Analysing the performance of automated transcription tools for indistinct audio recordings

Lauren Harrington[1] | Robbie Love[2] | David Wright[3]

[1]University of York | [2]Aston University | [3]Nottingham Trent University

Introduction

- The orthographic transcription of audio recordings can provide important evidence in a forensic case (Fraser, 2021), but producing transcripts is an extremely time-consuming task and is often a prerequisite to further analyses.
- Huge improvements in automatic speech recognition (ASR) have been observed throughout the past two decades, particularly with the recent development of deep learning (Xiong et al., 2016).
- The use of ASR could significantly decrease the amount of time and effort taken to produce a transcript and this could make such systems an attractive prospect to those in law enforcement (Loakes, 2022).
- The appropriacy of ASR for the transcription of indistinct forensic-like audio is worthy of investigation. This paper reports the design and results of a controlled transcription experiment in which twelve automated transcription tools produced transcripts for the same audio recording.

Context

- Many factors negatively affect the accuracy of automatic transcription systems, e.g. spontaneous speech and increased speech rate (Benzeghiba et al., 2007), overlapping speech (Raj et al., 2021), and background noise (Littlefield & Hashemi-Sakhtsari, 2002).
- These factors can be applied to forensic recordings which often involve multiple speakers and are of bad quality (Loakes & Fraser, 2021).
- Loakes (2022) tested two automatic transcription systems on a forensic-like poor-quality recording, and they found that performance was far worse than for a good quality recording, including issues such as consistently identifying non-speech sounds (e.g. drums, laughter) as speech and not transcribing large sections of the recording at all.
- In our study, we compare a longer indistinct recording across a larger set of automatic transcription systems.

ASR transcription systems studied

Descript | Google Cloud | HappyScribe

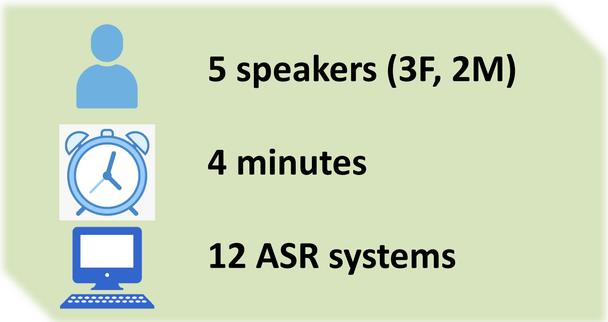
Konch | Microsoft Transcribe | Nvivo

Otter | Sonix | Temi | Transribear

Transcribe by Wreally | Trint

Data and Method

- A conversation between five adults in a busy restaurant was recorded on a smart phone. It shares many of the typical features of forensic recordings, including the presence of multiple speakers, background noise and use of non-specialist recording equipment. It has been found to pose a challenging transcription task for human transcribers (Love & Wright, 2021).



- The recording was processed by 12 free or commercially available ASR systems.
- 18 utterances were identified which are clear enough in their content to be confident of ground truth. We compared the output across the systems, identifying widespread gaps and common mistranscriptions (e.g. Figure 1).

Results

Table 1. Average proportion of matched words in each transcript compared to ground truth transcript across the 18 utterances.

ASR tool	% match	ASR tool	% match
Microsoft Transcribe	70.3	Temi	46.1
Konch	52.7	Transcribear	45.2
Descript	52.1	Transcribe by Wreally	26.4
Trint	51.2	HappyScribe	21.7
Nvivo	49.1	Google Cloud	14.9
Otter	48.2	Sonix	13.9

Examples of mistranscription:

- *ignore > nor, you know*
- *decipher > to say*
- *drunk > dropped*
- *see > say*
- *eyes > item*
- *chicken tikka masala > she can take*
- *calories > because we, characters*
- *samosa > small, similar*
- *worrying > varying*
- *supper > super*
- *deep fried > Deep Throat*

Discussion

- Initial analysis reveals a high level of variability across the twelve transcripts. Variation can be attributed to a few causes, including **phonetic similarity** and the interference of **inappropriate prediction from training data** (cf. Malik et al., 2021) (e.g. *deep fried* was transcribed by five systems, but two mistranscribed *fried* as *Throat*. In both cases, *Deep* and *Throat* are capitalised and seem to be a reference to the US Watergate scandal in the 1970s).
- Across a large sample of readily-available automated speech recognition (ASR) technologies, ASR does not appear to be suitable for the transcription of indistinct recordings for forensic contexts.
- As a result, our view is that, at present, it is more effective for humans to transcribe indistinct audio 'from scratch' as opposed attempting to manage and interpret the output of such systems.

Reference	I	can't	see	in	this	light	or	maybe	my	eyes	just	don't	see
Microsoft		Oh	see		this	place. Slice	well,	maybe	my	eyes	just	don't	see.
Descript		let's	see.				Well,	maybe	my	eyes	just	don't	see
Google Cloud								maybe	my	item.			
HappyScribe							Or	maybe	my	eyes	just	30	
Konch	I	would	say	in	this	place,	well,	maybe	my	eyes	just	don't.	
NVivo					This		well,	maybe	my.				
Otter							will	maybe	my	eyes	just	don't	see
Sonix													
Temi			see,				well,	maybe	my	eyes	just	don't	see
Transcribear							Well,	maybe	my	eyes	just	don't	see.
Transcribe by Wreally													
Trint	I		see		this	place.	Well,	maybe	my	eyes	just	don't	see.

Figure 1. An example of an aligned comparison of the transcripts produced by all twelve automated systems for the utterance "I can't see in this light or maybe my eyes just don't see".

Acknowledgements

This study was supported by the Aston University Institute for Forensic Linguistics (AIFL) seed-corn and network funds programme ("*Evaluating human and automated transcripts of speech recordings: implications for forensic linguistics*"). We are grateful to Dr Claire Dembry for granting us permission on behalf of Cambridge University Press for the re-use of pilot data from the Spoken British National Corpus 2014 project.

Contact

Lauren Harrington
PhD student, Forensic Speech Science
University of York, UK
lauren.harrington@york.ac.uk
@laurenrharr

Robbie Love
Lecturer, English Language
Aston University, UK
r.love@aston.ac.uk
@lovermob

David Wright
Senior Lecturer, Linguistics
Nottingham Trent University, UK
david.wright@ntu.ac.uk
@WrightDW

References

1. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C. and Rose, R., Tyagi, V. & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. Speech communication, 49(10-11), 763-786.
2. Fraser, H. (2021). Forensic transcription: the case for transcription as a dedicated area of linguistic science. in M. Coulthard, A. Johnson, and R. Sousa-Silva (eds.), The Routledge Handbook of Forensic Linguistics. Editors (2nd edn). London: Routledge, pp. 416–431.
3. Littlefield, J., & Hashemi-Sakhtsari, A. (2002). The effects of background noise on the performance of an automatic speech recogniser. DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO SCIENCES LAB.
4. Loakes, D. (2022) Does Automatic Speech Recognition (ASR) Have a Role in the Transcription of Indistinct Covert Recordings for Forensic Purposes? *Frontiers in Communication* Article 803452, Vol. 7, 1-13.
5. Loakes, D. & Fraser, H. (2021). Assessing the role of automatic methods for the transcription of indistinct covert recordings. In 29th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Marburg, Germany [online].
6. Love, R., & Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. Frontiers in Communication, 6:797448 (Research Topic: 'Capturing talk: The institutional practices surrounding the transcription of spoken language'). DOI: 10.3389/fcomm.2021.797448
7. Malik, M., Malik, M. K., Mehmood, K., and Makhdoom, I. (2021). Automatic speech recognition: a survey. Multimed. Tools. Appl. 80, 9411–9457.doi: 10.1007/s11042-020-10073-7
8. Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y. & Kanda, N. (2021, January). Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 897-904.
9. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D. & Zweig, G. (2017). Toward human parity in conversational speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 2410-2423.