Speech Recognition and Speaker diarization: A framework for Audio Segmentation

Chanchla Tripathi[1], Hemanshu Waghmare[2], Dhruv Dalvi[3], Rishabh Jain[4], Sanket Asole[5], Yuvraj Chavan[6]
[1-6]Yeshwantrao Chavan College of Engineering, Department of Computer Science and Engineering, Nagpur, Maharashtra, India
Email: {chanchalatripathi, hemanshu.waghmare, dhruvdalvi786, rishabhjain231102, sanketasole7755, yuvichavan968@gmail.com

Abstract*—With the exponential growth of digital audio data, Speech-to-Text (STT) conversion and text summarization have emerged as critical technologies within the domain of natural language processing (NLP). These advancements enable seamless transcription of spoken content into text and distillation of vast amounts of textual data into concise summaries. Applications span various fields, including meeting transcriptions, virtual assistants, podcast analysis, and automated content generation, emphasizing their increasing significance in both academic research and real-world implementation. In this paper, the strategies and findings of the first two of the five total phases in the project which develops a unified framework of the STT and summarization are described. Phase one deals with audio processing, which involves better noise reduction techniques such as use of audio spectral gating and use of long audio files into segments for easy processing. The second phase looks into speech recognition and diarization with the powerful WhisperX model to aspire for better transcription and successful speaker diarization in multi-speaker scenarios. This work's major contributions are: addressing specific steps in the preprocessing and transcription processes, as well as assessing the applicability of the used techniques, and discussing the studied issues. There is a strong synergy between these phases to support integration of transformer based text summarization models for further processing. This paper, in addition to showcasing the existing improvements, is a stepping stone toward reaching a more integrated pipeline for the processing, analyzing, and summarization of the passed audio data in various application areas.

Index Terms: Speech-to-Text, WhisperX, Noise Reduction, Audio Segmentation, Diarization, Text Summarization

I. INTRODUCTION

With the ever increasing era of big data and analytical decision making and the advancement in automation, audio is a key form of information. In business discussions, online classes, podcasts, interviews, and numerous other forms, audio data contains many insights to be gleaned that have to be captured and leveraged. In the context of knowledge extraction and analysis, the organizations of Speech-to-Text (STT) and text summarization have a crucial function of converting raw audio data into textual forms with practical relevance, and reducing extended information into comprehensible summaries. These technologies have proved as a tool of value in enhancing access to resources, performance and retrieval of information of use in different fields.

Text summarization complements STT systems by providing concise overviews of large textual data, saving time and effort in understanding the core content. This is especially useful for summarizing transcriptions from podcasts, speeches, or multi-hour recordings into a format that highlights key points [10].

The evolution of STT and summarization systems underscores significant advancements in computational methodologies:

Earlier STT systems relied on statistical models like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs).[1]. These approaches integrated acoustic, language, and pronunciation models to convert speech signals into text. However, they struggled with variations in accents, languages, and noisy environments, making them less reliable for real-world applications.

The advent of deep learning introduced neural network-based models, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks[3][6]. Contemporary models, such as transformer architectures like OpenAI's Whisper, have further enhanced transcription accuracy, processing speed, and adaptability to multiple languages and accents.

Early summarization techniques were rule-based, relying on manually crafted heuristics to extract or generate summaries. These methods often lacked contextual understanding and flexibility [4]. Modern summarization employs neural architectures like BERT, GPT, and T5, which enable abstractive and extractive summaries with contextual awareness, coherence, and grammatical correctness [8][5].
These advancements have paved the way for more efficient and accurate systems capable of addressing the growing demand for audio processing and information summarization[12].

II. METHODOLOGY

For this project, with regard to the methodology the research consisted of two major steps, namely audio preprocessing and speech recognition with speaker diarization. These were achieved in order to create manageable phases for audio data, for purposes of transcription and also to create foundation for text summarization phase. This paper focuses on audio preprocessing, segmentation and diarization and speech to text system implementation.

Figure 1: Workflow diagram

In the first phase, instances of 'pre-processing' of audio data were employed with the aim of smoothening the raw audio data through noise reduction and segmentation. Noise reduction endeavoured to increase the Signal to Noise Ratio (SNR) by eliminating interfering background noise while maintaining speech components. This process employed spectral gating, done using the noisereduce library which computes the Fourier transform to determine the noise frequencies to be eliminated. Audio files were managed using the pydub package which made it easier to load, manipulate and write data. The procedure included loading the audio file, showing the

spectra to discover the noise patterns, applying the spectra gating to eliminate these frequencies and saving the corresponding denoised audio for further processing [2]. By keeping the harmonics under control but letting through the necessary peaks and troughs of speech and other sounds, this approach eliminated many of the problems that typically arise in similar scenarios, such as overlapping speech and the presence of background noise in recordings made in the field.

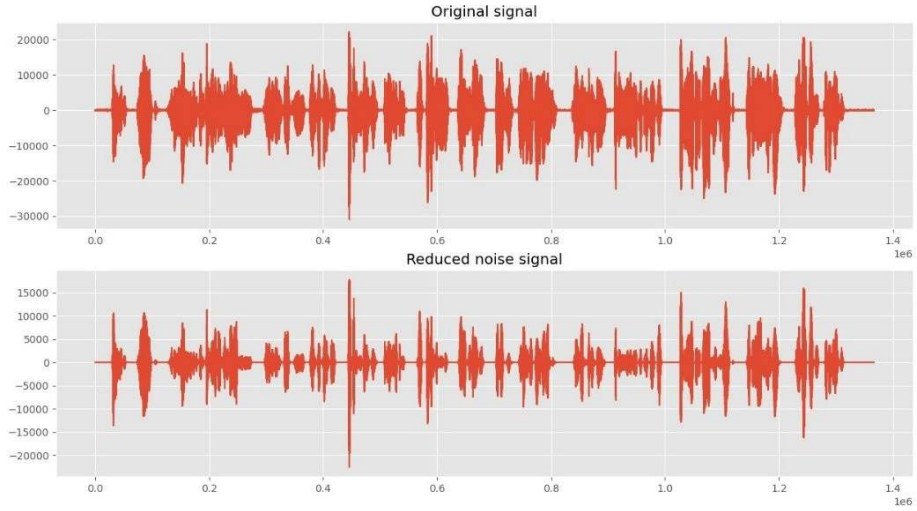

Figure 2: Original signal vs Reduced noise signal

Another important part of preprocessing was the audio segmentation. This step helped to split the received long files of audio data into more reasonable parts to improve memory utilization and the performance of the transcription models. For segmentation we used pydub to search for silences in the audio that were more than two consecutive seconds. The breaks were taken at such points so that every part contained complete thoughts which could be just a sentence long in some few instances. The flashes of segments were produced and these segmented files were stored specifically and waiting to be fed into the transcription model. This approach allowed minimizing the amount of calculations required and enhancing the processing of real-time transcription; it was applied in cases of long recordings or overlapping of speakers[11].

The second phase of the project was speech recognition and speaker diarization where the segmented audio files were translated into textual form but also demarcating different speakers. To this end, WhisperX, an advanced Speech-to-Text (STT) model based on the Whisper model of OpenAI, was used. To get the highest transcription quality and to work well for a range of accents, languages, and levels of noise, WhisperX utilizes transformer-based layers [9]. Given that the data was in segmented audio form, running them through WhisperX served to give time-aligned transcriptions. These outputs were further checked and verified by a manual process by comparing them with reference datasets.

Moreover, speaker diarization was incorporated into the WhisperX pipeline to address the problem of two or more speakers in a video. This process involved extraction of acoustic features from the audio, clustering of speeches into segments and labeling of speakers to the transcribed texts. Diarization helped to make the transcriptions informative on a contextual level to a significant extent, which makes them very valuable in tasks such as meetings summarization and podcast-episode analysis [10]. WhisperX for transcription and diarization provided accurate and context-specific results that created a foundation for the following steps in the project.

III. RESULTS AND DISCUSSIONS

Measures aimed to reduce noise contributed a positive improvement as far as transcription clarity was concerned. Moreover, whereas organic approaches to speech canonicalization tended to mute critical spectral constituents, the spectral gating confirmed such crucial bands to provide STT with a pristine input signal.

The authors also noted that segmentation was most useful in processing lengthy recordings for there was minimal strain on the memory. This step was necessary to guarantee WhisperX worked with manageable amount of audio segments plus did not slow down

WhisperX proved to provide very accurate transcriptions and to display the ability to recognize different speakers in complex environments. Diarization improved the usability of the transcriptions because it added the level of speaker-customary to the transcriptions. The Audio used for the testing had two speakers and the diarization model successfully identified the two speakers along with the duration they spoke for. Table 1 shows the result of the diarization.

Speakers	Start (in seconds)	End (in seconds)	Text
SPEAKER_00	1.6553480475382003	8.395585738539898	Mark? Hi! It's been ages since I last saw you. How are you and Jackie?
SPEAKER_01	8.463497453310698	9.617996604414262	Yeah, good thanks.
SPEAKER_00	10.00848896434635	12.419354838709678	And your new baby? George, isn't it?
SPEAKER_01	12.843803056027166	18.3616298811545	You've got a good memory. Yes, he's two now. What about you? Are you still working in the health centre?
SPEAKER_00	18.972835314091682	27.258064516129032	Yes, for the time being, but we're moving in a couple of months. Anyhow, I better go, I'm late for work. Lovely to see you again.
SPEAKER_01	27.818336162988118	29.83870967741936	Yeah, likewise. Keep in touch

Table 1. Diarization and Speech to text results for 30 seconds Audio

The diarization and speech-to-text conversion results presented in Table 2 were achieved using low-error-rate models on a 5-minute audio file.

Error Rates	Percentages
Word Error Rate (WER)	9.179 %
Character Error Rate (CER)	3.147 %

Table 2. Error rates for 5 minutes audio

IV. CONCLUSION

It is imperative to note that the current research presents an initial accomplishment in the development of the first two phases of STT and text summarization pipeline. Audio segmentation proved very useful to help with the transcription phase, as it helped managing memory loads and processing long recordings. It was highlighted that WhisperX successfully performs source-to-output speech-to-text conversion with high accuracy irrespective of the scenario concerning multiple speakers. The inclusion of the speaker diarization step improved the context by segmenting and labeling the speakers in the transcriptions.

Such improvements set a strong groundwork for the following phases that will deal with transformer-based summarization models and will include aspects of the system's scalability, faster processing time, and support for multiple languages.

These problems are essential for proper audio processing and this project being based on modern STT and summarization helps in the field of natural language processing. And when the system will be designed then the vast audio data can be dealt with effectively with this system ranging from accessibility, automation to content analysis in various fields. This way, eliminating the outlined challenges and following the proposed future directions, the project envisages the development of a scalable and versatile solution for the tasks connected with audio data processing, transcription and summarization.

V. CHALLENGES AND FUTURE DIRECTIONS

In the preliminary processes of the design implementation, there is a list of some challenges that emerged during the work that need additional enhancement for better effectiveness. One of the main difficulties was in handling recordings made directly from various outdoor sources with impairments. Quite its advantages, cacophonies with traffic, winds or even simultaneous talkers raised considerable problems. Noise in these circumstances was normally at frequencies similar to human speech, and attempts to exclude noise resulted in the exclusion of speech signals as well as reduced transcription, especially in field recordings or during public speeches. The next problem concerned overlapping speech in speaker diarization. While testing WhisperX it demonstrated good results in most cases with the disadvantage of segmentation and attributions when several speakers simultaneously spoke. This matter was most noticeable in discussions in dynamic groups, panel discussions, and ordinary meetings with interruptions. Further, the company aspect was seen to pose some challenges; this included scalability and efficiency – as the management of large datasets was computationally intensive especially during preprocessing and transcription.

Existing techniques, which work well on comparatively small dataset, do not scale up well for the enterprise grade or real time use. To respond to these threats the following potential future directions have been defined. Transformer-based text summarization models, including BART and T5, will be integrative since they have proven to have a massive impact. It is believed that these models will be capable of producing intuitive, accurate summaries of spoken content from transcriptions with the help of pre-trained models – fine-tuned on corresponding domain-specific datasets for improved accuracy and context relevance.

Further advancements in this segment of speaker diarization will be made based on eliminating the overlapping speech through superior clustering methods with the help of deep learning techniques like EEND which will manage multi-speakers well [9]. Further, the performance of the system using larger and more diverse data sets will improve the generalization of the data. These datasets will contain multiple languages, accents and audio qualities obtained from actually implemented and used systems like call center records, conference and interview records. The last step will focus on online processing and deployment on cloud environments such as AWS or Azure to achieve live transcription and summarization [7]. This will also help in scaling the above solutions for both the enterprise and the individual users. Such efforts are designed to bring improvements to the emergent characteristics as well as expand the range of contexts where systems may be useful in response to shifting needs and requirements.

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.

[11] Tripathi, C. A., Panchbhai, V. V., Damahe, L. B., Shirole, M. R., Tiwari, S., Rathi, R., & Varma, P. (2024, August). A review of techniques for semantic understanding of the text with term weighting. In AIP Conference Proceedings (Vol. 3139, No. 1). AIP Publishing.

[12] Tripathi, C., Sambare, A. S., & Mahakalkar, N. S. (2016). K-Nearest Neighbours (K-NN) Approach Based on Network Summarization.