A COMPREHENSIVE REVIEW OF SPEECH-TO-TEXT TECHNOLOGIES: INCORPORATING TRANSLATION, SUMMARIZATION, AND BEYOND

Alen Thomas, Rosu J Edanad, Sourav J Raju, Suchitra NT, Syeatha Merlin Thampy
Department of Computer Science
College Of Engineering Chengannur
Kerala, India
thomasalen634@gmail.com, rosujedanad@gmail.com, souravjraju333@gmail.com, suchitra.nt2003@gmail.com

***Abstract*—Speech-to-text technologies have transformed the way humans interact with computers by allowing spoken language to be transcribed effortlessly into written form. In addition to transcribing, incorporating translation and summarization features boosts the accessibility and usefulness of speech-to-text systems in various languages and fields. This paper investigates the changing STT technology landscape, focusing on its role in live communication, processing multiple languages, and creating content. The paper emphasizes the possibility of advanced applications in education, business, and accessibility services by tackling issues such as language diversity, and context awareness.**

***Index Terms*—Speech-to-Text (STT) Conversion, Language Translation Technologies, Text Summarization Tools, Accessibility in Digital Content, Remote Work and Online Learning Technologies, Text-to-Braille and Text-to-Speech (TTS)**

I. Introduction

The growth of digital content, online communication, and multimedia applications has led to a significant increase in the demand for technologies that enhance accessibility, interactivity, and comprehension. Efficient processing and conversion between speech, text, and other formats are crucial in fields such as business, education, healthcare, and entertainment. The demand for tools capable of managing large volumes of spoken or written data has risen sharply due to the rapid adoption of online learning, virtual meetings, and remote work. Speech-to-text, language translation, summarization, text-to-speech (TTS), and text-to-Braille generation are among the technologies that are increasingly vital for promoting inclusivity and boosting productivity, catering to the needs of diverse users including professionals dealing with complex information, non-native speakers, and individuals with disabilities.

A. Technological Components of Speech and Text Processing

1. Speech-to-Text (STT) Conversion: Conversion tools for speech-to-text transform spoken words into written text, allowing for instant transcription in various industries. Precise transcription ensures that information is retained in virtual meetings, online lectures, and webinars, enabling participants to refer back to discussions and notes at their convenience. The foundation of STT systems lies in Automatic Speech Recognition (ASR) models, which utilize deep learning techniques and extensive voice datasets to achieve high accuracy, even in noisy environments. However, challenges persist due to the complexity of spoken language, including accents, spontaneous speech patterns, and multiple speakers. The significant advancements in STT technology have led to improved access to digital information and enhanced human-computer interaction.

2. Language Translation: The use of language translation technology has become crucial for enabling cross-cultural communication in an increasingly globalized world. These tools facilitate the translation of written or spoken words between languages, making content understandable for individuals who are not native speakers. Real-time translation capabilities are particularly beneficial in multilingual virtual meetings, online courses, and international conferences as they enable smooth communication among participants with varying language backgrounds. These systems rely on machine translation and natural language processing (NLP) models, with translation services being accessible through APIs such as Google Translate. Despite the advancements made, challenges persist in maintaining contextual accuracy, handling linguistic complexities, and supporting languages with limited resources.

3. Summarization: Due to the large volume of digital content generated daily, it has become essential to use summarization tools to efficiently extract pertinent information. There are two main types of summarization methods: extractive and abstractive. Extractive methods identify important sentences from the original text, while abstractive methods use natural language to craft concise summaries. These tools are commonly employed for document summarization, generating lecture notes, and producing meeting minutes. Integrating speech recognition with summarization further enhances productivity by automating processes such as creating summaries from webinars or virtual meetings. However, overcoming transcription errors from spoken input and ensuring that summaries capture the key content remain challenges in this field.

4. Text-to-Speech (TTS) Conversion: Users with visual impairments or those who prefer audio formats can access digital content thanks to text-to-speech technologies, which convert

printed text into artificial speech. Screen readers, virtual assistants, audiobooks, and voice-activated gadgets all rely on text-to-speech (TTS) technologies to provide consumers with more options for how they consume information. Deep neural networks are used by modern TTS systems to generate voices that seem genuine and may change in pitch, pace, and tone in response to context or user choices. These technologies enable people with impairments to easily engage with technology, which is a critical component in enhancing accessibility and inclusivity.

5. Text-to-Braille Generation: Converting text to Braille ensures that individuals with visual impairments can access information independently. The technology for converting text to Braille enables people with visual impairments to engage with materials such as books, documents, and websites. These technologies are essential in public services, education, and professional environments. Although advancements in machine learning and natural language processing have accelerated and improved the quality of Braille conversion, challenges persist in adapting Braille layouts to different languages and accurately rendering complex content, such as mathematical notations.

II. Literature Review

Kazumasa Yamamoto et al. (2023) [1] devised a system for automated speech recognition, translation, and summarization of TED English lectures, with the objective of improving understanding for Japanese-speaking individuals. The system comprises three fundamental elements: DNN-HMM for speech recognition, a Transformer model for translation, and BERT-based summarization (BertSumExt) for the extraction of salient sentences. In the domain of speech recognition, the investigators attained a word accuracy of approximately 88% through the utilization of a combination of TED and Librispeech audio corpora. The Transformer-based translation system exhibited diminished efficacy with respect to speech inputs relative to text, yielding BLEU scores that were approximately 14% lower when processing recognized speech due to errors in recognition. Nonetheless, speech summarization demonstrated resilience to such inaccuracies, as the results of significant sentence extraction closely mirrored those of the original text summarization. This methodology offers effective instruments for subtitling English lectures accompanied by Japanese translations, emphasizing readability while ensuring robust summarization even amidst recognition challenges.

Nitesh Bharti et al. (2021) [2] devised a video-to-text summarization system with the objective of producing summaries derived from online meetings, webinars, and academic lectures. The system proficiently extracts audio components from video files, subsequently transcribing the audio into textual format utilizing Python-based methodologies, and employs text summarization algorithms to formulate concise notes. Furthermore, users are afforded the capability to revert the textual summaries back into audio form, thereby facilitating a more accessible mode of interaction with the content. This system, implemented with libraries such as SpeechRecognition, MoviePy, and spaCy, underwent evaluation utilizing YouTube videos. Although an established benchmark for accuracy is lacking, the outcomes were manually validated, demonstrating promising results when the audio quality was optimal, characterized by minimal background noise. This methodology possesses extensive applicability, including the generation of meeting minutes, lecture notes, and subtitles, thereby addressing the increasing demand for efficient content summarization in virtual environments.

Zhara Nabila et al. (2022) [3] undertook the development of a translation application leveraging the Google Translate API, intended to enhance accessibility and facilitate understanding of languages across multiple media formats. This application is designed to extract audio or textual content from YouTube videos and subsequently translate it into multiple languages, employing the Google Translate and Text-to-Speech (TTS) libraries. The system was engineered in Python and incorporates deep learning, machine translation, and natural language processing (NLP) elements to ensure the provision of precise outcomes. The primary objective of the translation tool is to aid users in engaging with content in foreign languages, particularly emphasizing educational applications, public outreach, and assistance for individuals with disabilities. During the evaluation phase, the application successfully translated 90.38% of videos into both text and audio formats, achieving a synchronization accuracy ranging from 89% to 97% between the produced text and audio outputs. The main limitations identified were instances where videos were subject to restricted public access, thereby obstructing translation efforts.

Prerna Mishra et al. (2023) [4] discuss research aimed at understanding the dynamics of video content and producing video summaries automatically, using machine learning techniques and Natural Language Processing (NLP) techniques. In the first stage, it implies transforming the video files into audio and converting the audio file into text. The converted text is further summarized using extractive and abstractive approaches. While the extractive approach takes some important lines from the text, in the case of the abstract approach, entirely new sentences pertaining to the context are made. The model uses the architecture of the BART transformer model from Facebook and adds NER to tag the summaries with relevant ids like person, place, and organizations, etc. Along with this, the use of pre-trained NLP models, MoviePy, and the speech recognition API from Google enables the model to handle the work on a corpus of videos, which ranges from news and speeches to entertainment, in both Hindi and English. The findings reveal that the model is capable of condensing the video contents and tagging the videos accordingly, which improves the video contents organization systems.

Atluri Naga Sai Sri Vybhavi et al.(2022) [5] discuss that in a system to summarize video transcripts, particularly from YouTube, using advanced NLP and Machine Learning techniques. This system is designed as per video content on platforms like YouTube grows rapidly, extracting key information from lengthy videos becomes challenging. This system addresses the issue by automatically retrieving video transcripts and generating concise summaries using tools like Hugging Face Transformers and pipelining, effectively condensing the text while preserving essential details. Although this growing content provides valuable educational and informational resources it presents difficulties in information extraction. The methods like Latent Semantic Analysis (LSA) can generate summaries without extensive computational power or large datasets. The system's workflow involves retrieving transcripts, processing them, and producing concise summaries using Hugging Face Transformers in Python. The summarization process employs pipelining to break down videos into smaller segments, extracting key content from each. Both abstractive and extractive summarization methods ensure that the most important information is retained. A user-friendly interface allows users to easily input video links and receive summarized content. This system offers an effective solution for summarizing YouTube video transcripts, reducing transcript length while maintaining accuracy and clarity. By using pipelining, the system efficiently processes transcripts, reducing them by about 70%, while retaining key ideas by eliminating unnecessary or repetitive phrases the approach is simple yet powerful, leveraging state-of-the-art NLP models to deliver concise, meaningful summaries. Future work could expand this research by including videos from other platforms and supporting multiple languages.

Golia and Kalita (2023) [6] presented a methodology that significantly improves the summarization of meeting transcripts through a concentrated focus on action items, thereby addressing the challenge of capturing long-term dependencies within transformer-based models. They introduced a recursive summarization algorithm, integrated with three distinct topic segmentation methodologies—namely chunked linear, simple cosine, and complex cosine segmentation—to systematically partition lengthy meeting transcripts into coherent sections. By employing BART for abstractive summarization and fine-tuning this model utilizing the XSUM and SAMSUM datasets, the proposed approach yielded coherent summaries concurrently, thereby ensuring contextual coherence and enhancing temporal efficiency. Their action-item extraction model, which underwent fine-tuning on a publicly available dataset, achieved a classification accuracy of 95.4%, thereby augmenting the quality of the summaries generated. In a comparative analysis against leading linear segmentation techniques, their chunked linear segmentation exhibited superior performance, achieving a 1.36% enhancement in BERTScore, while the overall model realized a 4.98% improvement over prior benchmarks. In spite of its achievements, the methodology encountered challenges including the management of redundant dialogue components and the necessity of providing meaningful context for the extracted action items.

Rajakumar B. et al. (2024) [7] introduced an innovative adaptive Speech-to-Text (STT) conversion system specifically designed for challenging noisy environments, aiming to achieve high accuracy in transcribing spoken language into written form. This system utilizes Python libraries such as SpeechRecognition and PyDub to effectively process and convert audio inputs into text, thereby serving various applications including voice-activated assistants, transcription services, and support tools for individuals with hearing disabilities. The system is capable of detecting spoken words and converting them into text in real time, while also accommodating a range of dialects and accents. Training was conducted using diverse datasets, including LibriSpeech, TED-LIUM, and Mozilla's Common Voice, enabling the model to manage a wide array of linguistic variations. Its architecture features audio preprocessing via PyDub, feature extraction through librosa, and speech recognition facilitated by multiple STT engines, such as the Google Web Speech API. The integration of sophisticated machine learning models, particularly deep learning methodologies, enhances transcription accuracy, with a particular emphasis on performance in noisy settings. Nevertheless, the system still faces challenges in effectively processing certain dialects and maintaining real-time performance.

Tyagi et al. (2023) [8] introduced a video summarization instrument that employs automatic speech recognition (ASR) alongside extractive text summarization methodologies to mitigate the burgeoning demand for succinct summaries of lengthy videos, especially within the contexts of education and professional environments. The apparatus leverages a convolutional neural network (CNN) for the conversion of speech into text, subsequently implementing extractive summarization methods to derive summaries from the transcribed material. The ASR model underwent training utilizing the LJSpeech dataset in conjunction with customized audio data, attaining a remarkably low word error rate (WER) of 0.10, with a model that demonstrates adaptability to various accents through the incorporation of pre-trained models provided by Google, Amazon, and Facebook. The text summarization phase utilized the Textrank algorithm; in instances where transcripts lacked punctuation, the Facebook/BART-large-CNN model was employed. The findings illustrated that this hybrid methodology proficiently generated precise, language-agnostic summaries while effectively addressing the complexities associated with spoken accents and the punctuation of transcripts. Furthermore, the tool proves advantageous for individuals with disabilities as well as professionals requiring expedited insights from video content.

Yi Ren et al. (2022) [9] present two models, FastSpeech 2 and FastSpeech 2s, which aim at enhancing the efficiency

as well as the performance of text-to-speech (TTS) systems. FastSpeech 2 resolves the shortcomings of the primary FastSpeech design by bunching the training process, eliminating the teacher-student distillation stage, and using ground-truth mel-spectrograms for training, which leads to better quality output with less information getting lost in the process. Furthermore, it uses more precise speech variation features specifically pitch, energy, and duration which are introduced via a variance adaptor in order to deal with the many-to-one mapping problem, thus improving the naturalness and expressiveness of the speech synthesizer. FastSpeech 2s builds on these refinements by performing text-to-speech directly into waveforms, thus skipping the mel-spectrogram part, achieving fully end-to-end synthesis even at faster inference speed. Both models provide significant advantages over earlier autoregressive and non-autoregressive TTS models, providing a threefold advantage in terms of training and faster quality speech synthesis at inference, thus making them revolutionary systems in the domain of fast and high-performance TTS systems.

Halitha Banu H et al. (2022) [10] discusses a system that is able to convert text into braille and audio based on the Speech Application Programming Interface (SAPI). The paper addresses the problem of accessing electronic and digital materials by visually impaired persons and provides a solution that transforms simple text into the braille system, enabling them to read through their fingers. This paper is an extension of previous works, including systems developed by Blenkhorn et al. for brailling word-processed documents, and by Singh et al. for the Shree Braille System, which converts English-Hindi Braille text with better efficiency. It also gives a brief overview of how normal English braille is coded in ASCII for easier and faster transferring of braille data over wire. The system is designed with a text-to-braille engine that accepts unformatted text and translates it into braille representation. Furthermore, the paper highlights the importance of audio output by utilizing SAPI so that the user can hear the converted braille text. This system demonstrates improvements in assistive devices by offering a way for blind persons to consume digital content.

III. Conclusion

The rapid growth of digital content, alongside the increasing demand for features that enhance interactivity, usability, and comprehension, has driven the advancement of assistive communication technologies such as speech-to-text (STT), translation, text summarization, text-to-speech (TTS), and text-to-Braille conversion. These technologies play a critical role in making digital information more accessible to a diverse range of users, including non-native speakers, professionals, and individuals with disabilities. Significant progress has been made in accurately transcribing spoken language, while translation tools have enhanced cross-cultural communication. Summarization techniques allow for more efficient handling of the vast amounts of digital information generated daily, and TTS and Braille conversion have greatly improved accessibility for those with visual impairments. However, challenges such as multilingualism, transcription accuracy, and content complexity remain. Addressing these issues in the next generation of information systems, through advancements in machine learning and natural language processing, will be essential for further expanding the reach and inclusivity of digital content in our increasingly interconnected world.

References

[1] K. Yamamoto, H. Banno, H. Sakurai, T. Adachi and S. Nakagawa, "A Study of Speech Recognition, Speech Translation, and Speech Summarization of TED English Lectures," 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 2023, pp. 451-452, doi: 10.1109/GCCE59613.2023.10315471.

[2] N. Bharti, S. N. Hashmi and V. M. Manikandan, "An Approach for Audio/Text Summary Generation from Webinars/Online Meetings," 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN), Lima, Peru, 2021, pp. 6-10, doi: 10.1109/CICN51697.2021.9574684.

[3] Nabila Z, Ayu HR, Surtono A. Implementation of Google Translate Application Programming Interface (API) as a Text and Audio Translator. Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi.;8(1):19-23.

[4] Mishra, Prerna , Garg, Kartik, Rathi, Naveen. (2023). Video-to-Text Summarization using Natural Language Processing. International Journal of Advanced Research in Science, Communication and Technology. 3. 2581-9429. 10.48175/IJARSCT-9160.

[5] A. N. S. S. Vybhavi, L. V. Saroja, J. Duvvuru and J. Bayana, "Video Transcript Summarizer," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 461-465, doi: 10.1109/MECON53876.2022.9751991.

[6] Golia L, Kalita J. Action-Item-Driven Summarization of Long Meeting Transcripts. InProceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval 2023 Dec 15 (pp. 91-98).

[7] B, Raja. (2024). Adaptive Speech-to-Text Conversion for Noisy Environments. 2455-6211.

[8] T. Tyagi, L. Dhari, Y. Nigam and R. Nagpal, "Video Summarization using Speech Recognition and Text Summarization," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-7, doi: 10.1109/INCET57972.2023.10169901.

[9] Ren Y, Hu C, Tan X, Qin T, Zhao S, Zhao Z, Liu TY. Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558. 2020 Jun 8.

[10] Halitha Banu H. Conversion Of Text to Braille and SAPI Based Audio Generation for Visually Impaired Peoples. JOURNAL OF ALGEBRAIC STATISTICS. 2022 Jun 1;13(2):1484-8.