Online Speaker Diarization Using Reinforcement Learning

by

Himanshu Gulechha (2348520)
Purusharth Malik (2348542)

Under the guidance of
Dr. Rajesh Kanna R.

Reinforcement Learning
Vth trimester, MSAIM
CHRIST (Deemed to be University)

10 December 2024

Abstract

This project addresses the complex task of speaker diarization, which involves separating and identifying speakers in audio streams. Traditional methods struggle with real-world challenges such as overlapping speech, dynamic speaker identities, and multilingual dialogues. To overcome these issues, the study proposes an online, reinforcement learning-based (RL) framework that conceptualizes speaker attribution as a sequential decision-making problem. The framework integrates feature extraction, embedding generation, and decision-making into a unified system, leveraging policy gradient techniques for dynamic adaptation. A multi-agent reinforcement learning (MARL) setup is employed, where each speaker operates as an independent agent.

Methodology

The system utilizes pre-trained Wav2Vec2 for feature extraction, transforming audio into high-dimensional speaker embeddings. Diarization is framed as a policy optimization problem, employing REINFORCE to maximize a reward signal based on diarization accuracy metrics or ground-truth alignment. The framework adapts dynamically to new speakers, resolves overlapping speech through specialized modules, and supports multilingual scenarios using embeddings trained on diverse datasets.

Evaluation

Experiments conducted on benchmark datasets like CALLHOME and the AMI Meeting Corpus demonstrate competitive performance, achieving low diarization error rates (DER) and effective adaptation to challenging conditions. The proposed system outperforms many state-of-the-art methods in terms of accuracy, robustness, and scalability.

Expected Outcomes and Applications

The framework aims to enhance diarization accuracy, adaptability in dynamic scenarios, and scalability across domains such as teleconferencing, media production, and legal transcription. Future extensions include incorporating graph attention mechanisms, self-supervised learning, and real-time processing capabilities.

Conclusion

By addressing the limitations of traditional methods and leveraging the strengths of RL, this study presents a scalable and versatile solution for speaker diarization, with promising applications in diverse real-world settings.