Leibar, Touch- and audio-based medication management service concept for vision impaired older people, 2011 IEEE Int. Conf. RFID-Technologies Appl. RFID-TA 2011. (2011) 244–251. https://doi.org/10.1109/RFID-TA.2011.6068645.

[37] B. Lučić, S. Ostrogonac, N. Vujnović Sedlar, M. Sečujski, Educational Applications for Blind and Partially Sighted Pupils Based on Speech Technologies for Serbian, Sci. World J. 2015 (2015) 1–14. https://doi.org/10.1155/2015/839252.

[38] A. Ali, S. Khusro, SA-MEAS: Sympy-based automated mathematical equations analysis and solver, SoftwareX. 25 (2024) 101596. https://doi.org/10.1016/j.softx.2023.101596.

[39] M.N.H. Khan, M.A.H. Arovi, H. Mahmud, M.K. Hasan, H.A. Rubaiyeat, Speech based text correction tool for the visually impaired, in: 2015 18th Int. Conf. Comput. Inf. Technol., IEEE, 2015: pp. 150–155. https://doi.org/10.1109/ICCITechn.2015.7488059.

[40] C. Edirisinghe, N. Podari, A.D. Cheok, A multi-sensory interactive reading experience for visually impaired children; a user evaluation, Pers. Ubiquitous Comput. 2018 (2018) 807–819. https://doi.org/10.1007/s00779-018-1127-4.

[41] D. Vander Wilt, M.M. Farbood, A new approach to creating and deploying audio description for live theater, Pers. Ubiquitous Comput. 25 (2021) 771–781. https://doi.org/10.1007/s00779-020-01406-2.

[42] F. Portet, M. Vacher, C. Golanski, C. Roux, B. Meillon, Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects, Pers. Ubiquitous Comput. 17 (2013) 127–144. https://doi.org/10.1007/s00779-011-0470-5.

[43] S.S. Singh, M. Agrawal, M. Eliazer, Collision detection and prevention for the visually impaired using computer vision and machine learning, Adv. Eng. Softw. 179 (2023) 103424. https://doi.org/10.1016/j.advengsoft.2023.103424.

[44] L. Hakobyan, J. Lumsden, D. O'Sullivan, H. Bartlett, Mobile assistive technologies for the visually impaired, Surv. Ophthalmol. 58 (2013) 513–528. https://doi.org/10.1016/j.survophthal.2012.10.004.

[45] R. Bin Islam, S. Akhter, F. Iqbal, M. Saif Ur Rahman, R. Khan, Deep learning based object detection and surrounding environment description for visually impaired people, Heliyon. 9 (2023) e16924. https://doi.org/10.1016/j.heliyon.2023.e16924.

[46] A. Charishma, A.A. Vaishnavi, D. Rajeswara Rao, T.T. Sri, Smart Reader for Visually Impaired, in: 2023 9th Int. Conf. Adv. Comput. Commun. Syst., IEEE, 2023: pp. 349–352. https://doi.org/10.1109/ICACCS57279.2023.10113122.

[47] U. Gawande, N. Rathod, P. Bodkhe, P. Kolhe, H. Amlani, C. Thaokar, Novel Machine Learning based Text-To-Speech Device for Visually Impaired People, in: 2023 2nd Int. Conf. Smart Technol. Syst. Next Gener. Comput., IEEE, 2023: pp. 1–5. https://doi.org/10.1109/ICSTSN57873.2023.10151637.

[48] T.M. Sivate, N. Pillay, K. Moorgas, N. Singh, Autonomous Classification and Spatial Location of Objects from Stereoscopic Image Sequences for the Visually Impaired, in: 2022 Int. Conf. Electr. Comput. Energy Technol., IEEE, 2022: pp. 1–6. https://doi.org/10.1109/ICECET55527.2022.9872538.

[49] F. Makhmudov, M. Mukhiddinov, A. Abdusalomov, K. Avazov, U. Khamdamov, Y.I. Cho, Improvement of the end-to-end scene text recognition method for "text-to-speech" conversion, Int. J. Wavelets, Multiresolution Inf. Process. 18 (2020) 2050052. https://doi.org/10.1142/S0219691320500526.

[50] I. Flores, G.C. Lacdang, C. Undangan, J. Adtoon, N.B. Linsangan, Smart Electronic Assistive Device for Visually Impaired Individual through Image Processing, 2021 IEEE 13th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2021. (2021) 1–6. https://doi.org/10.1109/HNICEM54116.2021.9731961.

[51] Memoona Mushtaq, Muhammad Munwar Iqbal, Ayesha Mariam, Aatka Ali, Muhammad Nabeel Asghar, Object Detection and Recognition for Virtual Vision: Using Text-to-Speech Conversion Technique, J. Comput. Biomed. Informatics. 4 (2022) 175–184. https://doi.org/10.56979/401/2022/82.

[52] H. Fernandes, P. Costa, V. Filipe, H. Paredes, J. Barroso, A review of assistive spatial orientation and navigation technologies for the visually impaired, Univers. Access Inf. Soc. 18 (2019) 155–168. https://doi.org/10.1007/s10209-017-0570-8.

[53] M.D. Messaoudi, B.A.J. Menelas, H. Mcheick, Review of Navigation Assistive Tools and Technologies for the Visually Impaired, Sensors. 22 (2022). https://doi.org/10.3390/s22207888.

[54] B. Kuriakose, R. Shrestha, F.E. Sandnes, Tools and Technologies for Blind and Visually Impaired Navigation Support: A Review, IETE Tech. Rev. (Institution Electron. Telecommun. Eng. India). 39 (2022) 3–18. https://doi.org/10.1080/02564602.2020.1819893.

[55] A.R. Façanha, T. Darin, W. Viana, J. Sánchez, O&M Indoor Virtual Environments for People Who Are Blind, ACM Trans. Access. Comput. 13 (2020). https://doi.org/10.1145/3395769.

[56] T.D. Chala, A.C. Guta, M.H. Asebel, Design and Development of a Text-to-Speech Synthesizer for Afan Oromo, SN Comput. Sci. 3 (2022) 1–7. https://doi.org/10.1007/s42979-022-01306-7.

[57] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R.A. Saurous, Tacotron: Towards End-to-End Speech Synthesis, in: Interspeech 2017, ISCA, ISCA, 2017: pp. 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452.

[58] M. Podsiadło, S. Chahar, Text-to-Speech for Individuals with Vision Loss: A User Study, in: Interspeech 2016, ISCA, ISCA, 2016: pp. 347–351. https://doi.org/10.21437/Interspeech.2016-1376.

[59] M. Gahlawat, A. Malik, P. Bansal, Natural Speech Synthesizer for Blind Persons Using Hybrid Approach, Procedia Comput. Sci. 41 (2014) 83–88. https://doi.org/10.1016/j.procs.2014.11.088.

[60] S. Lukose, S.S. Upadhya, Text to speech synthesizer-formant synthesis, in: 2017 Int. Conf. Nascent Technol. Eng., IEEE, 2017: pp. 1–4. https://doi.org/10.1109/ICNTE.2017.7947945.

[61] B. Asiedu Asante, H. Imamura, Speech Recognition and Speech Synthesis Models for Micro Devices, ITM Web Conf. 27 (2019) 05001. https://doi.org/10.1051/itmconf/20192705001.

[62] S.N. Kayte, M. Mundada, S. Gaikwad, B. Gawali, Performance evaluation of speech synthesis techniques for english language, Adv. Intell. Syst. Comput. 439 (2016) 253–262. https://doi.org/10.1007/978-981-10-0755-2_27.

[63] A. Valizada, S. Jafarova, E. Sultanov, S. Rustamov, Development and evaluation of speech synthesis system based on deep learning models, Symmetry (Basel). 13 (2021) 1–12. https://doi.org/10.3390/sym13050819.

[64] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, F. Wei, SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing, Proc. Annu. Meet. Assoc. Comput. Linguist. 1 (2022) 5723–5738. https://doi.org/10.18653/v1/2022.acl-long.393.

[65] Suno-ai, Bark, (n.d.). https://github.com/suno-ai/bark.

[66] Google, gTTS, (n.d.). https://gtts.readthedocs.io/en/latest/.

[67] N.M. Bhat, pyttsx3 2.90, GNU Gen. Public Licens. V3. (n.d.). https://pypi.org/project/pyttsx3/.