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.

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.

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.

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