Voice Carving and Automatic Speech Recognition applied to the Transcription of Police Interview Recordings

Johanna Morley[1], *Oscar Forth*[2], *and Anil Alexander*[2]
[1]*Strategic and Emerging Technology, London Metropolitan Police, UK*
johanna.morley@met.police.uk
[2]*Research and Development, Oxford Wave Research Ltd, UK.*
{oscar|anil}@oxfordwaveresearch.com

The London Metropolitan Police Service (MPS) holds around 850,000 detainee interviews per year at an average length of thirty minutes each. Although not all interviews are fully transcribed, a fair proportion of them are, and this is a costly end-to-end process. The impetus for this piece of work comes from the technology refresh of analogue interview equipment to digital interviewing, which allows potential for a range of tools to be applied to the audio recording. Technologies that can integrate with the digital interview equipment, with added capability of segmenting different voices on the recording and transcribing the interview accurately, have the potential to reduce the cost burden of interview transcription.

In these police interview recordings, there may be hundreds or, in some cases, thousands of ‘turns’ of conversation between the speakers, each of which may only be a few seconds or parts of seconds in length. This poses a real challenge to automatic speech transcription. Commercial state-of-the-art automatic speech recognition systems improve in word recognition accuracy when they have adapted to the speech of the person dictating the text. Once the recognition system is trained (internally adapts itself) with a speaker’s voice and the microphone used for recording, the recognition performance improves. Subsequently, if the system has to recognize the speech of another speaker, the performance will degrade unless the system has had the opportunity to train with this second speaker/recording device or can switch between its trained states. However, in an interview, the main speakers switch in time throughout the recording. A recognition system would thus have to continually switch between training states to maximize its recognition accuracy. Commercially available speech recognition systems have primarily been developed for the dictation and telephony market and are not all capable of performing this intelligent switching. In this work, we investigate the degradation in speech recognition accuracy because of changes in speakers and context, and whether this degradation can be reduced by applying speaker segmentation prior to the recognition, so that each instance of recognition only considers the voice of one speaker.

Proposed approach and test database

An automatic speaker segmentation system based on Alexander et al, 2010 is currently in use at the London Metropolitan Police Audio laboratory and is used as a pre-processing step to voice disguise. It allows the operator to quickly separate out the speakers in a recording into different channels. The separation allows for rapid processing of long interview recordings containing the speech of two speakers. We propose to use the output from such an automatic speaker segmentation system to interface with state-of-the-art speech-to-text systems. Each instance of the speech-to-text system is thus allowed to deal with only one speaker at a time, which should improve its accuracy. By intelligently combining the transcribed outputs of different instances of the speech-to-text system we can obtain the combined transcript of the interview. This automatically generated transcript will be provided to a human transcriptionist who would listen to the recording and manually correct errors in the transcript and make it ready for use as a transcript for use in court. The effort involved in correcting the transcript and the accuracy of transcription is evaluated and compared with human transcription.

The test database was recorded in a vulnerable witness interview room at a police station in London, using a new digital recording system that was under evaluation. This recording set-up was considered as typical for existing witness recordings made at the MPS. The room was reasonably well

sound-proofed with soft-furnishings, carpets and sofas. There was a small amount of electrical interference that was noted on the audio and some of the recordings were overdriven. There was a small amount of background noise, but this did not render the speech inaudible and was considered representative of real-world conditions. The database consisted of speech samples in two conversational styles, namely free speech and scripted speech**.** The free speech recordings consisted of simulated police interviews conducted by trained serving police officers as interviewers. The mock witnesses were police officers and staff, a member of the public, and two young children. The scripted recordings consisted of simulated interviews read from a script that was obtained from a collection of transcribed celebrity interviews in a national newspaper. They were asked to read them naturally, emoting if necessary. It is to be noted that the no explicit effort was made to match the microphones and the recording file characteristics with those of those of any commercial speech recognition system.

New Accuracy Metrics

While word error rates (WERs) provide an idea of how different the computer generated transcripts are against the actual speech produced by the speakers, they are they do not measure the actual effort involved by the transcriptionists to correct the transcripts or take into consideration the understanding of what is said (Wang et al 2003). In order to do this, we had to develop a metric that was proportional to the effort involved. When a human transcriptionist uses the automatically generated transcript as a starting point to create the final transcript, he/she will need to make a number of changes to the automatically generated transcript in order to make it correctly correspond to what is spoken in the recording. A new metric known as the transcription correction difficulty index (TCDI) is proposed to quantify the number of interactions the typist would need to perform to correct the transcript. The TCDI is based on the number of changes required in order to change the automatically generated transcript to a correct representation of what was actually said.

Conclusions

The main conclusions of our study were as follows:

- Separating the audio into the speech of individual speakers prior to sending it to the speech recognition systems increased the accuracy of the transcription and made the process of automatically producing a transcript easier.
- A practical speech recognition accuracy metric called Transcription Correction Difficulty Index (TCDI), relevant to our task, was proposed, that captures the effort required to correct the machine-generated transcript and make it ready for use in court.
- The transcripts produced were provided to human transcriptionists for feedback. Their feedback suggested that the total effort required in correcting the automatically generated transcripts was in excess of the time they would take to process the recording manually.
- In the existing interview recording setup, where no special modifications, such as the installation of bespoke microphones, were made to accommodate for the specific speech recognition systems, the tested systems did not provide sufficient recognition accuracy to deal with widely variable linguistic content of spontaneous speech in police witness interviews.
- The combination of speaker segmentation and the use of microphones and other recording conditions modified for specific speech recognition systems can potentially improve the transcription accuracy.

References

A. Alexander, O. Forth and R. How (2010) *'Voice carving in Police Dialogue: Forensic application of Automatic Speaker Segmentation'* in Proceedings of the AES 39th International Conference: Audio Forensics, Denmark

Wang, Y.Y., Acero, A., and Chelba, C (2003). *'Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy?'* Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop, 577–582.