A Tool for Automatic Speech Recognition in Intercepted Audios for Law Enforcement Agencies

DRAFT

Lindeberg Pessoa Leite

Federal Police Forensic Expert
Brazilian Federal Police
lindeberg.lpl@pf.gov.br

August 7, 2023

Abstract

Automatic Speech Recognition (ASR) is a critical tool utilized by Law Enforcement Agencies (LEAs) to transcribe intercepted telephone conversations for criminal investigations. However, accurately transcribing these conversations presents significant challenges due to low-quality intercepted telephone conversations. As a first step to address this issue, this study develop a practical tool to transcribe real intercepted telephone conversations. This tool can serve as a baseline for the development of more accurate models. Additionally, the study incorporates diarization techniques, which are essential for the identification and separation of different speakers in investigative processes.

Keywords ASR · Diarization · LEA

1 Introduction

Intercepted telephone conversations play a crucial role in criminal investigations, providing valuable evidence for LEAs. Automatically transcribing these conversations accurately is essential for analyzing the content and extracting relevant information. However, achieving accurate transcriptions poses significant challenges due to various factors that affect the quality of the recorded audio. Factors such as the conditions of the recording environment (e.g. open air and echo), properties of the audio transmitting medium (e.g. channel noise and crosstalk), and ambient effects (e.g. background noise) significantly impact the quality of intercepted telephone conversations [Roxanne, 2023]. Consequently, ASR systems designed for general speech recognition may struggle to achieve accurate transcriptions in this specific domain.

2 Related Work

This initial work builds upon the research conducted on Whisper by [Radford et al., 2022] for the transcription process. To achieve diarization, the tool utilizes the method developed by [Bredin et al., 2020]

3 Method

The approach consists of two main components:

- Transcription - This involves ASR performed on the audio files. It provides information about the segments of speech along with the recognized text for each segment.
- Diarization - This component performs diarization on the same audio file and offers information about the identified speakers and the time segments corresponding to each speaker's speech.

The tool combines the Transcription and Diarization components to produce a diarized transcription, where each segment of speech is associated with the corresponding speaker ID and the recognized text [Yin, 2023].

4 Future Work

A next step in this work is to fine-tune the Whisper model on intercepted telephone conversations. Considering the often low-quality nature of such audio recordings, incorporating denoising techniques before fine-tuning and inference stages could yield improvements in the results. By applying denoising techniques to the intercepted telephone conversations, the aim is to reduce background noise, distortions, and other unwanted audio artifacts that can decrease the accuracy of the transcription process. This preprocessing step holds the potential to enhance the overall audio quality, leading to improved performance during the inference phase.

In addition to fine-tuning and denoising techniques, conducting a comparative analysis with other state-of-the-art models such as Wav2Vec2 is of fundamental importance.

References

Project Roxanne. Automatic speech recognition setting: Benefits and limitations. ROXANNE Project Blog, 2023. URL `https://www.roxanne-euproject.org/news/blog/automatic-speech-recognition-setting-benefits-and-limitations#_ftnref1`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020.

Ruiqing Yin. pyannote-whisper. `https://github.com/yinruiqing/pyannote-whisper/tree/main`, 2023.