Colloquial Language Speech Converter API: A Comprehensive Survey

Muhammed Abnas *, Muhammed Imkan K M , Ajmal J S , Abhiram P Vasudevan , Shereena Thampi , Rosy K Philip

Posted Date: 30 December 2024

doi: 10.20944/preprints202412.2503.v1

Keywords: speech recognition; Wav2Vec 2.0; CTC/attention; SpeechBrain; machine translation; colloquial speech; malayalam dialects; speech-to-text; text-to-speech

Article

Colloquial Language Speech Converter API: A Comprehensive Survey

Muhammed Abnas [1,*], Muhammed Imkan K M [1], Ajmal J S [1], Abhiram P Vasudevan [1], Shereena Thampi [2] and Rosy K Philip [2]

[1] Computer Science And Engineering, College of Engineering Chengannur, Kerala, India
[2] Project Guide, College of Engineering Chengannur, Kerala, India
* Correspondence: abnas7511@gmail.com

Abstract: This literature survey focuses on various methodologies and frameworks that have been developed for speech recognition, accent recognition, and dialect translation. The research is aimed at building a comprehensive understanding of how existing technologies like Wav2Vec, SpeechBrain, and hybrid CTC/Attention models can be applied to create a speech converter API that translates colloquial Malayalam speech into standard Malayalam text and further into English. The goal of this API is to bridge communication gaps caused by the linguistic diversity within Malayalam dialects, thus improving accessibility in fields such as education, social media, and public services.

Keywords: speech recognition; Wav2Vec 2.0; CTC/attention; SpeechBrain; machine translation; colloquial speech; malayalam dialects; speech-to-text; text-to-speech

1. Introduction

The Malayalam language has a wide range of dialects across different regions of Kerala, making it difficult for speakers from one district to fully understand those from another. The aim of this project is to develop a Colloquial Malayalam Speech Converter API that uses advanced speech recognition models to convert dialectal Malayalam speech into standard Malayalam and subsequently into English. This will facilitate smoother communication across regions, benefiting sectors like education, tourism, and public services. In this literature survey, we review various methodologies and frameworks that can be leveraged for this purpose, such as Wav2Vec 2.0, SpeechBrain, and CTC/Attention-based models.

1.1. What Is a Colloquial Malayalam Speech Converter API?

The Colloquial Malayalam Speech Converter API is designed to address the communication gap between different Malayalam dialects by converting colloquial Malayalam speech into standard Malayalam text, followed by translation into English. This approach makes it easier for users from different regions of Kerala to communicate with each other and non-Malayalam speakers. The system leverages automatic speech recognition (ASR) and text-to-speech (TTS) technologies to accurately interpret and convert spoken language. By doing so, it ensures smoother communication, whether in social interactions, educational contexts, or tourism-related applications. This project aims to make communication more inclusive, accessible, and efficient for diverse audiences.

1.2. Technological Components Used in the Colloquial Malayalam Speech Converter API

The Colloquial Malayalam Speech Converter API relies on several specialized technologies:

1. Speech Recognition (ASR): This is the core component of the system, converting spoken Malayalam dialects into text. The API leverages Wav2Vec 2.0 and SpeechBrain models to accurately capture the various nuances in Malayalam dialects, ensuring the correct conversion of colloquial speech to standard Malayalam.

 doi:10.20944/preprints202412.2503.v1

2. Translation Engine: After the speech-to-text conversion, a translation engine (e.g., GoogleTrans API) is used to translate the standardized Malayalam text into English. This translation facilitates cross-linguistic communication and makes the output useful for non-Malayalam speakers.
3. Text-to-Speech (TTS): Once the translation is completed, the Google Text-to-Speech (gTTS) engine is employed to convert the translated English text back into speech. This ensures that the output is available in both text and auditory forms, providing accessibility for users who prefer spoken output.
4. Natural Language Processing (NLP): NLP models help enhance the system's understanding of different colloquial expressions and dialect variations. This component ensures that the API can accurately interpret a variety of regional Malayalam phrases and dialectal speech patterns.

1.3. Linguistic Diversity in Malayalam

Malayalam is spoken in Kerala, but its dialects vary significantly between regions such as Malappuram, Kozhikode, Thrissur, and Alappuzha. Each district has its unique pronunciation, vocabulary, and phrases, which often lead to misunderstandings even among native speakers. For example, the word "you" is spoken as X in Malappuram and X in Kozhikode, highlighting the necessity for dialect standardization. These linguistic differences make it essential to develop a system that can translate regional dialects into standard Malayalam, ensuring that communication is clear and consistent.

1.4. Relevance in Kerala and India

Given the linguistic diversity within Kerala, the Colloquial Malayalam Speech Converter API holds immense potential in various sectors, such as education, tourism, and media. Kerala's rapidly expanding digital infrastructure and high levels of internet penetration make it an ideal region for implementing such technologies. Furthermore, this API can serve as a tool for language preservation and documentation, ensuring that the dialects are understood and standardized, while also being accessible to a broader audience through English translations. With over 34 million Malayalam speakers across the globe, this API could also be extended to diaspora communities to help second and third-generation speakers reconnect with their native language and culture. This project aligns with India's broader initiatives for digital inclusion and language preservation, especially in rural areas where dialectal differences are more pronounced.

1.5. Global and Regional Impact

The ability to convert regional dialects into standardized language can benefit cross-cultural communication. With Malayalam being one of the 22 scheduled languages in India, this API could be extended to other regional dialects and languages as part of India's effort to improve linguistic inclusivity. Additionally, Kerala's significance as a global tourist destination means that the API can aid in real-time communication between locals and tourists, enriching the tourism experience.

2. Literature Review

Brydinskyi, V. et al. (2024) [1] proposed the enhancement of Automatic Speech Recognition (ASR) using personalized models based on wav2vec2. The system improves speech recognition by applying personalized fine-tuning on individual speaker datasets. Their study involved four key experiments: per-speaker tuning, per-subset tuning, fine-tuning on similar speakers, and dataset mixing, which collectively helped improve Word Error Rate (WER), especially for synthetic voices. Fine-tuning on similar datasets showed the highest improvement, reducing WER by up to 10%. The experiments were conducted on TedLIUM-3, CommonVoice, and GoogleVoice, utilizing 20 minutes of audio per speaker for training and testing. The study demonstrates the effectiveness of personalized ASR models, especially in low-resource environments with mixed natural and synthetic voice data.

Inaguma, H. et al. (2019) [2] introduced a multilingual end-to-end speech translation system using a sequence-to-sequence model. This system allows for direct speech translation from one language to another without separate ASR and translation components. The model was evaluated in

 doi:10.20944/preprints202412.2503.v1

one-to-many and many-to-many translation scenarios, using attention-based sequence-to-sequence architectures. The datasets included Fisher-CallHome Spanish, Librispeech, and TED corpus, and the model demonstrated notable improvements in zero-shot translation tasks. This system enables more efficient and effective multilingual speech-to-speech translation, making it suitable for multilingual communication applications, especially in low-resource language pairs.

Podila, R.S.A. et al. (2022) [3] developed a Telugu dialect speech dataset and evaluated it using deep learning techniques for dialect classification. The dataset consists of speech from 9 male speakers across three Telugu dialects, and models like BiLSTM with Attention Layer were used to classify the dialects. The best-performing model achieved an accuracy of 99.11%, highlighting the success of BiLSTM in recognizing regional speech variations. Despite the limited number of speakers, the results demonstrate the potential for deep learning methods in handling dialectal speech recognition, making it a valuable approach for identifying and processing regional speech patterns.

Yerramreddy, D.R. et al. (2022) [4] conducted a comparative evaluation of three ASR models—SpeechBrain, Whisper, and Wav2Vec2—to assess their performance in real-time speech recognition tasks. The models were tested on the Librispeech dataset, with SpeechBrain achieving the highest BLEU score of 79.32. Whisper followed with 73.34, and Wav2Vec2, though highly efficient in self-supervised learning, achieved a score of 66.45. The study revealed that SpeechBrain excelled in real-time transcription tasks with low Word Error Rate (WER), while Whisper demonstrated stronger multilingual transcription capabilities.

Penkova, B. et al. (2023) [5] developed an unsupervised neural machine translation (UNMT) system to translate Croatian dialects into modern Croatian. The system utilizes cross-lingual embeddings and a Transformer-based architecture to translate dialects like Shtokavian, Kajkavian, and Chakavian into standard Croatian. The model was trained on monolingual data and achieved a BLEU score of 12.8. The study demonstrates the potential of UNMT in handling low-resource dialects, despite the limited dataset size, by relying on unsupervised learning techniques and cross-lingual alignment for translation tasks.

Kaneb, Y. and Kodad, M. (2024) [6] reviewed NLP techniques for recognizing Arabic dialects. The study explored various deep learning models, including LSTM and Seq2Seq, and highlighted their success in dialect recognition tasks. Some models achieved accuracies as high as 98.65% on Arabic datasets, particularly when using bi-directional LSTMs. Despite the success of these models, the review identified the need for larger, more comprehensive datasets to fully capture the variety of Arabic dialects, which vary greatly between regions. This study's insights can be applied to Malayalam dialect recognition, emphasizing the use of deep learning techniques in low-resource languages.

Unni V. et al. (2020) [7] introduced a coupled training approach for sequence-to-sequence models to improve accented speech recognition. Their model pairs utterances of the same text spoken by different speakers with diverse accents. The L2 loss applied between paired context vectors helps the model develop accent-invariant representations. Tested on the Mozilla Common Voice dataset, the coupled training approach showed significant reductions in Word Error Rate (WER) across various accents, demonstrating its effectiveness in improving accent recognition within ASR systems.

AESRC 2020 (2020) by X. Shi et al. (2020) [8] presented the results of the Accented English Speech Recognition Challenge 2020, which involved accent recognition and accented speech recognition tasks. Models like TDNN with PPG features achieved an accuracy of 83.63% in accent recognition, while speech recognition models with CTC-LAS rescoring achieved a WER of 4.06%. The challenge datasets included 160 hours of accented English speech, with data augmentation techniques such as speed perturbation and volume augmentation used to enhance model performance.

Chen L. W. , and Rudnicky, A. (2021) [9] explored fine-tuning Wav2Vec 2.0 for speech emotion recognition tasks using Task-Adaptive Pretraining (TAPT) and pseudo-labeling techniques. The study showed a 7.4% improvement in accuracy for emotion recognition tasks, using the IEMOCAP dataset. The paper demonstrates the efficacy of Wav2Vec 2.0 in adapting to low-resource settings and improving performance in speech recognition tasks, especially when dealing with tasks like emotion detection.

 doi:10.20944/preprints202412.2503.v1

Gao Q. et al. (2021) [10] introduced an end-to-end speech accent recognition system that utilizes a hybrid CTC/Attention transformer-based ASR model. The model pre-trains an automatic speech recognition system by integrating Connectionist Temporal Classification (CTC) for output alignment and an attention mechanism for focusing on relevant audio features. By incorporating accent labels, the model simultaneously learns to recognize the content and accents of speech. Evaluated on the AESRC2020 dataset, which includes data from speakers across eight countries, the system achieved an accuracy of 80.98% on the development set and 72.39% on the test set. Despite its strong performance, it struggles with certain accents, such as Chinese and Japanese, and may benefit from further optimization for real-world applications. This hybrid approach is particularly relevant for our project, as it can be adapted to recognize Malayalam dialects effectively by using dialect labels to improve recognition accuracy.

3. Conclusions

The literature reviewed for the Colloquial Malayalam Speech Converter API highlights the advancements in speech recognition, machine translation, and natural language processing (NLP) technologies that are essential for addressing linguistic diversity in dialect-rich regions. Studies on hybrid CTC/Attention models, such as those proposed by Gao et al. (2021) [10], demonstrate the effectiveness of integrating Connectionist Temporal Classification (CTC) with attention mechanisms in improving speech recognition accuracy across various accents and dialects. Similarly, research on personalized models, such as the work by Brydinskyi et al. (2024) [1], emphasizes the role of fine-tuning Wav2Vec 2.0 in enhancing recognition for low-resource languages, a method highly applicable to our project. The application of multilingual end-to-end models as shown by Inaguma et al. (2019) [2], and deep learning techniques like BiLSTM with Attention Layer, exemplified in the work by Podila et al. (2022) [3], further underscore the potential for applying these models to dialect translation tasks. The comparative study of ASR models by Yerramreddy et al. (2022)[4] confirms that models such as SpeechBrain are well-suited for real-time speech-to-text applications, making them ideal for our use case.

Overall, the methodologies and technologies surveyed provide a solid foundation for the development of the Colloquial Malayalam Speech Converter API. By adapting these cutting-edge techniques, the API aims to bridge communication gaps between Malayalam dialects and provide accessible translation solutions, contributing to the broader efforts of language preservation and digital inclusion. Moving forward, continued refinement and optimization of these models will be essential to further improving accuracy and scalability in real-world applications.

References

1. V. Brydinskyi, D. Sabodashko, Y. Khoma, M. Podpora, A. Konovalov, and V. Khoma, "Enhancing Automatic Speech Recognition With Personalized Models: Improving Accuracy Through Individualized Fine-Tuning," IEEE Access, vol. 12, pp. 116649-116656, 2024, doi: 10.1109/ACCESS.2024.3443811.
2. H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual End-to-End Speech Translation," arXiv, 2019. doi: 10.48550/arXiv.1910.00254.
3. R. S. A. Podila, G. S. S. Kommula, R. K., S. Vekkot, and D. Gupta, "Telugu Dialect Speech Dataset Creation and Recognition Using Deep Learning Techniques," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10040194.
4. Yerramreddy D. R. , J. Marasani, P. S. V. Gowtham, G. Harshit, and Anjali, "Speech Recognition Paradigms: A Comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10544133.
5. Penkova B, M. Mitreska, K. Ristov, K. Mishev, and M. Simjanoska, "Learning Translation Model to Translate Croatian Dialects to Modern Croatian Language," 2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2023, pp. 1083-1088, doi: 10.23919/MIPRO57284.2023.10159848.

6. Kaneb and M. Kodad, "Literature Review: NLP Techniques for Arabic Dialect Recognition," 2024 International Conference on Circuit, Systems and Communication (ICCSC), Fes, Morocco, 2024, pp. 1-5, doi: 10.1109/ICCSC62074.2024.10617343.
7. Unni V. , N. Joshi, and P. Jyothi, "Coupled Training of Sequence-to-Sequence Models for Accented Speech Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 8254-8258, doi: 10.1109/ICASSP40776.2020.9052912.
8. X. Shi, et al., "The Accented English Speech Recognition Challenge 2020: Open Datasets, Tracks, Baselines, Results and Methods," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6918-6922, doi: 10.1109/ICASSP39728.2021.9413386.
9. Chen L. W. and Rudnicky A. , "Exploring Wav2Vec 2.0 Fine-Tuning for Improved Speech Emotion Recognition," arXiv, 2021, doi: 10.48550/arXiv.2110.06309.10.48550/arXiv.2110.06309.
10. Gao Q., H. Wu, Y. Sun, and Y. Duan, "An End-to-End Speech Accent Recognition Method Based on Hybrid CTC/Attention Transformer ASR," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7253-7257, doi: 10.1109/ICASSP39728.2021.9414082.