Insight Extraction Using Multimodal Analysis of Consumer Videos

Akshat Swaminath
School of Computer Science Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
akshat.swaminath2020@vitstudent.ac.in

Uma Priya D*
IEEE Member
School of Computer Science Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
umapriya.d@vit.ac.in
*Corresponding author

***Abstract*—In today's social media and public forums, consumers have a powerful platform to share their views on different products with a worldwide audience. For a product-based company, consumer data stands as an invaluable asset that facilitates a comprehensive understanding of consumer preferences, which products go well in the market, and where they can improve. This paper investigates consumer experiences using a multimodal analysis that involves video recordings from consumer interviews and group discussions. This paper goes beyond traditional qualitative research, examining non-verbal cues in consumer videos alongside transcribed speech. By thoroughly analyzing facial expressions, body language, vocal tone, and context in the video, our goal is to reveal deeper insights into consumer motivations, desires, and hidden pain points that are not always expressed verbally. The proposed work used Artificial Intelligence, specifically Automatic Speech Recognition (ASR) models, to solve the identified issue by creating accurate transcriptions of video content. Additionally, the utilization of the LLaMA2 framework facilitated the automatic extraction and interpretation of emotional cues embedded within the video recordings, thereby enhancing the depth and accuracy of the analysis process. The effectiveness of the proposed and existing work has been evaluated on standard datasets, and the results demonstrate that the proposed work outreaches the existing model with a good confidence score.**

***Index Terms*—Multimodal Analysis, Speech to Text, Multilingual Translation, Audio Processing, Transcription, Sentiment Analysis, Vocal Analysis, Automatic Speech Recognition**

I. Introduction

In today's consumer-focused industries, it's crucial for companies to grasp the intricate preferences, desires, and feedback of their audience to create products and services that truly connect. Conventional consumer research methods often focus solely on spoken or written feedback, overlooking the important non-verbal cues and contextual nuances like facial expressions, body language, and tone of voice. Therefore, there's a pressing need to explore new methods to tap into the wealth of information in diverse data sources.

The 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 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. Although some research works [3], [4] provide efficient video analysis, scalability and computational efficiency could be limiting factors, especially when dealing with large volumes of video data in real-time or near-real-time applications.

This paper solves the problem of insight extraction using Large Language Models (LLMs) to streamline the analysis process in consumer videos. By incorporating tools such as Whisper [5] and LLaMA 2 [6], the aim of this work 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 also 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.

The rest of the paper is organized as follows: Section II discusses the literature review, and Section III describes the methodology used for the problem discussed. Section IV discusses the results, and the paper is concluded in Section V.

II. 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 speech and emotions as they play a major role in the translation.

Huddar, M. G. et al. [7] analyses the sentiment and emotions in conversations using attention and RNN for multimodal data. Alex, S. B. et al. [3] uses attention and feature selection for speech emotion recognition based on sounds. García-Ordás et al. [8] analyses sentiment in audio recordings of varying lengths with a fully convolutional neural network. Trinh Van et al. [9] recognize speech emotions using deep neural networks. Chauhan, S., Saxena, S., & Daniel, P. [10] improves unsupervised neural machine translation for morphologically complex languages. Chen, G. et al. [16] maximizes cross-lingual transfer for zero-shot neural machine translation. Tang, Y. et al. [17] proposes a method for jointly pre-training speech and text representations to improve performance in speech translation and recognition tasks. The unified pre-training approach involves complex architectures or computationally intensive training procedures, making it challenging to deploy in resource-constrained environments or scale up to larger datasets. Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., & Li, L. [18] achieves end-to-end speech-to-text translation with a triple supervision approach. Sun, H. et al. [19] studies unsupervised neural machine translation for similar and distant language pairs. Pan, X., Wang, M., Wu, L., & Li, L. [20] applies contrastive learning for many-to-many multilingual neural machine translation. Sudhir, P., & Suresh, V. D. [21]

TABLE I
Summary of research works on Insight Extraction

Research Article	**Features**	**Methods used**	**Limitations**
Huddar, M. G. et al. [7]	Sentiment and emotion analysis in conversations	RNN with attention mechanisms	Specialized RNN architecture
Alex, S. B. et al. [3]	Speech emotion recognition	Deep neural networks with attention	Computational cost
García-Ordás et al. [8]	Sentiment analysis in variable-length audio recordings	Fully Convolutional Neural Network	Limited to audio data
Trinh Van et al. [9]	Emotion recognition in speech	Deep neural networks (CNN, CRNN, GRU)	Requires labeled data
Chauhan, S., Saxena, S., & Daniel, P. [10]	Enhancing unsupervised neural machine translation	Unsupervised neural machine translation	Limited generalization
Kim, C. M., Kim, K. H., Lee, Y. S., Chung, K., & Park, R. C. [11]	Real-time facial sentiment analysis	PP2LFA-CRNN model with image encryption	Limited to visual data
Park, C. et al. [12]	Speech-to-text post-processing using BTS	Back TranScription (BTS)	Domain specificity
Tang, Y., Pino, J., Li, X., Wang, C., & Genzel, D. [13]	Enhancing speech-to-text translation	Multitasking learning with speech and text translation tasks	Requires parallel text data
Zhang, Y. et al. [14]	automatic speech recognition	Semi-supervised learning with wav2vec 2.0 pre-training	Model size and language effectiveness
Indurthi, S. et al. [15]	End-to-end speech-to-text translation	Meta-learning with modality-agnostic multi-task model	Model complexity and data requirements

compares sentiment analysis approaches, applications, and classifiers. Kumar, V. S. et al. [4] analyze sentiment in speech signals using machine learning techniques. Park, C. et al. [12] mention improvements to speech-to-text conversion by back-transcribing text-to-speech output. Tang, Y., Pino, J., Li, X., Wang, C., & Genzel, D. [13] improves speech translation by leveraging auxiliary text translation tasks. Zhang, Y. et al. [14] discuss methods and techniques for training ASR models using semi-supervised learning approaches, which leverage labeled and unlabeled audio data to improve recognition accuracy. Indurthi, S. et al. [15] developed end-to-end models using modality agnostic meta-learning that directly translates spoken language into written text without relying on intermediate representations. Bell, P. et al. [22] review adaptation algorithms for neural network-based speech recognition.

While existing research has explored the potential of multi-modal 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.

III. METHODOLOGY

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 process necessitates human involvement.

The proposed work solves 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: 1) Focus group discussion preparation and 2) Post-recording. Each phase is described in detail below:

A. Focus group discussion preparation

Initially, the data related to the company products, customer feedback, market trends, and consumer preferences are gathered. Using the collected data, the product developer uses the Generative AI [23] to generate discussion questions. The model learns patterns and structures of language from the input data, enabling it to create relevant and coherent questions. Furthermore, the processes of selecting the candidates for the focus group discussion and recording the video discussions will 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 serves as the upload platform where discussion moderators upload recordings, discussion questions, and notes. This contextual information helps the model better understand the references made in the video.

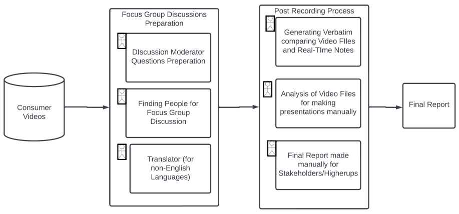
Consumer Videos
Focus Group Discussions Preparation
DIscussion Moderator Questions Preperation
Finding People for Focus Group Discussion
Translator (for non-English Languages)
Post Recording Process
Generating Verbatim comparing Video FIles and Real-TIme Notes
Analysis of Video Files for making presentations manually
Final Report made manually for Stakeholders/Higherups
Final Report

Fig. 1. Conventional Workflow of Insight Extraction

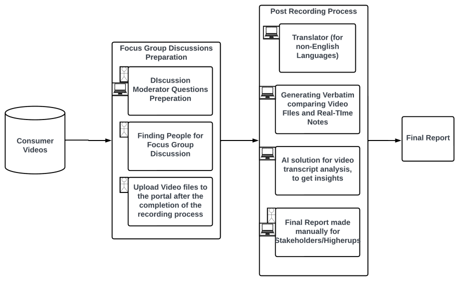
Consumer Videos
Focus Group Discussions Preparation
DIscussion Moderator Questions Preperation
Finding People for Focus Group Discussion
Upload Video files to the portal after the completion of the recording process
Post Recording Process
Translator (for non-English Languages)
Generating Verbatim comparing Video FIles and Real-TIme Notes
AI solution for video transcript analysis, to get insights
Final Report made manually for Stakeholders/Higherups
Final Report

Fig. 2. Proposed Workflow

B. Post-recording

The translator is substituted by the Whisper model [5], 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 [24] are used in the next two steps to draw insights from the video transcripts and make summarized points from the entire video file.

In general, input audio contains soft or whispered speech, which can be challenging for traditional speech recognition systems to accurately transcribe due to the lack of clear acoustic features. In this paper, the translation and transcription of the audio including whispered speech is taken care by the whisper-medium model [5] which is trained to recognize and transcribe soft speech by learning subtle acoustic cues and patterns associated with whispered speech. It also helps a lot in reducing the processing time. For the text summarization part, the Llama-2 model [6] is used, which generate the text by passing a small part of the transcript acquired by the whisper model. The Llama-2 model has the ability to understand and generate natural language makes it valuable for automating tasks that require human-like communication and creativity. 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.

IV. Experimental Analysis

For the analysis of the ASR models, Deepgram Playground [25] 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 [26], which was used as a standard to check the scores for all 5 selected models.

A. Existing Models for Comparison

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

1) **Nova-2 [27]:** The deepest-trained ASR model on the market, with 18% more accuracy than the previous Nova model.
2) **Whisper-medium [5]:** A family of encoder/decoder models trained on a large corpus of multilingual speech data by OpenAI.
3) **Nova-General [27]:** The deepest-trained ASR model is ideal for voice applications that require high accuracy in different contexts.
4) **General-Enhanced [27]:** Model useful for high accuracy timestamps, lower word error rates, and use cases that require keyword boosting.
5) **Base-General [27]:** Model useful for high accuracy timestamps and large transcription volumes.

B. Results and Discussion

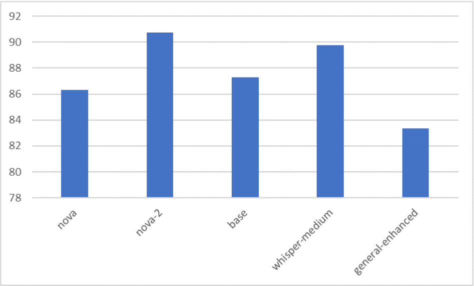

Fig. 3. Average Confidence Score for all the ASR models.

The figures provide insights into the performance of various models tested on the Dinner Party Corpus [26]. Figure 3 illustrate the average confidence scores of all the models that 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. Figures 4, 5, 6, 8, and 7, on the other hand, delve into the detailed confidence scores of the Nova-2, Whisper-Medium, Nova-General, General-Enhanced, and Base-General models 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. In addition, the execution time of the proposed work is less than that of the other models, which proves that the proposed model achieves better scalability.

V. Conclusions

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

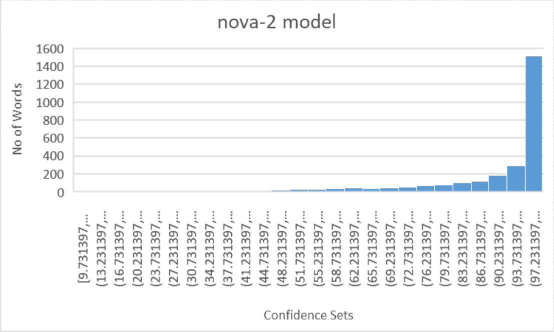

Fig. 4. Confidence Score of nova-2 model

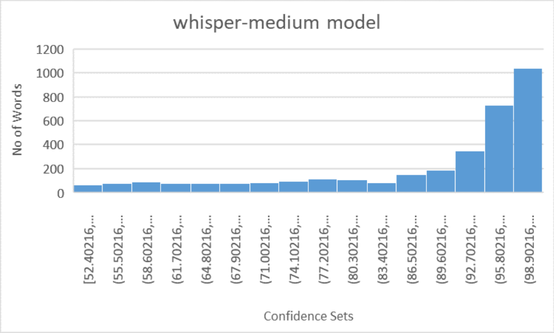

Fig. 5. Confidence Score of whisper-medium model

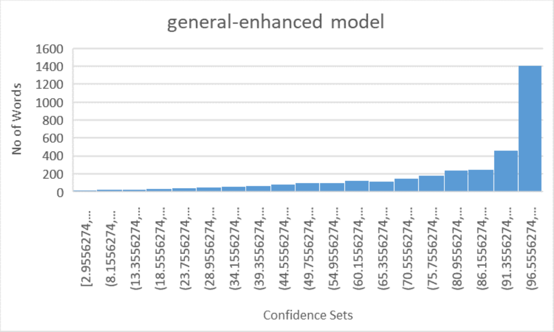

Fig. 7. Confidence Score of general-enhanced model

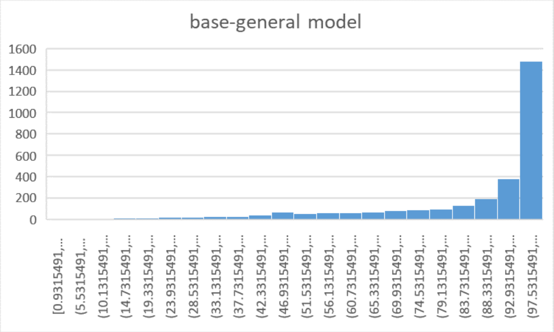

Fig. 8. Confidence Score of base-general model

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 particular 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.

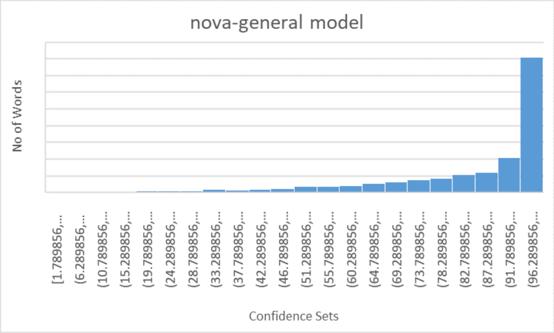

Fig. 6. Confidence Score of nova-general model

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.

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] 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.

[4] 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.

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

[6] 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.

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

[8] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-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.

[9] 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.

[10] 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.

[11] 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.

[12] 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-to-speech-to-text," in *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, 2021, pp. 106–116.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] D. gram, "Deepgram api play ground," 2024. [Online]. Available: https://playground.deepgram.com/?endpoint=listensmart$_{format} = truelanguage = enmodel = nova - 2$

[26] 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

[27] D. gram, "Deepgram api play ground," 2024. [Online]. Available: https://developers.deepgram.com/docs/models-languages-overview