Real-Time Phone Call Analysis: A Comprehensive Multimodal Approach

Daniele Ugo Leonzio, Daniele Moro, Paolo Bestagini, Stefano Tubaro
Politecnico di Milano

Abstract

We present a comprehensive real-time phone call analysis system that simultaneously performs emotion recognition, demographic estimation, deepfake detection, and speech transcription. This demonstration showcases an integrated pipeline capable of processing voice streams with minimal latency, providing critical insights for applications in security, customer service, and telecommunications monitoring.

The proposed system addresses the growing need for automated voice analysis in an era where synthetic speech generation and voice manipulation technologies are increasingly sophisticated. Our multi-task architecture processes incoming audio streams through parallel analytical pathways, enabling simultaneous extraction of multiple attributes without compromising processing speed or accuracy.

Emotion Recognition: Our system employs deep learning models trained on prosodic features, spectral characteristics, and linguistic content to identify emotional states including happiness, sadness, anger, fear, surprise, and neutral affect. The model achieves real-time classification with sub-second latency, providing temporal emotion trajectories throughout the conversation.

Age and Gender Estimation: Leveraging acoustic features such as fundamental frequency, formant distributions, and vocal tract characteristics, the system estimates speaker demographics. The age estimation module provides continuous age range predictions, while gender classification distinguishes between male and female vocal characteristics with high confidence scores.

Deepfake Detection: A critical component of our system analyzes subtle artifacts and inconsistencies characteristic of synthetic speech. Using a combination of artifact detection, consistency analysis across temporal segments, and neural vocoder fingerprinting, the module flags potentially manipulated or generated audio with detailed confidence metrics. This addresses the emerging threat of voice spoofing in authentication systems and fraud prevention.

Transcription: Real-time automatic speech recognition converts the audio stream to text with support for multiple languages and accents. The transcription engine employs state-of-the-art transformer-based models optimized for telephony audio, handling challenges such as background noise, channel effects, and speaker overlap.

The demonstration will showcase the system processing live phone calls and pre-recorded samples, visualizing all analytical outputs through an intuitive dashboard. The system architecture emphasizes modularity, allowing individual components to be deployed independently or as an integrated solution depending on application requirements.

This work has significant implications for telecommunications security, automated quality monitoring in call centers, fraud detection in banking systems, and elder care monitoring applications. The demonstration will include discussions of ethical considerations, privacy preservation techniques, and potential deployment scenarios across various domains.

Requirements

The demonstration is fully remote and requires minimal on-site setup. The system operates on a dedicated GPU-accelerated server for real-time processing of all analytical modules. Call handling is managed through a SIP (Session Initiation Protocol) server that routes voice streams to the analysis pipeline. We will bring a device capable of making VoIP calls and a laptop for viewing the demonstration dashboard through a web browser. We will need only a internet connection. All computational infrastructure is cloud-hosted, eliminating the need for local hardware requirements.