System Architecture

Our speaker diarization system leverages several highly optimized machine learning models and algorithms to allow diarizing hours of audio in a real-time streaming fashion with limited computational resources on mobile devices. The system mainly consists of three components: a speaker turn detection model that detects a change of speaker in the input speech, a speaker encoder model that extracts voice characteristics from each speaker turn, and a multi-stage clustering algorithm that annotates speaker labels to each speaker turn in a highly efficient way. All components run fully on the device.