INSIGHT EXTRACTION USING MULTIMODAL ANALYSIS OF CONSUMER VIDEOS

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology
in
Computer Science Engineering

by

Akshat Swaminath

20BCE2231

Under the Guidance of

Dr. Uma Priya D

Assistant Professor Sr. Grade 1

School of Computer Science Engineering (SCOPE)

VIT, Vellore

May, 2024

DECLARATION

I hereby declare that the thesis entitled “INSIGHT EXTRACTION USING MULTIMODAL ANALYSIS OF CONSUMER VIDEOS" submitted by me, for the award of the degree of *Bachelor of Technology in Computer Science Engineering,* to VIT is a record of bonafide work carried out by me under the supervision of Dr. Uma Priya D Ma’am.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date:10th May 2024

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled "INSIGHT EXTRACTION USING MULTIMODAL ANALYSIS OF CONSUMER VIDEOS" submitted by **Akshat Swaminath 20BCE2231**, School of Computer Science Engineering (SCOPE), VIT, for the award of the degree of *Bachelor of Technology in Computer Science Engineering*, is a record of bonafide work carried out by him/her under my supervision during the period, 01. 01. 2024 to 30.06.2024, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute orany other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion, meets the necessary standards for submission.

Place : Vellore

Date : **Signature of the Guide**

Internal Examiner **External Examiner**

Head of Department
Computer Science Engineering

ACKNOWLEDGEMENTS

I hereby declare that the report entitled “INSIGHT EXTRACTION USING MULTIMODAL ANALYSIS OF CONSUMER VIDEOS” submitted by me for the fulfillment of the requirement for the course CSE1904 Capstone Project is a record of bonafide undertaken by me under the supervision and guidance of Dr. Uma Priya D.I would like to thank my project guide. Without her support and suggestions, this project would not have been completed. I would also like to thank my superiors at Himalaya Wellness Company where I undertook this project as part of my non-PAT internship. Thanks to them I got to know about how the industry works and what are the actual challenges one faces during the application development lifecycle.

Name: Akshat Swaminath

Registration No: 20BCE2231

EXECUTIVE SUMMARY

The paper presents a methodology utilizing Large Language Models (LLMs) to streamline consumer video analysis, reducing processing time from days to a single working day. LLMs handle tasks like video transcription, summarization, and sentiment analysis, capturing both factual content and emotional tone. Experimental analysis compares Automatic Speech Recognition (ASR) models, highlighting LLM-based solutions' effectiveness. Future research may focus on enhancing accuracy and integrating insights into existing business intelligence platforms. This approach promises benefits in market research, product development, and customer service applications. The concise methodology enhances efficiency in extracting actionable insights from consumer videos, facilitating practical implementation.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1. ASR - Automatic Speech Recognition
2. LLM - Large Language Models
3. RNN - Recurrent Neural Network
4. IoT - Internet of Things
5. CRNN - Convolutional Recurrent Neural Network
6. HMM - Hidden Markov Models
7. GMM - Gaussian Mixture Models
8. SVM - Support Vector Machines
9. ANN - Artificial Neural Networks
10. AI - Artificial Intelligence
11. JSON - JavaScript Object Notation
12. META - Meta AI
13. HF - Hugging Face
14. GPU - Graphics Processing Unit
15. AV - Audio-Visual
16. API - Application Programming Interface
17. NLP - Natural Language Processing

INSIGHT EXTRACTION USING MULTI-MODAL ANALYSIS OF CONSUMER VIDEOS

By Akshat Swaminath

ABSTRACT

In the modern era of social media and public forums where consumers can discuss and share their views and opinions about certain products for all the world to see. This gives a huge advantage to the company manufacturing these products to get an idea of what the user wants and what are the areas where they are excelling and some areas where they can make some improvements. For a product-based company, the data from consumers is the most valuable resource that they can use. Companies can leverage this data to better their working and product development.

This project delves into the rich tapestry of consumer experiences by employing a multimodal analysis of video recordings from consumer interviews and group discussions. While traditional qualitative research often relies solely on transcribed utterances, this study transcends the limitations of text by investigating the non-verbal cues embedded within consumer videos. Through a comprehensive analysis of facial expressions, body language, vocal intonation, and contextual elements, we aim to unveil a deeper understanding of consumer motivations, desires, and pain points that may remain unexpressed or even disguised in their spoken words.

The project mainly deals with the transcription of Consumer focus group videos, summarizing the main points of the conversation, and then performing sentiment analysis on the extracted transcripts and deriving insights from the data extracted. Furthermore, after the basic goals are achieved then advanced analysis of the data can be performed such as facial expression analysis and vocal tone analysis.

Keywords – *Multimodal Analysis, Speech Text, Multi-Lingual Translation, Audio Processing, Transcription, Sentiment Analysis, Vocal Analysis*

1. INTRODUCTION

1.1. INTRODUCTION TO THE PROJECT DOMAIN

In today's consumer-focused industries, companies must grasp the intricate preferences, desires, and feedback of their audience to create products and services that truly connect. Traditional consumer research methods often rely on spoken or written feedback alone, missing out on the subtleties conveyed through non-verbal cues and contextual nuances such as facial expressions, body language, and tone of voice. Therefore, there's a pressing need to explore new methods that can tap into the wealth of information found in diverse data sources. Multi-modal analysis acts like detective work for communication, going beyond words to paint a richer picture [1]. Imagine listening with your eyes and feeling with your brain. That's the power of multi-modal analysis, an orchestra of information revealing a deeper understanding. In the context of consumer research, this translates to analyzing more than just the spoken word. This paper embarks on an exploration of consumer insights through the lens of multimodal analysis, focusing specifically on the utilization of consumer videos as a rich source of data. By integrating visual, auditory, and contextual cues, multimodal analysis transcends the limitations of traditional qualitative research methodologies, offering a holistic understanding of consumer behaviors, emotions, and preferences [2].

This research aims to explore the valuable insights from consumer videos using multimodal analysis. This work uses advanced technologies like Automatic Speech Recognition (ASR) and emotion detection to transcribe and analyze videos efficiently. By combining computational techniques with qualitative analysis, we uncover hidden patterns, sentiments, and motivations within consumer interactions captured on video. This paper introduces a solution to the problem of insight extraction that harnesses Large Language Models (LLMs) to streamline the analysis process. By incorporating tools such as Whisper [3] and LLaMA 2 [4], the aim is to reduce processing time significantly. The primary focus is on developing a user-friendly system that does not necessitate extensive technical expertise. The paper demonstrates how LLMs can expedite analysis while requiring minimal human intervention and computational resources.

The major contributions of this paper include:

- Exploring consumer insights through the lens of multimodal analysis, specifically focusing on utilizing consumer videos as a rich data source.
- Applying ASR and emotion detection to transcribe and analyze consumer videos efficiently.
- Demonstrating the efficient analysis of LLMs while minimizing human intervention and computational resources.

1.2. AIM OF THE PROJECT

The project aims to make a system that can assist the human agent to draw insights from Consumer Focus Group Videos.

1.3. OBJECTIVES OF THE PROJECT

This project addresses a need identified that analyzing consumer interaction videos is a time-consuming manual process, taking nearly a week to complete. This project aims to streamline this process by automating tasks and leveraging Large Language Models (LLMs). By utilizing publicly available models like OpenAI's Whisper for video transcription and Hugging Face's LLaMA 2 for summarization and sentiment analysis, the project seeks to reduce processing time and free up valuable human resources for more strategic tasks. The focus is on creating simple, user-friendly solutions that don't require extensive computer knowledge to operate.

1.4. NEED FOR THE PROJECT

The project came in as a requirement from the Global Research Center at the organization where I am undergoing my internship. The main requirement behind this project was to streamline the process of analysis of consumer interaction videos and reduce the processing time taken to do the work manually. According to the information provided by the project stakeholders, when done manually the complete process takes almost one week to analyze and draw insights from the data that can be used to make informed decisions regarding the products. So, the project aimed to automate all the manual tasks and reduce the time so that valuable human time can be put to better use.

1.5. SOLUTIONS

At an individual level, working on the project, I wanted to leverage the upcoming technologies of Large Language Models and make use of publicly available models to come up with simple solutions for the project that can easily deployed and used without the need for the user needing a lot of computer knowledge.

Many publicly available language models can perform the works and be modified based on the requirements of the project. For the Video Transcription part of the project, I utilized the LLM generated by OpenAI named Whisper which performs video transcription and translation. Furthermore, for the video summarization and sentiment analysis of the transcribed data I made use of the LlaMA 2 LLM that is available to use on Hugging Face via an Interface API call to the model.

2. PROBLEM DESCRIPTION

This project makes use of publicly available Large Language Models to reduce the man hours required to draw insights from the Consumer Focus Group Videos and make it easier for the concerned parties to make use of the valuable insights as soon as possible to make necessary changes in the product line based on the reviews from the customer. Keeping in mind the risk of using public LLMs and the privacy restrictions organization, the LLMs that are used for the project are not using the data to train themselves and are not saving any of the data that is being processed through them.

After extensive experimentation with pre-trained models for Audio-to-Text Conversion and Sentiment analysis on the video transcripts, a conclusion could be drawn that the use of pre-trained models will be less effective than fine-tuning an LLM to perform the task required. During the research, I tested out Models like BERT, SentiWRDNet, and VADER Lexicons to perform sentiment analysis on the transcripts with less than satisfactory results. After that for text summarization, I tested out the TensorFlow T5 model to test its capabilities in summarizing the transcripts and the results were like the other pre-trained models that I used for other requirements.

2.1. PROBLEM CHALLENGES

As the project dealt with a global audience and the insight from all the different types of consumers it more valuable for the company to draw specific decisions based on the insights drawn from conversations with certain focus groups. With such a diverse range from where the data can be generated and collected, it gives rise to several challenges.
The following are some of the challenges that drew my attention during the meetings that happened with the stakeholders.

- Long and informal discussions are not very specific and to the point, so sometimes the data is not very clean it needs to be cleaned to make some valuable insights from the data. In the case of the informal conversation part, there are parts of conversations that do not provide any valuable information regarding the product being discussed.
- As the sessions are being recorded in an international office, sometimes the sessions are conducted in the native languages, and the translation of the language to English becomes an additional task.
- Due to the presence of different nationality people, various ascents are being used in the videos, and having a system that transcripts the video with great accuracy is highly required having a low Word Error Rate gives a better advantage as it makes sure that there are not many words being misspelled and do not change the meaning of the sentences.
-

2.2. EXPECTED OUTCOMES

The project has very detailed and clear guidelines that outline what all are the outcomes that need to be met for the project to be considered a success. Being an active project and the development being under process the following are the expected outcomes from the project:

- Being able to transcribe and summarize hour or two-hour-long videos in a faster way, drastically reduces the processing time from data collection to generating and presenting reports to the required officials.
- Being able to perform the sentiment analysis on the conversations and derive insights from the transcripts to draw out conclusions and present them in a report

format.

- Being able to perform advanced analysis such as vocal analysis or facial expression analysis on the video files if possible.
- Getting keyframes in the videos of relevant clips from the hour-long video to clip the videos into smaller parts and present in the reports as is.
- Translation of the video files to English as there are videos recorded in the native language.

3. LITERATURE REVIEW

This section reviews the related works of insight extraction. Table I depicts the summary of the existing research works in the field of insight extraction. The review has considered features such as audio-only and emotions as they play a major role in the translation.

Huddar, M. G. et al. [5], analyses conversation sentiment and emotions using attention and RNN for multimodal data. Alex, S. B. et al. [6], Uses attention and feature selection for speech emotion recognition based on sounds. Garc´ıaOrdas et al. [7], analyses sentiment in audio recordings of ´ varying lengths with a fully convolutional neural network. Trinh Van et al. [8], recognize speech emotions using deep neural networks. Chauhan, S., Saxena, S., & Daniel, P. [9], Improves unsupervised neural machine translation for morphologically complex languages. Chen, G. et al. [10], maximizes cross-lingual transfer for zero-shot neural machine translation. Tang, Y. et al. [11], develops a unified system for speech translation and recognition through pre-training. Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., & Li, L. [12], achieves end-to-end speech-to-text translation with a triple supervision approach. Sun, H. et al. [13], Studies unsupervised neural machine translation for similar and distant language pairs. Pan, X., Wang, M., Wu, L., & Li, L. [14], Applies contrastive learning for many-to-many multilingual neural machine translation. Sudhir, P., & Suresh, V. D. [15], Compares sentiment analysis approaches, applications, and classifiers. Kumar, V. S. et al. [16], Analyze sentiment in speech signals using machine learning techniques. Kumar, A. [17], Uses hierarchical attention networks for sentiment classification in social IoT data. Ahmed, U., Jhaveri, R. H., Srivastava, G., & Lin, J. C. W. [18]creates an explainable deep learning system for sentiment analysis of mental disorders. Kim, C. M., Kim, K. H., Lee, Y. S., Chung, K., & Park, R. C. [19], Analyzes facial expressions in real-time for sentiment using a CRNN model. Park, C. et al. [20], mentions

improvements to speech-to-text conversion by back-transcribing text-to-speech output. Tang, Y., Pino, J., Li, X., Wang, C., & Genzel, D. [21], Improves speech translation by leveraging auxiliary text translation tasks. Zhang, Y. et al. [22], Pushes the limits of semi-supervised learning for automatic speech recognition. Indurthi, S. et al. [23], Perform end-to-end speech-to-text translation using modality agnostic meta-learning. Bell, P. et al. [24], Reviews adaptation algorithms for neural network-based speech recognition.

While existing research has explored the potential of multimodal analysis and advanced technologies in consumer research, there remains a gap in understanding how these approaches can be effectively integrated to derive actionable insights from consumer videos. This paper seeks to bridge this gap by presenting a comprehensive exploration of consumer insights through the lens of multi-modal analysis, while also introducing a solution that harnesses LLMs to streamline the analytical process and enhance efficiency

3.1. RESEARCH PAPERS:

Research Article	Features	Methods used	Limitations
Huddar, M. G. et al. [5]	Sentiment and emotion analysis in conversations	RNN with attention mechanisms	Specialized RNN architecture
Alex, S. B. et al. [6]	Speech emotion recognition using deep neural networks	Deep neural networks with attention	Computational cost
García-Ordás et al. [7]	Sentiment analysis in variable-length audio recordings	Fully Convolutional Neural Network	Limited to audio data
Trinh Van et al. [8]	Emotion recognition in speech using deep neural networks	Deep neural networks (CNN, CRNN, GRU)	Requires labeled data
Chauhan, S., Saxena, S., & Daniel, P. [9]	Enhancing unsupervised neural machine translation	Unsupervised neural machine translation	Limited generalization
Chen, G. et al. [10]	Effective cross-lingual transfer in NMT	Multilingual pretraining and fine-tuning	Requires multilingual data
Tang, Y. et al. [11]	Cross-modal learning for speech translation and recognition	Cross-modal learning with self-supervised and supervised subtasks	Task suitability
Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., & Li, L. [12]	Decoupling speech-to-text translation for accuracy	Triple supervision decoupling	Increased complexity
Sun, H. et al. [13]	Investigating unsupervised NMT effectiveness	Unsupervised neural machine translation	Effectiveness varies
Pan, X., Wang, M., Wu, L., & Li, L. [14]	Improving translation quality across multiple languages	Contrastive learning	Requires large training data
Sudhir, P., & Suresh, V. D. [15]	Exploring sentiment analysis methods	Rule-based, machine learning, lexicon-based approaches	Accuracy trade-off
Kumar, V. S. et al. [16]	Sentiment analysis using multi-modal data	Deep learning	Requires labeled data
Kumar, A. [17]	Sentiment analysis in social media data	Hierarchical Attention Network	Computational cost
Ahmed, U., Jhaveri, R. H., Srivastava, G., & Lin, J. C. W. [18]	Emotional analytics for mental disorder	Deep attention model with active learning	Domain expertise needed
Kim, C. M., Kim, K. H., Lee, Y. S., Chung, K., & Park, R. C. [19]	Real-time facial sentiment analysis	PP2LFA-CRNN model with image encryption	Limited to visual data
Park, C. et al. [20]	Speech-to-text post-processing using BTS	Back TranScription (BTS)	Domain specificity
Tang, Y., Pino, J., Li, X., Wang, C., & Genzel, D. [21]	Enhancing speech-to-text translation	Multitasking learning with speech and text translation tasks	Requires parallel text data
Zhang, Y. et al. [22]	Semi-supervised learning for ASR	Semi-supervised learning with wav2vec 2.0 pre-training	Model size and language effectiveness
Indurthi, S. et al. [23]	End-to-end speech-to-text translation	Meta-learning with modality-agnostic multi-task model	Model complexity and data requirements

TABLE 1: SUMMARY OF RESEARCH WORK

4. REQUIREMENT ANALYSIS

4.1. FUNCTIONAL REQUIREMENT

- Video Input: The system should be able to ingest consumer videos in various formats (e.g., mp4, avi).
- Multimodal Analysis:
- Visual Analysis: The system should be able to extract insights from the video content itself, such as objects, actions, and scene understanding.
- Audio Analysis: The system should be able to analyze the audio track, including speech recognition, sentiment analysis, and background noise identification.
- Text Analysis: The system should be able to process any text overlays or captions within the video for additional information.
- Insight Generation: Based on the multimodal analysis, the system should generate insights about consumer behavior, preferences, and opinions related to the video content.

4.2. NON-FUNCTIONAL REQUIREMENT

- Performance: The system should be able to process videos efficiently and generate insights within a reasonable timeframe.
- Scalability: The system should be scalable to handle a large volume of consumer videos.
- Accuracy: The extracted insights should be accurate and reliable.
- Security: The system should be secure and protect the privacy of consumer data.
- User Interface: A user interface could be developed to allow users to upload videos, view extracted insights, and filter results.
- Explainability: The system should be able to explain the reasoning behind the generated insights, allowing users to understand how the system arrived at its conclusions.

5. EXISTING SYSTEM DESCRIPTION

Automatic Speech Recognition has come a long way since its inception and now when we have Large Language Models (LLMs) at our disposal to fine-tune to make use of them to give us more accurate and fast responses. Before the increase in popularity of LLMs, other systems were used. Here are some of the methods that were used.

- Hidden Markov Models (HMMs): These were the foundation of early ASR systems. HMMs statistically model the sequence of sounds in speech, allowing the system to recognize words based on probabilities. While computationally efficient, HMMs struggle with complex variations in speech and require significant training data.
- Gaussian Mixture Models (GMMs): An improvement over HMMs, GMMs represent each phoneme (unit of sound) with a mixture of Gaussian distributions. This offers more flexibility in capturing variations in speech sounds. However, GMMs still face challenges with noise and require substantial data for training.
- Support Vector Machines (SVMs): SVMs can be used for ASR by classifying speech features into predefined categories corresponding to phonemes. They are advantageous for handling smaller datasets but might not perform as well with highly variable speech patterns.
- Artificial Neural Networks (ANNs): Early attempts at deep learning for ASR employed ANNs. These networks learn to map speech features directly to recognized words. However, the limited architecture of traditional ANNs restricted their accuracy compared to more advanced deep-learning techniques.

6. PROPOSED SYSTEM ARCHITECTURE

The integration of machines to aid in problem-solving represents a recent advancement in corporate business practices. Despite the prevailing belief in the superiority of human interaction for obtaining genuine insights from discussions, there’s an inherent challenge: it’s time-consuming. Manual processing of videos, for instance, can take days, with the traditional analysis of a single video extending over 2-3 weeks. This inefficiency not only hampers productivity but also ties up valuable time that could be better spent planning

various focus group discussions. Figure 1 illustrates the conventional workflow of insight extraction. It is observed that post-recording necessitates human involvement.

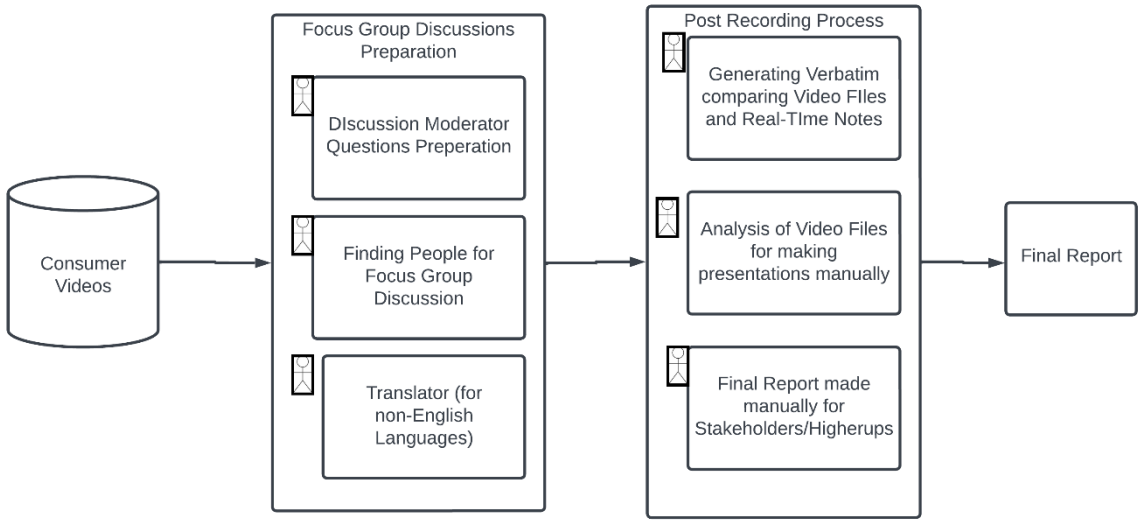

Figure 1: Conventional Workflow of Insight Extraction

The proposed works solve the time consumption problem and make it possible to draw analysis from the videos within a couple of hours of uploading the video, as illustrated in Figure 2. The proposed solution works in two phases:

Phase 1: Focus group discussion preparation

This phase works with the assistance of AI technology, in which the resource person can use Generative AI [25] solutions to make discussion questions. Furthermore, the processes of selecting the candidates for the focus group discussion and recording the video discussions are to be done manually. The candidates are selected by a survey shared with the consumers of the product and selected people in groups based on the answers. In the final deployment, the web portal is used as the upload section where the moderators of the discussion upload the recordings and the discussion questions and notes to give the machine some context to understand the references made in the video.

Phase 2: Post-Recording

The translator is substituted by the Whisper model [3], which can not only translate the video transcripts into English but also produce transcripts of the video file into various file formats such as text file, JSON file, .str file, etc. For getting the verbatim from the video files, we can use the subtitle files that have the embedded timestamps of the recording, which is used to get minute-long clips from the entire recording. LLMs [26] are used in the

next two steps to draw insights from the video transcripts and make summarized points from the entire video file. The translation and transcription now taken care of by the AI solutions, which is the whisper-medium model [3] also helps a lot in reducing the processing time. For the text summarization part, the Llama-2 model [4] is used, which is used for text generation by passing a small part of the transcript acquired by the whisper model. Getting verbatim from the videos is also made easy as it can be put as an extension to the chat interface which returns small video clips of the original discussions. The proposed system, in the final deployment, gives the human agent a chat interface where they can query the system on the video files and get generated information from the system, which has the context of the video being discussed.

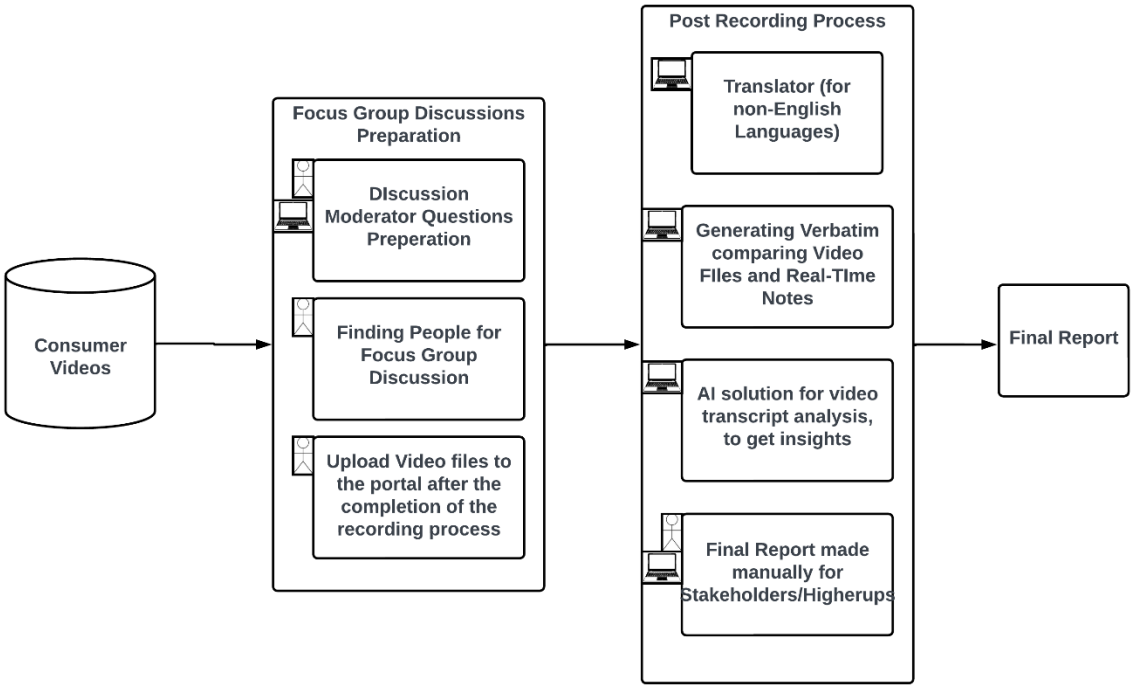

Figure 2: Proposed Workflow

7. MODELS USED

For the scope of this project, we are making use of two of the LLMs which are commercially available and are also available in the Public Domains without requiring any paid licenses. From institutions like OpenAI and META, the contribution that they are putting into the development of the models for public use is commendable. Without the steps from these organizational projects, this would not be possible.

The model used for this project is as follows:

- For video transcription, the model used was Whisper [3], published by OpenAI. Whisper is a speech recognition and translation model developed by OpenAI in late 2022. It supports over 90 languages and has a Word Error Rate in many languages which is acceptable to be considered as an accurate system. The model is trained on a large dataset which consists of 680,000 hours of audio and text from various languages. It has the capability of performing translation from non-English languages to English. It has a very robust system against noise and different accents that are spoken. The technical Architecture utilized by the Whisper is a Transformer-based architecture. A transformer-based architecture is a very powerful neural network model which can be used for sequence-to-sequence tasks. The best feature about this is its open-source nature which not only allows developers to further develop the capabilities of the system but also gives researchers access to explore the latest and the newest in the areas of ASR.
- For the summarization and sentiment analysis part of the project the LLM published by Meta AI is called Large Language Model from META AI or Llama 2 [4], to be more specific the model used for this project is called Llama-2-7b-chat-hf. Expanding upon the name of the model 7b specifies the number of parameters that are in the model and chat specifies that the model is fine-tuned for dialogue tasks. This means it is optimized to be informative and comprehensive in its response and to perform well in a conversation setting. The training data for Llama-2-7b-chat-hf is extensive and diverse, encompassing a broad range of publicly available text and code sources. The fine-tuning process further refines the model for dialogue tasks through dedicated datasets and human-annotated examples, though the specifics of

this data remain undisclosed. Like most of the chat models available that are used to generate response purposes, this model also has a cutoff date for the knowledge base which is September 2022 but this does not concern our project.

8. IMPLEMENTATION DETAILS

8.1. METHODOLOGY

Due to the resource intensiveness of this project to use two large language models, Google Colab Notebooks were used, as it provides a free online environment where we can make use of GPUs such as T4 for free which are essential for the processing of the LLMs.

So, the environment used in this project was a Colab Notebook which uses T4 GPU as an accelerator to make the processing fast. To use the Whisper model, we create a clone from the git repository and then install the packages from the source code within the cloned repository. Also, as the project deals with the use of video and audio files we install FFmpeg, which is a powerful command-line tool for multimedia processing that includes AV encoding, decoding, transcoding, muxing, demuxing, filtering, and more.

Once we have the transcripts, we can use them to get the summary of the transcripts in chunks of 1000 characters in bullet points with sentiment analysis for the given text chunk.

We do this using the LlaMA-2 model and we use some of the Python libraries to make use of this. First, we use transformers, which is a popular library used for working with pre-trained machine learning models, especially those for Natural Language Processing. It provides a function to load, fine-tune, and use these models for various tasks like task generation translation and question answering. Einops is a library used to simplify and optimize tensor manipulations in deep learning frameworks. Accelerate, as the name suggests, helps accelerate the training and interface of deep learning models by supporting features like mixed precision training, distributed training, and automatic organization. Lastly, langchain is a library that is specifically designed for building and managing pipelines that involve large language models. It allows one to chain together different code components like text processing, model inference, and post-processing to create complex workflows.

From here using the inference API provided by Hugging Face which is a service provider for these LLMs and other models which we can make use of for free when made available by the institutions. Hugging Face has the largest collection of pre-trained machine learning models, especially for NLP tasks like text generation, translation, and question answering.

These models are based on different architectures and are trained on various datasets some of which are publicly available. Overall, hugging face plays a significant role in democratizing machine learning by providing open-source tools, resources, and a collaborative environment.

Finally, we use the LangChain to interact with the LlaMA-2 model and use libraries like Prompt Template and LLMChain to give instructions to the model and get a generative response from the machine.

Lastly, Python Database connectivity was used to store the video transcripts and their summaries with the chunk data in an online hosted database (somee.com) where one can query and make insight derivation one step easier.

9. RESULTS AND EXPLANATION

For the analysis of the ASR models, Deepgram Playground [27] was used as it provided a simple interface to use most of the models we wanted to compare quickly. Using the API Playground, the confidence scores for all the models were tested on the video files from the Dinner Party Corpus Dataset [28], which was used as a standard to check the scores for all 5 selected models.

9.1. EXISTING MODEL COMPARISON

Here are the models that are used for the comparison study in this paper.

1) Nova-2 [29]: The deepest-trained ASR model on the market, with 18% more accuracy than the previous Nova model.

2) Whisper-medium [3]: A family of encoder/decoder models trained on a large corpus of multilingual speech data by OpenAI.

3) Nova-General [29]: The deepest-trained ASR model, ideal for voice applications that require high accuracy in different contexts.

4) General-Enhanced [29]: Model useful for high accuracy timestamps, lower word error rates, and use cases that require keyword boosting.

5) Base-General [29]: Model useful for high-accuracy timestamps and large transcription volumes.

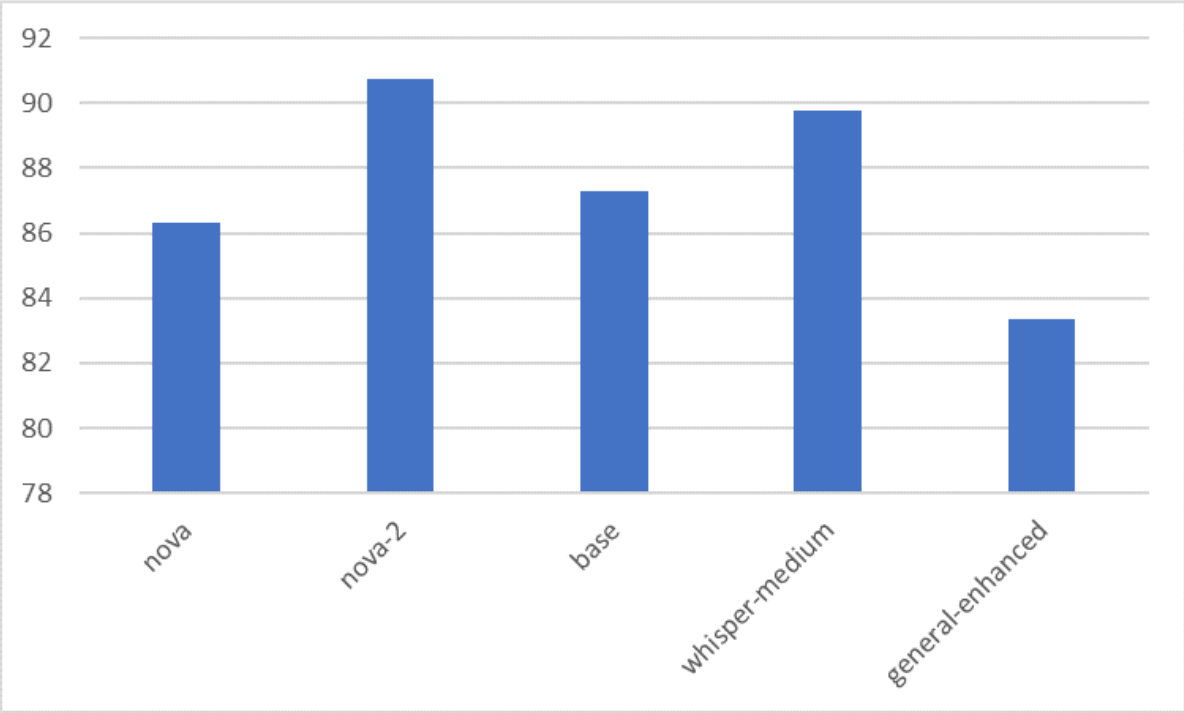

Fig 3: Average Confidence Score for all the ASR models.

9.2. RESULTS

The figures provide insights into the performance of various models tested on the Dinner Party Corpus [28]. The average confidence scores displayed in Figure 3 offer a broad overview of the models' performance across the dataset. These scores are calculated based on word-level confidence scores obtained from the Deepgram API Playground. Figure 4, on the other hand, delves into the detailed confidence scores of each model on a specific video file. This breakdown allows for a more granular analysis, showcasing how each model performs in different segments of the video. Upon analysis, it becomes evident that certain models, such as nova-2, whisper-medium, and nova-general, exhibit higher confidence scores and are more adept at generating accurate transcripts. These models are likely well-tailored for tasks requiring precise transcription, as indicated by their consistent performance across various segments of the video. In contrast, models like general-enhanced and base-general appear to excel in other aspects, such as accurately generating timestamps and handling large transcription volumes. While their average confidence scores might not be as high as those of other models, their specialized capabilities make them valuable for specific use cases, such as processing large volumes of audio data efficiently.

Here we have the detailed confidence scores of all the models on the video file selected, with the number of words in the given confidence range.

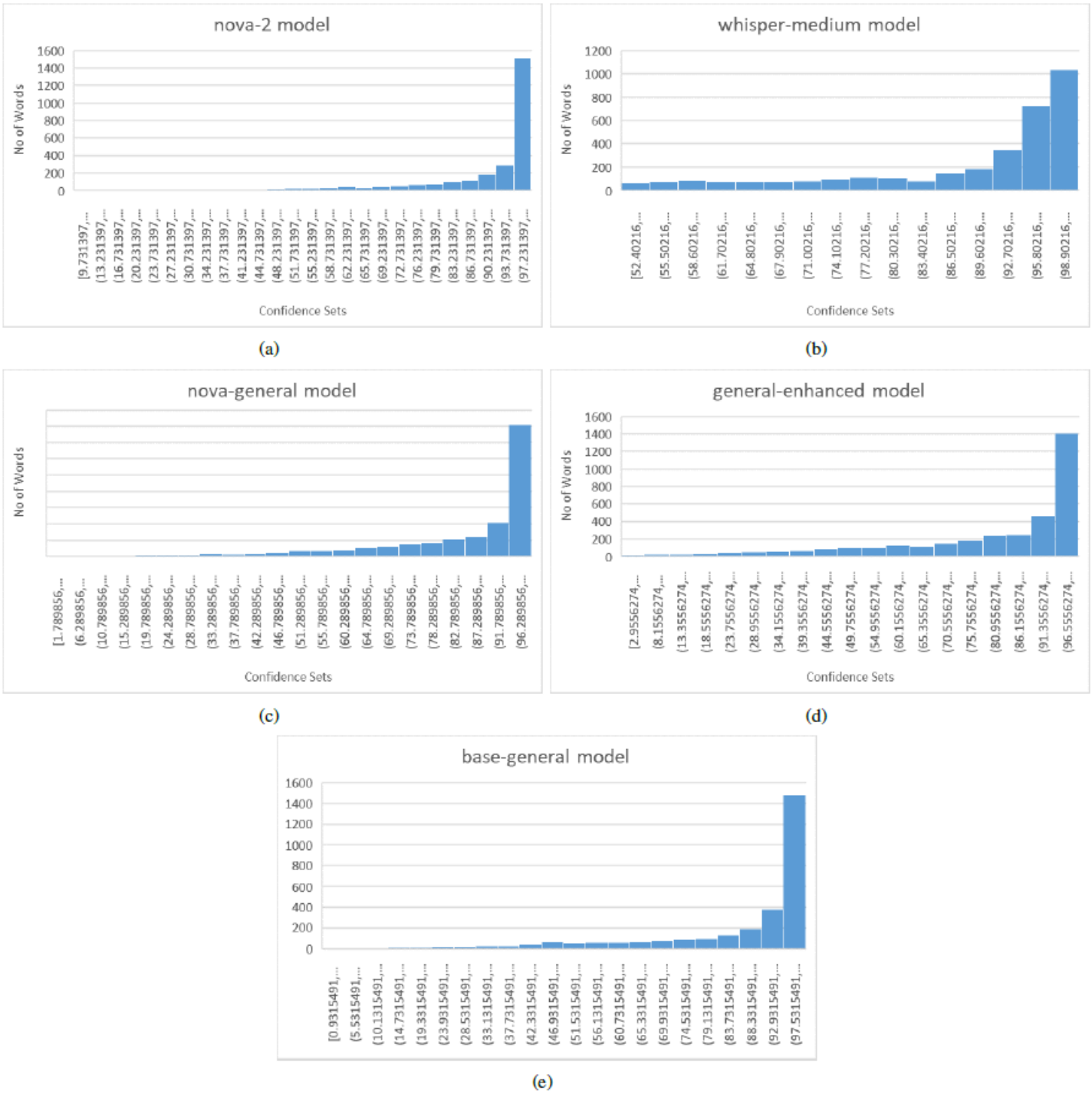

Fig. 4. (a) nova-2 (b) whisper-medium (c) nova-general (d) general-enhanced (e) base-general

10. CONCLUSION

The paper aimed to simplify the video translation process by using Large Language Models (LLMs), which reduced the time needed for various processes from days or hours to within a single working day. One successful use case involved reducing stakeholders' time generating reports after focus group meetings. By leveraging pre-trained LLMs, the project was able to compartmentalize tasks into smaller chunks, enabling the creation of specific models fine-tuned to problems. Specifically, the project focused on extracting insights from consumer videos through multimodal analysis, utilizing LLMs for tasks such as video transcript generation, summarization, and sentiment analysis. This novel approach offers several advantages over traditional methods, including the ability to handle unstructured video data and capture both factual content and emotional tone. The findings demonstrate the effectiveness of the method for extracting actionable insights, which can benefit applications such as market research, product development, and customer service. Future research directions may involve exploring more advanced LLMs and multimodal analysis techniques to improve accuracy and robustness. Additionally, integrating the extracted insights into existing business intelligence platforms could enhance practical applications.

11. FUTURE SCOPE

With the active development of Generative Pre-Trained Transformers and Large Language Models, the future scope for this project is very vast. With the high availability of these technologies to researchers and enthusiasts the developers are making sure that most of the people can get hands-on with the latest technology and people understand how the technology can be helpful to simplify the work. As most of the work used in this paper is open-source people can easily access the resources and work and further help in the development of these technologies.

12. REFERENCES

[1] Y. Guo, Z. Cheng, L. Nie, X.-S. Xu, and M. Kankanhalli, "Multi-modal preference modeling for product search," in Proceedings of the 26th ACM International Conference on Multimedia, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1865–1873. [Online]. Available: https://doi.org/10.1145/3240508.3240541

[2] X. Liu, Y. Liu, Y. Qian, Y. Jiang, and H. Ling, "Learning consumer preferences through textual and visual data: a multi-modal approach," Electronic Commerce Research, pp. 1–30, 2023.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[5] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multi-modal sentiment analysis and emotion detection in conversation using rnn," 2021.

[6] S. B. Alex, L. Mary, and B. P. Babu, "Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features," Circuits, Systems, and Signal Processing, vol. 39, no. 11, pp. 5681–5709, 2020.

[7] M. T. Garc´ıa-Ordas, H. Alaiz-Moret ´ on, J. A. Ben ´ ´ıtez-Andrades, I. Garc´ıa-Rodr´ıguez, O. Garc´ıa-Olalla, and C. Benavides, "Sentiment analysis in non-fixed length audios using a fully convolutional neural network," Biomedical Signal Processing and Control, vol. 69, p. 102946, 2021.

[8] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," Sensors, vol. 22, no. 4, p. 1414, 2022.

[9] S. Chauhan, S. Saxena, and P. Daniel, "Improved unsupervised neural machine

translation with semantically weighted back translation for morphologically rich and low resource languages," Neural Processing Letters, vol. 54, no. 3, pp. 1707–1726, 2022.

[10] G. Chen, S. Ma, Y. Chen, D. Zhang, J. Pan, W. Wang, and F. Wei, "Towards making the most of cross-lingual transfer for zero-shot neural machine translation," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 142–157.

[11] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli et al., "Unified speech-text pre-training for speech translation and recognition," arXiv preprint arXiv:2204.05409, 2022.

[12] Q. Dong, R. Ye, M. Wang, H. Zhou, S. Xu, B. Xu, and L. Li, "Listen, understand and translate: Triple supervision decouples end-toend speech-to-text translation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14, 2021, pp. 12 749–12 759.

[13] X. Liao, Y. Huang, P. Yang, and L. Chen, "A statistical language model for pre-trained sequence labeling: a case study on vietnamese," Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 3, pp. 1–21, 2021.

[14] X. Pan, M. Wang, L. Wu, and L. Li, "Contrastive learning for many-to-many multilingual neural machine translation," arXiv preprint arXiv:2105.09501, 2021.

[15] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," Global Transitions Proceedings, vol. 2, no. 2, pp. 205–211, 2021.

[16] V. S. Kumar, P. K. Pareek, V. H. C. de Albuquerque, A. Khanna, D. Gupta, and D. Renukadevi, "Multimodal sentiment analysis using speech signals with machine learning techniques," in 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon). IEEE, 2022, pp. 1–8.

[17] A. Kumar, "Contextual semantics using hierarchical attention network for sentiment classification in social internet-of-things," Multimedia Tools and Applications, vol. 81, no. 26, pp. 36 967–36 982, 2022.

[18] U. Ahmed, R. H. Jhaveri, G. Srivastava, and J. C.-W. Lin, "Explainable deep attention active learning for sentimental analytics of mental disorder," Transactions on Asian and Low-Resource Language Information Processing, 2022.

[19] C.-M. Kim, K.-H. Kim, Y. S. Lee, K. Chung, and R. C. Park, "Real-time streaming image based pp2lfa-crnn model for facial sentiment analysis," IEEE Access, vol. 8, pp. 199 586–199 602, 2020.

[20] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H.-S. Lim, “Bts: Back transcription for speech-to-text post-processor using text-tospeech-to-text,” in Proceedings of the 8th Workshop on Asian Translation (WAT2021), 2021, pp. 106–116.

[21] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, “Improving speech translation by understanding and learning from the auxiliary text translation task,” arXiv preprint arXiv:2107.05782, 2021.

[22] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” arXiv preprint arXiv:2010.10504, 2020.

[23] S. Indurthi, H. Han, N. K. Lakumarapu, B. Lee, I. Chung, S. Kim, and C. Kim, “End-end speech-to-text translation with modality agnostic meta-learning,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7904–7908.

[24] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, “Adaptation algorithms for neural network-based speech recognition: An overview,” IEEE Open Journal of Signal Processing, vol. 2, pp. 33–66, 2020.

[25] T. van der Zant, M. Kouw, and L. Schomaker, Generative artificial intelligence. Springer, 2013.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[27] D. gram, “Deepgram api playground," 2024. [Online]. Available: https://playground.deepgram.com/?endpoint=listensmartf ormat = truelanguage = enmodel = nova − 2

[28] M. V. Segbroeck, Z. Ahmed, K. Kutsenko, C. Huerta, T. Nguyen, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, “Dipco - dinner party corpus,” in Interspeech 2020, 2019. [Online]. Available: https://www.amazon.science/publications/dipco-dinner-party-corpus

[29] D. gram, “Deepgram api playground," 2024. [Online]. Available: https://developers.deepgram.com/docs/models-languages-overview