

Speaker-specific speech transcription for forensic investigative triage

Jadd Virji, Finnian Kelly, Anil Alexander

Oxford Wave Research Ltd., Oxford, UK
{jadd|finnian|anil}@oxfordwaveresearch.com

1. Introduction

In forensic and law enforcement investigations involving large volumes of multimedia, it may be necessary to rapidly triage hundreds or thousands of unlabelled files to find those containing speech, and from this subset to both identify speakers of interest and to gather intelligence from the spoken content. This application presents challenges on multiple fronts: firstly, forensic audio recordings are typically made in uncontrolled environments and often contain a high level of background noise, frequent non-speech events, and may use low-quality or compressed recording devices; secondly, with a large number of media files and a standard requirement for offline processing, an efficient and scalable pipeline is necessary. In this paper we present an initial approach for retrieval of speakers of interest from an unlabelled dataset, followed by speaker-specific speech transcription and analysis to enable rapid triage of the spoken content. We present a pilot approach for holistic assessment of the output transcriptions using a large language model (LLM), and demonstrate the potential of signal conditioning to improve transcription quality in degraded speech.

2. Approach

To retrieve a speaker of interest from an unlabelled set, we apply a segmental speaker recognition approach that compares a reference speaker embedding (x-vector) to embeddings extracted from short segments of the unlabelled recordings, enabling the recognition and localisation of speakers within long recordings. Speech from a speaker of interest is then passed to Whisper large-v3 for transcription and translation. Finally, we use an LLM to summarise the speaker-specific speech and apply NLP techniques (e.g. named entity recognition) to highlight content of interest. Given the aforementioned acoustic challenges with forensic audio, we optionally pre-process the audio with signal conditioning, including VAD and source separation (to extract the speech signal from the background). See Figure 1 for a process overview. A key requirement for automatic speech recognition in a forensic investigative setting is to produce a transcript that preserves the holistic meaning of the spoken content. To evaluate a transcript in this respect, we consider an LLM-based process, with a candidate transcription and ground truth as input, and a score on a Likert-style scale from 1 to 7 as output. We use few-shot in-context learning, i.e. providing the LLM examples of human-rated transcriptions, as well as descriptors of each quality level to achieve meaningful ratings.

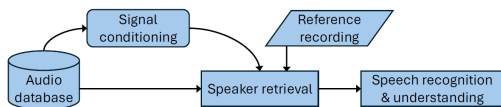


Figure 1: Speaker-specific triage process

3. Example Results

In our pilot experiment, we use a subset of NFI-FRIDA (Netherlands Forensic Institute - Forensically Realistic Inter-Device Audio), a database of speech recordings in Dutch acquired simultaneously by multiple forensically-relevant recording devices. Specifically, we use a distant microphone condition and include samples with added in-room noise. We simulate a retrieval scenario by comparing clean FRIDA reference recordings to distant noisy microphone recordings using VOCALISE [2] in segmental mode. The resulting scores are assessed according to the match rate at rank 1: a match rate of 70% is achieved for the raw recordings, rising to 100% after applying signal conditioning. We transcribe raw and conditioned recordings retrieved for speakers of interest using Whisper large-v3 and highlight key content using LLM summarisation, trigger words and wordclouds (e.g. Figure 2). Evaluating the transcriptions using our LLM approach on a 1–7 scale we observe a mean rating of 4.7 for raw recordings, rising to 5.1 for the conditioned recordings. Further analysis of the LLM ratings shows a correlation to perceptual audio quality and classic word error rate.

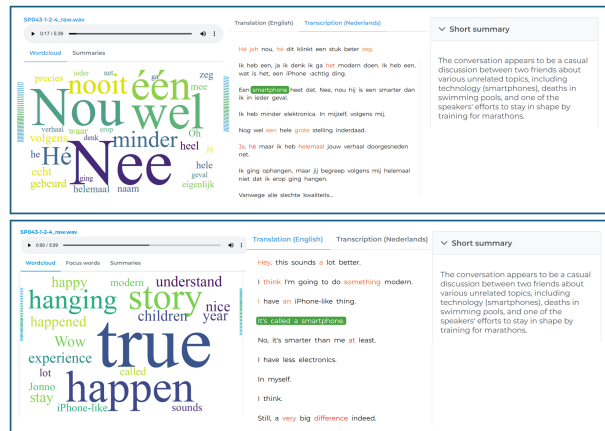


Figure 2: Transcription (top) and translation (bottom) for a retrieved recording of a FRIDA speaker of interest

4. References

- [1] D. van der Vloed, F. Kelly, and A. Alexander, “Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database”, *Odyssey 2020: The Speaker and Language Recognition Workshop*, Tokyo, Japan.
- [2] F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander. “Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors”, *Audio Engineering Society (AES) Forensics Conference 2019*, Porto, Portugal.