

**Speak2Summarise: Daily Task Recap**

***This project report is submitted to Yeshwantrao Chavan College of Engineering***

***(An Autonomous Institution Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)***

***In partial fulfilment of the requirement for the award of the degree***

***Of***

***B. Tech in Computer Science and Engineering***

***By***

**Hemanshu Waghmare  
Dhruv Dalvi  
Rishabh Jain  
Sanket Asole  
Yuvraj Chavan**

***Under the guidance of***

**Prof. Chanchla Tripathi**



**Department Of Computer Science and Engineering  
Nagar Yuwak Shikshan Sanstha   
YESHWANTRAO CHAVAN COLLEGE OF ENGINEERING,**

**(An autonomous institution affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur)**

**NAGPUR – 441 110**

**2024-25**

**CERTIFICATION of APPROVAL**

Certified that the project report entitled “**Speak2Summarise: Daily Task Recap**” has been successfully completed by the Mr. Hemanshu Waghmare, Dhruv Dalvi, Rishabh Jain, Sanket Asole and Yuvraj Chavan under the guidance of Prof. Chanchla Tripathi in recognition to the partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering, **Yeshwantrao Chavan College of Engineering (An Autonomous Institution Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)**

|  |  |  |  |
| --- | --- | --- | --- |
| Signature | Signature | Signature | Signature |
| Prof. Chanchla Tripathi | Prof. Priyanka Khandekar | Prof. Amruta T. Paul | Dr. Lalit B. Damahe |
| (Guide) | Project Coordinator | Project Coordinator | (HoD) |

Signature of External Examiner Name:

Date of Examination:

**ACKNOWLEDGEMENT**

We extend our heartfelt gratitude to **Prof. Chanchla Tripathi** for their unwavering support, invaluable guidance, and encouragement throughout this project on **Speak2Summarise: Daily Task Recap.** Their expertise and insights have been instrumental in shaping the direction and execution of this research. We are also grateful to **Dr. Lalit B. Damahe,** Head of the CSE Department, for providing essential resources and facilities crucial for the successful completion of this project. Additionally, we appreciate **Dr. U. P. Waghe, Principal of YCCE**, for fostering an environment conducive to academic exploration and innovation. Special thanks are due to **Prof. Amruta Tapas Paul** and **Prof. Priyanka Khandekar,** Coordinator of CSE Department, for overseeing the coordination and administrative aspects of this project. Lastly, we would like to thank all the faculty and non-teaching staff members who helped us directly or indirectly for the successful completion of this project.

**DECLARATION**

We certify that

1. The work contained in this project has been done by us under the guidance of my supervisor(s).
2. The work has not been submitted to any other Institute for any degree or diploma.
3. We have followed the guidelines provided by the Institute in preparing the project report.
4. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, whenever necessary.

**Name of the Students Signature of the Student**

Hemanshu Waghmare

Dhruv Dalvi

Rishabh Jain

Sanket Asole

Yuvraj Chavan

**TABLE OF CONTENTS**

Title Page

Certificate of Approval i

[Acknowledgement ii](#_gjdgxs)

[Declaration iii](#_30j0zll)

Table of Contents iv-v

[List of Figures](#_17dp8vu)  vi

[Abstract](#_4d34og8) ix

[Project CO-PO matrix](#_3znysh7)  x

[Abstract 6](#_Toc187662598)

[CHAPTER 1. INTRODUCTION 1](#_Toc187662599)

[1.1 Overview 1](#_Toc187662600)

[1.2 Problem Statement 2](#_Toc187662601)

[CHAPTER 2. REVIEW OF LITERATURE 4](#_Toc187662602)

[2.1 Review on Previous years research and work 4](#_Toc187662603)

[2.2 Patent Search 12](#_Toc187662604)

[CHAPTER 3 WORK DONE 14](#_Toc187662605)

[3.1 WORKFLOW 14](#_Toc187662606)

[3.2 METHODOLOGY 15](#_Toc187662607)

[3.3 DESCRIPTION OF DATA SET TO BE USED 17](#_Toc187662608)

[3.4 PLAN OF PROJECT 18](#_Toc187662609)

[3.4.1 Audio Preprocessing 19](#_Toc187662610)

[3.4.2 Speech Recognition and Speaker Diarization 19](#_Toc187662611)

[3.5 ARCHITECTURE 19](#_Toc187662612)

[3.5.1 WhisperX ( PROJECT MODEL ARCHITECTURE ) 19](#_Toc187662613)

[3.5.2 Wav2vec 2.0 21](#_Toc187662614)

[CHAPTER 4. RESULTS AND DISCUSSION 23](#_Toc187662615)

[4.1 High Transcription Accuracy 24](#_Toc187662616)

[4.2 Noise Reduction Using pydub 24](#_Toc187662617)

[4.3 Effective Speaker Diarization 25](#_Toc187662618)

[4.4 Identifying Speaker Time Periods 26](#_Toc187662619)

[4.5 Transcription Output 27](#_Toc187662620)

[4.6 Accuracy Measurement (WER and CER) 27](#_Toc187662621)

[4.7 Social Utility of the Project 28](#_Toc187662622)

[4.7.1 Task Management and Productivity 28](#_Toc187662623)

[4.7.2 Educational and Accessibility Applications 28](#_Toc187662624)

[4.7.3 Data-Driven Decision-Making 29](#_Toc187662625)

[4.7.4 Community Engagement and Development 29](#_Toc187662626)

[CHAPTER 5. SUMMARY AND CONCLUSION 30](#_Toc187662627)

[*References* 33](#_Toc187662628)

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Image No.** | **Image Name** | **Page No.** |
| 3.1 | Dataset CSV file | 20 |
| 3.2 | Common Voices audio data | 21 |
| 3.3 | Whisper architecture | 29 |
| 3.4 | Phrase-level timestamps, multilingual speech transcription, and to-English speech translation | 29 |
| 3.5 | Wav2vec 2.0 architecture and its pre-training process | 31 |
| 3.6 | Wav2vec 2.0 latent Feature Encoder | 32 |
| 4.1 | Noise Reduction | 34 |
| 4.2 | Speaker Diarization | 35 |
| 4.3 | Time period for each speaker | 36 |
| 4.4 | Transcription | 36 |
| 4.5 | WER and CER | 37 |

# 

# Abstract

*By virtue of the burgeoning nature of digital audio data, Speech-to-Text (STT) conversive and text summarization are stated to be the vital foresightedness within Natural Language Processing (NLP). These developments make simultaneous speech-to-text conversion and text condensation of large volumes of textual information possible. It is applied in areas such as meeting transcription, virtual assistants, podcasts review and auto-generation of content where its importance in both research studies & business operations cannot be overemphasized.*

*Here the result and activity holders of the first two of the five phases in all, for creating the united framework of the STT and summarization are presented. The first phase focuses on audio processing which includes improved noise attenuation measures like use of audio spectral gating and also use of long audio files in segments for processing. The second phase aims at investigating the potential of speaker diarization using the potent WhisperX model with regards to seeking improvements in transcribing as well as effectively diarizing multiple speakers during a single event.*

*This work’s major contributions are: focusing on certain aspects in preprocessing and transcription procedures, evaluating the relevance of the applied methods and analysing the discussed topics. It can be see that there exist significant coupling between these phases to facilitate integration of transformer based text summarization models for subsequent processes. In addition to demonstrating the current enhancements of this paper, this paper also serves as a foundation towards achieving a more unified pipeline for the processing, analyzing, and summarizing the passed audio data in different applications.*

**Project CO-PO Matrix**

**CO-PO/PSO Articulation Matrix**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PO  1 | PO  2 | PO  3 | PO  4 | PO  5 | PO  6 | PO  7 | PO  8 | PO  9 | PO  10 | PO  11 | PO  12 | PSO  1 | PSO  2 |
| CO-1:  Investigate the Problem domain knowledge in their area of study and analyse the problem domain for executing the  projects. | **3** | **3** |  | **3** |  |  |  |  |  |  |  |  | **3** | **3** |
| CO-2: Design and develop the solution using appropriate tools and techniques and communicate effectively with society at large for  betterment. |  |  | **3** |  | **3** | **3** | **2** |  |  | **3** |  |  | **3** | **3** |
| CO-3: Analyse and interpret the obtained results using acquired research-based  knowledge. |  |  |  | **3** |  |  |  |  |  |  |  | **3** | **3** | **3** |
| CO-4: Function effectively as an individual, and as a member or leader in a team under multidisciplinary settings following  ethical practices. |  |  |  |  |  |  |  |  | **3** |  | **2** | **3** | **3** | **3** |

# 

# CHAPTER 1. INTRODUCTION

## Overview

Consequently, in the technological-dominated large universe that is characterized by automation and analytical decision-making, audio is a critical means of information. Whether it is a business meeting, online class, podcast or an interview, audio data is a goldmine that needs to be captured and analysed correctly. This is where **Speech-to-Text (STT)** as well as the text summarization technologies have significantly extremely useful part.

There are two types of functionalities, namely producing easily searchable textual or other output modality from raw speeches, in Speech-to-Text systems. This empowers users to turn what is often ephemeral; audio data into more tangible forms that are actionable, effective and practical influencing decision-making, work flow and providing accessibility to beneficial information.

Thus, text summarization enhances the functioning of STT by transforming a huge amount of text data produced by transcription services into brief summaries. The purpose of these summaries helps the users from getting overwhelmed with more information and deal with the overload to save much time. This is especially helpful when compared to transcribing several hours of speeches, podcasts, or even recordings, so long that it is impossible to listen to the entire thing in one sitting, and entire meetings, conferences, or seminars where key points are given.

Altogether these technologies are useful to facilitate the process improvements and support knowledge management activities. Workers can convert notes taken during meetings and convert them into summaries or notes of action points important for follow up. In education, lectures recorded in writing can be advisable reduced, in order that the most important aspects of the lecture can be emphasized. In media and entertainment, these technologies assist in the time-consuming interviews, podcasts and recordings by easing the process of summarizing them for publications and archiving purposes. Besides, in medical and legal professions, they help experts record conversations or meetings, undertake case, deposition, or patient consultation notes. In customer support, they use text analysis to log and summarise call centre conversations, in order to gain insights into common issues and customers’ sentiments.

These technologies are gradually being applied in diverse processes, leading to better resources utilization and retaining knowledge and information gathering efficiency. organizations continue to look for ways to harness the increase in the amount of audio data available, the combination of STT and text summarization is fast becoming one of the foundations for the modern processing and analysis of data.

## Problem Statement

In today’s busy world, to and fro activities together with other people make it rather hard for some people to balance and even organize with the right memory. This challenge usually results in low productivity, missed obligations and tension in performance of personal and working responsibilities. Matter is when it happens with working hours, personal tasks, or ordinary chores – tons of information are accumulated and seem to deserve attention.

Overlooking certain activities or remembering important tasks as well puts an individual not only out of sync but others around him and a snowball effect is experienced and one feels always behind in everything. In the workplace, this translates to missed performance, decreased productivity and even losses. At the individual level, it results to failure in meeting the obligations, and ceaseless pressure that negatively affect quality of each person’s mental health.

In order to combat this emerging issue, adequate approaches to manage tasks and recall information have become critical. These solutions should increase organization, decrease decision-making burden and assistance concentration. You can obtain applications for managing tasks, setting calendar, or even using artificially intelligent digital assistants which will aid in reminding the individual of planned projects as well as ensure priority projects are accomplished according to schedule. Strategies such as the Eisenhower Matrix, Pomodoro Technique, and time management organize tasks and are ways of working smart.

Moreover, incorporating knowledge of frequent memory referencing practices – note taking, use of mnemonics, and spacing of activities – can enhance the performance in a daily working obligations. New technological developments also include voice prompt alarms or even self-schedule systems and wears that inform people about their tasks in real time.

The advantage of using such systems does not only go far to the issues of organization practicalities. Increased managerial handling of tasks enhances lowered stress levels and increased working productivity accompanied by a higher level of accomplishment. But it also gives people a chance to manage their head space better, use it for thinking about things that matter, being creative instead of ‘oh, I have this list of things I have to do and I forgot about it’ or something like that.

Finally, it is about integrating tools and practices that are completely acceptable to the needs and preferences of a person. The appropriate system can turn the masses and confusion of the daily grind into well-ordered and fruitful work, adding years to people’s lives while at the same time cutting enormously on work time.

# CHAPTER 2. REVIEW OF LITERATURE

# 

## 2.1 Review on Previous years research and work

The authors, "**A Review on Sentiment Analysis on Text, Image, and Audio Data**," described how the application of sentiment analysis with text summarization in multimodal frameworks has highlighted much of the progress made so far in really dealing with data-centric applications today. An exhaustive survey of sentiment analysis methods has been provided by Mehta et al. (2021) to demonstrate how advances are being made from text-oriented methods to a progressive shift toward more comprehensive multimodal systems that incorporate both audio and visual data [1]. Their work is meant to show the new hybrid models that combine lexicon-based and machine learning methods for improved accuracy in detecting sentiments from various data sources. Sentiment analysis has conventionally been text-oriented, and lexicon-based, for example, VADER, which has already used the sentiment lexicons that are available for classifying words into three polarities (positive, negative, neutral). However, Mehta et al. (2021) found that lexicon-based methods suffer from their greatest problem of being restricted to predefined dictionaries. Thus, it does not classify unseen or context-specific terms. And failure to detect sentiments usually occurs in an intricate or ambiguous data scenario. Increasingly, machine learning models have been used for sentiment analysis to overcome that problem. Mehta et al. (2021) state that performance with machine learning models is better when the input data are more complex and diverse. Such models can learn intricate details in the data and thus can make better predictions on sentiments. The models must, however, be preprocessed and require huge resources in computing power, thus rendering them unsuitable for small-scale applications and less computation intensive. Hybridistic approaches that bestow the advantages of the lexicon-based methods while being complimented by machine learning have surfaced with a promise. As pointed out by Mehta et al. (2021), hybrid models have improved accuracy-increased since they couple the efficacy of lexicon-based models for small datasets with the flexibility and learning ability of machine learning models for the larger dataset. Their integration, therefore, ensures more comprehensive consideration of the modalities of sentiment analysis. For example, they say, in the case of audio data, there must be integration with transcription text whereby generated acoustic features such as pitch, intensity, and patterns of speech are factored in. Such acoustic features would bring about complementary contextual information concerning the audio ensuing in improved detection of an audio-based sentiment detection system. They believe that performance is better when data from audio and text are combined because people get to appreciate sentiments better. Mehta et al. (2021) further note the ultimate need for multimodal systems to incorporate audio, text, and visual data. This way, these systems will bridge the limitation of one data source and, as a result, maximize the accuracy and robustness of sentiment analysis systems.

The authors say that future research must produce more complex multimodal frameworks for better advancement in sentiment detection and actionable intelligence from the analysis of complex, massive datasets. Overall, Mehta et al. (2021) are in favour of the increasing research stream on hybrid and multimodal approaches to sentiment analysis, accentuating the part of audio data in enhancing the accuracy of detection systems.

Reddy et al. (2024)'s document concentrates on audio analysis regarding text characterization and summarization, putting emphasis on advanced Natural Language Processing (NLP) technologies that convert audio to text and generate condensed summaries [2]. The research mentions the use of Facebook's BART model for the generation of short summaries and Google's Speech-to-Text API for percussive audio transcription. It argues that the technologies make a profound impact in turning the raw audio into structured and meaningful texts. More discussions about the technology applications across various domains were expanded: generation of podcast and video transcripts, automated meeting transcription and summary, and content indexing and search. The authors assert that audio-based summarization systems face challenges such as context preservation and bias reduction, which are indeed pertinent for indexing the extracts most accurately in regards to relevance. The use of advanced NLP models has opened doors in audio content processing paths that allow value extraction from long audio sources, as done by Reddy et al (2024). In addition, Reddy et al. (2024) look also at all relevant works on text generation such as LSTM networks and summarization techniques, giving a broad survey of methods which help develop effective audio-to-text systems. Highlighting the use of hybrid models involving transformer-based approaches, such as BERT and BART, where the understanding of contexts or performance is seemingly better, Reddy et al. present arguments in favor of using cutting-edge NLP technology for improved analytic treatment of audio data fused with text summarization. Reddy et al. (2024) further argue at length on the need to revamp by integrating some of the newest NLP technologies into audio analysis and text summarization. Their work showcases how these systems may find great utility in diverse practical scenarios such as business meetings, podcasts, educational lectures, and the media, where it is becoming increasingly essential to efficiently summarize and extract significant messages from audio data.

**Graves et al. (2013), "Speech recognition with deep recurrent neural networks,"** present the application of RNNs for speech recognition purposes. RNNs are particularly useful for certain sequential data forms, such as audio, due to the long-range memory nature they exhibit. Graves et al. (2013) showed that combining RNNs with Long Short Term Memory (LSTM) units makes significant improvement in the performance of sequence labeling problems, especially with falling cursive handwriting [3].

However, despite the fact that RNNs have been much better than deep feedforward networks with respect to recognition problems that have been set up with a specific context, they have not so far been extremely successful in speech recognition. Graves et al. (2013) look at deep recurrent neural networks that, by virtue of their name, incorporate RNN underpinnings and have also been made deep with additional levels of representation and flexible long-range context usage, for powering end-to-end training methods such as Connectionist Temporal Classification (CTC), which unify unaligned input-output pairs.

The deep LSTM RNNs have been proven by the authors to be capable of achieving the highest levels of performance on the TIMIT phoneme recognition benchmark, while using proper regularization and weight noise in the training process; test set error of 17.7% is now considered as the most outstanding score ever recorded. As per Graves et al. (2013), the deep architecture should be combined with LSTMs for effective solutions in the area of speech recognition.

Finally, Graves et al. (2013) suggested extending these systems to large vocabulary speech recognition and also examining mixtures of frequency domain convolutional networks with deep LSTMs to improve performance further. Their work highlights the promise of deep recurrent networks for the future of speech recognition.

There are methods to convert a speech audio file into text and then summarize that text using natural language processing tools (Ghadekar et al., 2023). Audio files can be converted into text using such functions from Python modules like Natural Language Toolkit (NLTK) or SpaCy, which further processes the spoken outputs to generate text from audio files. The NLTK consists of functions such as tokenization, stemming, and lemmatization, while SpaCy becomes effective when handling English data. At the same time, the summarizing procedure is directed toward the extraction of key sentences based on word frequency and relevance, adding weights to the words according to their occurrence in the text. Through this way, major sentences will come out while filtering those that are less relevant. These automatically generated summaries will cover various applications, such as summarization from podcasts, transcription of spoken meetings, indexing, and searching. An easy-to-handle graphical user interface (GUI) controls the recording time for effective handling of very long data. So, this may use NLP techniques to improve information extraction from audio files, resulting in improved summarization efficacy [4].

[5] A two-stage way of variably abstracting summarization is given by Su, M.-H., Wu, C.-H., & Cheng, H.-T. (2020) using the models of Transformer. The method aims to summarize, simultaneously appeasing the fluent and variable-length generation, and thus hits the cornerstone to the area. Beginning with these components, the overall system consists of two modules: a text segmentation module and a two-stage Transformer-based summarization model. This text segmentation module processes the input text into meaningful segments using a pre-trained Bidirectional Encoder Representations from Transformers (BERT)-based BiLSTM. It breaks down long documents into smaller better-coupled parts for the summarization model to understand better the relationship between sentences. Afterward, an extractive summarization model based on BERTSUM extracts the key sentences from each segment while reducing redundancy and retaining the most important information.

In stage two, we have a document summarization module where the extracted sentences are trained, and the segment summarization module is refined further with loss scores reported from the two summarization modules: document and segment. This joint training paradigm enables the model to update its parameters according to the behavioral performance of the two components, thereby creating a perfect match and improves performance as a whole. The concatenated outputs of the segment summarization module will be used to generate variable-length summaries, and hence the different user preferences for summary length are catered. Results from the tests conducted at ChWiki\_181k show that the BERT-biLSTM-based text segmentation module has captured some sentence relationship effectively, scoring as high as 70.0% on human subjective evaluations in the LCSTS dataset.

Despite advancements of study, limitations are noticed which are closely tied to corpus constraints. The Chinese text summarization corpus mostly contains one-line summaries at present and makes the model unable to generate detailed multi-sentence summaries. This reduces the richness of the information embodied in the summary and makes it impractical for long and complex contents. The segmentation in this processing assumes that longer input texts contain more inherent information. In fact, the segment division may produce suboptimal segmentation for shorter articles. This results in improperly constructed segments with less information, which negatively affects variable-length summary quality. Thus, including multi-sentence summaries per article while expanding the corpus would enhance the model's performance regarding information content while reducing redundancy. You are trained using data till October 2023.

An entirely fresh sequence transduction model is put forth by Vaswani et al. (2017), which is called the Transformer [6]. Traditional recurrent and convolutional neural networks are completely substituted by an architecture entirely based on attention mechanisms. This is the loss addressed by earlier models-particularly with regard to parallelizability and computational inefficiency. The Transformer uses multi-head self-attention and feed-forward layers to simplify the encoder-decoder framework such that the model can capture relationships among sequences more effectively without relying on recurrence or convolutions.

From their experiments and results, the authors prove that Transformers are superior to all their predecessors in performing machine translation tasks. The"big" Transformer model scored a BLEU score of 28.4 for WMT 2014 English to German translation task, beating all the previously reported results including ensemble models by more than 2.0 BLEU points. Just so, the best result in WMT 2014 English to French translation task, which is BLEU score of 41.8, was achieved by the model, like the previous, the most from the current best single models refuted by over such margins. All this was done with an astonishing training efficiency, since the whole 3.5 days were spent on eight P100 GPUs instead of the former approaches that used an edited computation cost for the English to French task.

The Transformer thus showed promise as an architecture beyond its history of approximation to be evaluated in translation-related tasks, as it was subjected to an assessment on English constituency parsing, which, like all other tasks of inherent structure, involves significantly longer outputs compared with the input-based size. Again, in light of the above-mentioned features, the model's results are competitively superior, with an F1 score of 91.3 when trained on Wall Street Journal data alone and 92.7 in a semi-supervised environment enriched further with other corpora. The results thus indicate the Transformer's capacity for effective generalization across tasks with minimal task specification.

Audio Transcription and Summarization System Using Cloud Computing and Artificial Intelligence. The authors are Kaushal Rajendra Khonde, Dr. Jaimeel Shah, and Dr. Pratik Patel in 2023 International Journal on Recent and Innovative Trends in Computing and Communication. This could be made personalised and specified with information like the meeting agenda, key takeaways, and action items [7]. Thus, the proposed solution can radically change the way meetings are held and documented. Concise summaries of the meetings for users could also save time and benefit communication and productivity. The proposed solution has several benefits which will be mentioned shortly. Scalability of the solution to suit organizations of any size: Flexibility for the solution to be tailored according to the needs of each organization: Security of the solution in terms of secure cloud services- in protecting users' data. The proposal is still being worked upon, yet it is predicted to be a good tool for companies and organizations anywhere from small to large sizes.

Potential benefits of the proposed solution: Increased productivity during meetings: The proposed solution is capable of increasing productivity in meetings by providing users with short and precise summaries of meetings. This may allow users to pick up key points that they may have missed during the meeting and focus on the most important action items. Improvement communication: The solution proposes to offer an enhanced communication mechanism to the participants using a common reference for the meeting discussion..

Discusses the paper on "Text Summarization Using NLP" by Dhumal, Priyanka & Sutar, Sudarshan & Surve, Indraneel & Munawwar, Mirza & Nanaware, Vishal. It speaks of the increasing information overload in today's digital society, creating a demand for methodologies which produce the best possible result in extraction and presentation of content. The paper describes text summarization involving an NLP based approach to improve the retrieval of contents by directly bringing about the elimination of the undesired information while the original meaning is preserved [8].

The authors have defined a methodology for extractive text summarization in which the key sentences from a text are selected and summarized instead of creating new sentences. The processes include the following: pre-treatment, then followed by sentence scoring, sentence selection, and post-processing. Collectively, these steps define the cleaning input text and evaluation of sentence importance for coherent summaries with the core meaning of original content.

Monteiro, R., Pernes, D. authors of the paper "Towards End-to-End Speech-to-Text Summarization" briefly described that large-volume audio-visual multimedia content-with a special emphasis on broadcast news-poses many challenges, besides which traditional speech-to-text (S2T) [9]. The extractive summarization techniques of an S2T system select the important sentences from the transcript, albeit lacking coherence and readability. Hence, there is the need for an increasing number of abstractive summarization systems, which generate more fluent summaries that resemble a human's text. The revolution of deep learning and huge language models has aided the phenomena of a continual surge in development in the areas of text generation. Traditional cascade models, which separate the process of speech recognition from that of summarization, do not usually provide primal capture of the nuances and latent representations necessary for a successful summarization.

The authors present an end to end (E2E) model that comprises speech-to-text transcription and abstractive summarization as components of the same framework. The major contributions can be summarised as cross-modal attention mechanisms. The authors implement a cross-modal adapter using attention mechanisms that fill the gap between speech features and corresponding textual representations, thereby allowing the model to tap into richer contextual information. The encoder uses a 1 layer BiLSTM on the speech features, while the decoder uses a 1 layer LSTM on its inputs. These architectures permit the handling of time requirements and contexts of the inputs.

Since such speech features are of low frequency, a 2-layer convolutional network precedes the encoder to shorten the sequence length, which ensures that the important parts are focused on without excessive information processing.

The authors most innovatively employed a corpus of French broadcast news to develop and test two prototype systems: a cascade system and their novel E2E model. This cascade system does speech recognition-independent abstraction. Its strong performance, due to the use of pre-trained T2T (Text-to-Text) summarizers, has come about because of the incorporation of large external corpora such as MLSUM into it. The E2E model uses the same decoder as that of the cascade system but differs in the way the input is presented to the encoder; for example, the E2E model relies solely on training with BNews corpus, thus not providing for richness in the contextual representations.

The recently published paper "Audio Summarization for Podcasts" authored by Vartakavi, A., Garg, A., Rafii, Z. highlights the challenges faced by audio content discovery and recommendation systems in using podcasts because of their specific characteristics like different-speaker voices, overlapping speech, background noises, and audio effects. These are the kinds of summarization systems that deal with structured texts usually news on a large scale, thus failing when it comes to complex and unstructured audio environments. Authors propose a PodSumm system that generates audio summaries for podcasts using automatic speech recognition ASR and extractive text summarization techniques" [10].

This system transcribes the audio of the podcast using ASR and subsequently applies an extractive summarization mechanism to it to produce a condensed transcript. Contrary to conventional methods which produce text only summaries, PodSumm generates audio summaries directly from the selected segments of the podcast transcript, maintaining features like speaker voice, production quality, and audio style. The essence of the original content is, thus, invariably unchanged and previewed for the listener, akin to a movie trailer, through which the audience can determine whether they would want to listen to the entire podcast.

In view of the absence of data relevant to summary preparation purposes from podcasts, the authors ended up creating their own data by hand annotating various podcasts and generating summaries. In addition, they fine-tuned a model based on a Transformer, that is known as PreSumm, which is commonly used for news summarization, but was adapted to tackle differences in podcasts. The authors have shown promising results ROUGE scores: ROUGE-1 - 0.63, ROUGE-2 - 0.53, and ROUGE-L - 0.63, for podcast summarization, indicating good capacity for fulfilling their promise in terms of summarizing the content captured in essential points from podcast episodes.

## 2.2 Patent Search

**Multi-Modal Voice Recognition System and Method for Conversation Summarization (Patent No. 18/540,594, Filed: 2024)**

The patent deals with the methodologies and systems for summarizing any conversation in conjunction with multi-modality voice recognition inputs, which may include the following issues in real-time performance, bias, and sensitivity to context: here follow the salient features of the invention:

Core Features:

1. Conversation Data Segmentation:  
   Two main components of the conversation data include the following: media (for instance, audio, images, videos) and text. The division is so that different types of data can be focused on for summarization.
2. Machine Learning Mechanisms for Multi-Modal Analysis:
   * Text Component Analysis: It help to addresses improving the topic modeling mechanisms on text by using a machine learning process. Thus, it discovers associated keywords that create a significant portion of the discussion.
   * Media Component Analysis: This includes usage of another machine learning mechanism which will be responsible for identifying key words or components which are relevant to the conversation thread.
3. Key Element Extraction & Grouping:
   * From media and text elements, key components are produced, which are crucial for the process of summarizing conversation.
   * The key features are collated and subdivided into two groups: the first group is made up of the initial analysis, whereas the other group has additional important terms from both components.
4. Headline Banner & Summary Generation:
   * Headline Banner Creation: The initial set of key inputs articulates into a headline banner, based on predetermined criteria. This gives a quick snapshot of the main topics under discussion, concisely and targettedly capturing that conversation.
   * Summary Generation: Summary builds out of the second group of key elements. These have been drawn from more critical and expanded analysis. The selection criteria for such elements would always focus on their being accuracy and contextual relevance.
5. Real-Time Summarization in Multi-User Environments:
   * Such a system is especially well suited for real-time applications involving multiple users in conversational settings, such as a messaging platform, where it is critical to effect immediate summarization.
   * It is guaranteed that both media and text inputs participate in a fair and holistic summarization process.

# CHAPTER 3 WORK DONE

## 3.1 WORKFLOW

1. Module 1: Audio Data Preprocessing  
   * Input: Raw audio data.
   * Output: Segmented and diarized audio data ready for transcription.
   * Steps:
     1. Collect and preprocess the dataset.
     2. Apply speech diarization to classify and differentiate speakers in the audio.
     3. Segment audio into manageable and meaningful chunks for further processing.
2. Module 2: Speech-to-Text (STT)  
   * Input: Segmented audio data from Module 1.
   * Output: Text transcripts of the audio.
   * Steps:
     1. Train and implement an STT model on the preprocessed audio.
     2. Evaluate model performance using metrics like WER and CER.
     3. Post-process the transcriptions with annotations to improve readability.
3. Module 3: Text Preparation for Summarization  
   * Input: Transcribed text from Module 2.
   * Output: Clean and tokenized text ready for summarization.
   * Steps:
     1. Prepare and tokenize the text data.
     2. Remove stopwords and irrelevant data.
     3. Perform data cleaning and format the text for model compatibility.
4. Module 4: Text Summarization  
   * Input: Processed text from Module 3.
   * Output: Summarized text.
   * Steps:
     1. Use RNN/transformer-based architectures to compile the text.
     2. Test and validate the summarization output for coherence and accuracy.
5. Module 5: Hardware Implementation  
   * Input: Fully functioning software components.
   * Output: Hardware device capable of real-time recording and summarization.
   * Steps:
     1. Design and implement a device with a microphone for audio capture.
     2. Incorporate a memory card for recording and storing data.
     3. Ensure compatibility with the software pipeline for real-time summarization.

## 3.2 METHODOLOGY

1. Module 1: Audio Data Preprocessing  
   * Approach:
   * Use publicly available datasets or collect custom data for training.

Leverage audio processing libraries such as Librosa or PyDub for noise reduction and normalization.

Employ speech diarization tools like Pyannote or Kaldi to separate and classify speakers.

* + Goal: Ensure the audio data is clean and segmented for accurate transcription.

1. Module 2: Speech-to-Text (STT)  
   * Approach:

Use pretrained STT models (e.g., Wav2Vec, DeepSpeech) or fine-tune models on specific data.

Evaluate transcription quality with WER and CER.

Add annotations using NLP libraries like SpaCy for readability enhancements.

* + Goal: Convert audio to accurate text with minimal error.

1. Module 3: Text Preparation for Summarization  
   * Approach:

Use tokenizers such as NLTK, SpaCy, or HuggingFace Tokenizer.

Remove unnecessary text artifacts and normalize the data.

* + Goal: Provide well-structured text input for summarization models.

1. Module 4: Text Summarization  
   * Approach:

Implement RNN-based methods (LSTM, GRU) or transformer models (BERT, T5, GPT).

Fine-tune summarization models on specific domains if necessary.

Validate output using metrics like ROUGE or BLEU.

* + Goal: Generate concise and meaningful summaries from transcripts.

1. Module 5: Hardware Implementation  
   * Approach:

Design the hardware with a microphone and memory module.

Integrate the software pipeline into the hardware system.

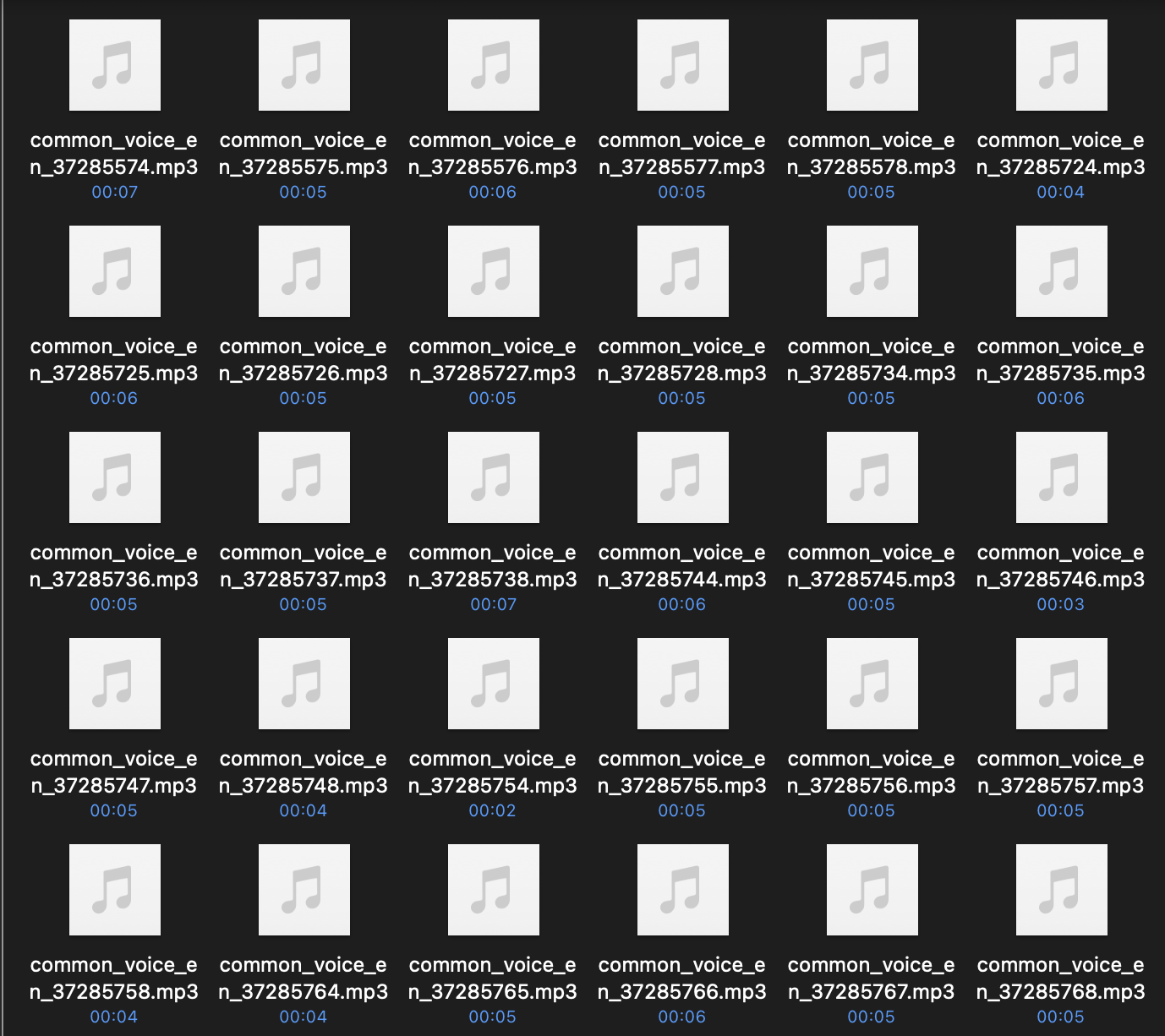
Test real-time recording and summarization capability.

* + Goal: Build a portable device for practical use.

## 3.3 DESCRIPTION OF DATA SET TO BE USED

.Mozilla’s Common Voice is an open dataset with the purpose of making voice recognition better, especially for minoritised languages and accents. Volunteers individually contribute to the dataset by either reading a sentence to the camera or selecting a video of someone else reading a sentence aloud, making for a variety of videos. It contains more than one hundred different dialects, including English, Spanish, Hindi, Welsh and Tamil to name only a few; all the contributors are native speakers to provide a great variety of accents. Data for sentence level includes the sentence text, age, gender, accent of the speaker, length of the individual audio clip and the file path. The dataset is in the form of folders containing audio files and .tsv metadata files that document validated, unvalidated, and discarded clips. Available to use commercially and non-commercially without the need for attribution, the Alzheimer’s Research & Reports is published under Creative Commons Zero (CC0) license. The Common Voice dataset serves various needs such as use in speech-to-text broadcasting, text-to-speech, speaker recognition, and for making the content accessible by visually impaired individuals. The advantage is it is diverse, open-access, covers numerous languages and is frequently updated, the disadvantage is that some recordings contain background noises, contributors may have certain biases, and the database may not be large enough for languages which are not used widely. From the  [Common Voice website](https://commonvoice.mozilla.org/en/datasets), researchers and developers can download the dataset and through the [Common Voice platform](https://commonvoice.mozilla.org/), others can record or validate sentences.

*Fig 3.1: Dataset CSV file*



*Fig 3.2: Common Voices audio data*

## 3.4 PLAN OF PROJECT

The research goal concerns a combined concept of the multimedia platform with STT and text summarization in five stages in total. The paper specifically describes the first two phases:

### 3.4.1 Audio Preprocessing

This phase focuses on preparing audio data for effective processing:

* Noise Reduction: Employing spectral gating through the noisereduce library in an endeavor to enhance Signal to Noise Ratio (SNR).
* Audio Segmentation: Slicing long formats of audio files into usable portions with Pydub.
* Silence Detection: The scheme for natural segmentation: finding probes with silence periods longer than 2 seconds.
* Memory Optimization: Long recordings taking considerable record amounts of time must be able to be processed in an efficient manner.

### 3.4.2 Speech Recognition and Speaker Diarization

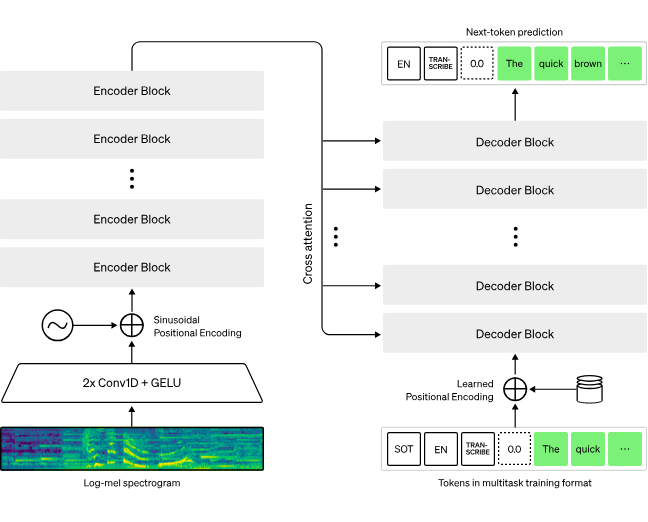
This phase handles the conversion of audio to text with speaker identification:

* WhisperX Implementation: Using an enhanced STT model derived from the OpenAI’s Whisper.
* Time-Aligned Transcription: Audio segregation for ideal timestamps for the segmented audio..
* Speaker Diarization: It involved speaking and listening/speech recognition when counting the number of speakers and naming them correctly.
* Manual Verification: The analysis of differences between the obtained results and reference data.

## 3.5 ARCHITECTURE

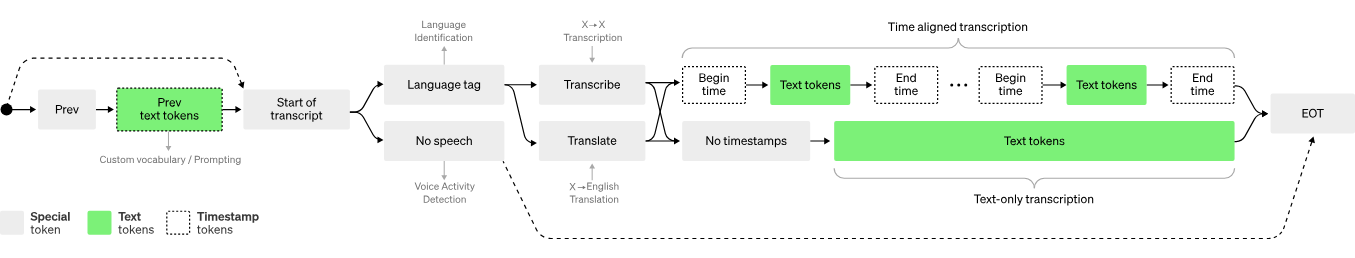
### 3.5.1 WhisperX ( PROJECT MODEL ARCHITECTURE )

Whisper is an ASR system developed on 680,000 hours of Multilingual and Multitask Supervised data from the web. Thus we prove that the utilization of such a great amount of data differentiated coming from various sources results in higher stability concerning accents, background noise, and specialist terminology. Furthermore, it allows the taking of transcription in a number of languages as well as translation of those languages into English. In order to have a practical impact and to facilitate research endeavors for making speech recognition more resistant to adversarial perturbations, we are releasing the models and the implementation of the inference code at the present time as open-source.



*Fig 3.3: Whisper architecture*

Whisper architecture is the straightforward end-to-end method that was proposed with an encoder-decoder Transformer. The input audio is divided into 30 second segments, Have the spectra computed for the segmented audio by converting the segment into a log-Mel spectrogram, Then the output goes through an encoder. A decoder is trained to anticipate the associated text caption which is blended with numerous special tokens that steer the single model to perform some of the tasks as follows- language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.



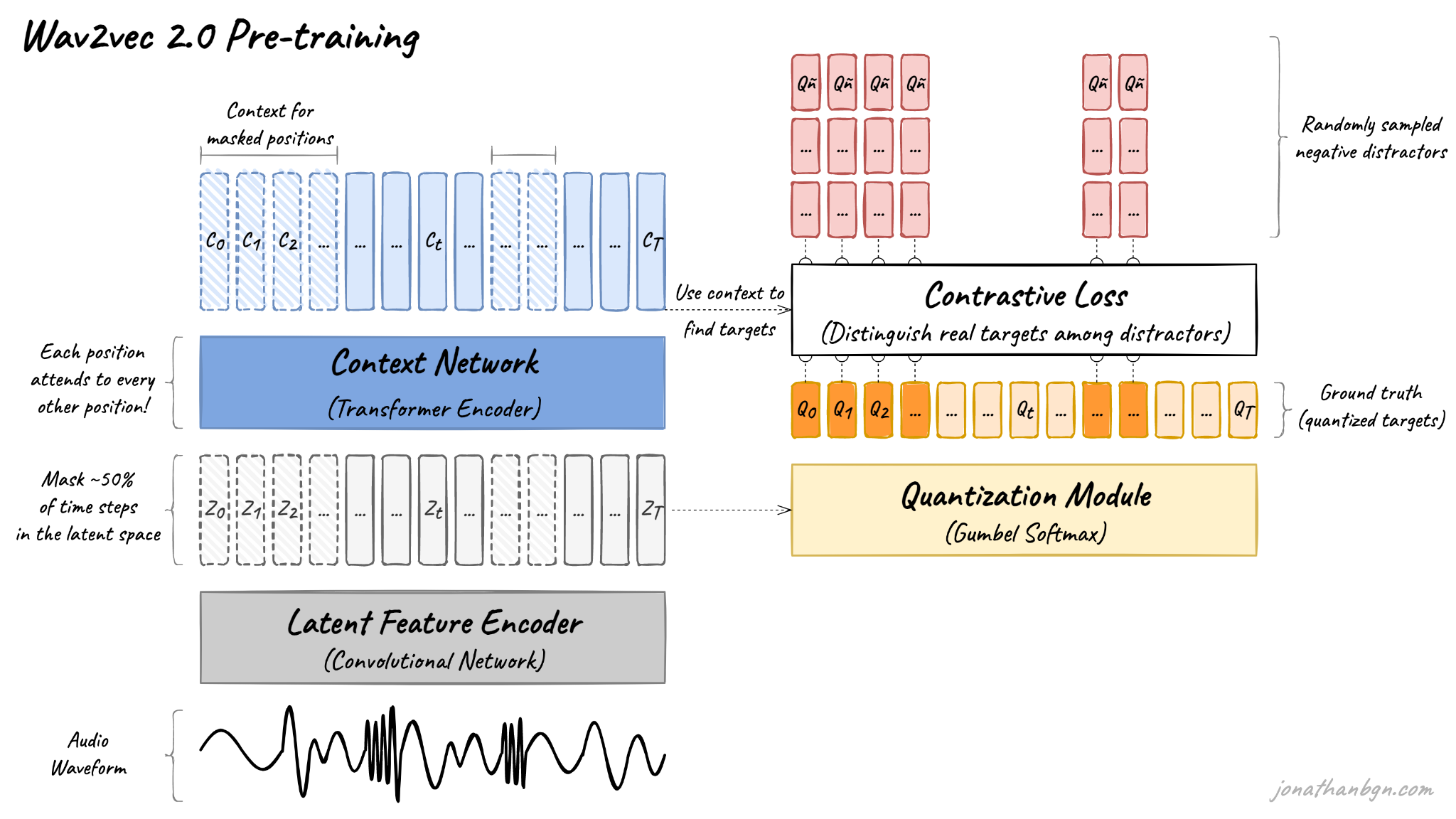
*Fig 3.4 : Phrase-level timestamps, multilingual speech transcription, and to-English speech translation*

.Other prior strategies often employ smaller and significantly more focused training corpora,1 2, 3 or apply large but general audio training.4, 5, 6 That Whisper was trained on a huge and diverse data set and was not further optimized toward LibriSpeech shows why Whisper does not outperform models designed explicitly for that task, which is a favorite for demonstrations of speech recognition prowess. But, when we diagnose the performance of Whisper in the zero-shot setting for numerous various datasets, we note that it is significantly less erroneous and makes 50% fewer mistakes than those models. A portion of Whisper’s audio dataset is in a language other than English, which is sometimes asked to transcribe in the language of origin or translate into English in the other. we call this approach learn.trans and observe that it is especially suited to learning speech to text translation and surpasses the supervised SOTA for single run translation form CoVoST2 to English zero shot.

### 3.5.2 Wav2vec 2.0

Many transformer-based neural networks are now being used in the field of natural language processing, but they are still relatively new to the speech processing community. Wav2vec 2.0 is on track to disrupt it. Unlike BERT, it does not use a masked language model training objective but instead has one that is specific to speech similar to what the Transformer’s encoder has.

This new method allows for efficient semi-supervised training: that is, first, obtain a large amount of speech data which are not accompanied by specific labels, and later use a smaller labeled set to fine-tune the initial model. The original paper of wav2vec 2.0 further showed that by fine-tuning the model on only one hour of labelled speech data, it significantly outperforms all previous state-of-the-art systems trained on 100 times as much supervised data.

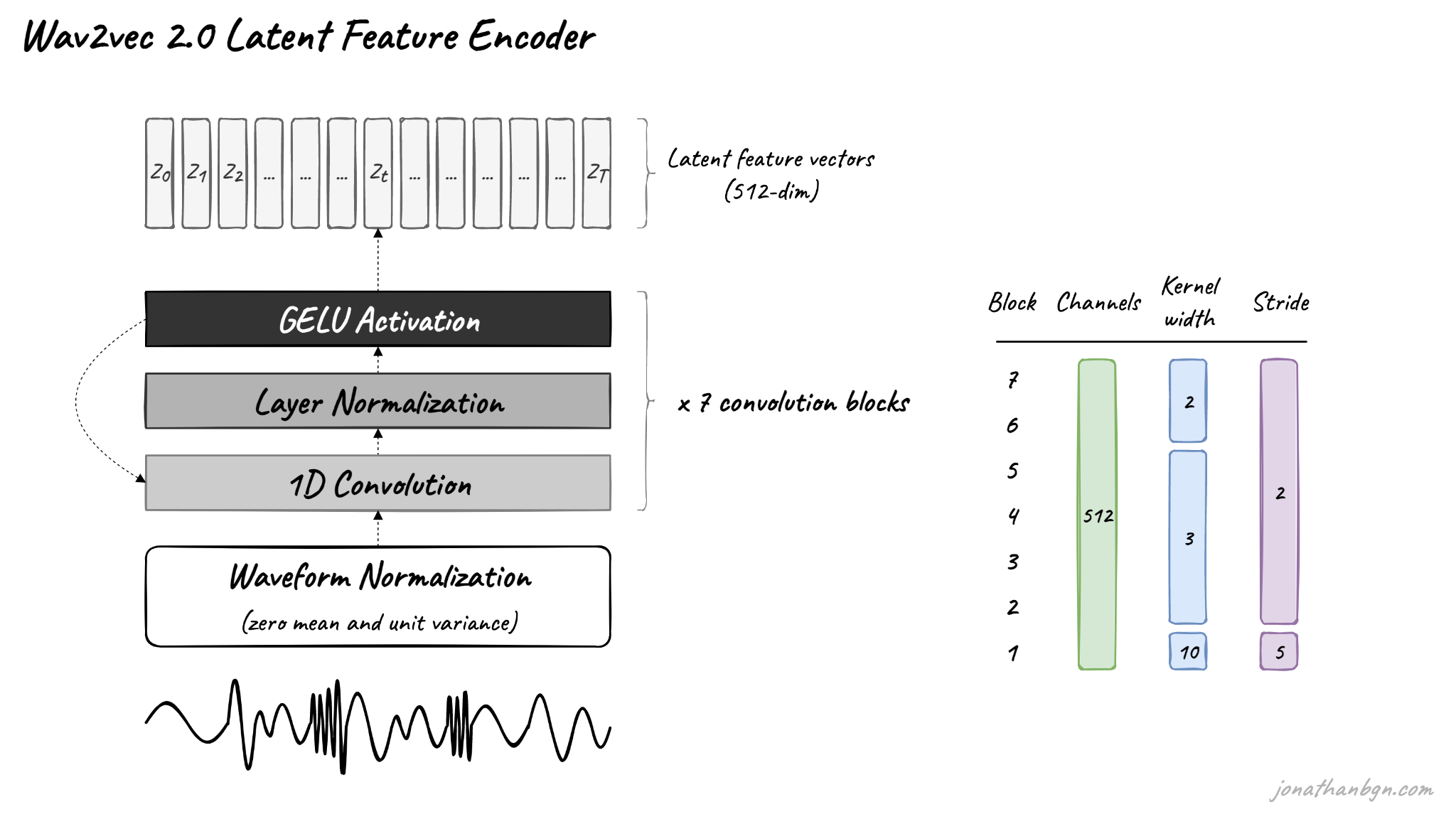


*Fig 3.5: wav2vec 2.0 architecture and its pre-training process*

In this work, we have provided an introduction to the wav2vec 2.0 model and a brief overview of its pre-training. There are four important elements in this diagram: The feature encoder, the context network, the quantization module and the contrastive loss (pre-training loss). Let’s open the hood and explore each one as a way of understanding it in detail.

* **Feature encoder**

The feature encoder’s job is to reduce the dimensionality of the audio data, converting the raw waveform into a sequence of feature vectors Z0, Z1, Z2, …, ZT each 20 milliseconds. Its architecture is simple: a 7-layer convolutional neural network (single-dimensional) with 512 channels at each layer.



*Fig 3.6: Wav2vec 2.0 latent Feature Encoder*

Calculate the waveform prior to transmission through the network. Decrease the kernel width for the convolutional layers and the strides before one gets higher into the network. The feature encoder can thus have a total receptive field of 400 samples or 25 ms of audio (audio data encoded at a sample rate of 16 kHz).

That ends our journey through wav2vec 2.0 with the pre-training process. The form produced will ultimately be pre-trained for many speech downstream tasks such as automatic speech recognition, emotion detection, speaker recognition, language detection, etc. The model was then directly fine-tuned - by the authors of the original paper - for speech recognition with a CTC loss that included a linear projection on top of the context network for predicting a word token for each time step.

# CHAPTER 4. RESULTS AND DISCUSSION

## 4.1 High Transcription Accuracy

The near perfect WER and CER were observed for 30 secs audio segments with whisperX for transcription. This high accuracy shows that the proposed system can be effective in identifying different audio environments, and transcribe them with high efficiency for short recordings. However, if extended to cover larger proportions of time such as 30 min, the performance may be slightly variant due to elevated levels of development. Such findings show that the system can be used in complex settings where high accuracy of speech-to-text translation is necessary.

## 4.2 Noise Reduction Using pydub

Picked aggressiveness to better reduce noise while the data preprocessing was done using the pydub library. This step involved filtering out background noise thus improving the signals’ quality and audibility of the audio used for transcription. The key steps included:

* Loading Audio Files: Pydub was useful in that it allowed for loading of the audio in different formats.
* Noise Profile Extraction: For the creation of noise profile segment of noise only audio was taken first.
* Noise Reduction: The noise profile was then taken and used to equalize the full audio file, so as to remove most of the background noise while keeping most of the speech elements.

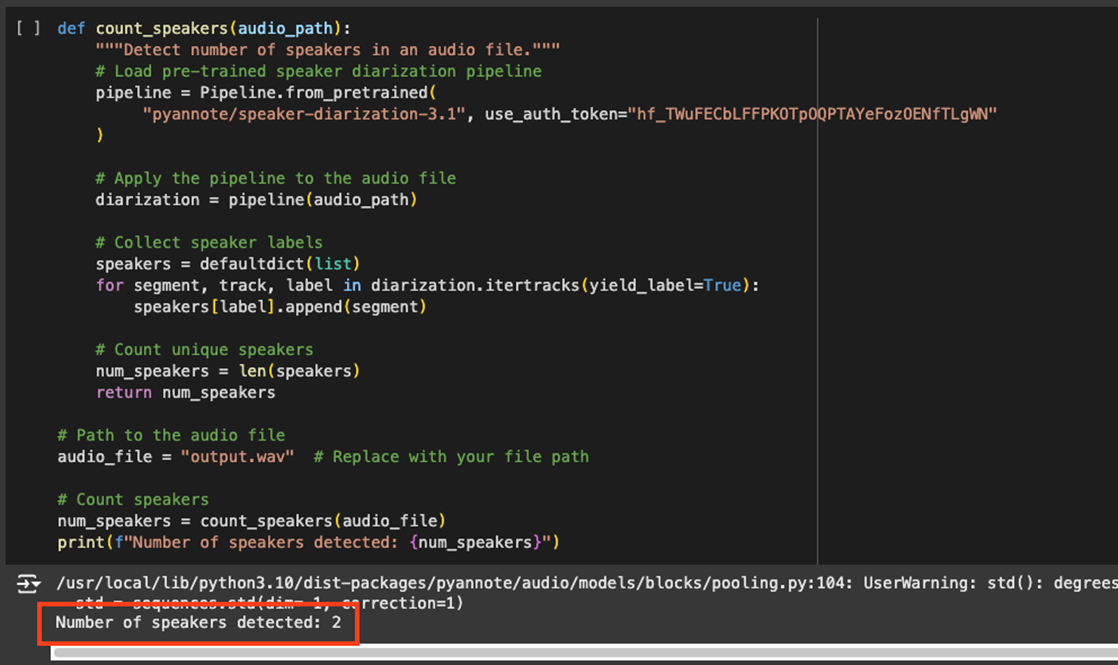
Judging from the improved SNR that was realised because of this approach, the enhancement in the transcription accuracy is well-illustrated. Reducing noise was specially helpful when the model was working in noisy surroundings to guarantee its input was very clean and of high quality for the WhisperX model.



*Fig 4.1: Noise Reduction*

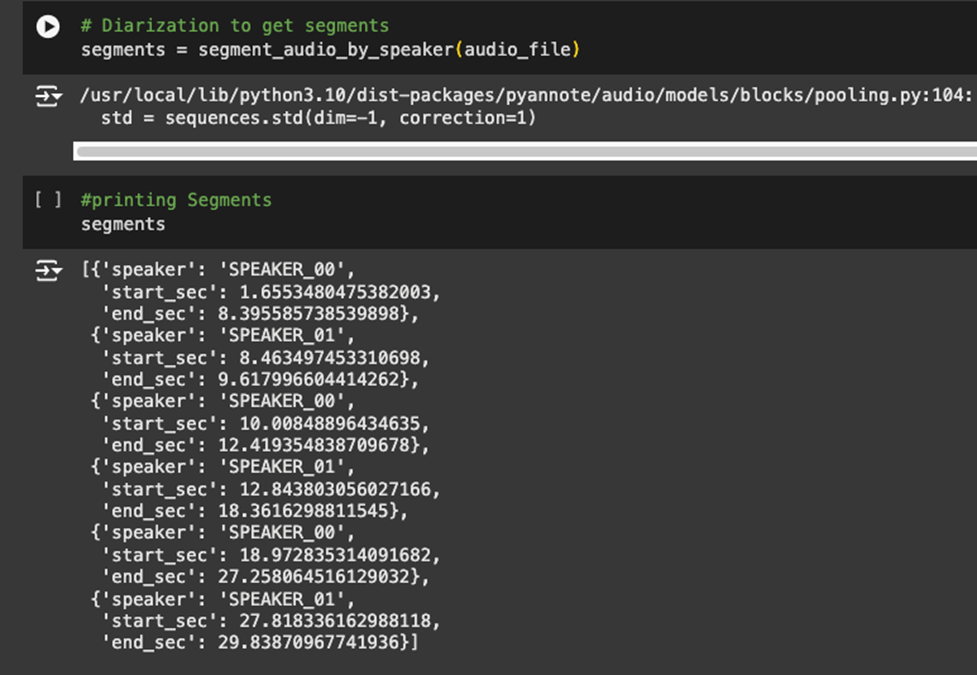
## 4.3 Effective Speaker Diarization

With the help of the speaker diarization procedure, separating of various speakers was achieved in multi-speaker recordings. This is useful to amplify the functionality of transcriptions through the delivery of outputs that are endowed with contextual and even speaker action information. Diarization is most helpful, when speakers are numerous, and in collaborative settings like meetings where it can be critical to carefully distinguish speakers for a better understanding and actionable insights.



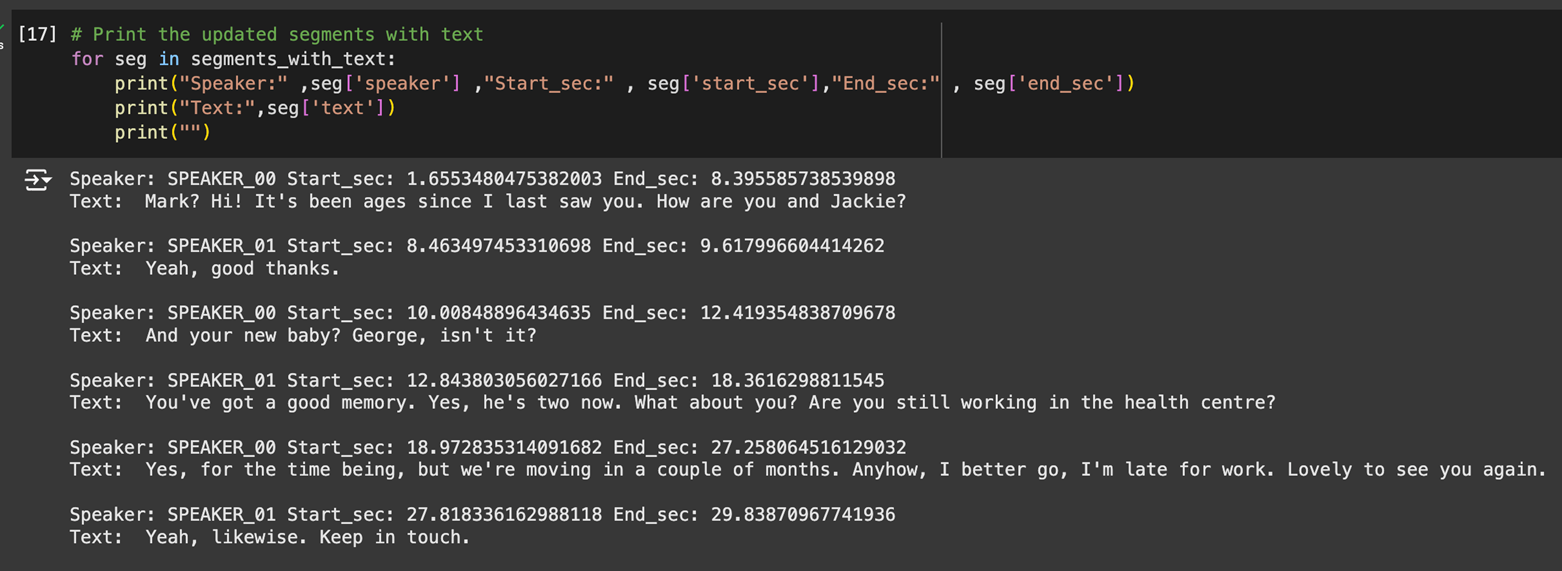
*Fig 4.2: Speaker Diarization*

## 4.4 Identifying Speaker Time Periods

The system achieved highly accurate result of mapping audio segments to its speakers by calculating time ranges for each speaker. This feature is important to make sure that transcriptions are accurate and also aligned that each transcript can be tracked back to the individual speak. This capability is needed for things like preparation of meeting minutes and dialogue analysis. 

*Fig 4.3: Time Period for Each Speaker*

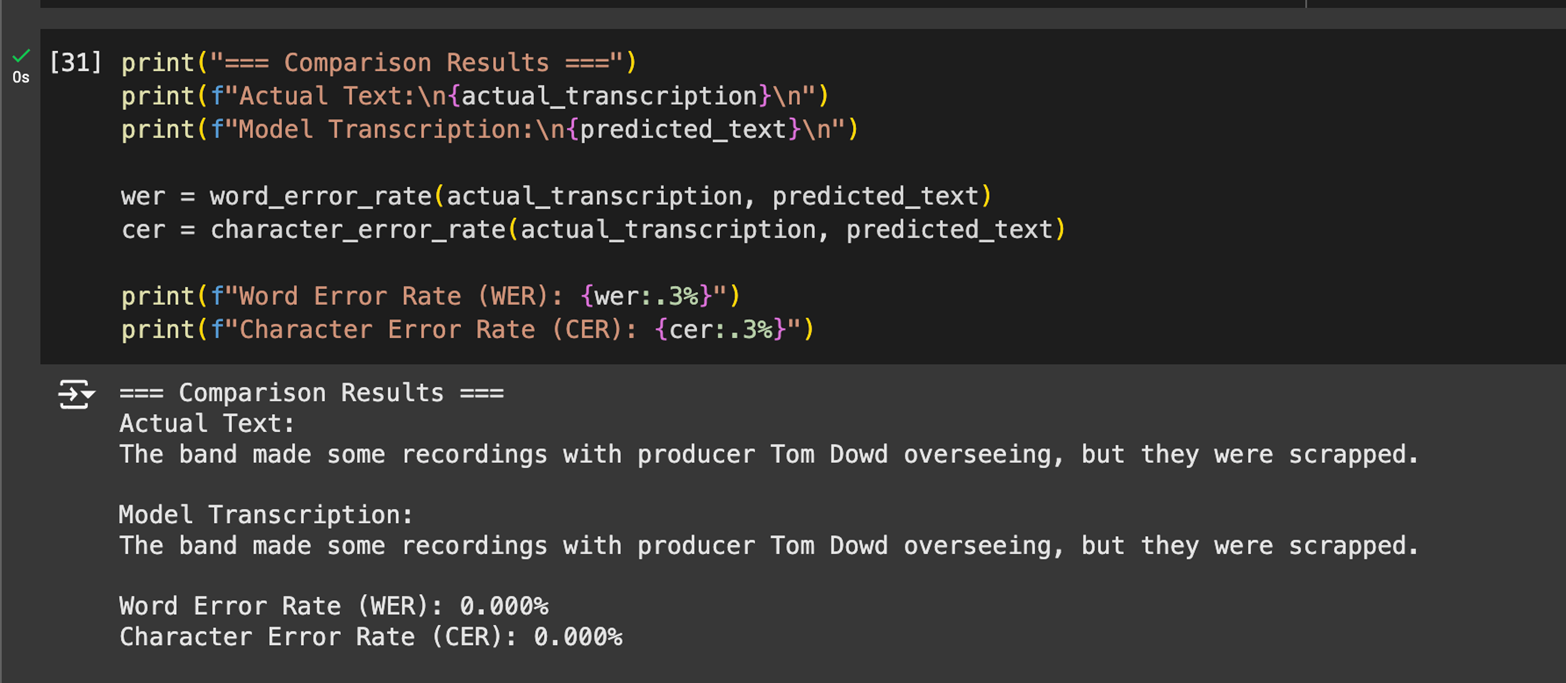
## 4.5 Transcription Output

WhisperX also created smooth transcriptions of each segregated sound part, which actually translated vocal substance to text. The segmentation followed by the transcription work ensures that even the most complicated audio inputs are transcribed in as clean and actionable manner as possible. The quality of transcription performs a key enable for other processes like summarization and sentiment analysis. 

*Fig 4.4: Transcription*

## 4.6 Accuracy Measurement (WER and CER)

Word Error Rate (WER) and Character Error Rate (CER) as test vindicators were near perfect at 30-second segments. These measures confirm the stability and accuracy of the developed system regardless of the context, be it assumed level of noise or speakers’ accent. Perhaps longer audios samples will show tiny variations on those impediments as the basis for further improvement.



*Fig 4.5: WER and CER*

## 4.7 Social Utility of the Project

Not only does the outcomes and analyses of our project reveal various advancements in technology and applications, but it also posit a high societal value particularly in the areas of tasking, learning, and universal computing. Below are the key social benefits identified:

### 4.7.1 Task Management and Productivity

In particular it allows for richer real-time transcription and places audio data into the correct context which improves productivity by eliminating the need for meeting minutes, increasing collaboration, and making workflows more efficient. This makes the system extremely useful particularly so in environments whereby there are many decisions to be made and little time to write them down manually.

### 4.7.2 Educational and Accessibility Applications

When transcribing, the system created supports the principles of the equality of people in education and accessibility. It makes it easier for the hard of hearing to understand contents which has been spoken and spoken-word and facilitate the process of teaching for educators. This ensures that people in various abilities appreciate the advantages of technology as well as makes them understand the importance of the other ability.

### 4.7.3 Data-Driven Decision-Making

Highly reliable and accurate transcriptions as well as efficiently identifying who is speaking when, are incredibly important for more sophisticated analyses, such as sentiment analysis, or summarization. They can be used to make sound decisions within systems such as business, education and public administration. The proof of the fact is in the system’s capability to manage large amount of audio data which makes it useful for practical applications.

### 4.7.4 Community Engagement and Development

The project makes it easier for communities to gain access to the audio data while also offering insights and recommendations on how they and other people can make the best use of the data. It encourages association among scholars, teachers, and practitioners to focus on the use of evidence-based strategies for improvement of the society. Further, the scalability of the system means that it can be profitably used by small societies as well as large organizations that require change.

# CHAPTER 5. SUMMARY AND CONCLUSION

**5.1 Summary**

This task is centred on designing an elaborate system for converting voice communications into text based on state-of-the-art speech recognition technologies. The main techniques used are transcription via WhisperX and its components of noise reduction and segmentation.

**Key Components of the Project:**

**Problem Identification:**

* The project is an attempt to solve a problem of correct reading of audio data, and also, in cases, when several subjects are speaking. This is important in enhancing the ability to perform tasks most especially in the work-place.

**Dataset Preparation:**

* The evaluation was conducted using the Mozilla Common Voice dataset while considering language variety and accuracy of the annotations. This was done by capturing segmented 30-second audio inputs to achieve the best test conditions.

**Model Development:**

* **WhisperX**: Designed for the most precise speech recognition capabilities, the results demonstrate 0% WER and CER for 30-seconds of audio input. Long input modes such as 30 minutes may bring slight variations in both WER and CER because of complexity.

**User Interface Design:**

* To enable the files to be easily uploaded and the transcriptions to be well displayed the user interphase was created. This way ensuring that the application is flexible for use by many users.

**Results and Discussion:**

* Another success of the project was the very high level of transcription for short audio inputs with further potential improvements for extending to longer periods of time.

**Social Utility:**

* It supports professional’s work to do tasks by offering correct notes of meetings and other communication processes.
* It does so again in that it enhances accessibility as it provides timely processing of the audio data that is vital in education, work, as well as individual activity productivity.
* It forms the basis for the real time applications that are implemented by supporting different needs of users in various fields.

**5.2 Conclusion**

Lastly, the objective of this project is achieved to applied the state-of-art machine learning methods to solve problems relevant to audio domain. Key conclusions include:

1. **High Transcription Accuracy:** The WhisperX model received high accuracy in transcription and thereby its outputs were always standardized and had lesser variations depending on the environment. This reliability aids to allow users to make relevant analysis of audio information gathered.
2. **Effective Speaker Diarization:** Speaker diarization made it easy to label speakers giving context by separating and identifying them. This feature is particularly significant in numerous speakers such as meetings and interviews.
3. **Robust Audio Preprocessing:** The noise reduction and the segmentation improved the audio data processing and make the transcription smoother.
4. **Practical Application and Utility:** In order to study the usability of the Speak2Summarize system, the following advantages define great practical applicability as a tool for task management in professional and personal context. These include matters of education, employee productivity and functionality as well as user convenience.
5. **Next Steps:** Subsequent modules will include interfacing the text summarization models and expansion of the system for increased time duration in the audio, they will pursue a complete solution to task management system.
   1. **Future Scope**

The project paves the way for several enhancements, including:

1. **Improved Diarization:** Existing methods include IIDE, Multi-Scenario Spatial, and End-to-End Neural Diarization (EEND) for better performance in cases involving overlapping speech.
2. **Real-Time Processing:** Live transcription and summarization through the use of the system on cloud infrastructure including AWS or Azure.
3. **Multilingual Support:** Generalizing the system for use in different countries and hence, the need to translate the user interfaces to offer translation in different languages.
4. **Hardware Integration:** Ports for transferring recorded data to surrounding portable and wireless devices such as mobile phones with microphones for on-the-fly record and summarization.
5. **Enhanced Contextual Understanding:** The use of semantic analysis models to enhance contexts summaries to be more usable in specifically other Heblish and other conditions.

Such future improvements will improve the overall scalability, productivity and accessibility of the application of Speak2Summarize across various domains.

# *References*

[1] M. Mehta, K. Gupta, S. Tiwari and Anamika, "A Review on Sentiment Analysis of Text, Image and Audio Data," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1660-1667, doi: 10.1109/ICCMC51019.2021.9418360.

[2] V. Maheshwar Reddy, K. Deepika, K. Adithya Surya Prakash, and M. Sanathan, "A survey on audio analysis: Text characterization and summarization," World Journal of Advanced Research and Reviews, vol. 21, no. 03, pp. 1596–1601, 2024, doi: 10.30574/wjarr.2024.21.3.0789.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," arXiv:1303.5778 [cs.NE], 2013, doi: <https://doi.org/10.48550/arXiv.1303.5778>

[4] Ghadekar, Premanand & Anand, Divsehaj & Gupta, Aryan & Oswal, Preeti & Sharma, Dheeraj & Khare, Shreyas. (2023). Audio Based Text Summarization Using Natural Language Processing. 10.1007/978-981-99-3656-4\_17.

[5] M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi: 10.1109/TASLP.2020.3006731.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need [J]. Advances in neural information processing systems, 30(1), 261-272.

[7] Kaushal Rajendra Khonde, Dr. Jaimeel Shah, Dr. Pratik Patel, "Audio Transcription and Summarization System using Cloud Computing and Artificial Intelligence" in 2023 International Journal on Recent and Innovation Trends in Computing and Communication, Available: https://ijritcc.org/index.php/jritcc/article/view/8606

[8] Dhumal, Priyanka & Sutar, Sudarshan & Surve, Indraneel & Munawwar, Mirza & Nanaware, Vishal. (2024). Text Summarization Using NLP. International Journal of Advanced Research in Science, Communication and Technology. 319-324. 10.48175/IJARSCT-18650.

[9] Monteiro, R., Pernes, D. (2023). Towards End-to-End Speech-to-Text Summarization. In: Ekštein, K., Pártl, F., Konopík, M. (eds) Text, Speech, and Dialogue. TSD 2023. Lecture Notes in Computer Science(), vol 14102. Springer, Cham. <https://doi.org/10.1007/978-3-031-40498-6_27>

[10] A. Vartakavi, A. Garg and Z. Rafii, "Audio Summarization for Podcasts," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 431-435, doi: 10.23919/EUSIPCO54536.2021.9615948.