International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2511125

A Review on Technologies for Group-Aware Malayalam Conversational AI

Husaima Mailanchy T K, Suha Narghees A S, Taniya T S, Vishnu M S, Dr. L.C. Manikandan

Department of Computer Science and Engineering, Universal Engineering College, Thrissur, Kerala, India

ARTICLEINFO

Article History:

Accepted : 08 Feb 2025

Published: 10 Feb 2025

Publication Issue

Volume 11, Issue 1

January-February-2025

Page Number

2037-2044

ABSTRACT

This review paper explores the foundational technologies required to develop a group-aware Conversational AI for the Malayalam-English bilingual community. The objective of the project is to create an AI system capable of interacting naturally in group settings, dynamically recognizing and responding to multiple speakers in real-time. The key components of this system include voice separation, which isolates individual speakers' voices in noisy environments, speech-to-text (STT), which accurately transcribes Malayalam speech that may contain English phrases, and text-to-speech (TTS), which synthesizes natural-sounding speech in Malayalam-English conversational patterns. This review covers recent advancements in each of these three areas by evaluating three core papers for each technology. Through this review, we aim to understand the current capabilities of these technologies and how they can be applied to build an accessible, scalable Conversational AI that bridges the language gap for Kerala's linguistically diverse population.

Index Terms—Conversational AI, Voice Separation, Speech-to-Text, Text-to-Speech, Malayalam-English, Multimodal Processing, Natural Language Processing, Bilingual Communication

Introduction

Conversational AI systems have evolved significantly, enabling seamless interaction between humans and machines. However, building AI that caters to bilingual communities, such as the Malayalam-English speakers in Kerala, presents a unique set of challenges. In everyday conversations, people in Kerala often switch between Malayalam and English, requiring AI systems to handle multilingual input effectively while maintaining natural conversational flow.

The goal of this review paper is to examine the foundational technologies required to develop a group-aware Conversational AI specifically designed for the Malayalam-English bilingual community. Such a system needs to be capable of real-time voice separation, accurately convert spoken language into text through speech-to-text (STT), and generate natural responses using text-to-speech (TTS) technology. Each of these components plays a critical role in creating a system that can identify multiple speakers, process their speech, and provide appropriate responses in both languages.

Voice separation is essential for distinguishing and isolating individual voices from a mixed audio input, especially in noisy environments where multiple speakers interact simultaneously. The review explores recent advancements in voice separation technologies, particularly those that enable efficient, real-time processing. Additionally, speech-to-text technologies allow the system to convert spoken Malayalam and English into text. In a multilingual environment like Kerala, accurate transcription of speech containing both languages is crucial for further processing and response generation. This review evaluates current STT models and their performance in handling bilingual speech. Furthermore, text-to-speech technology is responsible for generating spoken responses from text. To ensure the system delivers natural-sounding responses that reflect the conversational style of Kerala, this review examines the latest developments in TTS, focusing on models that can produce high-quality speech in both Malayalam and English.

By reviewing these key technologies, this paper aims to understand the current capabilities of voice separation, STT, and TTS, and how these advancements can be integrated to create a robust and scalable AI system tailored for the Malayalam-English bilingual population.

Literature Review

This literature review provides an overview of key technologies necessary for creating a group-aware conversational AI system tailored for Malayalam-English speakers. The discussion includes important areas such as voice separation, speech-to-text, and text-to-speech systems, looking at their progress, methods, and role in supporting bilingual communication. Understanding these technologies is important for enhancing how conversational AI operates in multilingual settings, particularly in situations where multiple people are speaking at once.

A. VOICE SEPARATION

Voice separation, in the context of our paper, refers to the process of distinguishing and isolating individual speakers' voices from a mixed audio input. This is particularly important when multiple speakers are present, as it allows each speaker's voice to be processed separately. In our project, voice separation is necessary for the conversational AI to handle group conversations effectively. It enables the system to improve the accuracy of subsequent tasks, such as speech-to text conversion, by processing each voice independently.

1) TOWARDS REAL-TIME SINGLE-CHANNEL SPEECH SEPARATION IN NOISY AND REVERBERANT ENVIRONMENTS: The paper introduces a method designed to efficiently separate voices from a single audio channel in real time, even in the presence of noise and reverberation [1]. The system uses a deep neural network (DNN) structured in three stages: noise suppression, voice separation, and de-reverberation [1].

The proposed model uses a causal DNN to handle real-time speech separation in challenging acoustic environments. It includes three modules—noise suppression (NS), speech separation (SS), and de-reverberation (DR)—based on a modified CRUSE architecture. This architecture consists of convolutional encoder-decoder layers and recurrent bottleneck layers, which help the model process temporal information more effectively. Additionally, convolutive transfer functions (CTFs), also referred to as deep filters, enhance the model's ability to outperform conventional time-frequency masking techniques. The model employs a subtractive separation (SUB) method, which iteratively estimates and removes each speaker's voice from the input, enabling separation of multiple speakers. A baseline end-to-end model with a three-layer LSTM is also proposed to handle noise suppression, separation, and de-reverberation in a single step. The system is trained on 16 kHz audio using the STFT for feature extraction, with AdamW as the optimizer and data augmentation to enhance generalization. Finally, its performance is evaluated against the SepFormer model, which serves as a benchmark for state-of- the-art speech separation [1].

This approach is well-suited for our project, where handling multiple speakers in real-time, especially in noisy environments, is crucial. The use of causal DNNs and subtractive separation ensures low latency, making it ideal for dynamically identifying and separating voices in a group aware conversational AI [1].

2) ATTENTION IS ALL YOU NEED IN SPEECH SEPERATION:This paper [2] presents SepFormer, a Transformer based model for efficient speech separation, replacing RNNs with multi-head attention to capture both short- and long-term dependencies. SepFormer achieves state-of-the-art performance on WSJ0-2mix and WSJ0-3mix datasets, offering faster processing and lower memory usage than RNN-based systems [2]

The SepFormer model is designed using a learned-domain masking approach, comprising three main components: an encoder, a masking network, and a decoder. The encode rutilizes 256 convolutional filters to process the input signal, creating an STFT-like representation [2]. The masking network applies dual-path transformers to mode l both short- and long term dependencies by dividing the input into overlapping chunks, where the IntraTransformer handles local processing and the InterTransformer captures global dependencies [2].

The decoder then reconstructs the separated audio signals using the masks generated by the masking network [2]. The model is Optimized with the Adam optimizer, incorporating dynamic mixing and speed perturbation, and achieves state-of-the-art result, with an SI-SNR improvement of 22.3 dB and SDR improvement of 22.4 dB on the WSJ0-2mix dataset[2].

This paper's Transformer-based approach to speech separa tion aligns with our project's need to efficiently handle multi- speaker interactions in noisy environments. By utilizing multi- head attention mechanisms and eliminating the need for RNNs, the SepFormer enables fast and accurate voice separation, which is essential for the real-time processing required in our group-aware Conversational AI system.

3) DUAL-PATH RNN: EFFICIENT LONG SEQUENCE MODELING FOR TIME-DOMAIN SINGLE-CHANNEL SPEECH SEPARATION: Luo et al. present the Dual-Path Recurrent Neural Network (DPRNN) [3], designed to handle long audio sequences in single-channel speech separation efficiently. This model divides the input sequence into smaller parts, allowing separate processing of local and global information. The dual-path approach ensures efficient and accurate separation of mixed speech by handling shorter chunks of data sequentially [3].

In detail, the model applies two distinct RNNs: one pro-cesses local relationships within each chunk, while the other captures broader dependencies across different chunks. This structure enables the DPRNN to handle long sequences with- out the performance drop commonly seen in traditional mod- els. When tested on the WSJ0-2mix dataset, the model demon strated notable improvements in resource usage, achieving a 49% reduction in computational requirements while enhancing speech separation performance by 4.6%. It also delivered a significant SI-SNR improvement of 18.8 dB, highlighting its robustness [3].

Given its efficiency and reduced resource demands, this model is well-suited to our project's requirements for real- time voice separation. Its ability to deliver strong performance while using fewer computational resources makes it a valuable option for applications in environments with limited process- ing capacity [3]

B. SPEECH-TO-TEXT

Speech-to-text technology converts spoken language into written text. In our project, this process is used to interpret the input from Malayalam speakers. After separating the individ-ual voices, the AI system converts the speech into text, which can then be used for further analysis or response generation. In the context of our project, the focus is on converting Malayalam speech, which may include some English words, into text, reflecting the typical bilingual speech patterns in Kerala.

1) WAV2VEC 2.0: A FRAMEWORK FOR SELF SUPERVISED LEARNING OF SPEECH REPRESENTA TIONS: The paper [4] introduces wav2vec 2.0, a framework for learning speech representations directly from raw audio in a self-supervised manner. The main takeaway is that by leveraging unlabeled speech data, the model significantly improves speech recognition, particularly in scenarios where labeled data is scarce [4].

In terms of methodology, the system uses a multi-layer CNN to transform raw audio into latent speech representa- tions. These representations are masked and passed through a Transformer to learn contextualized representations. The model employs product quantization and Gumbel softmax to produce discrete representations. The learning process is guided by contrastive loss and diversity loss, while fine-tuning is done with Connectionist Temporal Classification (CTC) loss for downstream tasks. This approach achieves a Word Error Rate (WER) as low as 1.8% for clean speech and 3.3% for noisy speech, surpassing previous state-of-the-art methods [4]. This paper's [4] relevance lies in its approach to building effective speech recognition systems using minimal labeled data. In the context of our project, this framework can be adapted to improve the accuracy and efficiency of Malayalam speech-to-text conversion, in low-resource settings [4].

2) GOOGLE USM: SCALING AUTOMATIC SPEECH RECOGNITION BEYOND 100 LANGUAGES: The paper [5]

introduces the Universal Speech Model (USM), a scalable ASR model designed to handle over 100 languages. It achieves state-of-the-art performance by pre-training on a large mul- tilingual audio dataset and fine-tuning with smaller labeled datasets [5].

In terms of methodology, the authors apply BEST-RQ, a BERT-based pre-training technique, to process over 12 mil- lion hours of unlabeled audio data from YouTube and other sources. This model is further refined using Multi-Objective Supervised pre-Training (MOST), which combines unlabeled speech, text, and paired audio-text data to improve alignment between speech and text. To achieve optimal performance on speech recognition tasks, the model incorporates Connection ist Temporal Classification (CTC) and Listen, Attend, and Spell (LAS) architectures. This approach allows the USM to achieve a Word Error Rate (WER) of 14.4% on long-form YouTube ASR tasks, outperforming other systems on multilingual benchmarks [5].

This research is relevant to our project because it demonstrates a scalable, multilingual approach to ASR that can be adapted to enhance Malayalam speech-to-text systems,especially in environments where labeled data is scarce [5].

3) WHISPER: ROBUST SPEECH RECOGNITION VIA LARGE-SCALE WEAK SUPERVISION: The paper introduces Whisper, a speech recognition model trained on 680,000 hours of weakly supervised multilingual audio. Whisper's primary objective is to generalize well across various languages and environments without requiring dataset specific fine-tuning, making it highly adaptable for broad use cases [6]

Whisper [6] is built on an encoder-decoder Transformer architecture. The input audio is processed into an 80-channel log-magnitude Mel spectrogram using 25-millisecond win- dows with a 10-millisecond stride. This is followed by sinu- soidal positional encoding and convolutional layers utilizing the GELU activation function. The processed data is passed through multiple pre-activation residual blocks to learn deeper speech representations. For text tokenization, Whisper em- ploys a byte-level BPE tokenizer adapted from GPT-2. The model uses multitask learning, handling transcription, translation, and language identification by predicting token sequences that specify which task to perform. To enhance transcription accuracy, Whisper predicts timestamps with a 20-millisecond resolution, improving synchronization between audio and text. Pre-training was done with AdamW optimization, gradient norm clipping, and FP16 mixed precision for more efficient training. The model was trained using a batch size of 256 for 220 updates, resulting in a Word Error Rate (WER) of 2.7% onthe LibriSpeech clean-test and 25.5% on the CHiME6 dataset for noisy speech. Whisper demonstrated a 55.2% relative error reduction compared to supervised models when tested on out-of-distribution datasets, showing exceptional robustness and generalization [6].

This paper is particularly relevant to our project, as Whis per's scalable zero-shot learning approach can help us develop a robust Malayalam speech-to-text system, offering high accuracy even in diverse, noisy environments [6]

C. TEXT-TO-SPEECH

Text-to-speech technology generates spoken language from

written text. In our project, this technology is used to produce responses in Malayalam and English based on the text the AI processes. The system is designed to generate speech that reflects the natural conversational style in Kerala, where Malayalam and English are often used interchangeably. This ensures that the AI's responses are clear and sound natural to the user.

1) NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS: This paper presents a neural network architecture aimed at generating natural-sounding speech from text inputs. It combines a sequence-to-sequence model to transform text into mel-spectrograms and utilizes a modified WaveNet vocoder to convert those spectrograms into time-domain waveforms [7].

The architecture is composed of two primary components: sequence-to-sequence network that predicts mel-spectrograms from input text, and a WaveNet vocoder that translates the spectrograms into audio waveforms. These mel-spectrograms, created using the short-time Fourier transform (STFT), serve as a simplified acoustic representation that streamlines the training process for both components [7]. The spectrogram prediction network features an encoder made up of convolutional layers followed by a bi-directional LSTM, while the decoder employs location-sensitive attention to produce the spectrogram frames step by step [7]. To further enhance accuracy, a post-net is employed to reduce the mean squared error (MSE) in the generated spectrogram [7]. The WaveNet vocoder uses dilated convolution layers and a mixture of logistics (MoL) model to generate high-quality audio wave-forms. The system incorporates regularization techniques such as dropout and zoneout, along with a stop token mechanism to ensure efficient inference. With these enhancements, the system achieves a mean opinion score (MOS) of 4.53, nearly matching that of human-recorded speech, and surpasses prior models like Tacotron with Griffin-Lim and WaveNet conditioned on linguistic features [7].

In our project, this system proves highly relevant for de-veloping a Malayalam text-to-speech solution. Its ability to produce high-quality, natural-sounding speech can be adapted to generate clear and accurate Malayalam speech, aligning well with the objectives of our project [7].

2) GLOW-TTS: GENERATIVE FLOW FOR TEXT-TO SPEECH VIA MONOTONIC ALIGNMENT SEARCH: : The Glow-TTS paper introduces a

flow-based generative model for parallel text-to-speech (TTS) that operates independently of external aligners [8]. By using dynamic programming to align text and speech representations, it enables faster and more robust speech synthesis.

Glow-TTS is a flow-based generative model designed for efficient, parallel text-to-speech (TTS) synthesis. Unlike tra- ditional autoregressive models, it aligns text and speech rep- resentations without external aligners by using dynamic pro- gramming for monotonic alignment, ensuring a non-skipping, linear progression from text to mel-spectrogram generation [8].

The model consists of a text encoder, duration predictor, and a flow-based decoder. The text encoder transforms phoneme sequences into latent representations, while the duration predictor learns to predict durations of phonemes using an MSE loss. The flow-based decoder utilizes invertible transformations to convert latent speech representations into mel-spectrograms [8]. Training is performed through an iterative Viterbi algorithm that maximizes the log-likelihood of the most probable alignment, and monotonic alignment search (MAS) is employed to find the best alignment during training. The system also employs multi-resolution STFT loss and adversarial training to improve waveform quality. Glow-TTS is capable of synthesizing high-quality mel-spectrograms up to 15.7 times faster than Tacotron 2. In evaluations, Glow-TTS achieved a Mean Opinion Score (MOS) of 4.01, demonstrating competitive voice quality compared to Tacotron 2's 3.88 MOS [8].

This study is highly relevant to our project as it emphasizes enhancements in speed and efficiency in TTS systems, which are essential for real-time applications. The ability of Glow- TTS to produce high-quality speech rapidly without external dependencies aligns well with our objective of developing an efficient, scalable TTS system for a bilingual conversational AI [8].

3) FASTSPEECH 2:FAST AND HIGH-QUALITY END-TO END TEXT TO SPEECH: FastSpeech 2 improves upon the authentic FastSpeech model, supplying faster and extra natural text-to-speech (TTS) synthesis by means of the use of unique pitch, length, and energy records [9]. This technique not best complements the expressiveness of synthesized speech however also reduces the complexity and variability in education, making TTS systems extra green and flexible.

FastSpeech 2 replaces the original model's teacher-pupil distillation with direct education on actual mel-spectrograms, retaining more particular facts and streamlining the training pipeline. A key component is the variance adaptor, which incorporates accurate length, pitch, and electricity predictors to cope with the ”one-to-many” mapping venture frequently found in TTS [9]. An superior variant, FastSpeech 2s, further speeds up the manner by generating waveforms at once from textual content, bypassing intermediate ranges for improved inference speeds. Results display that FastSpeech 2 provides terrific audio that now not handiest surpasses the unique Fast- Speech however is likewise corresponding to autoregressive fashions, reducing education time via threefold and attaining a 47x speedup in synthesis [9].

This version advances the field through enhancing TTS pace, exceptional, and schooling simplicity, making it feasible to generate greater herbal-sounding speech. Additionally, the efficiency profits position FastSpeech 2 as a strong candidate for real-time applications [9].

Comparative Analysis

The three models analyzed for speech separation—Real-Time Single-Channel Speech Separation [1], Dual-Path RNN [3], and SepFormer [2]—have distinct strengths and limitations. The Real-Time Single-Channel model uses a modular RNN-based approach, excelling in resource efficiency and low latency while maintaining competitive SI-SNRi and DNSMOS scores. The Dual-Path RNN efficiently balances dependencies with intra- and inter-chunk processing, achieving high SI-SNRi improvements and scalability, though it adds complexity for very long sequences. SepFormer, a Transformer-based model, provides state-of-the-art SI-SNRi scores and reduced memory demands but is less suitable for real-time use. The first model's real-time efficiency and modular design make it best for latency-sensitive scenarios.

The three speech-to-text models—Google USM [5], wav2vec 2.0 [4], and Whisper [6]—exhibit unique efficiencies.

Google USM's conformer-based architecture, trained on 12 million hours of audio, supports 100 languages but requires significant resources. wav2vec 2.0 excels in low-resource settings, cutting WER with minimal labeled data and using self-supervised learning, though it requires initial resource investment. Whisper's transformer encoder-decoder achieves robust zero-shot performance with 680,000 hours of training but is less efficient. Overall, wav2vec 2.0 provides a strong balance of resource use and performance.

The three text-to-speech systems—Tacotron 2 [7], Glow TTS [8], and FastSpeech 2 [9]—showcase varied efficiencies. Tacotron 2, an autoregressive model with a WaveNet vocoder, achieves a MOS of 4.53 but has slow inference and stability issues. Glow-TTS is 15.7 times faster than Tacotron 2 while maintaining quality but uses significant resources. FastSpeech 2, a non-autoregressive model, eliminates the teacher model, reducing training time threefold. It integrates pitch, duration, and energy predictors, achieving a MOS of 3.83 but needing extensive training data. FastSpeech 2 offers a good balance of speed and quality for real-time applications.

Conclusion

This review examined state-of-the-art technologies for the development of a group-aware Conversational AI designed for the Malayalam-English bilingual community, focusing on voice separation, speech-to-text (STT), and text-to-speech (TTS).

For voice separation, the Real-Time Single-Channel Speech Separation model by Neri et al. [1] emerged as the preferred choice due to its compact and resource-efficient DNN architecture, optimized for real-time processing in noisy and reverberant environments. In speech-to-text, the wav2vec 2.0 model by Baevski et al. [4] was selected for its self-supervised learning approach, which minimizes the need for large labeled datasets, making it ideal for low-resource settings like bilingual Malayalam-English speech. For text-to-speech, Fast Speech 2 by Ren et al. [9] was chosen for its speed, efficiency, and ability to generate high-quality speech with minimal computational demands.

In summary, these technologies—Real-Time Single-Channel Speech Separation, wav2vec 2.0, and FastSpeech 2—provide a solid foundation for building a scalable, efficient Conversational AI that meets the linguistic needs of the Malayalam-English bilingual population.

References

[1]. Neri, Julian, and Sebastian Braun. ”Towards Real-Time Single-Channel Speech Separation in Noisy and Reverberant Environments.” In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

[2]. Subakan, Cem, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. ”Attention is all you need in speech separation.”In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21-25. IEEE, 2021.

[3]. Luo, Yi, Zhuo Chen, and Takuya Yoshioka. ”Dual-path rnn:efficient long sequence modeling for time-domain single-channel speech separation.” In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech

and Signal Processing (ICASSP), pp. 46-50. IEEE, 2020.

[4]. Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. ”wav2vec 2.0: A framework for self-supervised learning of speech representations.” Advances in neural information processing systems 33 (2020): 12449-12460.

[5]. Seethala, S. C. (2017). Revolutionizing Data Warehouses in Manufacturing: Big Data-Infused Automation for ETL and Beyond. https://doi.org/10.5281/zenodo.14169254

[6]. Seethala, S. C. (2018). AI and Big Data: Transforming Financial Data Warehousing for Predictive Analytics. https://zenodo.org/record/14050624

[7]. Seethala, S. C. (2018). Leveraging AI in Cloud Data Warehouses for Manufacturing: A Future-Proof Approach. https://doi.org/10.5281/zenodo.14059537

[8]. Seethala, S. C. (2019). Data Warehouse Modernization with AI: A Strategic Path for the Retail Industry. https://doi.org/10.5281/zenodo.14168854

[9]. Seethala, S. C. (2019). AI-Enhanced ETL for Modernizing Data Warehouses in Insurance and Risk Management. https://doi.org/10.5281/zenodo.14059551

[10]. Seethala, S. C. (2019). Scaling Financial Data Warehouses with AI: Towards a Future-Proof Cloud-Based Ecosystem. International Journal of Scientific Research & Engineering Trends, 5(6). https://doi.org/10.61137/ijsret.vol.5.issue6.575

[11]. Zhang, Yu, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhe-huai Chen, Nanxin Chen et al. ”Google usm: Scaling automatic speech recognition beyond 100 languages.” arXiv preprint arXiv:2303.01037 (2023).

[12]. Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. ”Robust speech recognition via large-scale weak supervision.” In International conference on machine learning, pp. 28492-28518. PMLR, 2023.

[13]. Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen et al. ”Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.” In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779-4783. IEEE, 2018.

[14]. Kim, Jaehyeon, Sungwon Kim, Jungil Kong, and Sungroh Yoon. ”Glow- tts: A generative flow for text-to-speech via monotonic alignment search.” Advances in Neural Information Processing Systems 33 (2020): 8067-8077.

[15]. Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. ”Fastspeech 2: Fast and high-quality end-to-end text to speech.” arXiv preprint arXiv:2006.04558 (2020).