International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

A Review of Cross-Platform Document File Reader Using Speech Synthesis

Ogenyi Fabian Chukwudi[1], Val Hyginus Udoka Eze[1], Ugwu Chinyere N[1]

[1] *Department of Publication and Extension, Kampala International University. Uganda.*

Article History
Received:
17.10.2023

Revised:
09.11.2023

Accepted:
27.11.2023

***Corresponding Author:**
Val Hyginus Udoka Eze
Email:
ezehyginusudoka@gmail.com

Abstract: Document files are files used to store documents on storage devices primarily for computer use. Software is used to view these files, displaying their text content in a legible way. However, it is essential to have programs for transforming electronic files into versions usable by those who suffer from specific disabilities. This paper reviewed fifteen published articles in the field of document file reading. It was observed from the review that various attempts have been made by different researchers in order to develop a software cable for converting document files that consist of text to an audio format. Text may now be easily translated into natural-sounding voice across many platforms using different software. It was observed from the systematic review that the use of AI such as the GPT-3.5 and GPT-4 Turbo Large Language Model (LLM) technologies has the best performance because it does not end at producing a vocal sound that is similar to human own, but it also translates different languages. In conclusion, cross-platform document file reader (text-to-speech) synthesis has improved user experiences in a variety of applications such as language learning, audiobooks and virtual assistants.

Keywords: Cross-Platform, Document and File, Reader, Speech and Synthesis, Text.

International Journal of Artificial Intelligence, vol. 10, no. 2, pp. 103-110, December 2023. **DOI: 10.36079/lamintang.ijai-01002.569**

1. Introduction

Document files are text or binary files for storing documents on a storage media, especially for use by computers. Currently, there exists a multitude of document file formats with file extensions such as .doc, .docx, .pdf, .txt, .xls and .xlsx. Such files can be viewed by using special software which displays the text contents of these files in a readable format. The aim of this paper is to review some various stages at which several cross-platform and platform-dependent systems for reading document files were modified to have text-to-speech synthesizers (TTS) for reading out the content of document files in audio format. In order to solve problems encountered by various individuals with different challenges such as; learning disabilities, visual impairments, literacy difficulties and People who multitask. According to [1], materials such as JDK8 (Java Development Toolkit version 8), Java Programming Language, JavaFX framework (GUI development tool), NetBeans (Integrated Development Environment) and Laptop (Development Machine) were used. Finally, a conclusion was drawn that the use of text-to-speech synthesizers (TTS) as software has a great impact not only on people with different disabilities and impairments but on society at large. The introduction of the LLMs technology such as; GPT-3 turbo, GPT-3.5 and GPT-4 has made it easier, not only to conversion text to speech in a human-like voice, but translate texts to different languages [2].

2. Literature Review

There are several cross-platform and platform-dependent systems for reading document files but most of these document file readers do not have text-to-speech synthesizers for reading out the content of document files in audio format. Portable Document Format (PDF) is a file format developed by Adobe Systems Inc. for electronic document delivery, viewing, and printing. Version 1.0 of the PDF specification was introduced in 1992, and Adobe released tools to create and view PDF files in 1993. Sixteen years later, the most current specification version is 1.7, and it has been approved as the ISO 32000-1 standard (PDF) [3]. The basic Acrobat Reader, available for several desktop and mobile platforms, is freeware; it supports viewing, printing and annotating of PDF files.

Microsoft Office is a family of software developed by Microsoft. Initially a marketing term for an office suite, the first version of Office contained Microsoft Word, Microsoft Excel, and Microsoft PowerPoint. Microsoft Office prior to Office 2007 used proprietary file formats based on the OLE Compound File Binary Format [4].

Foxit Reader is a multilingual freemium PDF tool that can create, view, edit, digitally sign, and print PDF files [5]. Foxit Reader is developed by Fremont, California-based Foxit Software. Early versions of Foxit Reader were notable for startup performance and small file size [6].

Ghostscript is a suite of software based on an interpreter for Adobe Systems' PostScript and Portable Document Format (PDF) page description languages. Its main purposes are for the rasterization or rendering of such page description language files, for the display or printing of document pages, and for the conversion between PostScript and PDF files [7].

3. Methodology

3.1. Review of Speech Synthesis

According to [8]. The basic idea of text-to-speech (TTS) technology is to convert written input to spoken output by generating synthetic speech. There are several ways of performing speech synthesis [9]:

1. Simple voice recording and playing on demand;
2. Splitting of speech into 30-50 phonemes (basic linguistic units) and their recombination in a fluent speech pattern;
3. The use of approximately 400 diphones (splitting of phrases at the center of the phonemes and not at the transition).

3.2. Mechanical to Electrical Synthesis

The earliest efforts to produce synthetic speech were made over two hundred years ago [10] [11] [12]. In St. Petersburg 1779 Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially. He constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds like in musical instruments.

International Journal of Artificial Intelligence, vol. 10, no. 2, pp. 103-110, December 2023. **DOI: 10.36079/lamintang.ijai-01002.569**

A few years later, in Vienna 1791, Wolfgang von Kempelen introduced his "Acoustic-Mechanical Speech Machine", which was able to produce single sounds and some sound combinations [12] [13]. In fact, Kempelen started his work before Kratzenstein, in 1769, and after over 20 years of research he also published a book in which he described his studies on human speech production and the experiments with his speaking machine. The essential parts of the machine were a pressure chamber for the lungs, a vibrating reed to act as vocal cords, and a leather tube for the vocal tract action. By manipulating the shape of the leather tube, he could produce different vowel sounds. Consonants were simulated by four separate constricted passages and controlled by the fingers. For plosive sounds, he also employed a model of a vocal tract that included a hinged tongue and movable lips. His studies led to the theory that the vocal tract, a cavity between the vocal cords and the lips, is the main site of acoustic articulation. Before von Kempelen's demonstrations, the larynx was generally considered a center of speech production. Kempelen received also some negative publicity. While working with his speaking machine, he demonstrated a speaking chess-playing machine. Unfortunately, the main mechanism of the machine was concealed, legless chess-player expert. Therefore, his real speaking machine was not taken as seriously as it should have [11] [12].

In about mid 1800's Charles Wheatstone constructed his famous version of von Kempelen's speaking machine which is shown in Figure 1. It was a bit more complicated and was capable of producing vowels and most of the consonant sounds. Some sound combinations and even full words were also possible to produce. Vowels were produced with vibrating reed and all passages were closed. Resonances were affected by deforming the leather resonator like in von Kempelen's machine. Consonants, including nasals, were produced with turbulent flow through a suitable passage with reed-off.

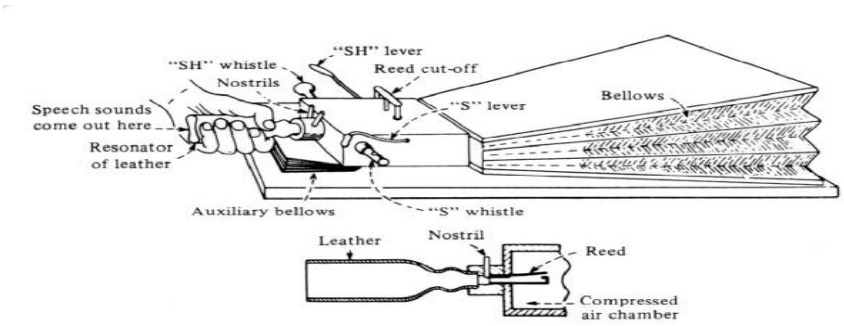

Figure 1. Wheatstone's Reconstruction of von Kempelen's Speaking Machine

The connection between a specific vowel sound and the geometry of the vocal tract was found by Willis is as shown in figure1 [12]. He synthesized different vowels with tube resonators like organ pipes. He also discovered that the vowel quality depended only on the length of the tube and not on its diameter. In the late 1800's Alexander Graham Bell with his father, inspired by Wheatstone's speaking machine, constructed the same kind of speaking machine. Bell made also some questionable experiments with his terrier. He put his dog between his legs and made it growl, and then he modified the vocal tract by hand to produce speech-like sounds [10] [12]. The research and experiments with mechanical and semi-electrical analogues of vocal systems were made until 1960's, but with no remarkable success. The mechanical and semi-electrical experiments made by famous scientists, such as Herman von Helmholz and Charles Wheatstone are well described in [10] [11] [12].

3.3. Development of Electrical Synthesizers

The first full electrical synthesis device was introduced by Stewart in 1922 (Amezaga & Hajek, 2022) [13]. The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with the two lowest formants, but not any consonants or connected utterances. The same kind of synthesizer was made by Wagner [10]. The device consisted of four electrical resonators connected in parallel and it was excited by a buzzlike source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra. In 1932 Japanese researchers Obata and Teshima discovered the third formant in vowels [12]. The three first formants are generally considered to be enough for intelligible synthetic speech.

International Journal of Artificial Intelligence, vol. 10, no. 2, pp. 103-110, December 2023. **DOI: 10.36079/lamintang.ijai-01002.569**

The first device to be considered as a speech synthesizer was VODER (Voice Operating Demonstrator) introduced by Homer Dudley at New York World's Fair in 1939 [10] [13]. VODER was inspired by VOCODER (Voice Coder) developed at Bell Laboratories in the mid-thirties. The original VOCODER was a device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct the approximation of the original speech signal. The VODER consisted of a wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten bandpass filters whose output levels were controlled by fingers. It took considerable skill to play a sentence on the device. The speech quality and intelligibility were far from good but the potential for producing artificial speech was well demonstrated. The speech quality of VODER is demonstrated in the accompanying CD.

In the late 1970s and early 1980's, a considerably amount of commercial text-to-speech and speech synthesis products were introduced [13]. The first integrated circuit for speech synthesis was probably the Votrax chip which consisted of cascade formant synthesizer and simple low-pass smoothing circuits. In 1978, Richard Gagnon introduced an inexpensive Votrax-based Type-n-Talk system. Two years later, in 1980, Texas Instruments introduced linear prediction coding (LPC) based Speak-n-Spell synthesizer based on low-cost linear prediction synthesis chip (TMS-5100). It was used for an electronic reading aid for children and received quite considerable attention. In 1982, Street Electronics introduced Echo low-cost diphone synthesizer which was based on a newer version of the same chip as in Speak-n-Spell (TMS-5220). At the same time Speech Plus Inc. introduced the Prose-2000 text-to-speech system. A year later, first commercial versions of famous DECtalk and Infovox SA-101 synthesizer were introduced [13].

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. One of the methods applied recently in speech synthesis is hidden Markov models (HMM). HMMs have been applied to speech recognition from the late 1970s. For speech synthesis systems, it has been used for about two decades. A hidden Markov model is a collection of states connected by transitions with two sets of probabilities in each: a transition probability which provides the probability for taking this transition, and an output probability density function which defines the conditional probability of emitting each output symbol from a finite alphabet, given that the transition is taken [14] [15] [16].

According to [8], presents a theoretical framework on technologies used in speech synthesis. The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The two primary technologies generating synthetic speech waveforms are:

1. Concatenative synthesis
2. Formant synthesis

Each technology has strengths and weaknesses, and the intended uses of a synthesis system will typically determine which approach is used.

1. Concatenative synthesis

Is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output.

2. Formant synthesis

This does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components. Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech [15][17][18]. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High-speed synthesized speech is used by the visually impaired to quickly navigate computers using a screen reader. Formant synthesizers are usually smaller programs than concatenative systems because

they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice [15] [19]. Some milestone in the development of speech synthesis is as illustrated in Figure 2.

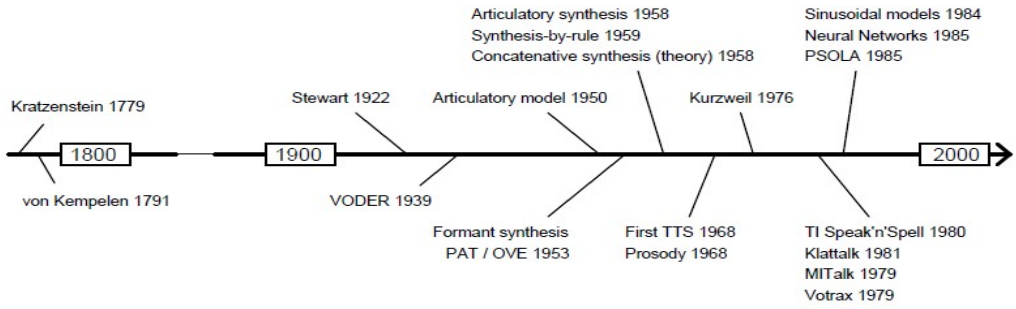

Figure 2. Some Milestones in Speech Synthesis.

3.4. Flowchart of a Speech Synthesizer

The flowchart of the working principles of speech synthesizer is as shown in Figure 3.

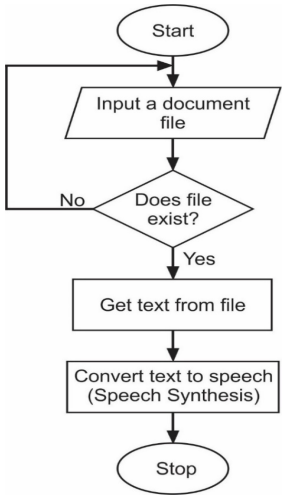

Figure 3. Flowchart of Speech Synthesizer

3.5. Block Diagram of Speech Synthesizer

The block diagram of the Speech Synthesizer and it step by step working procedures is as shown Figure 4.

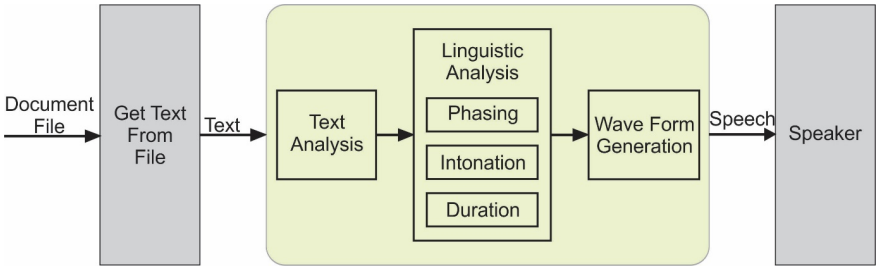

Figure 4. Block Diagram of the Speech Synthesizer

International Journal of Artificial Intelligence, vol. 10, no. 2, pp. 103-110, December 2023. **DOI: 10.36079/lamintang.ijai-01002.569**

1. Text analysis: consists of normalization of the text wherein the numbers and symbols become words and abbreviations are replaced by their whole words or phrases etc. The most challenging task in the text analysis block is the linguistic analysis which means syntactic and semantic analysis and aims at understanding the context of the text.
2. Phonetic Analysis converts the orthographical symbols into phonological ones using a phonetic alphabet. For e.g., the alphabet of the International Phonetic Association contains phoneme symbols, their diacritical marks and other symbols related to their pronunciation.
3. Prosody is a concept that contains the rhythm of speech, stress patterns and intonation. At the perceptual level, naturalness in speech is attributed to certain properties of the speech signal related to audible changes in pitch, loudness and syllabic length, collectively called prosody.
4. Speech Synthesis block finally generates the speech signal. This can be achieved either based on parametric representation, in which phoneme realizations are produced by machine, or by selecting speech units from a database. The resulting short units of speech are joined together to produce the final speech signal.

From the diagram of Figure 4, the main input that the system requires is a document file. The text content of this file is extracted using a suitable algorithm. The speech synthesis engine takes in the text as input and then pre-processes and analyzes it into phonetic representation which is usually a string of phonemes with some additional information for correct intonation, duration, and stress. A sound wave is then generated and passed to the computer's speaker. The speaker enables a user to hear the audio output.

The display interface of Voice-Based Document File Reader Window without GPT turbo series embedded is as shown in Figure 5. This interface has the ability to fetch text from a particular file and converts it to audio format but lacks the LLM technologies.

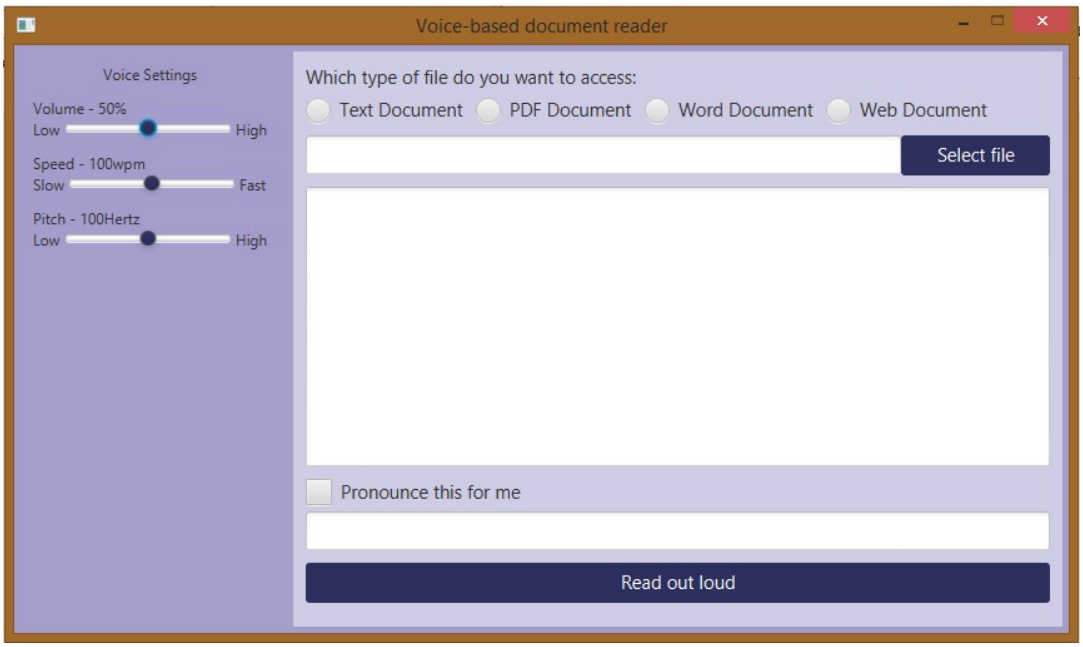

Figure 5. Voice-Based Document File Reader Window

Advantages of Cross Platform Document File Reader using Speech Synthesis

1. People with learning disabilities: Some people have difficulty in reading large amounts of text due to dyslexia and other learning disabilities. The development of this software has brought an ease to reading large amounts text and how is being pronounced.
2. People who have literacy difficulties: Some people have basic literary levels. They often get frustrated trying to read content because so much of it is in text form. With this speech synthesizer the time to read those contents are reduced to minimal.

3. People who multitask: A busy life often means that people do not have time to do all the reading. Having a chance to listen to the content instead of reading it allows them to have their work done or carried out simultaneously.
4. People with visual impairment: Speech synthesis can be a very useful tool for the mild or moderately visually impaired. Even for people with visual capability to read, the process can often cause too much strain.

4. Finding and Discussion

In search for the solution of some problems encountered by people with different challenges such as people with learning disabilities, literacy difficulties, visual impairment, and for those who multitask (The Busy bodies). Many Researchers have come up with their different views by employing different technologies at which document files can be read. This paper reviewed the sequential order of evolvement of text to speech from mechanical to electrical synthesis. The inability of other platform-dependent to universally run in every platform has limited its efficiency thereby giving cross-platform an edge as it runs in every platform.

5. Conclusion

Efforts have been made by many researchers on how to develop a technology that is capable of converting documents consist of text to an audio format as information is widely spread in a text formats. The conversion of text document to speech through the use of Artificial intelligence has advanced significantly as a result of text-to-speech synthesis. The inability of other platform-dependent to universally run in every platform has limited its efficiency thereby giving cross-platform an edge as it runs in every platform. Hence, among the existing cross-platform synthesis the GPT-3.5 and GPT-4 Turbo LLM are the best-developed AI technologies that not only convert text to a human-like voice but also translate texts to different languages.

References

[1] F. C. Ogenyi, H. Udeani, "Design and Implementation of a Cross Platform Document File Reader using Speech Synthesis," *Newport International Journal of Engineering and Physical Sciences*, vol. 3, no. 1, pp. 1-11, 2023.

[2] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5734-5738, June 2021.

[3] K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, M. Stella, "Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. Big Data Cogn," *Comput,* vol. 7, no. 3, pp. 124, 2023.

[4] S. Rautiainen, "A look at Portable Document Format vulnerabilities," *Information Security Technical Report,* vol. 14, no. 1, pp. 30-33, 2009.

[5] Microsoft Office File Formats. MSDN Library. Microsoft. [Accessed: February. 2, 2013].

[6] B. Popa. Foxit Reader. Softpedia. SoftNews. 2014.

[7] Foxit Software Foxit Reader 3.0. PC World. Archana. [Accessed: July 7, 2015].

[8] H. Ingo. *Open Life: The Philosophy of Open Source*. Lulu.com – via Google Books. 2006

[9] J. Sodnik and S. Tomažic, "Spatial speaker: 3D Java text-to-speech converter," *In Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2009.

[10] R. A. Cole. *Survey of the State of the Art in Human Language Technology*, 1996.

[11] J. Flanagan. *Speech Analysis, Synthesis, and Perception*. New York: Springer-Verlag, Berlin-Heidelberg, 1972.

[12] J. Flanagan, and L. Rabiner. *Speech Synthesis*. Pennsylvania: Dowden, Hutchinson & Ross, Inc., 1973.

[13] M. Schroeder, "A Brief History of Synthetic Speech," *Speech Communication*, vol. 13, pp. 231-237, 1993.

[14] D. Klatt, "Review of Text-to-Speech Conversion for English," *Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737-793, 1987.

[15] K.F. Lee. *Automatic Speech Recognition: The development of the SPHINX system*. Boston: Kluwer Academic publishers, 1989.
[16] N. Kaur, P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," *Artif Intell Review*, vol. 56, no. 7, pp. 5837–5880, July 2023.
[17] C. N. Ugwu and V. H. U. Eze, "Qualitative Research," *IDOSR of Computer and Applied Science*, vol. 8, no. 1, pp. 20–35, 2023.
[18] V. H. U. Eze, "Qualities and Characteristics of a Good Scientific Research Writing; Step-by-Step Approaches," *IAA Journal of Applied Sciences*, vol. 9, no. 2, pp. 71–76, 2023.
[19] C. N. Ugwu, V. H. U. Eze, J. N. Ugwu, F. C. Ogenyi, and O. P. Ugwu, "Ethical Publication Issues in the Collection and Analysis of Research Data," *Newport International Journal of Scientific and Experimental Sciences*, vol. 3, no. 2, pp. 132–140, 2023.