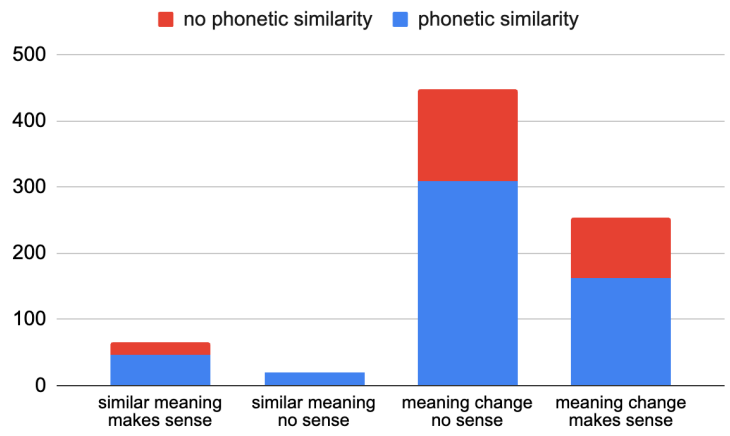
Automatic transcription in policing contexts: a linguistic evaluation of performance

Lauren Harrington[1], Jessica Wormald[1] & Tallulah Buckley[2]
[1]University of York, [2]University of Cambridge

The use of automatic speech recognition (ASR) software has become increasingly accessible to those working with recordings of spoken data, including police forces in England and Wales. ASR offers an attractive solution to the time-consuming nature of transcription, but the performance of ASR in police settings remains underexplored. The aim of this study is to better understand how ASR performs on real spoken data and to evaluate the impact of errors from a forensic linguistic perspective.

Recordings containing sociolinguistic interviews were extracted from the Punjabi English in Bradford and Leicester corpus (Wormald 2016). A subset of 13 male speakers aged between 19 and 29 from Bradford and Leicester was used for this study. Files were edited to contain the voice of a single speaker and limited to 5 minutes in duration, and were transcribed using CrisperWhisper (Wagner et al. 2024). Performance was evaluated at the word-level, such that pairings of the uttered word and ASR transcription were categorised as either correct or as an error (substitution, deletion, insertion).

Given the forensic focus of our work, we further investigated the substitution errors (i.e. mistranscribed words) by annotating each error according to (a) whether the word *"made sense" in context*, (b) whether there was a considerable *change in meaning*, and (c) the degree of *phonetic similarity* between the reference and ASR-transcribed word. This allows us to interpret the potential impact of these errors within a human verification process, whereby a human (e.g. police transcriber) is tasked with checking the content of an ASR transcript.



The ASR system has an overall error rate of 10.3%, and around 60% of the errors involve the mistranscription of a word (i.e. a substitution error). Figure 1 shows the number of substitution errors within 4 categories, grouped according to whether the mistranscribed word has a similar meaning to the reference word and whether it makes sense in context. Overall, we find that two thirds of substitution errors (67%) have a moderate to very high level of phonetic similarity with the target word. Furthermore, 32.4% of substitutions fall into the most potentially harmful category - whereby there is a change in meaning but the error makes sense when read in context - and 63.4% of these errors are phonetically similar to the reference word. We theorise that errors which are phonetically similar and make sense in context will be most challenging to identify, and through this work aim to raise awareness of the prevalence of these errors within ASR transcripts and equip police transcribers with the skills to effectively recognise and correct them.

References

Wagner, L., Thallinger, B., & Zusag, M. (2024). CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions. *Proceedings of Interspeech*, 1265-1269.

Wormald, J. (2016). *Regional Variation in Panjabi-English*. PhD thesis, University of York.