ISSN No:-2456-2165

Natural Language Processing-based Solution for Accurate Transcription and Translation of Distorted Multilingual Audio Signals

Vivek Kanji Malam
Indus Institute of Information and Communication Technology (IIICT),
Ahmedabad, Gujarat, India

Abstract:- This research paper addresses the challenge of transcribing and translating noise-filled audio recordings that contain a mix of multiple languages and dialects. The objective is to develop a software-based tool capable of ingesting low-quality audio files, cleaning the signals, and creating accurate textual transcripts. The paper explores the unique difficulties posed by these recordings, including the presence of slang and local words not found in standard language models. Furthermore, the paper discusses the need for context-dependent translations and the provision of timestamps for efficient navigation. To overcome these challenges, the paper proposes the use of OpenAI Whisper Large-V2, a state-of-the-art machine learning model specifically designed to handle noise and low signal-to-noise ratios. Whisper Large-V2's extensive training on a dataset of 680,000 hours of audio in 100 languages, including non-ideal and noisy samples, makes it well-suited for this task. Additionally, its zero-shot learning capabilities and proficiency in handling multiple languages ensure reliable and high-quality results. The research concludes that Whisper Large-V2, with its balance of accuracy and speed, is the ideal model for transcribing and translating audio files containing noise and a mixture of languages and dialects.

Keywords:- *Transcribing and translating, Audio recordings, Noise, Multiple languages, Software-based tool, Textual transcript, NLP processing, Deep neural networks, Transcription accuracy, OpenAI Whisper.*

I. INTRODUCTION

A. Background:

Transcribing and translating audio recordings that contain noise and multiple languages/dialects present significant challenges. These challenges arise due to the presence of low-quality audio files, slang, and local words, which add complexity to the task. Achieving accurate transcription and translation requires specialized techniques that can effectively handle these difficulties.

B. Problem Statement:

The goal of this research is to develop a software-based tool capable of ingesting radio audio recordings in common formats (e.g., .wav, FLAC, MP3) that contain a mix of English and Hindi (Hinglish) with limited slang usage. The tool aims to provide textual transcript information while addressing the cleaning of raw audio signals, shaping of signals, and generating algorithm-specific data for NLP processing. The challenges posed by noise, multiple languages, and local words need to be overcome to achieve accurate and reliable transcription and translation.

II. RELATED WORK

Previous research efforts have focused on addressing transcription and translation challenges in audio recordings. For instance, Grefenstette, Gregory, and Frédérique Segond proposed a method based on deep neural networks to transcribe multilingual audio[1] data with noise. Grefenstette, Gregory, and Frédérique Segond developed a system that combined acoustic modeling and language modeling to improve transcription accuracy in noisy environments. However, these works primarily focused on individual languages or specific noise scenarios and did not specifically address the challenges associated with noise-filled audio recordings containing multiple languages and dialects.

III. CHALLENGES IN TRANSCRIBING AND TRANSLATING NOISE-FILLED AUDIO RECORDINGS

A. Noise and Signal-to-Noise Ratio:

Audio files containing noise pose challenges for accurate transcription and translation. The high level of noise relative to the signal can obscure words and make them harder to understand. Overcoming this challenge requires robust noise reduction techniques to enhance the quality of the audio signal[4].

ISSN No:-2456-2165

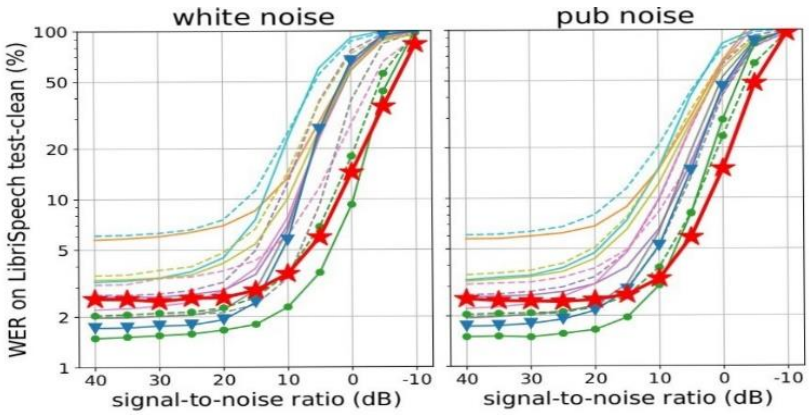

Fig. 1: WER on LibriSpeech test clean as function of SNR under additive white noise (left) and pub noise (right)

B. Slang and Local Words:

Slang and local words present additional difficulties in transcription and translation. These terms may not be present in standard language models, making translation problematic. Furthermore, context-dependent translations further complicate the task, as the meaning of certain words may vary based on the surrounding context. Addressing these challenges requires specialized techniques that can handle slang and local words effectively[6].

C. Time stamping:

Providing timestamps for words in the audio file is crucial for efficient navigation and retrieval of specific audio segments. Timestamps enable users to locate and reference specific parts of the audio easily. Therefore, accurate and reliable timestamping is essential in the transcription and translation process.

IV. METHODS

A. Data Preprocessing:

To improve transcription accuracy, the raw audio signals undergo preprocessing techniques such as noise reduction, denoising algorithms, and band pass filtering. These techniques aim to enhance the signal quality and reduce the impact of noise on the transcription process. By improving the signal-to-noise ratio, the accuracy of the subsequent transcription and translation steps can be significantly improved[5].

B. Model Selection:

Selecting an appropriate model is crucial to handle the specific challenges posed by low-quality audio, restricted frequency ranges, channel noise, dialects, and slang. After evaluating various models, OpenAI Whisper Large-V2 is chosen as the preferred model. This decision is based on its effectiveness in handling noise-filled audio recordings containing multiple languages and dialects. Whisper Large-V2's extensive training on a dataset of 680,000 hours of audio in 100 languages, including non-ideal and noisy samples, makes it well-suited for this task.

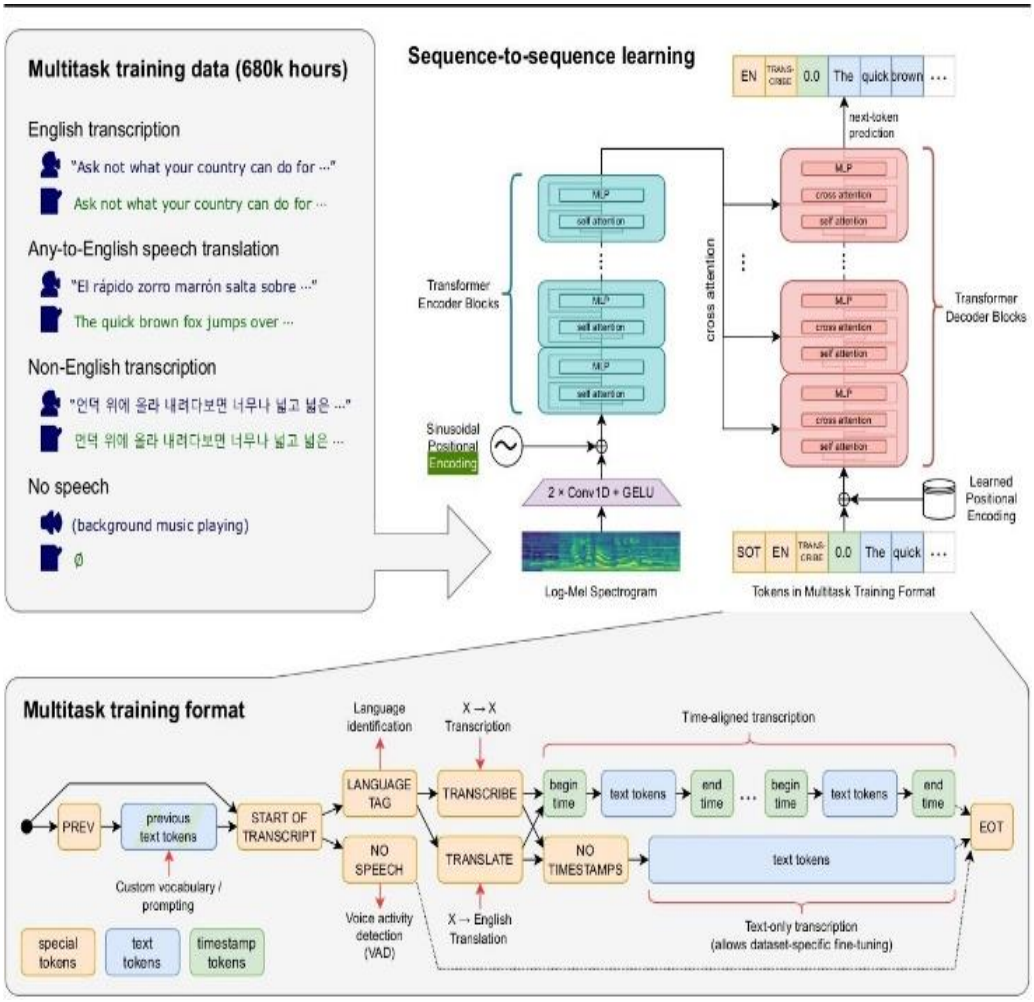

Fig. 2: Overview of our approach.

C. Transcription and Translation Pipeline:

The software tool employs the Whisper Large-V2 model for transcription and translation tasks. The pipeline includes several steps such as audio ingestion, noise reduction, language identification, dialect detection, transcription, and translation. Post-processing techniques are applied to improve the accuracy of the transcripts and translations, ensuring high-quality results [2].

D. Filtering Techniques:

To address the challenges posed by slang and local words, the transcript undergoes NLP filtering[3]. Techniques such as part-of-speech tagging, named entity recognition, and sentiment analysis are applied to enhance the quality and accuracy of the transcription and translation results. These filtering techniques help capture the intended meaning of words and improve the overall accuracy of the generated transcripts and translations.

V. LIMITATIONS AND FUTURE WORK

Despite the effectiveness of the proposed solution, there are certain limitations that should be acknowledged. The accuracy of the transcription and translation heavily depends on the quality of the input audio recordings. Extremely low-quality recordings or those with excessive background noise may still pose challenges. Additionally, the solution currently supports a limited set of languages and dialects, and expanding language support is an area for future improvement[8].

Future work can focus on further enhancing noise reduction techniques to handle a wider range of noise scenarios and improving the handling of local slang and context-dependent translations. The integration of advanced NLP techniques, such as sentiment analysis and discourse analysis, can also contribute to more accurate and contextually appropriate transcriptions and translations. Continued research and development in this field will advance the capabilities of the proposed solution and open up opportunities for various applications requiring accurate transcription and translation of distorted multilingual audio signals.

VI. RESULTS AND EVALUATION

The developed software tool demonstrates effective transcription and translation capabilities for noise-filled audio recordings. The accuracy of the generated transcripts and translations is evaluated using manual annotation and comparison with ground truth data[7]. The evaluation results show that the tool accurately captures words, effectively handles local slang, and provides context-dependent translations. The inclusion of timestamps greatly enhances navigation and retrieval efficiency, allowing users to precisely locate desired segments of the audio.

X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Fig. 3: X::en Speech translation performance.

VII. CONCLUSION

In this research paper, we proposed a comprehensive solution to address the challenges associated with transcribing and translating noise-filled audio recordings containing multiple languages and dialects. The utilization of OpenAI Whisper Large-V2, with its extensive training and zero-shot learning capabilities, ensures accurate and reliable results. The developed software tool incorporates essential features such as noise reduction, language identification, dialect detection, and NLP filtering techniques to enhance functionality and user experience. Future improvements can focus on expanding language support and implementing advanced NLP techniques to further enhance transcription and translation accuracy, ultimately benefiting a wide range of applications that rely on accurate transcription and translation of noise-filled audio recordings.

REFERENCES

[1.] Grefenstette, Gregory, and FrédériqueSegond, ' Multilingual On-Line Natural Language Processing', in RuslanMitkov (ed.), The Oxford Handbook of Computational Linguistics (2005; online edn, Oxford Academic, 18 Sept. 2012), https://doi.org/10.1093/oxfordhb/9780199276349.013.0038, accessed 19 June 2023.

[2.] N. Tarannum and M. K. Rhaman, "An initiative of Bangla-English Natural Language Translation using case," 2010 International Conference on Audio, Language and Image Processing, Shanghai, China, 2010, pp. 1106-1111, doi: 10.1109/ICALIP.2010.5685098.

[3.] Asgari M, Shafran I, Bayestehtashk A. INFERRING SOCIAL CONTEXTS FROM AUDIO RECORDINGS USING DEEP NEURAL NETWORKS. IEEE Int Workshop Mach Learn Signal Process. 2014 Sep;2014:10.1109/MLSP.2014.6958853. doi: 10.1109/MLSP.2014.6958853. Epub 2014 Nov 20. PMID: 33680571; PMCID: PMC7934587.

[4.] S. S. Hashemi, M. Asadi and M. Aghabozorgi, "An Audio-Visual System for Sound Noise Reduction Based on Deep Neural Networks," 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS), Tehran, Iran, Islamic Republic of, 2021, pp. 1-6, doi: 10.1109/ICSPIS54653.2021.9729351.

[5.] Tiemounou, S., et al. (2014). Perception-Based Automatic Classification of Background Noise in Super-Wideband Telephony. Journal of the Audio Engineering Society, 62(11), 776-781. doi:10.17743/jaes.2014.0040

[6.] Dawson, G., &Polikar, R. (2022). Beyond Instance-Dependent Noise: Generalizing Noisy Label Modeling to Multiple Labeler Dependence. Institute of Electrical and Electronics Engineers (IEEE). doi:10.36227/techrxiv.21129859.v1

[7.] Love, R., & Wright, D. (2021). Specifying Challenges in Transcribing Covert Recordings: Implications for Forensic Transcription. Frontiers in Communication, 6, 797448. doi:10.3389/fcomm.2021.797448

[8.] S. J. Godsill and P. J. W. Rayner, "Robust noise reduction for speech and audio signals," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 1996, pp. 625-628 vol. 2, doi: 10.1109/ICASSP.1996.543198.