Transcription in noise: investigating how errors produced by lay listeners vary across increasing noise levels

Lauren Harrington
Department of Language and Linguistic Science, University of York, York, UK
lauren.harrington@york.ac.uk

Transcripts of speech evidence are used for investigative and evidential purposes in the UK criminal justice system. A range of people may be responsible for producing transcripts of forensic audio materials, i.e. recordings which typically suffer from poor quality audio conditions. These include forensic phonetics experts, private transcription agencies, police detectives or other police employees such as interview transcribers (Tompkinson et al., 2022). Outside of linguistics, transcription is widely viewed as an 'administrative burden' and its complexities are often underappreciated. Given this, and the practicalities and costs associated with hiring an expert, it is likely that the majority of forensic audio materials are transcribed by lay people (i.e. transcribers without linguistic or phonetic training), without systematic methods for dealing with challenging audio. It is well documented that transcription accuracy declines for poorer quality audio (e.g. Lange et al., 2011). This study investigates *how* transcription performance declines across conditions, by exploring the types of errors produced by transcribers, the impact of those errors, and how their distribution varies across different noise conditions. It extends previous work, which tends to focus on accuracy rates rather than taking into account the actual content of the transcripts.

135 participants from across the UK were recruited via Prolific to take part in an online transcription experiment. Each participant was presented with 24 stimuli, comprising short extracts from the mock police interview task from the DyViS corpus (Nolan et al., 2009). Each extract contained between 3-6 seconds of speech from an individual male speaker of Standard Southern British English. The stimuli were mixed with speech-shaped noise in Praat (Boersma & Weenink, 2022) at signal-to-noise ratios (SNR) of 6 dB, 0 dB and -3 dB, which represent 'fair', 'moderate' and 'poor' intelligibility respectively. Participants were presented with 8 stimuli at each SNR and were instructed to transcribe all of the speech in the audio file. Participant transcripts were compared at the word level with a reference transcript, which was produced by the author using the studio quality version of the files. Word pairings were marked as a match or assigned an error category - substitution, insertion, or deletion.

As expected, performance declines with increased noise levels (i.e. as the SNR decreases). An average word accuracy rate of 84.4% is achieved by participants in the 'fair' condition (6 dB SNR), which drops to 69.2% in the 'moderate' condition (0 dB SNR) and 49.4% in the 'poor' condition (-3 dB SNR). An ANOVA comparing a null model to a model including SNR as a fixed effect revealed a statistically significant effect of noise on transcription accuracy ($\chi^2 = 177.72$, $df = 4$, $p < .001$).

To better understand the nature of the decline in performance, the specific types of transcription errors were analysed. Deletions account for the majority of transcription errors in all conditions, comprising an increasing proportion of the total errors as noise increases (57.2% in 6 dB SNR, 64.1% in 0 dB SNR, 73.1% in -3 dB SNR). Substitution errors, whereby the listener has transcribed an incorrect word, are also frequently produced in all noise conditions, accounting for around a quarter to a third of the total errors (34.4% in 6 dB SNR, 29.8% in 0 dB SNR, 22.8% in -3 dB SNR). Insertion errors are relatively rare and account for only 4-9% of errors within each condition. Overall, these findings suggest that listeners adopt a more conservative approach to transcription when noise levels are high, opting to omit words entirely rather than offering an interpretation. However, it should be noted that although deletions make up an increasingly bigger proportion of errors as the noise level increases, substitution errors are still produced more frequently in the noisier conditions.

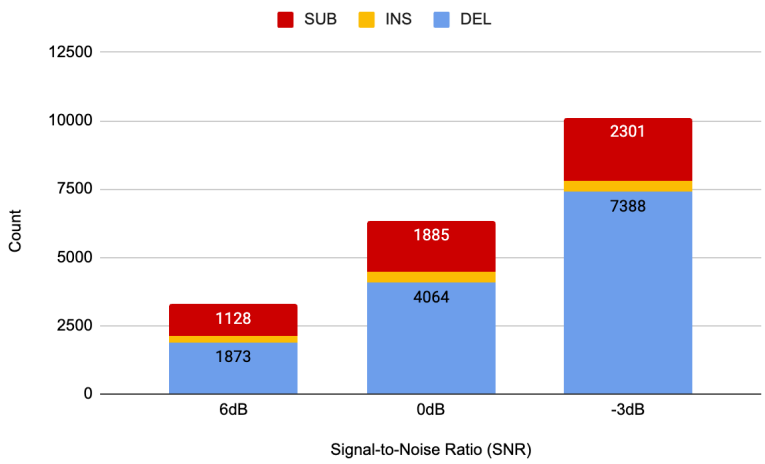

Figure 1. Raw counts of each error type (substitution, insertion deletion) within each noise condition.

Substitutions are of particular interest given that they may prime a reader of the transcript to hear the mistranscribed content (e.g. Fraser & Kinoshita, 2021). This poster will further explore the types of substitution errors produced across noise conditions, and the impact of those errors on the resulting transcription. As demonstrated by the examples in Table 1, many of the participant transcriptions in the least noisy condition tend to be either entirely correct or contain relatively minor grammatical substitutions, but as noise increases the transcriptions can become more nonsensical or result in a change in meaning. These findings will offer critical insights into the specific types of transcription errors lay listeners make when attempting to transcribe poor quality audio materials.

SNR	*Participant transcriptions*
6 dB	I've been in there a couple of times but I've never got to know any of the people that work there
	I've been in there a couple of times but I've never got to know anybody who works there
0 dB	I've been in there a couple of times but I didn't particularly know any of the people who work there
	I've been [...] there a couple of times but I think it's [...] only the people who work there
-3 dB	I really don't have time but I do [...] know people that work there
	[...] It's hard when you spend all day working

Table 1. Example participant transcriptions within each noise condition for the utterance "*I've been in there a couple of times but I've never got to know any of the people who work there*". Substitution errors are highlighted in grey and deletion errors are marked with [...].

References

Boersma, P., & Weenink, D. (2020). Doing phonetics by computer [Computer program] (Version 6.1.30).

Fraser, H., & Kinoshita, Y. (2021). Injustice arising from the unnoticed power of priming: How lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio. *Criminal Law Journal, 45*(3), 142–152.

Lange, N. D., Thomas, R. P., Dana, J., & Dawes, R. M. (2011). Contextual biases in the interpretation of auditory evidence. *Law and Human Behavior, 35*(3), 178–187.

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL, 16*(1), 31-57.

Tompkinson, J., Haworth, K., & Richardson, E. (2022). *For the record: assessing force-level variation in the transcription of police-suspect interviews in England and Wales*. iIIRG Conference, Winchester.