Localized Speech-to-Text for Inclusive Learning

Pranav H
Dept. of Computer Science
Amrita School of Computing
Amrita Vishwa Vidhyapeetham,
Bengaluru, India

Mayank Pandey
Dept. of Computer Science
Amrita School of Computing
Amrita Vishwa Vidhyapeetham,
Bengaluru, India

I. Introduction

The most natural way for communication as humans is speech, and there have been constant efforts to make speech a viable and effect way of communication with computing devices, this is achieved with the ever-evolving toolset known as speech to text. These tools constantly evolving as they are , have failed to account for variations in dialects [3] and pronunciation of words along with being available for regional and vernacular languages which varies the impact and effectiveness of these tools from region to region [4]. In this context, it is empirical that we strive to improve these tools to overcome this disparity, making them available to all irrespective of region or language.

In light of the pressing need for these technological tools to be not just widely accessible, but also effectively utilized across diverse regions and languages, this paper puts forth a novel proposition. It suggests the deployment of an on-device, self-supervised [5], speech-to-text module that has been localized specifically for the multitude of regional languages spoken across India. The primary objective of this module is not merely to transcribe speech to text, but to serve a greater purpose - to act as a facilitator in the learning process. By catering to the unique linguistic nuances of regional Indian languages[6,7], this module aims to bridge the gap between technology and effective learning, thereby making education more inclusive and comprehensive.

Here we strive to implement the following tools to facilitate our goal towards a smart and inclusive classroom :

- Discussion Logs : Leveraging Audio Fingerprinting and text independent speaker recognition systems [8] to keep an individualised, speaker separated log of the class room discussions, which will enable a quick and easy review of the classroom discussions.
- Summarization : Using the discussion logs made, prepare a short summary to assist in quick review.
- Individualised Remainders and Summary : Leveraging the discussion logs to make a personalised summary tailored to the individual comprising of individual assigned tasks and conversations.
- Live Captioning and Translation : Enable Seamless Real Time Captioning [9] and Translation of the classroom to the user's chosen language.
- Evaluation Aid : Identify the key words used in response to a question and cross referencing [10] it with the expected key words and helping the evaluator during a Viva.

The Scope for the use of this tool is vast and the speech to text module proposed in this paper can be leveraged to assist beyond these tools implemented herein.

II. Literature Review

This collection of research papers provides a comprehensive overview of advancements in the field of Speech-to-Text (S2T) technologies. The authors explore various aspects of S2T, including neural basis of speech production and comprehension, personalized federated learning, voice synthesis, text-to-speech conversion, and keyword extraction techniques. These studies highlight the ongoing efforts to improve the efficiency, accuracy, and personalization of S2T systems, while also addressing challenges such as communication overhead, performance degradation, aperiodic distortion, and the impact of punctuation on keyword extraction. The research underscores the potential of S2T technologies in facilitating seamless human-computer interaction and advancing the field of natural language processing.

Ariel Goldstein et al. [11] discuss the neural basis of real-world speech production and comprehension. They used a deep multimodal speech-to-text model named Whisper to predict neural responses to both acoustic and semantic aspects of speech. The study reveals a distributed cortical hierarchy for speech and language processing, with sensory and motor regions encoding acoustic features of speech and higher-level language areas encoding syntactic and semantic information.

Yichao Du et al. [12] propose a personalized federated Speech-to-Text (S2T) framework to address the challenges of extensive communication overhead and performance degradation caused by data heterogeneity in S2T tasks. They introduce FedLoRA, a lightweight LoRA module for client-side tuning and interaction with the server to minimize communication overhead, and FedMem, a global model equipped with a k-nearest-neighbor (kNN) classifier that captures client-specific distributional shifts to achieve personalization and overcome data heterogeneity. Their approach significantly reduces the communication overhead on all S2T tasks and effectively personalizes the global model to overcome data heterogeneity.

Praveena Nuthakki et al. [13] discuss the challenges of traditional concatenation speech synthesis technologies and propose a new approach for voice synthesis. They suggest that their model's minimal aperiodic distortion makes it an excellent candidate for a communication recognition model. The authors also highlight the need for more robust network foundations and optimization methods for their proposed algorithm to perform at its best. They believe that their approach is as close to human speech as possible, despite the fact that speech synthesis has a number of audible flaws.

Swaroopa Shastri et al. [14] explore the conversion of English to Hindi, first to text, and subsequently to speech. They use two approaches for text character recognition from images: a maximally stable extensible region (MSER) and grayscale conversion. The paper also deals with geometric filtering in combination with stroke width transform (SWT). After detecting text sequences and fragmenting them into words, a 96 percent accurate spell check is performed using naive Bayes and decision tree algorithms. Finally, they use optical character recognition (OCR) to digitize the text and a text-to-speech synthesizer (TTS) to convert it to Hindi.

Nuhu Kontagora et al. [15] evaluate the suitability of conventional keyword extraction methods on a speech-to-text corpus. They collected a new audio dataset for keyword extraction using the World Wide Web (WWW) corpus. The performances of Rapid Automatic Keyword Extraction (RAKE) and TextRank are evaluated with different Stop lists on both the originally typed corpus and the corresponding Speech-To-Text (STT) corpus from the audio. They considered metrics of precision, recall, and F score for the evaluation. From the obtained results, Text Rank with the FOX Stoplist showed the highest performance on both the text and audio corpus, with F scores of 16.59% and 14.22%, respectively. Despite lagging behind text corpus, the recorded F score of the TextRank technique with audio corpus is significant enough for its adoption in audio conversation without much concern. However, the absence of punctuation during the STT affected the F score in all the techniques.

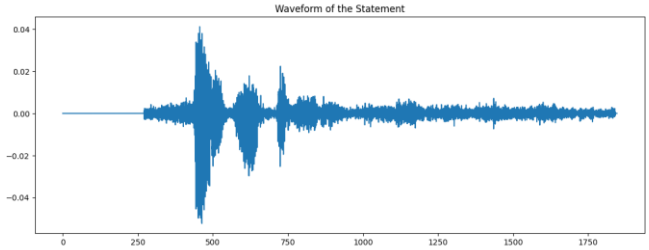
III. METHODOLOGY

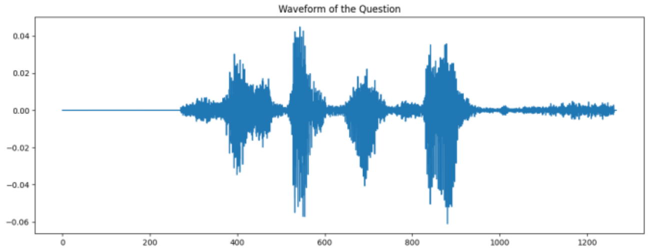
We conducted the below experiments :

1. The first experiment involved comparing the derivative signal with the original speech signal.
2. In the second experiment, the points of zero crossing in the first derivative signal were detected. The average length between two consecutive zero crossings for speech and silence regions was compared, and the pattern was observed.
3. The third experiment required speaking five favourite words. The length of the speech signals was observed and compared with those of a project team-mate.
4. The fourth experiment involved selecting a sentence that could be used for making a statement or asking a question, such as "You are going to college on Sunday(./?)". Two signals were recorded - one making the statement and the other asking the question. The two signals were then studied and compared.

IV. Results

1. The average length between two consecutive zero crossings for speech regions was found to be 1.0. However, for silence regions, the average length could not be calculated due to the absence of zero crossings.
2. The length of the 'United(1).wav' file was measured to be approximately 3.56 seconds, while the 'United(2).wav' file was slightly shorter, with a length of approximately 3.18 seconds.
3. The Below figures show the waveforms for the Statement and Question.





V. Conclusion

In conclusion, the conducted experiments provided valuable insights into the characteristics of speech signals. The average length between two consecutive zero crossings was determined to be 1.0 for speech regions, highlighting the distinct patterns that differentiate speech from silence. The inability to calculate the average length for silence regions due to the absence of zero crossings further emphasizes this distinction. Additionally, the comparison of the lengths of the 'United(1).wav' and 'United(2).wav' files demonstrated the variability in speech signal lengths. Finally, the analysis of the waveforms for the Statement and Question underscored the nuanced differences in speech patterns. These findings collectively contribute to a deeper understanding of speech signal analysis and its potential applications in speech-to-text technologies.

REFERENCES

[1] Y. Wei, J. Xiong, H. Liu, Y. Yu, J. Pan, and J. Du, “AdaStreamLite: Environment-adaptive Streaming Speech Recognition on Mobile Devices,” in Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 4, Association for Computing Machinery, New York, NY, USA, December 2023, Art. No. 187, pp. 1-29. doi: 10.1145/3631460

[2] J. Laures-Gore, C. R. Rogers, H. Griffey, K. G. Rice, S. Russell, M. Frankel, and R. Patel, “Dialect identification, intelligibility ratings, and acceptability ratings of dysarthric speech in two American English dialects,” in Clinical Linguistics & Phonetics, Taylor & Francis, pp. 1-12. doi: 10.1080/02699206.2023.2301337.

[3] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, “Towards inclusive automatic speech recognition,” in Computer Speech & Language, vol. 84, 2024, 101567, ISSN 0885-2308. doi: 10.1016/j.csl.2023.101567.

[4] V. Karthikeyan and S. Suja Priyadharsini, “Modified layer deep convolution neural network for text-independent speaker recognition,” in Journal of Experimental & Theoretical Artificial Intelligence, vol. 36, no. 2, Taylor & Francis, 2024, pp. 273-285. doi: 10.1080/0952813X.2022.2092560.

[5] P. Gambhir, A. Dev, P. Bansal, and D. K. Sharma, “End-to-end Multi-modal Low-resourced Speech Keywords Recognition Using Sequential Conv2D Nets,” in ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 23, no. 1, Association for Computing Machinery, New York, NY, USA, January 2024, Art. No. 7, pp. 1-21. doi: 10.1145/3606019.

[6] Devare, M. ., & Thakral, M. . (2023). Enhancing Automatic Speech Recognition System Performance for Punjabi Language through Feature Extraction and Model Optimization. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(8s), 307–313.

[7] F. Wu et al., “Wav2Seq: Pre-Training Speech-to-Text Encoder-Decoder Models Using Pseudo Languages,” in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5. doi: 10.1109/ICASSP49357.2023.10096988.

[8] Y. Wei, J. Xiong, H. Liu, Y. Yu, J. Pan, and J. Du, “AdaStreamLite: Environment-adaptive Streaming Speech Recognition on Mobile Devices,” in Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 4, Association for Computing Machinery, New York, NY, USA, December 2023, Art. No. 187, pp. 1-29. doi: 10.1145/3631460.

[9] L. Liu, L. Liu, and H. Li, “Computation and Parameter Efficient Multi-Modal Fusion Transformer for Cued Speech Recognition,” arXiv preprint arXiv:2401.17604, 2024. Primary Class: cs.CV.\

[10] R. Shukla, “Keywords Extraction and Sentiment Analysis using Automatic Speech Recognition,” arXiv preprint arXiv:2004.04099, 2020. Primary Class: eess.AS.

[11] Goldstein, Ariel, et al. "Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations." *bioRxiv* (2023): 2023-06.

[12] Du, Yichao, et al. "Communication-Efficient Personalized Federated Learning for Speech-to-Text Tasks." *arXiv preprint arXiv:2401.10070* (2024).

[13] Nuthakki, Praveena, et al. "Deep Learning based Multilingual Speech Synthesis using Multi Feature Fusion Methods." *ACM Transactions on Asian and Low-Resource Language Information Processing* (2023).

[14] Shastri, Swaroopa, and Shashank Vishwakarma. "An Efficient Approach for Text-to-Speech Conversion Using Machine Learning and Image Processing Technique."

[15] Nuhu Kontagora, Bello, et al. "Performance Evaluation of Keyword Extraction Techniques and Stop Word Lists on Speech-To-Text Cor." (2023).

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.