ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 3, pp: (65-69), Month: July - September 2023, Available at: www.researchpublish.com

uTranscribe: OCR and Voice-to-Text Transcriber using Long-Short Term Memory Algorithm

[1]Kyle Christian D. Hingpit, [2]Dan Andrew B. Mendoza, [3]Matthew Steven E. Ortiz, [4]Allistair Gerard R. Vinoya, [5]Jerian R. Peren

Affiliation: Lyceum of the Philippines University Cavite

DOI: https://doi.org/10.5281/zenodo.8199253

Published Date: 31-July-2023

Abstract: **In this study, the researchers focused on the design and development of "uTranscribe: An OCR and Voice-to-Text Transcriber using Long-Short Term Memory Algorithm" to address the existing gap in transcriber system availability. Additionally, the uTranscribe system was compared to other existing transcribing systems in terms of accuracy performance. The researchers followed the design science research methods for the research practice, incorporating the agile process for the system development. A comparative analysis of uTranscribe's accuracy was performed based on Word Error Rate (WER) metric, comparing its result to the Word Error Rate (WER) baseline standard of transcription systems. A descriptive survey method was employed to assess its acceptability based on ISO IEC 25010 criteria with the evaluation conducted by IT experts and End Users. The study findings indicate that uTranscribe's accuracy can be improve by increasing training time and incorporating additional datasets. Moreover, the study revealed positive result regarding the application's usability, demonstrating excellent user experience and ease of learning for respondents. However, the system needs to enhance its portability, particularly its adaptability to different platforms. Based on the conclusions drawn, respondents provided several recommendations, including improving the system's capability to recognize various accents in videos, incorporating a lasso tool for OCR snipping, implementing file validation during uploads, listing fonts recognized by the application, and making minor improvements to the user interface. Overall, this study contributes to the fields of OCR, Voice-to-Text technology, and supports the LSTM algorithm. It provides valuable insights that can guide future advancement in these technologies, with the end goal of enhancing the overall user experience when interacting with online content.**

Keywords: **OCR, Voice-to-Text Technology, LSTM Algorithm, ISO IEC 25010.**

I. INTRODUCTION

Optical Character Recognition (OCR) technology has been in development since the 1960s as a way to digitize printed text by converting it into machine-encoded text [1]. The inventor, Ray Kurzweil, developed the first OCR device in the 1970s, which was primarily used to digitize printed text [2]. Today, OCR technology is widely used in various industries, including finance, healthcare, and e-commerce. On the other hand, speech-to-text technology, also known as voice recognition, has a long history dating back to the 1950s and 1960s [3]. Early speech recognition systems were limited in their capabilities, only able to recognize a small number of words in a specific context. However, with the advent of machine learning and deep learning techniques, speech-to-text technology has advanced significantly, and is now widely used in various applications such as dictation software and speech-to-text transcription services, as well as in other applications such as voice commands systems [4].

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 3, pp: (65-69), Month: July - September 2023, Available at: www.researchpublish.com

Optical Character Recognition technology has been widely used in various industries for the past few decades. The technology enables the conversion of printed text into machine-encoded text which can be edited, searched, and stored digitally [5]. On the other hand, Speech-to-Text technology, also known as Automatic Speech Recognition (ASR), has been in development since the 1950s and has advanced significantly with the integration of machine learning algorithms The technology enables the conversion of spoken language into written text, which can be used for various applications such as voice commands systems, dictation software and speech-to-text transcription services [6]. It is also used in industries such as finance, healthcare, and customer service to automate processes and improve efficiency [7].

The current applications for OCR, such as Google Lens, Adobe Scan, Fine Reader, and others, as well as Speech-to-Text applications like Express Scribe, The FTW Transcriber, Verbit, and more, require users to download and install two separate apps to utilize both functionalities. Considering this, this study aims to develop an integrated application that combines both OCR and Speech-to-text capabilities, eliminating the need for multiple app installations.

The findings of this study are beneficial to students, teachers, general users, and future researchers. Firstly, students can utilize the developed system to study specific parts of a video lesson and tutorial, by way of text transcription, without spending excessive time. Secondly, teachers can save time and effort by extracting information from videos to enhance their teaching materials. The developed system also improves the overall experience and simplifies access to information for general users while watching videos. Lastly, future researchers with similar goals can use the study as a helpful resource for further improvements.

II. METHODOLOGY

A. System Architecture

Fig. 1: System Architecture of uTranscribe.

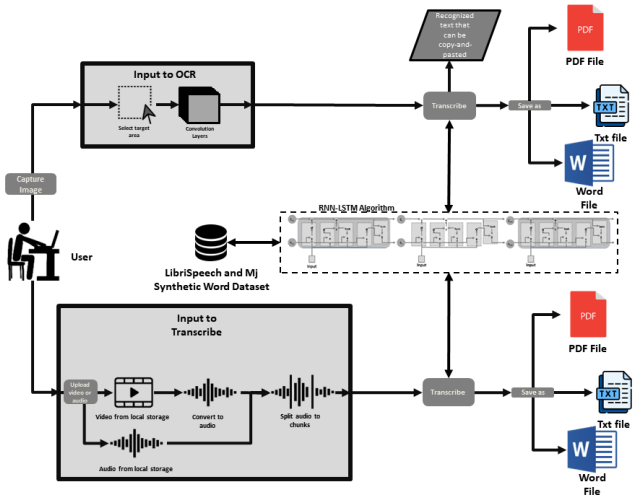

Figure 1 shows the system architecture of the study which consists of two paths for input processing. In the first path, the user interacts with the capture image button and selects a target area containing the desired text from the video. This selected area undergoes quick processing through a convolution layer to extract sequential feature representation. The extracted features are then fed into the LSTM algorithm, trained on the MJ Synthetic Word Dataset. The system alerts the user when letter/characters are recognized, allowing for easy copy-and-paste, and can also be saved into various file formats such as PDF, TXT, and Docx. The second path involves uploading a video or audio file, which is converted to audio format and split into chunks for faster processing. The LSTM algorithm, trained on the LibriSpeech dataset, processes these audio chunks. The resulting transcribed audio can be saved in various file formats, such as PDF, TXT, and Docx.

B. Software

To be able to transcribe the voice from a video or audio into text, the system must recognize the differences between words and accents. By training the LSTM algorithm using the LibriSpeech datasets, comprising approximately 1000 hours of English speech, the system can differentiate the spoken words that is uploaded into the system.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 3, pp: (65-69), Month: July - September 2023, Available at: www.researchpublish.com

Fig. 1I: Available datasets of LibriSpeech

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Figure II shows the available data in the LibriSpeech dataset, which are divided into training, development, and testing datasets. The training and the testing data are used for our system.

III. RESULTS

A. Result of Testing

Table I: Functionality Test Result

Test respondents	Pass	Fail	Test criteria	Percentage
Technical adviser	13	0	13	100.00%
IT expert 1	13	0	13	100.00%
IT expert 2	13	0	13	100.00%
IT expert 3	13	0	13	100.00%
IT expert 4	13	0	13	100.00%
IT expert 5	13	0	13	100.00%

Table 1 shows the result of Functionality Testing of the uTranscribe system. Based on the data, it can be concluded that during the functionality testing, none of the six (6) respondents encountered any errors or bugs in the system. All functions successfully executed the required inputs and outputs, demonstrating a high level of performance. As a result, the system received a total of 13 "pass" ratings and 0 "fail" ratings, providing compelling evidence that the system is not only functional but also highly acceptable in terms of its functionality performance.

TABLE II

System/ Research	Conditions	Input					Word Error Rate (WER)	Baseline (20%)	Remarks
		Type	Size	Length	Audio Bit rate	Audio Quality			
uTranscribe using LSTM	2 layer, 100 hidden states	.mp4	18.1MB	00:05:47	127kbps	Quiet, no background sfx, American accent	17%	<20%	Passed
		.mp4	45.7MB	00:16:33	127kbps	Noisy, a lot of background noise/sfx, mixed accents	44.30%	>20%	Failed
		.mp4	222MB	00:40:31	127kbps	Noisy, a lot of background noise, American accent	11%	<20%	Passed

Note: A Word Error Rate (WER) of less than 20% is considered acceptable, greater than 20% is considered Unacceptable. Datastore, 2023

Table 2 shows the result of Accuracy Performance test result. Based on the data, our system achieved a Word Error Rate (WER) below the baseline standard of 20% in two test cases. The 5-minute video achieved a WER of 17%, and the 40-minute video achieved a WER of 11%. However, the system failed to meet the standard for the 16-minute video, with a WER of 44.30%. This outcome can be attributed to the audio quality of the test video. The videos that passed the test only contained American Accents, while the video that failed had a mixture of multiple accents.

B. Results of Evaluation

Table III: Summary of End User's Evaluation by Iso Iec 25010 Criterion

ISO IEC Criterion	Mean	Verbal Interpretation
Functionality Suitability	4.60	Highly Acceptable
Performance Efficiency	4.4	Highly Acceptable
Compatibility	4.68	Highly Acceptable
Usability	4.55	Highly Acceptable
Reliability	4.49	Highly Acceptable
Integrity	4.65	Highly Acceptable
Portability	4.6	Highly Acceptable

Note: For interpretation, the following remarks apply to mean interval: 4.21 – 5.00 for Highly Acceptable, 3.41 – 4.20 for Moderately Acceptable, 2.61 – 3.40 for Acceptable, 1.80 – 2.60 for Fairly Acceptable, and 1.00 – 1.79 for Poorly Acceptable.

Table 3 shows a summary of the End User's Evaluation by ISO IEC 25010 Criteria. The table shows the average mean and verbal interpretation of the result for the End Users. The is a total of twenty (20) respondents who evaluated the system and gave their comments and suggestions.

All criterion received a Highly Acceptable interpretation. The table shows that Compatibility shows the highest mean of 4.68 and Performance Efficiency received the lowest mean with 4.4.

Table IV: Summary of It Experts' Evaluation by Iso Iec 25010 Criterion

ISO IEC Criterion	Mean	Verbal Interpretation
Functionality Suitability	4.53	Highly Acceptable
Performance Efficiency	4.27	Highly Acceptable
Compatibility	4.50	Highly Acceptable
Usability	4.53	Highly Acceptable
Reliability	4.70	Highly Acceptable
Integrity	4.20	Highly Acceptable
Portability	4.40	Highly Acceptable
Maintainability	4.40	Highly Acceptable

Note: For interpretation, the following remarks apply to mean interval: 4.21 – 5.00 for Highly Acceptable, 3.41 – 4.20 for Moderately Acceptable, 2.61 – 3.40 for Acceptable, 1.80 – 2.60 for Fairly Acceptable, and 1.00 – 1.79 for Poorly Acceptable.

Table 4 shows a summary of the IT Experts' Evaluation by ISO IEC 25010 Criteria. The table shows the average mean and verbal interpretation of the result for the IT Experts. The is a total of five (5) respondents who evaluated the system and gave their comments and suggestions.

All criterion received a Highly Acceptable interpretation. The table shows that Reliability shows the highest mean of 4.70 and Performance Efficiency received the lowest mean with 4.4.

IV. CONCLUSION

Based on the results of testing and evaluations, it can be concluded that the researchers successfully implemented the Long-Short Term Memory Algorithm in the "uTranscribe: OCR and Voice-to-Text Transcriber using Long-Short Term Memory Algorithm". The results suggest that the system's accuracy can be further enhanced by increasing training duration and incorporating additional dataset with diverse English accents, on top of the LibriSpeech dataset.

Feedback from the evaluation respondents highlighted the need to improve the system's capability to recognize different accents, add an OCR snipping lasso tool feature, implement file validation for uploads, provide a list of recognized fonts, and minor improvements to the User Interface.

Overall, this study contributes to the fields of OCR, Voice-to-Text technology, and supports the LSTM Algorithm. It provides valuable resources that can guide future research in these technologies, with the end goal of enhancing overall user experience when interacting with video contents.

REFERENCES

[1] Darko, "The History of Optical Character Recognition," GorillaPDF Blog. https://gorillapdf.com/blog/the-history-of-optical-character-recognition/

[2] "What Is Optical Character Recognition (OCR)?," IBM, Feb. 2022, [Online]. Available: https://www.ibm.com/cloud/blog/optical-character-recognition

[3] "A brief history of speech recognition | Sonix," Sonix. https://sonix.ai/history-of-speech-recognition

[4] "The History of Voice Recognition Technology," Blog, Apr. 2022, [Online]. Available: https://www.condecosoftware.com/blog/a-history-of-voice-recognition-technology/

[5] S. Khurana, "Applications of OCR You Haven't Thought Of - The Startup - Medium," Medium, Apr. 28, 2018. [Online]. Available: https://medium.com/swlh/applications-of-ocr-you-havent-thought-of-69a6a559874b

[6] "What is Speech to Text? - Speech to Text Transcription Explained - AWS," Amazon Web Services, Inc. https://aws.amazon.com/what-is/speech-to-text/

[7] Eric-Urban, "Speech to text overview - Speech service - Azure Cognitive Services," Microsoft Learn, May 12, 2023. https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/Speech-to-Text