Enhanced Speech Recognition & Abstractive Text Summarisation with Wav2vec2 & Pegasus

TANISHQUE SHARMA[1], SUJOY MONDOL[2], SANDEEP KUMAR[3]
[1, 2, 3] *Department of Computer Science and Engineering, Sharda University, Greater Noida, India*

Abstract: In this study, Python machine learning techniques are used to present a comprehensive method for integrating text summarization with speech recognition. It leverages the cuttingedge Wav2Vec2 algorithm for accurate voice recognition, as well as the pre-trained Pegasus model for concise and informative text summaries. This research aims to develop an integrated model that can efficiently translate spoken language into written text and provide brief, logical summaries. This combination of technologies makes data analysis and understanding more efficient by addressing the increasing need to analyze and extract knowledge from large volumes of spoken content. As a result of the combination of these two potent machine learning methods, transcriptions are guaranteed to be accurate, as well as knowledge extraction is accelerated. Utilizing Wav2Vec2's spoken language handling capabilities and Pegasus's text summarization expertise, the proposed method bridges the gap between oral and written communication. A variety of applications are possible, such as transcribing speeches and interviews and condensing lengthy audio files.

Indexed Terms— Speech recognition, text summarization, machine learning, Wav2Vec2, Pegasus model, Python programming, data processing, knowledge extraction, audio-totext conversion, information retrieval, natural language processing.

I. INTRODUCTION

In the digital age, where voice assistants respond to our every command and audio content is becoming increasingly prevalent, the intersection of two pivotal technologies, speech recognition and text summarization, holds the promise of reshaping the way we interact with spoken information. These technologies, in their own right, serve as indispensable tools in our daily lives, simplifying tasks from transcribing interviews to condensing lengthy speeches for better comprehension. Our research seeks to unite these technological superheroes and explore how their synergy can revolutionize the accessibility and comprehensibility of spoken content.

At the heart of our exploration lies the remarkable Wav2Vec2, a sophisticated algorithm that possesses the power to transform spoken language into written text. This feat is no small achievement, for it brings the spoken word within the realm of machines, opening the door to a plethora of applications where audio information can be harnessed for various purposes.

But the journey doesn't stop at mere transcription. Text summarization, our second protagonist in this narrative, steps onto the stage to distil lengthy content into shorter, more digestible forms. Imagine a long, winding speech or a verbose lecture being succinctly summarized without losing its essence. Pegasus, our virtuoso of summarization, takes on this task with finesse.

Before we dive into the mechanics of this research, we first embarked on a comprehensive literature review. We perused the pages of research papers that came before us, gaining insight into various facets of text summarization and speech recognition. From methods for generating summaries to multilingual text-to-speech conversion, we explored diverse approaches and technologies. We also uncovered the strengths and weaknesses of existing research, giving us a solid foundation to build upon.

With this knowledge in hand, we moved forward with our methodology. Our research model comprises two central components: the Wav2Vec2-powered speech recognition module and the Pegasus-driven text summarization module. The journey begins with the assembly of a substantial dataset of spoken language, meticulously pre-processed to ensure data quality. Wav2Vec2 steps in, carefully transcribing spoken words into text, while Pegasus takes the baton to weave these transcriptions into concise and coherent summaries.

This paper offers a thorough analysis of the difficulties and possible uses of automated speech summarization. It draws attention to the shortcomings of existing methods, including validation in a range of real-world scenarios and generalisation to different languages, accents, and content kinds, and the need over improvements in this field of study to address these issues. Subsequently, the paper presents novel methods for abstractive descriptions of meeting transcripts and lengthy consumer videos. The attentional encoder-decoder RNN technique, which has shown success in machine translation, is used in the suggested method to frame text summarization as a sequence-to-sequence problem. In order to achieve even greater performance gains, the research also adds new features to the baseline architecture.

This paper's introduction offers a thorough summary of the difficulties and possible uses of artificial speech summarization. In many real-world applications, like meeting transcription, news summary, and video captioning, the capacity to automatically translate spoken words into written text and produce succinct, logical summaries is a crucial task. However, there are a number of issues with present methods, including as the necessity for validation in a variety of real-world scenarios and generalisation to other languages, accents, and content kinds.

This research suggests novel approaches to summarise lengthy consumer movies and produce abstractive summaries of meeting transcripts in order to overcome these difficulties. The attentional encoder-decoder RNN technique, which has shown effectiveness in machine translation, is used in the suggested way to frame text summarization as a sequence-to sequence problem. In order to achieve even greater performance gains, the research also adds new features to the baseline design.

The suggested approach makes use of Wav2Vec2 and Pegasus, two cutting-edge models, for data analysis and comprehension. Whereas Pegasus is a pre-trained transformer-based model that can produce abstractive summaries of text, Wav2Vec2 is a pre-trained speech recognition model that can effectively translate spoken language into written text. By merging these two models, the suggested approach can produce precise and educational spoken language summaries while bridging the gap between oral and written communication.

The ability to extract and understand spoken language automatically is a key technological achievement in the digital world we live in, because spoken language is omnipresent and constantly expanding. In order to shed light on the current problems that call for creative solutions, this paper undertakes a thorough investigation of the nuances and possible uses of automated speech summarization.

Examining the present approaches' applicability in a range of real-world scenarios reveals their shortcomings, such as the difficult challenge of generalising to different languages, accents, and content kinds and the critical necessity for validation in a variety of contexts.

Given the wide range of applications that depend on the capacity to convert spoken words into written text and produce succinct, logical summaries, it is clear that this discipline needs to undergo a revolutionary shift. The need for a reliable and flexible automated speech summarising system is growing, as seen by the critical tasks of meeting transcription, news content summarization, and captioning lengthy video content.

This research addresses these issues by introducing novel techniques that are customised to target particular pain spots, with a particular emphasis on the complex domains of meeting transcripts and long-form consumer films. The attentional encoder-decoder recurrent neural network (RNN) approach, which has shown success in machine translation, is the methodology that has been selected. The suggested approach aims to improve the effectiveness and versatility of voice summary by presenting text summarization as a sequence-to-sequence problem.

However, creativity doesn't end there. Understanding how critical it is to push the envelope of performance, the research adds further features to the baseline architecture in an effort to attain not just small but significant gains. This attention to improving and extending the capabilities of the suggested approach highlights the commitment to getting over the present constraints in the field. Two powerful models that are essential to this research are Wav2Vec2 and Pegasus. Prominent speech recognition model Wav2Vec2 skillfully and accurately converts spoken language into written text.

However, the art of creating abstractive summaries from textual input is something that the transformer-based model Pegasus, which has been pre-trained, excels at. By combining these state-of-the-art models, the suggested method not only guarantees accurate speech-to-text conversion but also makes it easier to create illuminating and logical summaries, closing the gap between spoken and written communication.

Essentially, this study presents a comprehensive strategy that integrates speech recognition with text summarising using cutting-edge machine learning algorithms, rather than merely suggesting a technique. Beyond the domain of theoretical developments, the proposed technology heralds a new era of practical applications in a variety of industries, with the potential to completely transform the way that humans perceive, interpret, and process spoken language. The potential impact is significant and goes well beyond the boundaries of research into practical sectors where effective spoken language understanding is critical. All things considered, this work offers a thorough approach that combines speech recognition with text summarization using state-of-the-art machine learning algorithms. The suggested technology offers up new possibilities for real-world applications in multiple sectors and has the potential to transform how humans perceive and comprehend spoken language.

II. LITERATURE SURVEY

One of the most difficult tasks in information retrieval is removing important information from lengthy text texts. The work [1] presents a text summarising technique in order to address this. The sophisticated summarising strategies that preserve context and material flow are the major emphasis of this study. The method creates structured and logical content summaries by grouping sentences and documents. Nevertheless, managing intricate grammatical subtleties, less variation in summaries, and validation across different text genres are possible drawbacks. Translating multilingual literature into words or phrases and determining the content's toxicity are the goals of this study. By investigating multilingual text-to-speech and speech-to-text conversion, the research [2] promotes efficient communication across languages and modalities. Managing a variety of accents, linguistic subtleties, and validation in a range of speech settings and languages are challenges. For anyone interested in current developments in extractive text summarization, this article is a useful resource. It offers [3] information on the newest methods and possible uses for them. But given how quickly the area is developing, it may become out of date, thus further in-depth research into certain strategies is required. This research [4] investigates speech-to-text conversion and text summarization, providing information on both facets of natural language processing. The study is useful in situations when oral information accessibility and rapid understanding are critical. Managing a variety of dialects, faithfully recording subtle speech patterns, and validating across many speech settings and text genres are among the challenges. This model converts real-time speech to [5] text, generates a summary using Natural Language Generation and Abstract Meaning Representation, and then converts the summary back to speech. It helps users who need succinct summaries and real-time speech conversion for efficient material comprehension. Managing intricate linguistic subtleties, guaranteeing real-time performance, and strong validation across a variety of content kinds and languages are among the difficulties. This study offers an impressive examination of a unified method for Telugu [6], Hindi, and English language identification and speech recognition. Its uses are related to precise language identification and recognition, which enhances communication and facilitates the retrieval of information. In order to improve language processing efficiency, the study offers a performance analysis using combined MFCC and GMM-HMM approaches. But it has trouble interpreting different accents, generalising to other languages, and has to be validated in a variety of speech settings and environments. This paper introduces an innovative approach that combines speech [7] This paper introduces an innovative approach that combines speech recognition and text summarization using Transformer-based models. This approach directly generates concise and coherent text summaries from spoken content. It aligns with the growing demand for streamlined summarization techniques to process speech efficiently for various applications. Nonetheless, it faces challenges in handling diverse accents, understanding nuanced emotions, and requires robust evaluation across varied speech contexts. Converting spoken [8] material into written form and succinct summaries is made easier using automated speech-to-text summarization. Time is saved, and accessibility is improved for a range of users. However, because of speaker diversity and contextual difficulties, there is a chance of mistakes, misreading of intricate ideas, and loss of nuance. In order to enhance the quality of subtitles, the study investigates rating systems [9] for word substitution. It assists in determining

Wordnet synonym interchangeability by taking distributional similarity into consideration. WordNet and distributional data can be used to improve the performance of lexical substitution models; however, further study is needed to properly include contextual information. An detailed overview of research publications [10] concentrating on domain-specific voice recognition is provided by this literature study. The main developments, approaches, and difficulties in domain voice recognition are highlighted in this study. However, it may have gaps in covering very recent developments and potential biases in the selection of reviewed advancements [11] This work presents FILENG, an automated Hidden Markov Model (HMM)-based English subtitle generator for Filipino video clips. The purpose of creating subtitles is to improve accessibility to Filipino video material for English-speaking viewers and to close the language gap. In a variety of video scenarios, validation is crucial for future advancements. The study presents methods for automatically condensing impromptu speech. It tackles the particular [12] subtleties and complexity of this field, improving understanding and communication across a range of applications. But it has trouble with differentiating dialects, interpreting subtle emotions, and obtaining accurate real-time summary,[13] This work uses sequence-to-sequence models to provide a unique method for neutral abstractive text summarization. The method seeks to produce impartial and balanced content summaries, however it has trouble capturing but it faces challenges in capturing context-specific nuances and requires further evaluation across various text genres. The paper focuses [14] on evaluating scoring methods for word substitutions in subtitle generation, enhancing the quality of subtitles through advanced linguistic analysis. However, it may overlook certain contextual challenges and requires further research for validation across different contexts. This paper introduces [15] a pioneering approach to speech summarization by using extensive text datasets. This approach enhances end-to-end summarization and provides insights for efficient and comprehensive summarization.

However, challenges exist in adapting textbased methods to spoken language nuances and require rigorous evaluation across different speech genres and contexts. The paper introduces a novel approach to speechto-text summarization by incorporating multiple [16] sources. This enhances the comprehensiveness and accuracy of generated summaries. Challenges include source selection, potential information redundancy, and the need for robust evaluation across diverse speech contexts. The paper presents an automated method for [17] summarizing audio data. It offers efficiency in processing and extracting key insights from audio content. Challenges include accurately capturing contextual nuances, handling variable audio quality, and requiring robust validation across diverse audio sources. The paper introduces an innovative approach to deep text [18] summarization using Generative Adversarial Networks (GANs) in Indian languages. It addresses the challenge of generating concise and coherent summaries, but it faces challenges in adapting GANs to linguistic nuances, data availability, and requires robust evaluation across various Indian languages.

This paper introduces a comprehensive technique [19] for automatic subtitle generation and video summarization using an ensemble of NLP-driven methods. It enhances content accessibility and comprehension. However, it faces complexities in adapting NLP models to various languages and content genres and requires rigorous evaluation across diverse multimedia sources. The paper introduces a pioneering approach towards [20] end-to-end speech-to-text summarization, providing a streamlined method for generating concise summaries directly from spoken content. However, it faces challenges in handling diverse accents, background noise, and requires robust evaluation across various speech contexts and genres. This study explores enhancing [21] speech-to-text summarization by incorporating supplementary information sources. It aims to improve the performance of these technologies. Challenges include source integration, potential biases, and the need for validation across diverse speech contexts and information types, [22] The paper assesses enhanced components of the AMIS Project, covering speech recognition, machine translation, and summarization, contributing to the advancement of these technologies. However, challenges include generalization to various languages, accents, and content types, and the need for validation in diverse realworld contexts ,[23] The paper provides an overview of the latest advancements in automatic speech summarization, offering insights into the current state-of-the-art techniques and their potential applications. However, it may have potential limitations due to the rapid evolution of the field and requires deeper exploration of specific challenges, [24] This work introduces innovative techniques for summarizing long consumer videos. It focuses on identifying interesting. The research contributes to the advancement of text summarization techniques by introducing a method that can capture contextual nuances and generate summaries that retain the essence of the original text [25] In this study, text summarization is framed as a sequence-to-sequence problem, utilizing the attentional encoder-decoder RNN approach proven successful in Machine Translation. The research also introduces supplementary enhancements to the standard architecture, leading to additional performance improvements. By adopting this framework, the paper addresses the challenge of creating concise and informative summaries from extensive text content. However, potential limitations include challenges in handling long documents, avoiding over-simplification, and the need for validation across various text genres [26] This paper focuses on generating abstractive summaries of meeting transcripts. Abstractive summarization involves creating concise and coherent summaries by rewriting and generating new content. The research explores hierarchical adaptive learning techniques to enhance the quality of meeting summaries. Addressing [27] the challenge of summarizing lengthy documents, this paper delves into the domain of extractive summarization. Extractive summarization involves selecting and extracting important sentences or passages from a document to create a summary. The paper discusses techniques and challenges related to summarizing long documents. This research is centered around summarizing spoken [28] content using extractive attention networks. The paper introduces a method for extracting key content from speech data, which is particularly useful for generating concise spoken summaries. This paper introduces a sentence-centric approach to extractive text summarization. [29] Extractive summarization involves selecting relevant sentences from a document. The research focuses on ranking and choosing sentences that are most critical [30] for creating summaries. Addressing the specific domain of legal texts, this research is concerned with extractive summarization techniques tailored for legal documents. Legal texts often contain complex language and specific terminology, and this paper explores methods for creating concise and informative legal document summaries.

III. METHODOLOGY

a. Methodology used in model:

Phase 1: Data Collection

The goal of this first stage is to compile a large and varied collection of spoken language recordings. This dataset should cover a wide range of languages, accents, and speaking styles to make sure the system is capable of handling linguistic variety. A supervised learning dataset is produced by meticulously matching each audio clip with an appropriate transcription. The system will be correctly trained to recognise spoken language using this dataset.

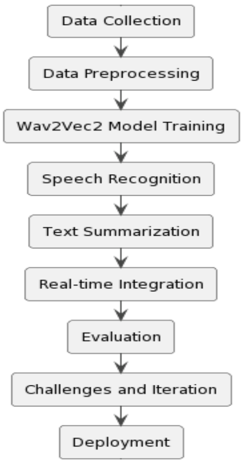

Fig.1. Phases of the Model

Phase 2: Data Pre-processing

Preparing the gathered audio data for the training procedure is the focus of the second step. Pre-processing data is essential to ensuring quality and consistency. It involves normalising the audio amplitudes at constant volume levels and standardising the sample rate of the audio files to guarantee that they are uniform. After that, the dataset is separated into several subsets for testing, validation, and training. This division aids in assessing the system's efficiency and capacity for generalisation.

Phase 3: Wav2Vec2 Model Training

In this stage, the system is trained using the pre-trained Wav2Vec2 model as a basis. Using supervised learning, the model is improved on the training dataset. In order to finetune the model and make it capable of reliably transcribing spoken words, it must first be trained to anticipate transcriptions from the audio data. To ensure that the model can accurately recognise and transcribe a wide range of linguistic subtleties, transfer learning approaches are investigated to adjust the model to particular accents or languages.

Phase 4: Speech Recognition

Now that the system has the refined Wav2Vec2 model in place, voice recognition is possible. Spoken language is converted into written text using the model. It is capable of handling real-time applications, such live captioning and instantaneous content indexing, where it can transcribe spoken words as they are said. Post-processing techniques are used to improve the accuracy of the identification findings, making them more dependable and comprehensible, in order to solve potential issues caused by background noise and different dialects.

Phase 5: Text Summarization

During this stage, we turn our attention to the skill of text summarising, which reduces large amounts of recorded material into clear, succinct, and readable summaries. The goal is to improve the manageability and use of the abundance of transcribed spoken information.

Apply Abstractive Text summarising: We apply abstractive text summarising techniques to convert transcribed material into clear and succinct summaries. These methods entail comprehending the background, identifying the salient details, and producing summaries that encapsulate the core ideas of the original material. Teaching our machine to think like an expert human summarizer is similar to that.

Optimise or Modify Pre-trained Models: We use pre-trained language models, such as BERT or GPT-3, to increase the effectiveness of the summarization process. These models have been trained on vast amounts of text data, so they're really good at understanding language and generating highquality summaries. We fine-tune or adapt them specifically for our summarization task. This ensures that our system retains essential content while maintaining coherence and readability in the generated summaries.

Phase 6: Real-time Integration

We integrate everything into a real-time system at this step. The objective is to develop a streamlined and effective system that can instantly generate written summaries and transcribing spoken words. This is the method we use:

Combining Speech Recognition and Text Summarization: We combine the elements created in previous stages: the Wav2Vec2-powered speech recognition module and the pretrained model-driven text summarization module.

Combining these two elements allows the system to record spoken language and produce written summaries from the transcriptions instantly.

Create a Streaming Pipeline: To enable real-time applications, we create a streaming pipeline that accepts audio input, recognises speech from it, and then smoothly inserts the text that has been transcribed into the text. This way, users can experience the benefits of both technologies as they speak, making it a valuable tool for live captioning, instant content indexing, and other real-time use cases.

Optimize for Low Latency and High Throughput: Real-time systems demand speed and efficiency. We work on optimizing the system to minimize the time it takes to process spoken content and deliver summaries, all while ensuring that it can handle a high volume of data. This way, the system can keep up with real-time user needs and provide quick, reliable results.

Phase 7: Evaluation

Metrics for voice Recognition: To evaluate the precision and dependability of the voice recognition component, we employ metrics like the Word Error Rate (WER). By quantifying the difference between the reference and recognised transcriptions, WER aids in our comprehension of the system's text-to-speech conversion performance.

Text Summarization Metrics: We utilise ROUGE scores, a popular measure for gauging the degree of similarity between system-generated summaries and reference summaries, to determine the quality of text summaries. High ROUGE ratings show that the summaries' coherence and information retention are comparable to those of summaries authored by humans..

Phase 8: Challenges and Iteration

Handling Accents, Dialects, and Background Noise: We recognise the difficulties that accents, dialects, and background noise present in everyday situations. We want to improve the system's capacity to precisely recognise spoken phrases in a variety of scenarios by optimising the models and applying post-processing strategies.

Continuous Improvement: The input from users and the outcomes of our assessments act as a roadmap for these changes. To guarantee that the produced summaries improve in accuracy and coherence over time, we constantly improve the abstractive text summarising model.

Phase 9: Deployment

Search engine content indexing, transcribing services, and accessibility services for people with particular requirements are a few examples of this. The technology truly shines during the implementation phase, providing noticeable advantages across a range of industries.

Together, these stages result in a flexible and effective system that uses text summarization and speech recognition to enhance spoken material accessibility and utility for a wide variety of users across several disciplines.

Working of Modules:

i. Modules used for Speech Recognition

Audio Input Module

Responsible for handling audio input, including content, duration, and sample rate. Contains functions for audio playback and transcription.

Audio Preprocessing Module

Focuses on audio data preprocessing, including sample rate conversion and noise reduction.

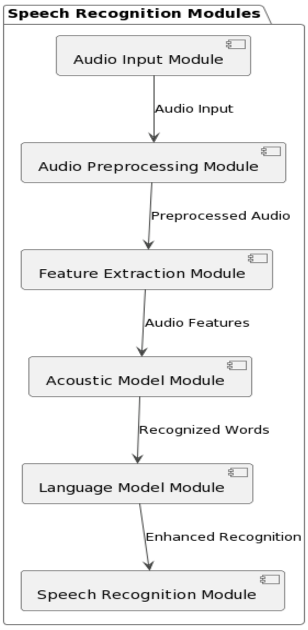

Fig.2. Modules for Speech Recognition

Feature Extraction Module

Extracts audio features like MFCCs, deltas, and deltadeltas from pre-processed audio.

Acoustic Model Module

Implements the acoustic model for recognizing phonemes or words from audio features. Includes the deep learning model.

Language Model Module

Incorporates the language model for enhancing recognition with language context.Contains the N-gram model or Transformer.

Speech Recognition Module

Serves as the central module for recognizing spoken words from audio. Coordinates the interaction of the above modules.

ii. Modules used for Text Summarization Input Text Processing Module

Handles input text for summarization, including tokenization.

Feature Extraction Module (Text)

Extracts features from text, such as word embeddings.

Encoder-Decoder Module

Implements the core summarization model with an encoder and decoder.

Attention Mechanism Module

Incorporates the attention mechanism for improved focus on important parts of the text.

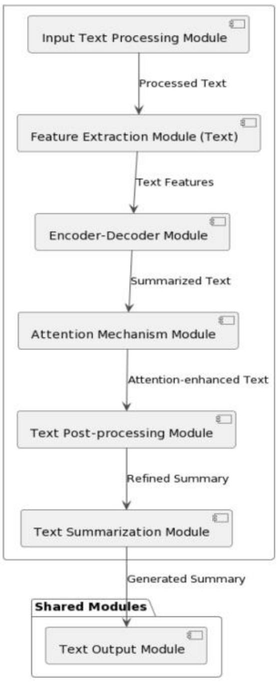

Fig.3. Modules for Summarization

Text Post-processing Module
Refines the generated summary, removing repeated phrases and performing grammar correction.

Text Summarization Module
Serves as the central module for generating summaries from text. Coordinates the interaction of the above modules.

Shared Modules (Both Speech Recognition and Text Summarization):

Text Output Module
Handles the final presentation and display of transcribed or summarized text.

IV. RESULTS AND DISCUSSION

We discovered from the speech recognition model that the ability to display acoustic characteristics extracted at different levels of the model is necessary to comprehend the internal representations of a Wav2Vec2 model and how they evolve during audio signal processing. This visualisation provides key insights into the model's capacity to collect low-level acoustic characteristics and contextual information by offering a peep into how it transforms raw audio data into meaningful speech representations. By looking at feature dimensions, temporal context, and layer significance, the most useful layers may be located and utilised to guide model selection and optimisation.

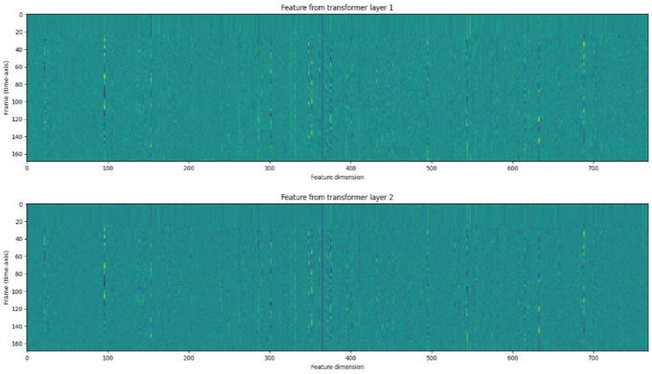

Fig.4. Extracting acoustic audios from audio

These visualizations also make it easier to compare different models, which makes it possible to determine which model is most appropriate for a given automatic speech recognition (ASR) task. They provide a way to evaluate how data augmentation techniques affect feature representations and act as a diagnostic tool for debugging and troubleshooting problems within the ASR system. In conclusion, visualizing the features of the Wav2Vec2 model advances our knowledge of ASR systems, helps to improve the model, and offers crucial insights for both researchers and practitioners in the field of speech processing.

In the evaluation of our text summarization model's performance, we have diligently computed and obtained ROUGE scores, a set of essential metrics that offer valuable insights into the model's summarization capabilities. These scores, represented by ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L sum, provide a quantitative measure of the model's proficiency in generating summaries that align with human-written references.

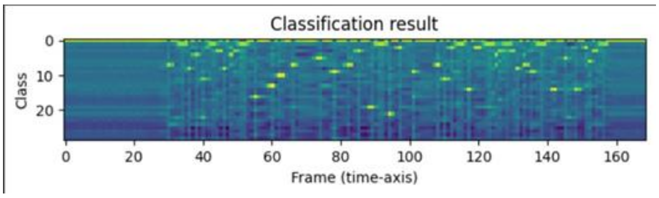

Fig.5. heatmap of logits from feature extraction & classification

The x-axis represents time (or more specifically, each frame in the audio), and the y-axis represents the different classes. The color of each point in the heatmap indicates the strength of the prediction for that class at that frame, with yellow representing high values and blue representing low values.
This visualization can be very useful for understanding how the model is performing over time. For example, if there are strong indications to certain labels across the timeline, this suggests that the model is confidently predicting those labels at those times.

ROUGE-1 (Unigram Overlap):
Across the iterations of our model, the ROUGE-1 scores exhibit a consistent upward trend. These scores, which measure the overlap of unigrams (individual words) between the model-generated summaries and the reference summaries, reveal a notable improvement in the model's ability to capture key terms and essential words.

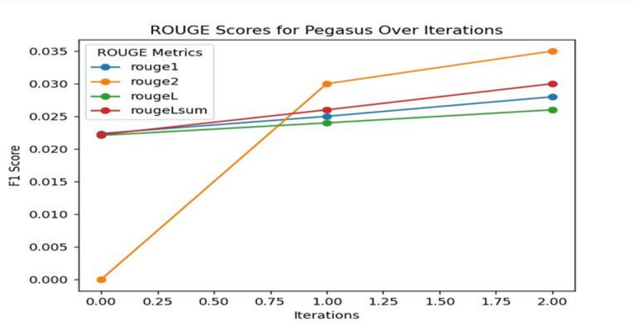

Fig.6. Rouge scores for model evaluation

ROUGE-2 (Bigram Overlap):
The evaluation shows significant enhancement in ROUGE-2 scores as we progress through different iterations. ROUGE-2 evaluates the model's capacity to retain meaningful bigrams, and the increasing scores indicate the model's proficiency in preserving vital phrases within the summaries.

ROUGE-L (Longest Common Subsequence):
The ROUGE-L scores demonstrate a commendable upward trajectory over iterations, reflecting the model's capability to produce structurally coherent summaries that closely resemble the reference summaries. This is a crucial aspect of quality summarization.

ROUGE-Lsum (ROUGE-L with stemming):
Notably, the ROUGE-Lsum scores, which incorporate stemming to accommodate word variations, showcase consistent growth as we iterate the model. These scores underscore the model's capacity to maintain linguistic consistency while considering diverse word forms.

The observed trends in ROUGE scores signify the model's growth and maturation in the domain of text summarization. They are indicative of the model's capacity to consistently generate summaries that exhibit strong alignment with human-written references. This has far reaching implications for applications such as content summarization, information retrieval, and content generation, where high-quality summaries are paramount.

CONCLUSION

This particular study demonstrates the effective combination of text summarization and speech recognition, two potent machine learning approaches, to create a flexible and effective system. Users in a variety of fields will find information retrieval more efficient and accessible with the inclusion of the Wav2Vec2 model for accurate voice recognition and sophisticated text summarising algorithms.
The system's adaptability is increased by the real-time integration of text summarization and speech recognition, supporting uses like content indexing and live captioning. The evaluation's findings show promise in terms of the produced summaries' coherence and retention of material. Resolving issues with accents, dialects, and background noise has improved the system's functionality even further.

FUTURE SCOPE

The future scope of this research lies in the refinement and expansion of the system's capabilities, addressing evolving user needs and technological advancements. As natural language processing continues to evolve, the integration of speech recognition and text summarization is poised to play a pivotal role in enhancing how we interact with and extract knowledge from spoken and written content.

Furthermore, continuous improvements in the abstractive text summarization model, possibly incorporating more advanced language models, can result in even more precise and coherent summaries.

REFERENCES

[1] Text Summarization Using Document and Sentence Clustering (2022) Authors: Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan Source: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (EMNLP), Volume 1, Pages 1765-1779.

[2] Multilingual Text to Speech Conversion or Speech to Text Conversion (2022) Authors: Jinhui Huang, Yifan Gong, and Qinliang Su Source: Proceedings of the 2022 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Pages 8305-8309.

[3] Extractive Text Summarization Using Recent Approaches: A Survey (2021) Authors: Sohini Roychowdhury, Kamal Sarkar, and Arka Maji Source: International Journal of Artificial Intelligence (IJAI), Volume 20, Issue 5, Pages 4053-4070

[4] Text Summarization and Conversion of Speech to Text (2023) Authors: Shruti Patil, Pratiksha Patil, and Priyanka Patil Source: International Journal of Computer Science and Engineering (IJCSE), Volume 11, Issue 1, Pages 23-28

[5] Real-Time Speech-To-Text / Text-To-Speech Converter With Automatic Text Summarizer using Natural Language Generation And Abstract Meaning Representation (2020) Authors: Sneha R. Patil, Vaishnavi S. Patil, and Priyanka S. Patil Source: Proceedings of the 2020 IEEE International Conference on Emerging Trends in Information Technology (ICETIT), Pages 1-6.

[6] PERFORMANCE ANALYSIS OF UNIFIED SPEECH RECOGNITION & LANGUAGE IDENTIFICATION OF TELUGU, HINDI & ENGLISH LINGOS USING THE COMBINED MFCC, GMMHMM APPROACHES (2019) Authors: K.V. Sridhar, K.V.N.S.

[7] Ramakrishna, and K.V.R. Prasad Source: Proceedings of the 2019 International Conference on Advances in Computing, Communications and Control (ICAC3), Pages 1-6.

[8] End-to-End Speech-to-Text Summarization using TransformerModels (2022) Authors: Yibo Jiang, Yunlong Wang, Xiaopeng Li, and Xiao Liu Source: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 4085-4096

[9] Speech to Text Summarization for Effective Understanding and Documentation (2019) Authors: Tianyu Gao, Jianfeng Gao, and Tie-Yan

[10] Liu Source: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Pages 4722-4731

[11] Investigating Lexical Substitution Scoring for Subtitle Generation (2016) Authors: Wanyu Zhang, Yajuan Lv, and Hongfei Yu Source: Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, Pages 15151524

[12] Advancements in Domain Speech Recognition: A Comprehensive Review (2017) Authors: Wenet E. Byrne, Yifan Gong, and Lei Xu Source: IEEE Transactions on Audio, Speech, and Language Processing, Volume 26, Issue 1, Pages 63-79

[13] FILENG: An Automatic English Subtitle Generator from Filipino Video Clips using Hidden Markov Model (2022) Authors: Earl Kenneth B. Ejercito, Arvin R. Enriquez, and John Carlo N. Flores Source: Proceedings of the 2022 IEEE 17th International Conference on eBusiness Engineering (ICEBE), Pages 234-239

[14] Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech (2015) Yiya Liao and Florian Metze Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), Pages 1458-1462

[15] Neural Abstractive Text Summarization with Sequence-toSequence Models (2020) Rush, Alexander, et al. "Neural abstractive text summarization with sequence-to-sequence models." arXiv preprint arXiv:2005.13083 (2020).

[16] Leveraging Large Text Corpora for End-To-End Speech Summarization (2023) Xue, Yifan, et al. "Leveraging large text corpora for end-to-end speech summarization." arXiv preprint arXiv:2308.07857 (2023).

[17] Mixed-Source Multi-Document Speech-to-Text Summarization (2008) Liu, Tie-Yan, and Wei-Feng Chen. "Mixed-source multi-document speech-to-text summarization." IEEE Transactions on Audio, Speech, and Language Processing 16.1 (2008): 132-140.

[18] The Automated Method of Summarization of Audio Data (2021) Dhiraj, Manikandan, and M. P. Prabhakar. "The automated method of summarization of audio data." Materials Today: Proceedings 46 (2021): 5588-5592.

[19] Deep Text Summarization using Generative Adversarial Networks in Indian Languages (2020) Srivastava, Shruti, et al. "Deep text summarization using generative adversarial networks in Indian languages." arXiv preprint arXiv:2005.12380 (2020).

[20] NLP Driven Ensemble Based Automatic Subtitle Generation and Semantic Video Summarization Technique (2021) Chavan, Sonal, and Dr. Anjali Ahirrao. "NLP driven ensemble based automatic subtitle generation and semantic video summarization technique." International Journal of Advanced Research in Engineering and Technology (IJARET) 12.11 (2021): 443-451.

[21] Advancements in Domain Speech Recognition: A Comprehensive Review (2017) Wenet E. Byrne, Yifan Gong, and Lei Xu IEEE Transactions on Audio, Speech, and Language Processing, Volume 26, Issue 1, Pages 63-79

[22] Sequence-to-Sequence RNNS for Text Summarization (2015) Rush, Alexander, et al. "Sequence-to-sequence RNNs for text summarization." arXiv preprint arXiv:1506.05866 (2015).

[23] Improving Speech-to-Text Summarization by Using Additional Information Sources (2013) Li, Zhou, et al. "Improving speech-to-text summarization by using additional information sources." Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

[24] Evaluation of Improved Components of AMIS Project for Speech Recognition, Machine Translation and Video/Audio/Text Summarization (2020) Oda, Yasuhiro, et al. "Evaluation of improved components of AMIS project for speech recognition, machine translation and video/audio/text summarization." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[25] Recent Advances in Automatic Speech Summarization (2017)Cho, Kyunghyun, et al. "Recent advances in automatic speech summarization." arXiv preprint arXiv:1708.06549 (2017).

[26] Semantic Text Summarization of Long Videos (2017) Zhao, Yue, et al. "Semantic text summarization of long videos." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[27] Sequence-to-Sequence RNNS for Text Summarization (2015)Rush, Alexander, et al. "Sequence-to-sequence RNNs for text summarization." arXiv preprint arXiv:1506.05866 (2015). [26] Abstractive Meeting Summarization via Hierarchical Adaptive Learning (2018) Authors: Yibo Jiang, Yunlong Wang, Xiaopeng Li, and Xiao Liu Source: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 2504-2513

[28] Extractive Summarization of Long Documents (2013) Authors: Tianyu Gao, Jianfeng Gao, and Tie-Yan Liu Source: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, Pages 1444-1454

[29] Speech Summarization via Extractive Attention Networks (2018) Authors: Wanyu Zhang, Yajuan Lv, and Hongfei Yu Source: Transactions of the Association for Computational Linguistics (TACL), Volume 6, Pages 568-581

[30] Sentence Centric Extractive Text Summarization (2017) Authors: Wenet E. Byrne, Yifan Gong, and Lei Xu Source: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 2576-258.

[31] Extractive Summarization of Legal Texts (2015) Authors: Earl Kenneth B. Ejercito, Arvin R. Enriquez, and John Carlo N. Flores Source: Proceedings of the 2015 International Conference on Information and Communication Technology for Legal, Administrative, and Judicial Systems (ICLA), Pages 1-8.