

Towards improving transcripts of audio recordings in the criminal justice system

Lauren Harrington

PhD

University of York

Language and Linguistic Science

February 2024

Abstract

Transcripts of speech evidence are often presented to juries for the purpose of providing a written record of speech within an audio recording and/or aiding members of the court in understanding the speech content of poor quality recordings. The accuracy and impartiality of these transcripts is paramount, but such qualities require systematic and scientifically-informed methods. There are three main aims of this thesis: (1) to gain a better understanding of current practices when transcribing poor quality evidential audio, (2) to explore how background noise and regional accents affect the content of transcripts, and (3) to develop a method for evaluating transcripts for forensic purposes. The first aim is explored through a survey of expert transcription practices and a focus interview regarding common issues encountered in non-expert transcripts; both studies demonstrate variability in the methods employed to produce transcripts. A push towards standardisation is recommended, encouraging (a) further research on transcription methods, focusing on method validation and proficiency testing, and (b) the production of standards and/or guidelines. The second and third aims of the thesis are addressed in two studies in which human and automatic transcription performance is compared across different audio qualities and accents. Findings reveal that transcripts are significantly worse for audio with increased background noise and substantially worse for unfamiliar accents. A new forensically-motivated method of evaluating transcripts is employed in these studies, focusing on substitution errors and their potential impact on meaning; this can be used (or developed) for further research and proficiency testing. The work in this thesis shows a huge research gap concerning the production of transcripts for use in the criminal justice system, which needs to be addressed by further empirical testing of transcription methods, human and automatic performance, and human-automatic hybrid approaches to transcription.

Table of Contents

Abstract.....	2
Table of Contents.....	3
List of Figures.....	4
List of Tables.....	6
Acknowledgments.....	9
Author's declaration.....	10
1. Introduction.....	12
2. Research background.....	31
3. Thesis aims.....	58
4. Article 1 - Forensic transcription practices: survey results from experts in Europe and North America.....	62
5. Additional resource - Focus interview: non-expert transcripts in England and Wales.....	95
6. Article 2 - A forensic approach to transcription errors: factors affecting human performance.....	111
7. Article 3 - Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance.....	156
8. Discussion.....	199
9. Appendix.....	230

List of Figures

Article 1

Figure 1: Countries or jurisdictions in which the respondents primarily work, ordered from most to least frequently answered.....	70
Figure 2: How frequently respondents carry out transcription as part of their work duties....	71
Figure 3: How often respondents employ phonetic and/or acoustic analysis during the production of a transcript.....	74
Figure 4: Responses concerning whether audio recordings are typically enhanced before transcription.....	75
Figure 5: Distribution of respondents' ratings for each factor.....	78
Figure 6: Responses concerning the existence of protocol or precures to mitigate the effects of cognitive bias and priming.....	79
Figure 7: Responses concerning reference to transcripts produced by the instructing party.....	80

Article 2

Figure 1: Alignment of reference transcript (Truth) and participant transcript (Hypothesis) using a custom-built tool. Cells highlighted in yellow indicate a non-match, i.e. an error.....	123
Figure 2: Percentage of words correctly transcribed across accent groups (SSBE speakers, non-SSBE speakers) and different signal-to-noise ratios (SNR; 6 dB, 0 dB and -3 dB). The level of background noise increases from left to right.....	128
Figure 3: Percentage of words correctly transcribed across accent groups (SSBE speakers, non-SSBE speakers) and different difficulty ratings (where 1 represents least difficult and 5 represents most difficult).....	129
Figure 4: Raw number of errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of deletions (yellow), substitutions (orange) and insertions (red), and percentages within each noise condition are included. The level of background noise increases from left to right on each plot.....	130
Figure 5: Raw number of substitution errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of minor errors (purple) and major errors (pink). The level of background noise increases from left to right on each plot.....	133
Figure 6: Raw number of substitution errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of function words that were	

substituted (dark green) and content words that were substituted (light green). The level of background noise increases from left to right on each plot..... 134

Article 3

Figure 1: Average word error rate in each of the four conditions (SSBE studio, SSBE SSN, WYE studio, and WYE SSN) for all three ASR systems (Amazon, Rev, and Google). ASR systems are ordered from left to right according to lowest to highest average WER..... 174

Figure 2: Number of substitution errors produced by each ASR system in each accent, in studio condition (left) and speech-shaped noise condition (right). ASR systems are ordered from left to right according to lowest to highest average WER..... 178

List of Tables

Introduction

Table 1: Example transcript of a call to the emergency services in which there are three speakers: the emergency services operator, the caller and a male speaker in the background.....	12
Table 2: Typical situations where experts are instructed to prepare transcripts. Text copied from section 1 of Article 1.....	15
Table 3: Examples of different types of cognitive bias, as defined by the UK Forensic Science Regulator in FSR-G-217 ‘Cognitive Bias Effects: Relevant to Forensic Science Examinations’.....	23

Research background

Table 1: Factors that may affect the creation of transcripts according to Fraser (2022, pp. 5-6).....	47
Table 2: Comparison of two transcripts with a reference transcript, with errors highlighted. A shaded red cell shows a deletion, bold red text shows a substitution, and bold blue text shows an insertion.....	51

Article 1

Table 1: Working arrangements or affiliations of respondents.....	71
Table 2: Number of transcribers that typically work on a transcript.....	72
Table 3: Number of drafts typically produced.....	73
Table 4: Responses concerning the use of automatic speech recognition during the production of a transcript.....	75
Table 5: Ways in which sections of lower confidence are represented within transcripts.....	76
Table 6: Ways in which unintelligible speech is represented within transcripts.....	77
Table 8: Responses concerning respondent’s views on whether cognitive bias could play a significant role on the transcription of forensic audio materials.....	78

Article 2

Table 1: Comparison of two transcripts with a reference transcript, with errors highlighted. A shaded red cell shows a deletion, bold red text shows a substitution, and bold blue text	
---	--

shows an insertion.....	118
Table 2: Classifications of different types of substitution errors.....	125
Table 3: Output of model comparisons between the full model (including accent background and level of background noise/SNR) and the null models which excluded each of the independent variables in turn.....	128
Table 4: Multinomial logistic regression analysis of the distribution of error types across conditions. Results from the model output (beta and Standard Error) were used to calculate the odds ratio (OR) and upper and lower confidence intervals (CI). Statistical significance is demonstrated with asterisks according to the degree of significance.....	132

Article 3

Table 1: Examples of linguistic content of stimuli from each speaker.....	169
Table 2: Two potential transcriptions of the utterance “packet of gum in the car”. Deletions are represented by a shaded red cell and substitutions are represented by bolded red text.....	171
Table 3: Phonetic realisations of four vocalic variables across the two varieties of British English analysed in this study, Standard Southern British English (SSBE) and West Yorkshire English (WYE). Variables are defined using Wells’ (1982) lexical sets.....	172
Table 4: Counts of each error type (insertions, deletions, and substitutions) produced by each system for Standard Southern British English speech. SSN refers to the audio quality with added speech-shaped noise.....	176
Table 5: Counts of each error type (insertions, deletions, and substitutions) produced by each system for West Yorkshire English speech. SSN refers to the audio quality with added speech-shaped noise.....	177
Table 6: Examples from the data of substitution errors involving pronouns. Words involved in substitution are highlighted in bold red text.....	184

Discussion

Table 1: Blank transcript template that could be used by non-expert transcribers. See Appendix A for guidance on how to use the template.....	202
---	-----

Appendix

Table 1: List of conventions that can be used in a transcript.....	233
Table 2: Blank template of transcript. The transcript is a table made up of 4 columns.....	235

Table 3: Blank template of transcript that can be used in cases of overlapping speech. The transcript is made up of six columns.....	235
Table 4: Transcript template that has been filled in.....	236
Table 5: Transcript template for overlapping speech that has been filled in.....	237

Acknowledgments

I am so grateful to have worked with two of the most wonderful supervisors to exist (a bold claim, I know!). Vince and Rich - thank you so much for your support and your guidance over the last three and a half years. You have taught me so much and I really do feel proud of what I have achieved with your help. And thank you both for being so *chill*. Your calm vibes have pulled me out of a mental breakdown more times that I can count...

There are so many others in the department at the University of York to thank. Paul Foulkes for introducing me to forensics all those years ago. Phil Harrison for your (frequent!) technical help. Eleanor Chodroff (now at the University of Zurich) and Shayne Sloggett for your guidance as TAP members. James Tompkinson for sharing my interest in police interview transcription and reviewing parts of this monster. Thanks also to Kate Earnshaw and Bryony Nuttall at the Forensic Voice Centre for helping out at various points.

I was lucky enough to travel to Melbourne for a five week research project with Helen Fraser and Debbie Loakes at the Research Hub for Language in Forensic Evidence. Thank you for welcoming me with open arms and making my trip down under so interesting! I feel privileged to have been able to take part in your masterclasses and to have given talks alongside you both. Helen - your guidance on issues surrounding transcription has been invaluable to my research, as I'm sure you'll see from the number of references to your work in this thesis!

To my friends. Alice, you deserve the number one spot - thank you for listening to hundreds of hours of voice notes over the last four years, and for providing support and feedback when I needed it most. I'm so grateful for our friendship and I'm excited for us to develop our academic careers together. Ben, you have been by my side for so many experiences during this PhD - it really wouldn't have been the same without you. To Leah, Chloe, Grace, Vic and Jamie - thank you for being part of this experience.

Finally, to Ryan - my biggest supporter and my best friend. Thank you for the time and effort you put into developing software for this PhD (!), for countless Excel formulae, for proofreading so many drafts, and for listening to me endlessly ramble about transcription. Thank you for the coffees, and the rice pudding, and the comfort of a hug when the stress got too much. I love you.

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author unless otherwise indicated in the 'Research Degree Thesis Statement of Authorship' that precedes each paper. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Article 1 "Forensic transcription practices: survey results from experts in Europe and North America" has been submitted for publication in the *International Journal of Speech, Language and the Law* and is currently under review.

Article 3 "Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance" has previously been published:

- Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication*, 8, 1165233.

Signed: 

Date: 29th February 2024

For Grandma & Grandpa

*I'm so happy you're here
to see me finish this*

1. Introduction

This thesis concerns the production of transcripts within the criminal justice system. Before exploring the ways in which transcripts are used and produced within this context, it is first necessary to define what a ‘transcript’ is. A transcript is a document that presents, in written form, the speech content of an event such as an interview or meeting. In most situations, transcripts will be produced by using an audio or video recording of the speech. Transcripts often contain an indication of who is speaking and most often attempt to capture the speech in a verbatim manner. An example of a transcript is presented in Table 1:

Speaker	Speech
Operator	Emergency, which service?
Caller	Ambu- no, police please.
Male 1	Hey, put that phone down, now.
Caller	Reg, stop it.

Table 1: Example transcript of a call to the emergency services in which there are three speakers: the emergency services operator, the caller and a male speaker in the background.

Evidence presented in court cases can often feature recordings of speech; perhaps the suspect confessed to the crime in a police interview, or CCTV footage captured the audio from an off-camera altercation at the crime scene, or the offender’s voice can be heard in the background of a telephone call to the emergency services. In England and Wales, if a particular statement made within an audio recording “demonstrates the commission of the offence or has other evidential value” (Crown Prosecution Service, 2020, Annex 3) and is deemed admissible as evidence, then the audio recording can be presented to the court to prove the authenticity of the statement. A transcript of the speech content may be produced as “an administrative convenience” (Health and Safety Executive, n.d.). The transcript itself is not the evidence; it is viewed as an ‘aid’ to the court, though in some cases, the transcript may be presented without the audio and instead read aloud by a member of the court, such as a barrister, a judge or a witness.

Recordings used for evidential purposes often suffer from poor intelligibility due to the way in which the audio has been collected; the speech may be captured using covert recording devices planted in a location such as a suspect’s car, and so the noise of the engine or radio may distort the speech; or the speech of interest is taking place in the background of a

phone call to the emergency services; or the speaker may be moving around and the microphone only clearly picks up short portions of what they are saying. Multiple speakers, overlapping speech, background noise, and channel distortion are all factors which can contribute to this type of recording being extremely challenging to understand and to follow. In the worst case scenario, without high quality listening equipment, it can be difficult to make out any words at all.

First and foremost, the purpose of a transcript is to serve as a written record of speech, acting as a helpful reminder of what was said. It is much easier to refer back to a written transcript than an audio recording, particularly in settings where there are time and technological constraints that make repeated instances of audio playback challenging. Another important purpose of a transcript is to help the reader follow the speech within an audio recording, which can be challenging during an interaction involving multiple speakers; or even to help them understand the speech in cases of poor intelligibility, such as those described in the paragraph above, where it may be extremely challenging for the listener to understand any of what is being said without the aid of a transcript.

In the criminal justice system, transcripts are used in a range of different ways and at different stages of the legal process. Police detectives may use a transcript of, for example, a suspect's police interview or audio from a surveillance device to aid their investigation. A transcript may be relevant in determining whether the Crown Prosecution Service will pursue a prosecution, and eventually a transcript may be admitted as evidence to the court during a trial. In the case of forensic speech experts, who are often called upon to give expert testimony in cases involving speech or voice evidence, transcripts may be produced on the instruction of the prosecution or defence to provide an expert opinion about what was (or wasn't) said, or the expert may produce a transcript for personal use to aid their investigation in speaker comparison casework.

The end user of these transcripts therefore varies according to the purpose of the transcript. Perhaps the transcript is produced and used only by the transcriber in cases of expert casework or police investigations; or the transcript is considered by lawyers and a judge in a pretrial investigation for a case which does not end up going to trial; or the transcript is presented to the jury alongside an audio recording as key evidence during a trial. This thesis focuses on the final context, where the transcript is used for evidential purposes and the end user is members of the court and, in particular, the jury. It is in this context that the effects of inaccurate or misleading transcripts are perhaps most sharply felt, given the extremely powerful influence that a transcript can have over the way in which a listener interprets the

speech within an audio recording; once an inaccurate transcript is presented, members of the jury may never be able to ‘unhear’ its contents (Fraser et al., 2011). Should those contents be falsely incriminating, this could lead to a miscarriage of justice. These issues will be discussed in greater detail in section 2.3.2.

There are two main types of transcripts that could be presented to a jury: those representing the speech within a poor quality audio recording containing forensic evidence, or those recording the contents of a police interview. However, the ways in which these transcripts are produced are very different, and each has its own issues. These issues are explored in the following subsections.

1.1 Transcripts of evidential recordings

Speech evidence can be extremely powerful; hearing someone commit or even confess to a crime on tape could completely change a court case. Table 1 presents a range of typical recordings for which forensic practitioners are instructed to produce a transcript. Those marked with an asterisk comprise the most common types of recording to be transcribed by forensic practitioners in England and Wales (Richard Rhodes, personal communication). Throughout this thesis, “forensic practitioner” is a term used to describe a forensic phonetics expert who (a) carries out forensic transcription casework as part of their work duties, (b) has specialist qualifications, training and/or experience in forensic speech science, and (c) who would be considered an expert witness in court.

Type of recording	Description
Telephone calls to the emergency services*	All emergency calls are recorded, and can contain key evidence. The speech of interest is often in the background of the call, and may be simultaneous to interactions between the caller and the emergency operator.
Other types of telephone call	For example, bank or insurance calls in fraud cases.
Covert or undercover officer recordings*	Recording devices planted in locations or operated by undercover officers can capture relevant conversations from target speakers.
Recording of conversations made (non-)deliberately by a participant	Some speakers accidentally or deliberately/covertly record themselves or others making admissions or discussing offences; this might be done using a smartphone voice recording application.
CCTV recordings*	CCTV recordings might show a violent incident, a burglary or discussions about offending, for example. These can include fixed CCTV camera systems in houses or commercial properties, video doorbells (such as <i>Ring</i> or <i>Nest</i> devices), or portable devices such as body-worn cameras which are increasingly being used by police and security staff.
Social media videos or recordings	Material showing, or relating to, offences may be posted online or found on seized digital devices, often smartphones. These can include videos from social platforms like <i>Facebook</i> , <i>X</i> or <i>YouTube</i> , voice messages from <i>WhatsApp</i> , <i>Snapchat</i> and similar applications.

Table 2: Typical situations where experts are instructed to prepare transcripts. Text copied from section 1 of Article 1.

Telephone calls to the emergency services are one of the most commonly transcribed types of recordings, but it is not always the caller-operator speech that is of interest. A simultaneous interaction may be taking place in the background of the phone call; perhaps the caller is requesting police presence due to a drunken ex-partner banging on their door

and shouting abuse; or the caller is requesting an ambulance due to a verbal and physical fight that has broken out at a pub. The speech of a suspect could be captured in the background of the phone call, placing them at the scene and taking part in an altercation. However, the words that are being said can be extremely challenging to make out due to overlapping speech, the distance of the speaker from the microphone, and issues related to telephone transmission.

In cases where speech evidence plays a key role in a trial, the standard procedure in England and Wales is for the audio recording to be played aloud for the court to hear, and members of the court will be provided with a transcript to help them follow along with the speech. The audio playback procedures in English and Welsh courts often involve loudspeakers in a corner of the courtroom, far away from the jury and in a room filled with wooden panelling and glass, i.e. two extremely sound-reflective materials. When these speakers fail, lawyers have been known to resort to playing the recording aloud on their laptop and holding a courtroom microphone next to the laptop speaker (see Section 5: “Additional resource”). Suffice to say, the jury is very rarely presented with the audio evidence in a high quality manner that is comparable with how an expert would listen when transcribing, and instead deal with audio playback procedures that decrease the intelligibility of speech within an evidential recording even further.

In some contexts the speech will be extremely challenging to understand without the aid of a transcript; perhaps there are too many people speaking at once to follow what is going on, or there is so much noise that it is hard to make out what is speech. In such cases the transcript is viewed as a ‘listening tool’ to help the jury understand the words being spoken in the recording, while the audio recording remains the main evidence. However, the poor intelligibility of the speech means that it would be very easy for a jury to be misled by an inaccurate transcript (see section 2.3 on priming for a more in depth exploration of this issue). It is therefore very important that transcripts provided to juries are accurate and reliable.

1.1.1 Expert transcribers

When the speech is particularly challenging to understand, experts in forensic speech science may be approached by the police or lawyers to transcribe recordings of forensic interest. Experts are well-placed to carry out such a task due to their expertise in phonetics and because they have a higher level of awareness of the limitations of transcription, particularly with regard to poor quality audio, and of factors that may affect their ability or

influence their perception. However, there is a lack of general consensus within the forensic speech science community regarding the optimal way to produce transcripts of evidential audio recordings, and there has been very little research conducted on the methods that experts do employ.

Forensic transcription, i.e. the transcription of (often poor quality) evidential audio recordings, is a perfect example of a forensic science that has developed in an 'ad hoc' way; each forensic speech laboratory seems to have developed their own practices according to what they believe to be best practice (see Article 1 for the results of a survey about expert transcription practices), often using findings from a range of fields, such as psycholinguistics and the forensic sciences in general, to guide their methods. Other than a small-scale proficiency test developed by Tschäpe and Wagner (2012), the findings of which suggest that experts perform better than lay people when transcribing poor quality evidential audio and that groups of experts perform better than individual experts, there has not yet been any detailed investigation into which practices produce the best, most reliable transcripts.

This lack of standardised and well-researched methods may cause issues for forensic speech experts, particularly in the UK context due to the Forensic Science Regulator's recent focus on the validation of methods. According to the UK Forensic Science Regulator (FSR), validation is the "process of providing objective evidence that a method, process or device is fit for the specific purpose intended" (Code of Practice v1, p. 361). Validation therefore involves the testing of documented procedures in order to understand the performance of the method and to show that the method is appropriate for the task at hand.

The FSR defines Forensic Science Activities (FSA) and determines which of those require accreditation to ISO/IEC 17025 and the Statutory Code of Practice. Questioned Content Analysis, which encompasses expert transcription and disputed utterance analysis, is a recognised FSA but currently does not require accreditation¹ (Code of Practice v1, Part F - DIG401); however, the FSR sets standards and offers guidance² even for fields that do not require accreditation, and forensic providers are expected to follow the Code regardless of accreditation requirements. For forensic speech experts, this is reinforced by many clients now enquiring about the accreditation status of forensic speech and audio laboratories and lawyers asking in court about this regulation (Richard Rhodes, personal communication).

¹ This has recently been explicitly stated in a new draft of the FSR Code of Practice released in February 2024; see point (b) of 96.3.1 of this document: [Draft Code of Practice](#)

² Relevant guidance includes Legal Obligations on expert witnesses, Cognitive Bias Effects, Validation, Reports, and Expert's meetings.

Some of the main questions that forensic speech experts are asked, and are obligated to answer in their reports, concern the following areas of the Code:

- Competence of the practitioners involved in the work
- Validity of the methods employed
- Documentation of the methods employed
- Whether the equipment/software used is tested and is fit for purpose
- Whether the work is undertaken in a suitable environment

There is clearly a need to demonstrate that expert transcribers are competent, the methods employed are valid and follow a standard procedure, and the right equipment and software are used. Without being able to clearly show these things, the following issues could arise:

- Police, the jury or the court may lack confidence in the expert or the type of expertise
- Evidence may be deemed inadmissible³ or afforded less weight
- Evidence may not be procured in the first place, due to the customer identifying a lack of validation as an issue
- Expert could be subjected to criticism

However, as it stands, very little of what this actually entails is known; there are no guidelines for the training required by expert transcribers, no standardised proficiency testing, and no research into the validity of the methods used and which equipment is best suited for the task.

1.1.2 Non-expert transcribers

Although forensic experts are subject to these factors, evidential audio can equally be transcribed by non-experts, such as police officers, to whom none of the above applies. This is because transcription by non-experts does not fall under the remit of the FSR and is excluded from the Speech and Audio analysis FSAs⁴, meaning there is no requirement for regulation or accreditation. Given the amount of recordings that require transcription, as well as financial and time constraints, it is likely that the majority of evidential recordings will be

³ According to section 4.2.6 of the FSR guidance on validation, the Criminal Practice Directions suggest that validity should be the first thing considered by the courts in cases of questioned admissibility (FSR-G-201, 2020).

⁴ This has recently been explicitly stated in a new draft of the FSR Code of Practice released in February 2024; see point (e) of 96.5.1 of the draft FSR Code document: [Draft Code of Practice](#)

transcribed by non-experts, such as police detectives or others employed by the police⁵. Little is known about the production of transcripts of evidential audio recordings by non-experts, though the huge variability in the quality of such transcripts suggests that they are produced in an ‘ad hoc’ manner with no guidelines or standards in place.

Two forensic speech experts, who have a combined total of over 30 years’ experience with forensic casework, were interviewed as part of this thesis about some of the most common issues that are encountered within non-expert transcripts (see Section 5: “Additional resource”). These include⁶:

- Lack of time points, making it very difficult to locate whereabouts in the recording a transcribed utterance occurs
- No standard formatting, such that some transcripts are laid out in a clear manner with one line per speaker, while others consist of a single block of text
- Inappropriate or incorrect speaker attribution, whereby a transcriber attributes speech to named speakers despite questions about their status, or misattributing speech of two individuals for a portion of the transcript before switching back to correct attribution
- No indication of a switch between verbatim and summarised speech, making a transcript very challenging to follow alongside the audio
- Attempts to transcribe unintelligible portions of recordings, or marking clear speech as ‘inaudible’ due to lack of understanding of dialectal or slang forms

The main problem with non-expert transcripts that the interviewed experts recounted was the extent of the variability found within such transcripts. While some of the transcribers do a relatively good job at representing the speech within the recording, others produce transcripts of a much lower quality. Many of the issues found in non-expert transcripts stem from (a) a lack of understanding on the part of the transcribers regarding speech perception and audio factors, (b) inexperienced transcribers who have no guidelines to follow, and (c) a lack of resources and investment in this task.

⁵ Results of a Freedom of Interest request reported in Tompkinson et al. (2022) reveal that some transcribers employed to produce transcripts of police interviews (also known as ROTI clerks) also transcribe other types of recordings, including those which may be evidential in nature.

⁶ Richardson et al. (2022) highlight a number of similar issues in (non-expert) police-suspect interview transcripts.

1.2 Transcripts of police-suspect interviews

A person suspected of being involved in a crime will be interviewed by the police in order to obtain “accurate and reliable accounts... about matters under police investigation” (College of Policing, 2022, Principle 1) from the interviewee. All interviews conducted with a person suspected of committing a crime must be audio-recorded according to Code E (the revised Code of Practice on Audio Recording Interviews with Suspects) of the Police and Criminal Evidence Act 1984 (PACE). Prior to PACE, the interviewing officer would produce from memory an official interview record some time after the event. Such procedures allowed for the corruption or even fabrication of interview records by police officers (Haworth, 2018); the act of recording the audio is supposed to ensure that an accurate record of the interview can be made. Interviews can be conducted in a range of locations, such as at a police station or in a prison, but should most importantly take place in a quiet and controlled environment so that the recording device can pick up the speech of each of the participants, all of whom are aware of its presence.

Following the recording of a police-suspect interview, a ‘Record of Taped Interview’ (ROTI) can be produced by a ROTI clerk (a police employee whose role is to transcribe police interviews) as a written record of the interview contents, and this will be used by police officers in their investigation and by the defence in advising their client. If the case goes to court then the audio recording and accompanying ROTI “become part of the prosecution case” (Haworth, 2018, p. 433) and the contents of the interview can be presented to the jury as evidence. Although the audio recording of the interview is considered the ‘real’ evidence, the ROTI is admissible as a ‘copy’ of the audio, which often leads to a complete reliance on the ROTI as an official record of the contents of the police-suspect interview.

However, a prominent issue with ROTIs is the fact that they can be little more than a summary of the police interview, with some parts, which are chosen by the ROTI clerk, transcribed verbatim. In many instances this is not an issue as the transcript is produced as a formality and is not used for any other purposes; for instance, it may be agreed that no further action will be taken against the interviewee or the offence may be relatively minor with no facts in dispute and, as such, a record of the police interview is not required (Richardson et al., 2022). However, if the case goes to trial, as described above, the ROTI will often be presented to the court as evidence (Haworth, 2018), and if sections of interest have been omitted, mistranscribed or not accurately portrayed in a summarised section, the content of the ROTI could be used against the interviewee; according to section 34 of the Criminal Justice and Public Order Act 1994, if a suspect fails to mention something in their

police interview that is later relied on as part of their defence, then the court is entitled to 'draw inferences' as to why it was not mentioned sooner (Haworth, 2018). Inconsistencies between the content of the transcript and what the interviewee says in court, which could potentially arise as a result of a ROTI clerk's decisions about what to transcribe and what to summarise, could therefore be extremely damaging. An additional issue with summarising sections of the police-suspect interview is the requirement for reporting verbs in this style of writing. The choice of verb is yet another way in which subjectivity and transcriber interpretation (Haworth et al., 2023) can permeate ROTIs. For example, words such as "admitted" or "denied" have very different connotations to a more neutral verb like "said" or "stated".

Furthermore, it seems that the production of ROTIs is predominantly viewed as an administrative role with GCSE-level qualification requirements and a focus on audio and copy typing skills (Tompkinson et al., 2022). However, ROTI clerks are tasked with selecting which parts of a police-suspect interview may later be relevant for investigative or evidential purposes. Haworth highlights the problematic nature of this allocation of tasks, stating that it seems "entirely unrealistic to expect those with no legal qualifications to understand and apply" (Haworth, 2018, p. 15) working legal knowledge regarding each crime and what may be deemed relevant for the prosecution's case; this leaves room for error and the misapplication of knowledge.

A team of researchers at Aston University, led by Kate Haworth, has investigated the production of ROTIs in recent years (see Haworth et al., 2023) and found, among many other issues, a huge amount of variability in the transcripts produced by ROTI clerks. Transcribers receive no training on problematic aspects of the transcription process and very little guidance to follow when producing transcripts (Haworth, 2018) to such a point that individual transcribers within a police force end up developing their own practices. This variability was not seen as concerning among the ROTI clerks in a focus group interview conducted by Haworth (2018) and, in fact, was praised as transcribers having 'individual style and flare'. An example provided by Haworth (2018) demonstrates three different ways in which ROTI clerks would represent silence on the part of the interviewee following an interviewer's question: "no audible reply", "defendant remained silent", and "defendant refuses to reply". Each interpretation of the silence potentially portrays the interviewee in a different light: the first suggests that the interviewee may have either (a) said something that couldn't be heard on the recording or (b) responded in a non-audible way, e.g. shrugging their shoulders; the second suggests prolonged silence on the part of the interviewee, who the transcriber assumes will be charged with a crime through their use of the term

‘defendant’; and the third, and most problematic of the interpretations, suggests an active process of refusal by the interviewee. Such inconsistency across transcribers is compounded by the fact that police forces in England and Wales operate individually.

1.3 Standardisation of methods

Most of the issues described above stem from a lack of standardised methods.

Standardisation is one step towards the ultimate goal of better, more accurate, and more impartial transcripts being presented to juries, achieved through the use of robust methods founded on linguistic and psychological knowledge.

For experts in England and Wales, the lack of standardisation can cause issues because they are expected to follow the FSR Code of Practice which requires the methods employed to be valid⁷ and to follow a standard procedure⁸. However, in order to establish a standard procedure and to test its validity, it is first necessary to find out what experts are currently doing. The technical method tends to be similar across practitioners: use good-quality headphones; listen repeatedly to sounds, words and phrases using audio playback software in a quiet listening environment; write down what is heard orthographically; carry out acoustic-phonetic analysis if necessary. But little is currently known about the procedural differences, such as how many transcribers contribute to the production of a transcript, how many drafts are made, and when and how contextual information is used. Once these have been established, it will be possible to conduct further research examining each part of the process in order to determine which combination of practices result in the most reliable transcripts. Article 1 will present the results of a survey on international expert transcription practices, with the aim of collating detailed data about the current ways in which transcripts are produced by experts, and suggests a number of areas for future research with the ultimate aim of developing standardised procedures.

As explained above, there are no requirements for accreditation or regulation, and equally no guidelines, for non-experts transcribing evidential recordings or police interviews; this leads to an ‘ad hoc’ approach to transcription, and the complexity of this task is often underestimated by those unfamiliar with linguistics and issues related to bias. Cognitive bias is a prevalent issue in the forensic sciences, and is defined by the UK FSR as “a pattern of deviation in judgement whereby inferences about other people and situations may be drawn in an illogical fashion” (Code of Practice v1, p. 342). Essentially, cognitive bias can influence

⁷ Section 30.3.1 of FSR Code of Practice (2023).

⁸ Section 30.5.1 of FSR Code of Practice (2023).

the way in which a person perceives information and it can also affect the methods employed and the conclusions formed. There are many different types of cognitive bias, such as those defined in Table 3 which are particularly relevant to transcription:

Cognitive bias	Definition according to FSR	Example (relevant to transcription)
Expectation bias	The expectation of what an individual will find affects what is actually found	A listener hears what they expect to hear, even if the auditory and acoustic evidence does not align with that interpretation
Confirmation bias	People test hypotheses by looking for confirming evidence rather than for potentially conflicting evidence	A listener believes the speaker in an audio recording is confessing to a crime and will search for speech content that aligns with a confession
Contextual bias	Information aside from that being considered influences (either consciously or subconsciously) the outcome of the consideration	Knowing that the audio recording is part of a drug trafficking case causes the listener to hear drug-related terminology

Table 3: Examples of different types of cognitive bias, as defined by the UK Forensic Science Regulator in FSR-G-217⁹ 'Cognitive Bias Effects: Relevant to Forensic Science Examinations'.

Little is known about non-expert transcription of poor quality evidential recordings, though there is a large amount of variability in the quality of such transcripts. It is unlikely that appropriate consideration is given to factors that can affect transcription performance, such as the level of background noise and the regional accents involved in the transcription process (i.e. the accents of both the speaker and the transcriber, as well as the transcriber's knowledge of the speaker's accent). Both of these factors can have an impact on transcription, and therefore the content of transcripts, yet are most probably overlooked in the transcription practices of non-experts given a general lack of knowledge about the problems with transcription. Article 2 will explore the content of transcripts produced by lay people and the way in which the level of background noise and accent background of the transcriber can affect the types of errors made.

⁹ See section 1.2.1 of FSR-G-217: [Cognitive Bias Effects](#)

For police interview transcribers, a lack of guidelines and standards leads to individual ROTI clerks developing their own practices regarding the representation of certain features. This individualistic approach combined with the practice of summarising large portions of the interview, thereby incorporating the transcriber's personal interpretation of the speech content into the ROTI, leads to transcripts that vary massively across transcribers and across police forces and do not necessarily accurately reflect the speech content. A transcript provided to a jury should be impartial, representing the speech content in full so that the jury can interpret the speech in their own way, rather than from someone else's interpretation (i.e. a summary). In many cases, ROTIs do not make it to trial and their potentially subjective nature is not an issue. However, those that do end up being presented to a jury should be impartial, which may be more successfully achieved through verbatim transcription, given that it would avoid the inherent subjectivity that comes with summarising materials (the issues with ROTIs, in particular the problems with summarising, were presented in section 1.2).

Transcription is a time-consuming process and it is understandable that verbatim transcription of all police interviews¹⁰ is not viable. However, verbatim transcription of those that will be presented to a jury could potentially be achieved in a relatively timely manner through the use of automatic transcription software within the production process. Article 3 will investigate the viability of incorporating an automatic recognition system (ASR) into the transcription of police interview recordings, whereby an ASR transcript acts as a 'first draft' of a verbatim transcript which is then post-edited by a human transcriber.

It is necessary to note that verbatim transcription is not necessarily desirable in all cases; as previously discussed, not all ROTIs go on to play an investigative or evidential role and so verbatim transcription for all police interview records would be an ineffective use of police resources. Similarly, long evidential recordings (e.g. multiple days of audio material from a covert recording device) will likely contain only small sections of speech that are of interest within an investigation or for evidential purposes. Richardson, Haworth and Deamer (2022) highlight that the most meaningful question to ask is whether a transcript is 'fit for purpose'. They propose that transcripts should be measured against their specific intended purpose, with the following questions in mind:

¹⁰ The results of a FOI request sent to all of the police forces in England and Wales (Tompkinson et al., 2022) show that the amount of transcription work varies per force. The largest police force, the Metropolitan Police Service based in London, reported that they carry out an average of 144 transcription tasks (equalling 6530 minutes) in an average week. The Metropolitan Police Service's response to the FOI request is publicly available on the following webpage: [Record of Taped Interview \(ROTI\) and Record of Video Interview \(ROVI\) transcripts | Metropolitan Police](#)

- How close is the transcript to the original (i.e. audio recording)?
- Who has agency over what goes into the record?
- Who has ownership of the record?
- How useable is it?
- How much resource (e.g. staff, time) does it require?

Richardson et al. (2022) suggest scoring transcripts for the five concepts above (accuracy, agency, ownership, usability, and resource efficiency). They highlight that the original audio recording of a police-suspect interview scores very highly for accuracy but achieves a low score for resource efficiency and an even lower score for usability. While ROTIs achieve a comparatively lower score for accuracy, usability is maximised and resource efficiency also scores highly.

The issue with verbatim transcription, and one of the primary reasons for instead summarising the speech content of an audio recording, is that increasing accuracy tends to decrease usability and resource efficiency. In the case of ROTIs, for example, transcripts are often used by detectives to quickly understand the content of the police-suspect interview and to identify lines of inquiry to pursue, and so a summary is preferable. However, if the speech content itself (i.e. the specific words that the interviewee uttered) is of interest, as is the case with poor quality evidential recordings, then accuracy becomes very important. The view taken within this thesis is that verbatim transcription is not necessary or even desirable in all contexts within the criminal justice system; however, verbatim transcription is preferable, and indeed necessary in the interests of justice, in cases where the specific words spoken in an audio recording go on to play an evidential role¹¹.

1.4 Lack of research

The primary motivation for conducting research on the topic of transcription within the criminal justice system is the huge gap in the literature, particularly concerning the transcription of evidential audio recordings. There is currently very little academic research addressing the production of transcripts of evidential audio recordings, even within the

¹¹ It may be the case that non-verbatim (i.e. summarised) transcription is appropriate for some sections of poor quality evidential recordings; for example, an audio recording that is over 20 hours in length is unlikely to require verbatim transcription for the entire duration; however, the view taken in this thesis is that any parts which are of evidential interest should be transcribed in a verbatim manner, as even one word inaccurately transcribed or substituted could have a major impact on the case.

forensic speech science community which includes many forensic practitioners who carry out this task as part of their work duties. The Research Hub for Language in Forensic Evidence, headed by Helen Fraser at the University of Melbourne, highlights the need for “accountable evidence-based methods” (Fraser, 2022, p. 2) for the transcription of poor quality evidential audio recordings, though what these methods entail has not yet been established.

Very little is known about the methods employed by non-experts to produce transcripts of evidential audio recordings, though it seems clear to forensic experts (who have seen many of such transcripts) that there are no official procedures or guidelines to follow. It is likely that transcription is viewed as a simple task of ‘just writing down what the person said’ by those unfamiliar with the linguistic research surrounding transcription. But transcription is an extremely complex process (which is explored in greater detail in section 2) that is made even more challenging in cases of poor quality audio. Given that there are no standards for transcripts produced by non-experts, it is likely that no consideration is given to the person who actually transcribes the audio, and whether they are well suited to the task of transcription in general or transcribing a particular recording that features, for example, accented speech.

While the transcription of police interviews has been explored in more depth, particularly by researchers on the For The Record project headed by Kate Haworth at Aston University, an area that has not yet been investigated is the automation of police interview transcription. It is not currently feasible to incorporate automatic speech recognition (ASR) into the transcription of poor quality evidential recordings; multiple studies have shown that automatic systems are not yet capable of producing reliable transcripts of poor quality audio (Loakes, 2022; Harrington et al., 2022¹²; Loakes, 2024), often producing nonsensical transcriptions or omitting large portions of speech. However, the audio quality in police interview recordings is often much better given multiple factors such as the question-and-answer format of the interview and the fact that speakers know they are being recorded. As of early 2022, when a Freedom of Information request was sent out by researchers at Aston University (Tompkinson et al., 2022), no police forces¹³ in England and Wales reported the use of automatic transcription systems, though three forces indicated

¹² This study was carried out independently of the doctoral research presented in this thesis, with colleagues at Aston University and Nottingham Trent University and funded by the Aston Institute for Forensic Linguistics seed-corn and network funds programme.

¹³ It should be noted that 36 police forces out of a total of 43 in England and Wales responded to the FOI. Of the remaining seven, there is a possibility that one or more of the forces have incorporated automatic methods into their transcription procedures, although this seems unlikely.

plans to use them in the future. Given the constantly improving performance of ASR technology and ever-increasing public awareness of the power of Artificial Intelligence¹⁴, it is likely that more forces will turn towards automatic transcription in the next few years. However, without further research into the feasibility of incorporating automatic methods into the transcription process, which takes into account the problems associated with ASR and transcription in general, there is a significant risk that police forces will employ automatic transcription systems in a linguistically uninformed way which lacks transparency and potentially undermines the quality of evidence used in the criminal justice system.

1.4.1 International context

It should be noted that the research within this thesis primarily concerns the transcription of legal and forensic audio recordings in England and Wales. This is for a number of reasons, including doctoral supervision by a UK forensic practitioner, the developing regulatory context in the UK and the (albeit relatively small) amount of existing work concerning police interview transcription by researchers at Aston University. The scope of this thesis is necessarily limited to the UK context in order to maintain a narrow research focus; however, there is also a distinct lack of research on the topic of forensic transcription on an international level, particularly with regard to non-expert transcription about which there is very little documentation. Fraser discusses the Australian legal context, in which transcripts of poor quality evidential recordings tend to be transcribed by police investigators who receive ‘ad hoc’ expert status through repeated listening of an audio recording (French & Fraser, 2018). Transcription of poor quality audio materials by police officers seems to be a common theme, with documentation of such practices taking place in Italy (Cenceschi & Meluzzi, 2023) and multiple references to “police transcripts” in the survey responses presented in Article 1. However, very little is known about *how* the police officers produce such transcripts and, for example, how much guidance and/or training they receive.

There is also relatively little documentation available (or easily discoverable/accessible) about the production of police interview transcripts in other countries; Komter (2022) details the contemporaneous transcription practices in the Netherlands, whereby the questioning officer (or a second reporting officer) produces a record of the police-suspect interview during the event; Byrman and Byrman (2018) report a similar process in Sweden, whereby interviews are summarised during the event and utterances are only transcribed in a verbatim manner when it is considered important to do so. Whether there are other countries

¹⁴ Artificial Intelligence (AI) has received a lot of media attention in recent years, particularly the chatbot Chat GPT and image generator DALL·E by Open AI.

who carry out a process similar to the production of ROTIs in England and Wales (i.e. transcription of police-suspect interview audio recordings after the fact) is unknown.

Though other countries may have different operating procedures or regulatory contexts, it is hoped that the research and themes presented within this thesis can also help to inform practice, or inspire further research and discussion, outside of the UK.

1.5 Summary

This introduction has provided the necessary context for understanding the issues surrounding current methods of transcript production within the criminal justice system in England and Wales. The two types of transcript that are of interest in this thesis are those representing the speech within poor quality evidential recordings, and those representing the content of police-suspect interviews. Because of the way in which these transcripts are used, i.e. in an evidential capacity contributing to the prosecution or defence of criminal suspects, it is extremely important that they are both accurate and impartial, i.e. free from bias. The methods used to produce transcripts contribute to their accuracy and impartiality, and therefore deserve much more academic attention and empirical research in order to determine the best practices to achieve these qualities; for forensic practitioners in England and Wales, such qualities are also a requirement of the UK Forensic Science Regulator. This thesis provides an exploration into the practices currently employed as well as potential novel methods.

1.6 Structure of thesis

This thesis is composed of five main chapters: a research background, three articles, and a discussion. There is also an 'additional resource' included between Article 1 and Article 2.

- The research background provides a comprehensive overview of literature related to transcription in the criminal justice system, addressing issues with transcription and the way in which transcripts are used in court, as well as factors which can affect transcription performance.
- Article 1 presents the results of a survey of forensic transcription practices employed by practitioners across Europe and North America, revealing what experts' methods currently look like and highlighting areas of disagreement which require further research.

- An additional resource is provided between Article 1 and Article 2, which details the content of a focus interview with two forensic practitioners concerning the quality of non-expert transcripts of poor quality evidential recordings. The interview was conducted as a result of comments made by practitioners within the expert survey, and provides context for the following article on non-expert transcription.
- Article 2 focuses on non-expert transcription of poor quality audio recordings, with a particular interest in the effects of increased background noise and degree of familiarity with an accent.
- Data analysis is motivated by the forensic implications of transcription errors in both Article 2 and Article 3, the latter of which explores transcription by automatic speech recognition systems and how the content of transcripts is affected by the introduction of background noise and different speaker accents. The experimental findings in Article 3 are then interpreted in the context of police-suspect interview transcription, exploring whether automatic methods could be incorporated into the interview transcription process.
- Finally, the discussion returns to the main aims of the thesis, reviewing how these have been achieved and what the combination of findings means for the future of transcription within the criminal justice system.

1.7 References

- Cenceschi, S. & Meluzzi, C. (2023). Transcription and voice comparison of noisy interceptions: remarks from an audio forensics report. *STUDI AISV*, 10, 99-111.
- College of Policing. (2022). *Investigative interviewing*. College of Policing.
<https://www.college.police.uk/app/investigation/investigative-interviewing/investigative-interviewing#:~:text=The%20aim%20of%20investigative%20interviewing,without%20any%20omissions%20or%20distortion>.
- Crown Prosecution Service. (2020). *Charging (The Director's Guidance) - sixth edition, December 2020, incorporating the National File Standard*. Crown Prosecution Service.
<https://www.cps.gov.uk/legal-guidance/charging-directors-guidance-sixth-edition-december-2020-incorporating-national-file>
- Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, 7.
<https://doi.org/10.3389/fcomm.2022.898410>
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2). <https://doi.org/10.1558/ijsl.v18i2.261>

- French, P., & Fraser, H. (2018). Why “Ad Hoc Experts” should not Provide Transcripts of Indistinct Audio, and a Better Approach. *Criminal Law Journal*, 298–302.
- Harrington, L., Love, R., & Wright, D. (2022, July). *Analysing the performance of automated transcription tools for covert audio recordings*. Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, 22(4), 428–450.
- Haworth, K., Tompkinson, J., Richardson, E., Deamer, F., & Hamann, M. (2023). “For the Record”: applying linguistics to improve evidential consistency in police investigative interview records. *Frontiers in Communication*, 8.
<https://doi.org/10.3389/fcomm.2023.1178516>
- Health and Safety Executive. (n.d.). *Sound and videotape recordings*. Health and Safety Executive. Retrieved February 8, 2024, from
https://www.hse.gov.uk/enforce/enforcementguide/court/physical-sound.htm#P12_1788
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7.
<https://doi.org/10.3389/fcomm.2022.803452>
- Loakes, D. (2024). Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare? *Frontiers in Communication*, 9.
<https://doi.org/10.3389/fcomm.2024.1281407>
- Richardson, E., Haworth, K., & Deamer, F. (2022). For the Record: Questioning Transcription Processes in Legal Contexts. *Applied Linguistics*, 43(4), 677–697.
- 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*. Conference of the International Investigative Interviewing Research Group, Winchester, UK.
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test*. Conference of International Association for Forensic Phonetics and Acoustics, Santander, Spain.
- UK Forensic Science Regulator. (2020). FSR-G-201: Validation (Issue 2) [URL:
https://assets.publishing.service.gov.uk/media/5f6b1a3de90e077ca292204f/201_-_FSR-G-201_Validation_Guidance_Issue_2.pdf]
- UK Forensic Science Regulator. (2023). Code of Practice (Version 1 - March 2023) [URL:
https://assets.publishing.service.gov.uk/media/64da431cc8dee4000d7f1c1e/FINAL_2023.1.18_Code_of_Practice.pdf]

2. Research background

The complexity of transcription is often underestimated; contrary to popular opinion, it is not as simple as ‘writing down what someone says’. There are subconscious processes, as well as many conscious decisions, that take place during transcription which can have a substantial impact on what is transcribed and how that content is later perceived.

Section 2.1 explores numerous issues with the act of transcription and section 2.2 explores issues with transcripts used in the criminal justice system. Section 2.3 addresses the psychological phenomenon of priming, i.e. being influenced to hear certain things through exposure to information, and how priming affects multiple stages of the transcript’s journey through the criminal justice system. Section 2.4 explores factors which can affect a transcriber’s ability, and section 2.5 introduces the concept of automatic speech recognition; a brief overview of these issues is provided in this section as they are explored in greater depth in Article 2 and Article 3.

2.1 The act of transcription

Transcription is often considered to be a simple process of transforming spoken data into written data; however, these types of data constitute two completely different media which are not directly equivalent (Biber, 1988). Writing conventions are unable to express some of the paralinguistic (e.g. intonation, emphasis) and extralinguistic (e.g. head-nodding, raised eyebrows) signals that are relied on by speakers to get their meaning across (Walker, 1990). As such, transcripts can never be “entirely accurate representations of spoken discourse” (Jenks, 2013, p. 259) given their inability to “totally capture the complexity of the interaction” (MacLean et al., 2004, p. 113).

However, in the criminal justice system, transcripts are often used interchangeably with the audio, particularly in the case of police interview transcripts (ROTIs), and are treated as a direct ‘copy’ of the speech content. This is particularly worrying given recent research by Deamer et al. (2022) that demonstrated significant differences in the way in which an interviewee is perceived across different modalities (audio recording versus transcript). Participants in their study were presented with either a 3-minute audio clip of a publicly-available police interview or the corresponding transcript, and were asked a series of questions regarding the emotional state of the interviewee as well as their credibility, plausibility, sincerity and innocence. Readers of the transcript were more likely to attribute

emotional properties to the interviewee, finding them more anxious and fearful and less relaxed than those who listened to the audio recording. The interviewee was also perceived as significantly more agitated, aggressive, defensive and nervous by those who had read the transcript compared with those who had listened to the audio recording. The direction of these results will likely vary across interviews due to differences in factors such as conversational topic, interview content and emotional differences as well as different speakers; but the crucial finding is that the transcript and audio recording have the potential to generate significantly different perceptions of speakers, further highlighting that a transcript and the speech content of an audio recording are not equivalent.

Furthermore, in producing a transcript, a transcriber is necessarily required to make a number of different representational decisions, e.g. what counts as language and what is meaningful. Such decisions are shaped by the lived experiences of the transcriber (Jenks, 2013) and their cultural knowledge of the language's discourse practices (Green et al., 1997). Representational decisions that a transcriber must make may concern, among other things, ungrammatical constructions, dialectal forms and sections of disfluent speech (e.g. false starts, repetitions, filled pauses); how these features are represented can affect the way in which a speaker is portrayed, which demonstrates the power that transcribers hold over the way in which speakers are ultimately perceived by the reader of the transcript. Many transcribers are not necessarily conscious of every interpretive choice that they make or of its representational consequences (Bucholtz, 2000).

Tompkinson et al. (2023) demonstrated that a transcript with pauses included led to significantly different perceptions of a speaker than a transcript with pauses omitted. Those who read the transcript without pauses were more likely to perceive the speaker as more aggressive, assertive and contemptuous than those who read a transcript with pauses included. This shows that a simple choice, like including some representation of a pause, can have a significant impact on the way in which a speaker is perceived (whether that is through the use of an ellipsis '...' or standard Jeffersonian conventions representing the length of the pause¹⁵). In legal contexts, where the judge or jury-members form opinions of the speaker from the evidence with which they are presented, such decisions are of great importance. Those producing transcripts for use in the criminal justice system should therefore be extremely conscious of the decisions that they make (e.g. how to represent certain features, what to include, how to format the transcript) and the ways in which such decisions could affect the reader's perception of the speaker. Many forensic experts carrying

¹⁵ The Jeffersonian conventions (Jefferson, 2004) for pauses are "(.)" for pauses between 0.08 and 0.2 seconds in length, and the absolute time in seconds in brackets for longer pauses, e.g. "(3.2)".

out transcription take a careful approach to the way in which they represent speech, but it is unlikely that those untrained in linguistics, such as police interview transcribers, will be aware of the weight given to their transcription decisions.

Another challenge that transcribers face is the dichotomous requirement for both readability and accuracy (Gibbons, 2003). Readability concerns the ease with which a reader can understand the transcript; this often entails adhering to standard writing conventions (Jenks, 2013) and formatting the transcript in such a way that a reader can intuitively follow the discourse. Readability is a fundamental principle of many transcription systems (Kowal & O'Connell, 2004), and is extremely important in the context of transcripts within the criminal justice system. The principal aim of an evidential transcript is to 'aid' the user (i.e. the jury) in understanding the speech content, and so making a transcript more challenging to follow, e.g. through the use of complex conventions and inclusion of every speech error, places greater demands on readers (Gibbons, 2003).

However, in achieving a more readable transcript, information is often lost and the content of the transcript ends up being a less accurate representation of the speech. Such lost information may comprise features such as disfluencies (false starts, hesitation words, filled pauses, repetitions), non-standard forms (e.g. "cause" in place of "because") and dialectal forms (e.g. "yous" in place of "you"). The accuracy of transcripts for use in the criminal justice system is extremely important; if speech evidence is playing a key role in a trial then it is crucial that such speech is represented accurately so as to not misconstrue what was said. However, increasing 'accuracy' could come at a cost. Including disfluencies such as filled pauses ('er' and 'erm') in a transcript has been shown to generate perceptions of uncertainty on the part of the speaker (Collins et al., 2019); likewise, speakers that produced a filled pause at the beginning of an answer were judged as less knowledgeable than those who produced a silent pause of the same length in that position (Brennan & Williams, 1995). Disfluencies are often subconsciously repaired by listeners (e.g. Levelt, 1983) to such an extent that they do not realise the speech was not completely fluent; the inclusion of this feature in a transcript may unnecessarily draw attention to the disfluency, thereby creating a different perception of the speaker than that generated by a transcript without disfluencies or the audio recording itself.

2.2 Issues with transcripts in court

2.2.1 Reliance on transcripts

A major issue with transcripts concerns the way in which they are often relied upon in place of the audio, i.e. the 'real evidence'. Within the criminal justice system, transcripts are considered an 'aid' to the court in understanding the speech content of a recording, yet they often become "interchangeable" (Haworth, 2018, p. 434) with the audio evidence in the eyes of the court. Haworth (2018) highlights this issue with regard to police interview transcripts, which are often presented without the original audio recording and are therefore relied upon as an official record of the police interview. Haworth (2018) discusses a chain of 'contamination' that takes place, whereby (a) the original version of the discourse is transformed into an audio recording, thereby losing contextual information and cues; (b) the audio recording is then transformed into a written format, which causes para- and extralinguistic information to be lost and is essentially the transcriber's interpretation of the speech content; and (c) the transcript may then be read aloud to the court, often by a barrister, thereby injecting new para- and extralinguistic cues. The jury can therefore end up being presented with, and consequently using as part of their overall judgement, a lawyer's (likely biased) interpretation of the transcriber's interpretation of the speech content of a discourse that has been removed from its original context.

Even in cases where the audio recording is played to the court and the transcript is provided alongside it as an 'listening tool', e.g. in the case of poor quality evidential recordings, the court often heavily relies upon the transcript as a result of the poor intelligibility of the speech. The transcript therefore plays a critical role in the way in which listeners interpret the speech, given that many of the words are often unintelligible without the help of a transcript. Research by Fraser and colleagues (e.g. Fraser & Kinoshita, 2021) has demonstrated that the provision of a transcript alongside poor quality audio can be majorly problematic; listeners can confidently accept an inaccurate interpretation offered within a transcript, even if they had previously understood no words at all, and exposure to an inaccurate transcript can distract from more plausible interpretations (see section 2.3.2 for a more detailed account of these findings).

This problem is further compounded in English and Welsh courts by inadequate audio playback procedures (as discussed in section 1.1). According to forensic experts interviewed as part of this thesis (see Section 5: "Additional resource"), jurors are rarely given the opportunity to carefully listen to the audio with headphones, and are instead most commonly

presented with the audio through loudspeakers in the courtroom. This is problematic for a number of reasons: (a) the courtroom is filled with materials such as wood and glass which cause reverberation and therefore make the audio more challenging to understand, (b) the speakers are not necessarily placed near to the jury so members of the jury may struggle to hear, and (c) the audio is played a limited amount of times, with someone, likely untrained in audio and speech comprehension issues, deciding how the audio is presented (e.g. in short sections or in a large block) and how many times it is played. In such cases, it is unreasonable to expect members of the jury to independently understand the speech content within a poor quality audio recording.

2.2.2 Content of transcripts

Another issue with transcripts concerns what is contained within them, given that the court often has to rely on the transcript as an accurate record of the speech content. What a transcript looks like and contains primarily depends on the purpose of the transcript. In academic contexts, such as within the field of conversation analysis, transcripts contain verbatim speech marked with complex conventions (most commonly following the Jeffersonian transcription system) for the representation of everything deemed potentially linguistically relevant, such as pauses, emphasis, prolonged sounds, intonational patterns, inbreaths and outbreaths, and speaking rate. A court transcript will look very different, given that the aim is to produce an official record of trial proceedings and therefore the priority lies within the readability of the speech content; complex conventions and phenomena such as false starts and hesitations are not included and grammar is frequently 'cleaned up' (Walker, 1990) so that the end user (often members of the public, i.e. lay people) can easily and intuitively follow the content of the transcript.

Problems can arise when the content of a transcript forms part of the evidence in a case, and therefore the exact words uttered are of great importance. The purpose of a police interview transcript, often referred to as a Record of Taped Interview (ROTI), is to produce a record of the interview as a formality, and this may then be used as part of an investigation or eventually as evidence in a court case. ROTIs generally comprise a summary of the interview with only certain parts of the interview transcribed in full (Haworth, 2018). In the absence of specific instructions or guidance, the ROTI clerk (the transcriber) must decide which parts of the speech content are most evidentially relevant and should therefore be transcribed verbatim, despite having no legal training or expertise; the other parts of the interview are paraphrased or summarised in the transcriber's own words, which can be particularly problematic for a number of reasons. Firstly, in summarising the speech it is

necessary to use reporting verbs and the choice of word can generate differing perceptions of the speech (Haworth et al., 2023); for example, “he **admitted** he was at the crime scene” implies some admission of guilt, whereas “he **said** he was at the crime scene” is more neutral. Secondly, the credibility of defendants can be called into question if they state something in court that was not previously mentioned at the interview stage (Haworth, 2018); however, it may be that the ROTI clerk did not consider verbatim transcription to be necessary for a particular section of the interview and so there is not an accurate record of the speech content. These issues stem from the fact that the ROTI is not designed to be an accurate record of the speech content of the interview recording, yet can be used as such in court.

The purpose of a transcript of an evidential recording is to accurately represent the speech content, as some parts of the recording will be of great interest to the court. It is extremely important that the content of the transcript is reliable, given that a mistranscription of one word could change an utterance from completely innocent to falsely incriminating. This type of transcript will be somewhere in between an academic transcript and a court transcript, such that all speech content, including disfluencies, is retained within the transcript but conventions are kept simple so that the end user (often the jury) can follow along. However, it is often not possible to transcribe the recordings in full due to poor intelligibility as a result of audio-related factors (e.g. technical issues, background noise, distance from microphone, etc.), and so these transcripts often contain sections where the speech is marked as ‘unintelligible’. Transcripts of poor quality evidential recordings are sometimes produced by experts in forensic speech science, who understand the issues related to transcribing poor quality audio and, crucially, the limitations of such a task. In most cases, though, such transcripts will be produced by non-experts, such as police officers, and forensic speech scientists have identified many issues with the content of these transcripts (see Article 1 and Section 5: “Additional resource”).

The interview presented in section 5 of the thesis (“Additional resource”), conducted with two forensic practitioners who are often presented with transcripts produced by the police or other members of instructing parties, revealed that there is huge variability in the quality of non-expert transcripts; sometimes the transcripts are of relatively good quality in terms of the accuracy of the content and the representation of inaudible sections of speech. However, there are many cases where non-expert transcribers will attempt to ‘fill in’ sections of unintelligibility with their best guesses, even when such sections may not actually contain speech. Conversely, there are many cases where non-expert transcribers will be extremely conservative in their transcription, and mark sections as unintelligible despite an expert

being able to make out some of the speech content. French and Fraser (2018) also highlight issues with non-expert transcripts, giving some well-known examples of police officers' wrongful transcriptions of utterances in poor quality evidential audio, e.g. "he died after wank off" in place of "he died after one cough". But the main issue highlighted by French and Fraser is the knowledge of contextual information that police have when producing transcripts, which can prime the way in which they influence the speech.

2.3 Priming in poor quality audio

In psychology, priming is typically defined as "facilitative effects of an encounter with a stimulus on subsequent processing of... a related stimulus" (Tulving et al., 1982 p. 336). In the context of this thesis, priming can be understood as a process whereby the perception of speech (in an audio recording) can be subconsciously influenced by exposure to some type of 'information'. Both transcribers and readers of the transcript can be influenced by such knowledge. In the case of transcribers, the influential information could be contextual information about, for example, the circumstances of a case, the speakers in the recording, or other information about the content of the recording. In the case of the transcript reader, it could be an interpretation put forth in a transcript. The following subsections address the way in which both of these types of information can prime listeners and what the consequences of that priming can be; section 2.3.1 addresses the priming power of contextual information, focusing on the way in which transcribers can be affected by their knowledge and expectations, and section 2.3.2 addresses the priming power of the words within transcripts and the consequences on readers' perceptions of the speech.

In forensic contexts, priming may not be 'facilitative' but rather quite harmful, particularly if the information is unreliable or incorrect. However, priming is not always a problem; having relevant background information can aid listeners in interpreting speech in poor quality audio recordings (Fraser, 2021), and the same is true for everyday conversations. If we consider priming as, essentially, the use of top-down information (such as contextual information) to guide our understanding of speech, then this is present in most speech-based interactions. Without knowing that the speaker is referring to the weather, an utterance like "bit nasty, isn't it?" will likely make little sense to a listener; but if the listener has been 'primed' to think about the weather, e.g. the speaker has pointed at the rainclouds out of a window, then there will probably be less confusion.

A phonemic restoration study by Warren (1970) clearly exemplifies the way in which priming subconsciously influences speech perception, even in the case of good quality audio

recordings. Participants were presented with short sentences in which the final word of the sentence was the 'prime' and one phoneme earlier on in the sentence was replaced by white noise, e.g. "the *noise*eel was on the [prime]". By changing the prime, the researcher was able to manipulate the phoneme that participants would perceive in place of the white noise. For example, when the prime was the word 'axle', participants perceived the missing phoneme as /w/ such that they heard the sentence 'the wheel was on the axle'; when the prime was the word 'orange', participants heard the sentence 'the peel was on the orange', thereby perceiving /p/ in place of the white noise. Participants were generally surprised that the recordings were not complete, highlighting the way in which listeners can understand sounds even when they are not present, while not even realising that such sounds were not present.

The study by Warren (1970) demonstrates how listeners can be primed to hear speech sounds that are not present by exposure to a single word. Bruce (1958) demonstrates how listeners can be primed to hear words that are not present by exposure to a conversational topic. Participants (unknowingly) heard the same five sentences presented in noise on five occasions, and each time the sentences were preceded by a different word revealing the topic of the sentence ("sport", "food", "weather", "travel" or "health"). Participants were asked to repeat aloud as much of the sentence as possible and to guess if uncertain. The main aim of the experiment was to show that appropriate primes would lead to higher intelligibility of the speech content, and this was found by the highest levels of accuracy being achieved for each sentence when the keyword was matched. However, it was also found that 'inappropriate' keywords (i.e. those that did not match the content of the sentence) had a substantial influence on the way in which the sentences were perceived. Participants misinterpreted many words in line with the topic they had been primed with; for example, the sentence "I tell you that our team will win the cup next year", for which the appropriate prime was "sport", was interpreted in five different ways by one participant:

- (a) Weather: I tell you that I see the **wind** in the south next year.
- (b) Travel: ----- next year.
- (c) Sport: I tell you that our **team** will **win the cup** next year.
- (d) Food: I tell you that our **tea** will be something to do with **beer**.
- (e) Health: I tell you that our team has been **free from injury** all this year.

In all cases except (b), where the majority of the sentence was not interpreted at all, each interpretation has clearly been influenced by the presented topic. It should be noted that such an effect was only present in degraded listening conditions. The signal-to-noise ratio

(SNR) for the sentences in this experiment ranged between -11 dB and -17 dB; anything below a SNR of 0 dB means that the noise is louder than the speech content, and is therefore extremely challenging.

It is well established that listeners are more susceptible to priming in poorer listening conditions (Lange et al., 2011), and the reason for this concerns the different types of processes involved in speech perception. ‘Bottom-up’ processing involves auditory information, i.e. the sounds present, while ‘top-down’ processing involves the listener’s knowledge and expectations (Fraser, 2003). When the audio is of good quality, listeners are able to use both bottom-up and top-down information to decipher what is being said, but in poorer listening conditions the amount of bottom-up information is reduced due to issues with the signal, e.g. background noise, channel degradation, and overlapping speech. Listeners must therefore rely more heavily on top-down information that they can gather from their knowledge of the situation and their expectations and assumptions about what is being said and how it is being said (Fraser, 2020).

2.3.1 Priming power of contextual information

A forensically-motivated study by Lange et al. (2011) reveals how listeners can be influenced to hear words that are not present in poorer quality audio by exposing them to contextual information regarding the circumstances of a recording. Participants were told that the sentences to be transcribed were taken from either (a) criminal suspects’ interviews or (b) job candidates’ interviews, or (c) no information was given. All sentences were benign in nature, such as “I got scared when I saw what it’d done to him”, and researchers were interested in incriminating mistranscriptions, such as “I got scared when I saw what *I’d* done to him”. Sentences were presented in three different levels of degradation: no degradation, a low pass filter at 1000 Hz (less degraded) and a low pass filter at 670 Hz (most degraded). It was found that listeners in the suspects’ interview condition were significantly more likely to make incriminating misinterpretations of the speech, and this was driven by a higher proportion of misinterpretations occurring in the less degraded condition (low pass filter at 1000 Hz).

When there was no audio degradation, participants achieved word recognition accuracy rates of around 80% in each of the contextual bias conditions, and despite a slight increase in the number of incriminating misinterpretations in the suspects’ interview condition, the majority of errors were unrelated to a crime. In the most degraded audio condition, with the low pass filter at 670 Hz, word recognition accuracy rates approached 0% and almost all of

the errors were unrelated to crime, even in the suspects' interview condition. In the less degraded audio condition, with the low pass filter at 1000 Hz, a pattern emerged whereby slightly higher rates of incriminating misinterpretations were observed in all contextual bias conditions, and a huge increase in incriminating misinterpretations was observed in the suspects' interview condition. For the 'no bias' and 'job candidate' conditions, incriminating misinterpretations accounted for around 10% of the transcription results, but this soared to almost 30% in the suspects' interview condition.

These results suggest that contextual information, as subtle as the fact that the speech came from a criminal suspect's interview, can cause listeners to misinterpret completely innocent speech as incriminating in moderately poor listening conditions. This effect is not observed in extremely poor quality audio, when there is little bottom-up information to work with, which suggests that general contextual biases tend to influence listeners' interpretation when there is some bottom-up information available (Lange et al., 2011). Participants provided confidence ratings which demonstrated that in the less degraded audio condition, where a huge increase in the proportion of incriminating misinterpretations was observed, participants were almost as confident with their incriminating misinterpretations as they were with their accurate transcriptions.

Giroux (2022) carried out a replication of the study by Lange et al. (2011) with very similar results. Participants were presented with sentences in degraded audio at similar levels to the above study (via the use of low pass filters at 600 Hz-1600 Hz) and were either given no context about the recordings or told that the recordings came from wiretapped conversations with criminal suspects. Those in the 'criminal suspect' condition made significantly more incriminating misinterpretations than participants who had not received any contextual information, further demonstrating that having a pre-existing belief about the context of recordings can lead listeners to be more likely to interpret the recordings "in a manner that is consistent with this belief" (Giroux, 2022, p. 39).

Access to contextual information can therefore have a substantial impact on the way in which speech in an audio recording is understood. Simply knowing that the police interview concerns drug trafficking could lead transcribers to hear terminology related to drugs in poorer quality sections of the recording; or a police detective's expectation that the suspect uttered a confession to the crime under his breath in a telephone call recording could lead the detective to believe that he heard such a thing, despite that section of the recording not containing speech sounds. The latter example is a real situation that took place in New Zealand in the early 2000s, in which a detective believed that he could hear a confession of

murder within a phone call made to the emergency services. The case has been subject to much discussion in the field of forensic speech science and will be summarised below (see Innes (2011) for a full account of the case).

In June of 1994, a young man named David Bain returned home from his paper round to find that his parents and three siblings had all been shot, and in a clear state of distress, he made a phone call to the emergency services asking for help. In 1995 he was convicted on five counts of murder and sentenced to life imprisonment, and after multiple appeals, the case was finally sent to retrial in 2009. During this time, the police digitised the recording of the phone call to the emergency services, and a detective believed he could hear Bain say “I shot the prick” during the call, a confession of guilt that had not been picked up on during his first trial. The operator who had taken the emergency services call was approached and amazed to now be able to hear the suggested utterance, and then experts in forensic speech science were contacted by both the prosecution and defence to assess the recording. Experts on both sides agreed that the section of interest did not contain the utterance “I shot the prick”, though there was no agreement on what was actually said or if the section even contained speech sounds. As a result, that portion of the phone call was removed from the recording that was played to the jury in the retrial, and so the police detective’s interpretation did not form part of the prosecution’s evidence.

The above scenario is an example of a ‘disputed utterance’, in which different interpretations of a word or phrase within a very challenging section of audio are put forth by the prosecution and defence, one of which is often incriminating. Disputed utterances are an extreme example which demonstrates the importance of reliable transcripts; the utterance in question may be considered a crucial piece of evidence against the suspect according to one interpretation, but an irrelevant or even exonerating piece of evidence according to another. The jury’s acceptance of an incriminating yet inaccurate interpretation of an utterance could lead to a wrongful conviction and therefore a miscarriage of justice, which highlights the magnitude of potential issues arising from an unreliable transcript. When a disputed utterance features in a case, experts in forensic speech science are often called in to carry out detailed phonetic and acoustic analysis on the speech to determine whether the interpretation(s) put forth could be plausible. Fraser (2020) details her experience with a number of disputed utterance cases, with one particular example leading to a conviction of accessory to murder largely on the basis of the disputed utterance. Disputed utterances occur when bottom-up information is substantially reduced and, in some cases, when top-down information is (too) heavily relied upon; the assumption of guilt on the part of the

police detective in the Bain case is an example of how powerful a listener's expectations can be on how they perceive speech.

All of this suggests that the knowledge of contextual information can be extremely powerful and also extremely misleading if that information is inaccurate or unreliable. This poses the question of whether transcribers should have access to contextual information. Fraser (2022) posits that relevant, reliable contextual information is essential for transcribing poor quality evidential recordings, but that its knowledge must be managed carefully. This mirrors the practice of many expert transcribers, who apply a practice sometimes referred to as 'linear sequential unmasking' in their transcription process, whereby information is slowly and carefully revealed to the transcriber; this is explored in greater depth in Article 1.

It is unknown how much contextual information is known by non-expert transcribers, although evidential recordings are often transcribed by detectives working on the case and ROTI clerks are likely given details about the police interview and the suspected offence. It is unlikely that any consideration is given to psychological phenomena such as priming or personal biases in these contexts, given a lack of widespread knowledge among lay people regarding the complexity of transcription.

2.3.2 Priming power of transcripts

As is evident from the literature reviewed above, priming can have a very significant impact on transcribers and the way in which they perceive speech and therefore transcribe speech. However, priming is also a concern at a later stage of the transcript's journey through the criminal justice system. The end user of a transcript in many cases is the jury, i.e. the triers of fact who must decide on a defendant's innocence, and the transcript itself can prime listeners to hear the words contained within a transcript alongside a poor quality audio recording, even if those words are not actually said in the audio.

Giroux (2022) demonstrates that exposure to transcripts is even more powerful than exposure to contextual information about the circumstances of the recording. Participants were asked to transcribe the speech in a series of degraded audio recordings and were given varying degrees of information; participants either (a) received no contextual information at all, (b) were told that the utterances were all taken from wiretapped conversations with criminal suspects, or (c) were presented with transcripts containing incriminating misinterpretations (i.e. inaccurate transcriptions of innocuous utterances) prior to transcribing each utterance. Those in the 'criminal suspect' condition made significantly

more incriminating misinterpretations in their transcriptions than those in the 'no context' condition, and those in the 'incriminating transcript' condition made significantly more incriminating misinterpretations than those in the 'criminal suspect' condition. The explicit suggestion of an incriminating misinterpretation (i.e. via the presentation of a transcript) can therefore be extremely powerful in its influence over listeners.

Fortunately, in the Bain case discussed above, the police detective's interpretation was not put to the jury, but Fraser and colleagues wanted to explore the consequences that could have occurred should the jury have been presented with a transcript containing the detective's interpretation. There are multiple safeguards put in place within the Australian legal system, which much of the current research on forensic transcription (mostly conducted by Fraser) concerns, and these aim to mitigate the risk of a jury being misled by inaccurate transcripts. The first of these safeguards is the 'aide memoire instruction', which calls for judges to instruct the jury that they must "listen carefully and reach their own conclusion about the content of the audio, using the transcript only as an aid" (Fraser, 2021, p. 142).

In one study using the Bain crisis call, Fraser et al. (2011) explored the effectiveness of the 'aide memoire instruction' by carrying out an experiment in which two groups of participants were presented with alternative interpretations of the section of interest. Group A was presented with the incriminating interpretation "I shot the prick" that the police detective heard, while Group B was presented with "he shot them all" which placed guilt on the father. Information about the case and interpretations of the section of interest were gradually revealed and participants were asked questions at each stage in order to track changes in their perception of the content. Only 4 of the 190 participants (1 in Group A, 3 in Group B) heard the phrase "I shot the prick" at the beginning of the study, when they listened to the audio 'cold' (without a transcript), but 30% of Group A were confident that they heard this phrase after it was suggested and this number remained relatively consistent until the point at which details of the full real-life story was presented towards the end. Worryingly, after the full story and even expert testimony from a phonetician that "I shot the prick" was implausible, around half of the 30% who had heard this phrase were still confident that "I shot the prick" had been uttered by the caller. This is only a small percentage (around 17% of participants in Group B) but it shows that it would be possible for members of the jury to be primed so greatly by the inaccurate transcript that other evidence does nothing to persuade them. The results from this experiment demonstrate that listeners can be greatly influenced by an inaccurate transcript, even after being presented with evidence against such an interpretation, and the 'aide memoire instruction' cannot successfully protect against the priming effects of a transcript.

A second safeguard is that, in cases of multiple interpretations, lawyers on both sides are expected to review the transcripts and put forth an agreed-upon version for the court, and if no such agreement can be reached then multiple versions may be presented to the jury (Fraser & Kinoshita, 2021). In another study using the Bain audio, Fraser and Kinoshita (2021) explored this safeguard which is concerning for two reasons: first of all, lawyers and judges are no better placed to be unaffected by the priming power of a transcript than members of the jury; and secondly, the order in which the jury is presented with the different versions of the transcript may have an impact on which interpretation they ultimately believe they can hear. Participants in the experiment were divided into two groups and listened to the audio in three stages: at the first stage they listened 'cold' (i.e. with no transcript) and in the second and third stages they were presented with both the incriminating interpretation "I shot the prick" and a more plausible interpretation "I can't breathe". One group was first presented with the "prick" interpretation (hence referred to as the PB group), and the other group was first presented with the "breathe" interpretation (hence referred to as the BP group). At the first stage participants were asked whether they could hear words in the section of interest and, if so, to transcribe what they could hear. In subsequent stages they were asked whether they agreed with the interpretation put forth and, if not, were asked to transcribe what they believed they could hear instead.

Results showed that as participants were exposed to different interpretations, their perception of the section of interest varied significantly. The percentage of participants that could not hear speech in the section of interest decreased substantially from the beginning to the end of the experiment, suggesting that sections of a recording previously considered to not contain speech can be 'made intelligible' to listeners through the use of a transcript. There was also a highly significant priming effect of the interpretations put forth to participants. Not a single participant heard "I shot the prick" at the beginning of the experiment when listening 'cold', but exposure to this interpretation led to 48% of the PB group and 56% of the BP group indicating that they could hear these words. What is most interesting, however, is how the groups changed their opinion after exposure to a second prime. At the first stage, when listening 'cold', 28% of the BP group transcribed the words "I can't breathe" and after this interpretation was suggested, 84% of the group indicated that they could hear these words. However, when the group was then exposed to the "prick" interpretation, this percentage drastically decreased to 12%, with 56% of the group now accepting the "prick" interpretation. 48% of the PB group indicated that they could hear "I shot the prick" following its suggestion but this percentage decreased to only 4% after "I can't breathe" was suggested, which 68% of the group now believed they could hear.

The authors highlighted the fact that fewer participants in the group that was exposed to “I can’t breathe” after “I shot the prick” were convinced by this interpretation than in the group who heard “I can’t breathe” first (68% versus 84%). The PB group, who heard “I can’t breathe” as the second prime, also accepted the “breathe” interpretation with lower confidence. These findings suggest that the perception of an objectively more plausible interpretation can be substantially impacted by previous exposure to a persuasive but implausible interpretation, in terms of both the possibility of listeners remaining confident in the implausible interpretation and the confidence with which listeners accepted the more plausible interpretation. Furthermore, participants’ ratings of their confidence did not align with their actions. For example, at the first stage many listeners in the PB group expressed confidence in hearing an interpretation other than “I shot the prick” (i.e. no words, “I can’t breathe”, or other words), but half of the group drastically changed their opinion following exposure to the “prick” interpretation, with almost all of them expressing confidence in their hearing. This again changed drastically after exposure to “I can’t breathe”, where all but one who had previously heard “I shot the prick” then changed their mind. The findings taken as a whole suggest the ineffectiveness of the safeguards, given the fact that unreliable transcripts can be extremely distracting from more plausible versions, as well as the issues related to the order in which the jury are presented with different versions of a transcript.

Exposure to a transcript can create expectations about the speech content of an audio recording, and this in turn can lead listeners to believe that the speech is more intelligible than it actually is. Lange et al. (2011) demonstrated this effect by asking participants to estimate how many words within a poor quality audio recording could be accurately transcribed by someone listening without the aid of a transcript. By comparing these estimations against transcription performance by a group of participants in a transcript-absent condition, the authors showed that participants who were exposed to a transcript overestimated the proportion of words correctly transcribed without a transcript by over 50%. In the most degraded audio condition (with a low pass filter at 670 Hz), word recognition accuracy was below 5%, yet those who had been presented with a transcript estimated accuracy rates of well over 50%. In the least degraded audio condition (with a low pass filter at 1000 Hz), word recognition accuracy approached 40% on average, but participants who had seen a transcript estimated accuracy rates of over 80%. Such findings suggest that listeners provided with a transcript may fail to notice the poor quality of an audio recording and struggle to imagine not being able to hear the interpretation they have been exposed to.

Transcripts provided alongside poor quality audio recordings can therefore lead to a false sense of confidence in the given interpretation as well as an underestimation of the strength of audio degradation. Given that listeners can be heavily primed by exposure to a transcript, it is unlikely that errors within transcripts would be identified by members of the court, including lawyers, judges, juries, etc. It is likely that such errors would only be picked up on in rare cases of disputed utterances, but once a misleading interpretation has been put forth, it could be extremely difficult for listeners to hear anything else. For this reason it is extremely important that the transcripts that are presented to the jury in an evidential capacity are reliable (Fraser, 2022).

2.4 Other factors affecting transcription

The previous sections of this research background have demonstrated that there are many things that can substantially impact transcription and therefore the content of transcripts. For example, the transcriber has to make many decisions about what to include (e.g. pauses) and how to represent it, as well as balancing the requirements of readability and accuracy. More specifically to recordings of forensic interest, the knowledge of contextual information (e.g. that the speech is from a police interview) can have a significant impact on how the speech is perceived and therefore transcribed. There are many other factors that can affect the creation of transcripts, a number of which are highlighted in Fraser (2022); such factors can apply to the audio, the speaker and the transcriber (Table 1).

Type	Examples
Audio factors	<ul style="list-style-type: none"> • The understanding of the person recording the audio of the purpose and context of the recording • The equipment used to record the audio, and knowledge of the person recording the audio of how to use it • Processing applied to the audio, either at the time of recording or later
Speaker factors	<ul style="list-style-type: none"> • The language and variety used • The register and style of the speech content • The formality of the speech • The pragmatic nature of speech, e.g. intonation, voice quality
Transcriber factors	<ul style="list-style-type: none"> • The level of training for the style of transcript required • Their personal aptitude for transcription • Their understanding of the transcript's purpose • Their knowledge of the language, variety and register used • Their knowledge and expectations of the content and context of the recording

Table 1: Factors that may affect the creation of transcripts according to Fraser (2022, pp. 5-6).

With regard to evidential recordings of forensic interest, it is very often the case that these have been captured in a way that is not conducive to good audio quality, e.g. by a covert recording device hidden in a stationary location or on an undercover officer, interactions in the background of a telephone call, audio extracted from a CCTV recording, etc. The equipment used may not have originally been intended to capture the speech evidence, or the person recording the speech may not have known that the recording would later be of forensic interest. Police-suspect interviews can also feature sections of poor quality, despite the requirement for these to be audio recorded; this may result from the rustling of papers, the whirring of laptop fans, or reverberation from the room (Richard Rhodes, personal communication). Poor audio quality is therefore a factor that is present in many audio recordings transcribed for use within the criminal justice system, and often results in lower intelligibility of the speech content. This makes the task of transcription much more challenging, given that there is less auditory material available which can be used to decipher what is being said, and often results in much worse performance (e.g. Lange et al.

2011; Clopper & Bradlow, 2008). This issue is explored in more depth in section 2.1 of Article 2.

Another factor which is prevalent in evidential recordings is different regional accents; transcribers may be tasked with transcribing speech in an accent other than their own, e.g. another regional variety of British English or non-native-accented English. This may cause issues in some circumstances, particularly when the transcriber is unfamiliar with the accent of the speaker(s). Research in the field of psycholinguistics has shown that listeners perform worse for unfamiliar accents than familiar accents in a range of speech processing tasks; a significant delay in reaction times for unfamiliar accents has been observed in lexical processing tasks (Adank & McQueen, 2007; Floccia et al., 2006, Sumner & Samuel, 2009) and sentence processing tasks (Adank et al., 2009). Furthermore, lower word recognition accuracy has been observed for unfamiliar accents in transcription tasks (Smith et al., 2014; Clopper & Bradlow, 2008) when the quality of the audio is less than optimal. These findings suggest that the combination of speech spoken in an unfamiliar regional accent and the poor audio quality often found in evidential recordings or police-suspect interview recordings will lead to worse transcription performance, i.e. transcripts of poorer quality.

Within these psycholinguistic studies, both the native accent of the listener and the standard variety of their country are judged as 'familiar', with no significant differences in performance observed across these two accents within the lexical processing studies cited above. With regard to transcription, Clopper and Bradlow (2008) found that General American was the most intelligible accent in noise for all listeners in their study, such that speakers of a Northern variety of American English actually performed better for General American than their own accent, and that there were no significant differences in performance as a result of listeners' accents. However, Smith et al. (2014) explored the transcription of two varieties of British English in noise and found disparities in performance for Standard Southern British English, whereby native speakers of that accent had an advantage over speakers of a non-standard regional variety, Glasgow English. This effect is worth exploring in the context of transcription within the criminal justice system, as transcribers will likely have to deal with speakers from all over the United Kingdom and it may be the case that some transcribers are better suited than others for certain tasks, given their experience with a particular variety.

The factors discussed above - audio quality and regional accent - should be taken into account when developing robust methods for the transcription of audio materials for use within the criminal justice system, given that they have been shown to affect transcription performance. However, it is currently only known that these factors *can*, in principle,

negatively impact performance rather than *how* they negatively affect performance. Many of the previous studies on transcription in noise (e.g. Clopper & Bradlow, 2008; Smith et al., 2014) evaluate the data in terms of how many words are corrected transcribed, with no consideration of what is happening in the parts of the transcripts that are not correct. From a forensic perspective, the incorrect words in a transcript are of huge importance, given that they could potentially be the difference between a perfectly innocuous utterance and an incriminating one. A novel approach must therefore be taken when analysing transcripts for use in forensic contexts (see Section 2.6), and this is explored in Article 2 and Article 3.

2.5 Automatic transcription

Until recent years, transcription has been a process mostly carried out by humans, particularly in circumstances where the content of the recording is considered important. However, the rapidly advancing technology of automatic speech recognition (ASR) systems means that many fields are now relying on automatic methods for the production of transcripts. For example, ASR systems have been incorporated into the production of transcripts of meetings of the Icelandic Parliament (Fong et al., 2018) and the Japanese Parliament (Mimura et al., 2021). ASR systems are able to produce transcripts much more quickly than human transcribers, and can process huge amounts of data in a relatively short amount of time. This can be hugely effective (and cost-efficient), particularly in cases of routine transcription, where records of the speech content of an event need to be formally logged, such as the parliamentary proceedings detailed above.

However, many of the issues discussed in the previous section, i.e. those that can have a substantial impact on human transcription, also affect transcription by automatic speech recognition systems. Poor quality recordings, particularly those which resemble evidential recordings, pose significant challenges for automatic systems (e.g. Littlefield & Hashemi-Sakhtsari, 2002; Loakes, 2022; 2024; Harrington et al., 2022). Non-standard regionally-accented speech and non-native-accented speech also proves to be more challenging for automatic systems than speech uttered in a 'standard' variety (Lima et al., 2019; Markl, 2022; DiChristofano et al., 2022). Furthermore, factors relating to gender (Tatman, 2017), ethnicity (Koenecke et al., 2020) and variety of English (Meyer et al., 2020) have also been shown to impact automatic transcription performance.

The use of artificial intelligence is becoming more established in police practices (Science & Technology in Policing, 2023), and although automatic systems are not currently implemented in the transcription of police interviews (Tompkinson et al., 2022), it is not

difficult to imagine more police forces becoming interested in this approach within the next few years. Given that ASR systems are susceptible to biases and the transcripts produced always require checking and/or editing by humans (a process which could be affected by priming), it is important that the incorporation of automatic technologies into the transcription of such forensically-important recordings is carried out in a scientifically-informed manner. Widespread implementation of this technology by those unaware of the issues related to automatic systems, as well as broader issues related to transcription such as priming, could lead to poor quality evidence, in terms of accuracy and impartiality, making its way through the criminal justice system. This issue is explored in greater depth in Article 3, along with an overview of automatic transcription and its potential incorporation into the transcription of audio recordings of forensic interest.

2.6 Transcription accuracy metrics

There are many ways in which transcription performance can be measured and these often vary according to the specific application of the analysis. Previous work on the effects of ‘familiarity’ tends to measure performance by calculating the number or percentage of words correctly transcribed (e.g. Burda et al., 2006; Derwing & Munro, 1997; Jones et al., 2019; Smith et al., 2014), or more specifically *content words* (Clopper & Bradlow, 2008) or selected *keywords* (Walker, 2018) correctly transcribed. Linguistic research tends to focus on success rates, i.e. words correctly transcribed, rather than error rates. Conversely, the industry standard used for measuring transcription accuracy in automatic transcription systems is word error rate (WER). This measure is calculated by counting the number of errors (insertions, deletions and substitutions) within a transcript and dividing by the number of words uttered. This metric is usually presented as a percentage, where 0% demonstrates a perfect match between the content of the transcript and the speech uttered, and it can surpass 100% in scenarios where there are more errors in the transcript than words contained within the speech content (i.e. the reference transcript).

From a forensic perspective, it is much more important to consider transcription errors than the percentage of words transcribed correctly. However, quantified overall error rates employed in automatic transcription assessment, such as word error rate, can be distracting in forensic contexts and obscure details about performance. For example, two systems could achieve the same WER where one has deleted most of an utterance but the other has substituted key words, significantly changing the meaning. Consider the example in Table 1, where two possible transcriptions of the utterance “he was having the dragon themed curry” are presented. Both transcripts contain five errors and achieve a WER of 71%, despite huge

differences in their length, meaning and level of incrimination; the substitutions in Transcript 1 change a completely innocent utterance into an incriminating one, and if combined with the court's expectations about the crime (e.g. if the speaker was suspected of drug-related crimes) and an extremely poor quality audio recording, such mistranscriptions could be accepted by the court and the content of the transcript could falsely act as incriminating evidence against the speaker. The deletions in Transcript 2 lead to a transcript that is noticeably incomplete but the central theme of the utterance is retained (i.e. it is about a curry) and, crucially, these errors do not create a falsely incriminating interpretation of the speech content. In a forensic context, therefore, it is crucial to consider not only the number of errors but also the type and magnitude of those errors.

Reference	he	was	having	the	dragon		themed	curry
Transcript 1	he	was	hiding		drugs	in	the	curry
Transcript 2				the				curry

Table 2: Comparison of two transcripts with a reference transcript, with errors highlighted. A shaded red cell shows a deletion, bold red text shows a substitution, and bold blue text shows an insertion.

There is very little work concerning transcription performance within the field of forensic speech science. One attempt at analysing the accuracy of forensic transcripts was carried out by researchers (Tschäpe & Wagner, 2012) from the Department of Speaker Identification and Audio Analysis at the German Federal Criminal Police Office (Bundeskriminalamt; BKA). A small-scale proficiency test was devised to investigate the validity of their transcription methods and to compare the performance of multiple experts, individual experts and non-linguists. The test material used in this experiment consisted of a spontaneous conversation between two German adult males during a car journey that had been recorded on an undercover microphone located between the car's speakers and transmitted via GSM (Global System for Mobile Communications, i.e. by mobile telephone). This style of (covert) recording replicates a common type of sample used by forensic phonetic practitioners in casework. In order to create a reference transcript against which the participant responses could be compared, a high-quality reference recording of the conversation was made simultaneously with microphones attached to each speaker's clothing.

Participants in the proficiency test were 12 German-speaking forensic phonetic experts (including one group of 2 and one group of 3) and 8 non-phonetician BKA employees. The test involved transcribing 211 seconds of speech which was then compared with the

reference transcript on a syllabic level and marked for substitutions, insertions or omissions (deletions). When presenting the experts' results, Tschäpe and Wagner offer three measurements: the percentage of correct identifications, the percentage of omissions and the percentage of substitutions and insertions combined. The authors offer an assumption that within forensic casework, reduced information causes less damage than false information. This view coincides with the work on the 'priming' power of transcripts; once an interpretation has been put forth, it can be almost impossible to dismiss (e.g. Fraser et al., 2011; Fraser & Kinoshita, 2021). Taking this assumption into account, Tschäpe and Wagner offer a different measurement when comparing the performance of experts and non-linguists, which assigns two error points to each substitution or insertion and only one error point to each deletion. The average number of error points within each of the experts' transcripts and the non-experts' transcripts is then compared.

The approach taken by Tschäpe and Wagner (2012), where the focus lies on the errors within the transcripts, is a much more appropriate way to analyse transcripts from a forensic perspective. The differentiation made between the types of errors where information is added (insertions and substitutions) and those where information is omitted (deletions) is a crucial part of interpreting the quality of transcripts. One of the issues with Tschäpe and Wagner's method of evaluating transcripts is that there is no consideration of the different types of (sub-)errors within the three main error categories (insertion, substitution, deletion); for example, all substitution errors are treated equally, despite the fact that some substitutions may make little to no difference to a transcript (e.g. "**the** curry" → "**a** curry") while others could have a very substantial impact on the interpretation of the speech (e.g. "**dragon**" → "**drugs**" in the example within Table 2).

A more nuanced method of evaluating transcripts for forensic purposes, which accounts for the different types of errors as well as their potential impact, is needed. The experimental work in this thesis aims to develop such a method, which can be tailored to small-scale and large-scale data analysis and used for further testing of transcribers and transcription methods. The method will be employed to analyse transcription performance in Article 2 and Article 3.

2.7 Summary

This chapter has covered a number of issues that are relevant to the research carried out in this thesis. Each of the following articles also contains its own review of existing research related to the central themes of the study.

By now, it should be clear that transcription is not a neutral process, but one that involves numerous (often subconscious) decisions by the transcriber, and there are multiple factors that should be taken into consideration when producing transcripts, such as:

- The transcriber's knowledge of contextual information about the circumstances or content of the recording
- The audio quality (e.g. level of background noise) of the recording
- The accent of the speaker in the recording and the transcriber's level of familiarity with that accent

Given the powerful influence that transcripts can have over the way in which speech in poor quality recordings is perceived by the transcript reader, it is essential that the transcripts presented to courts alongside, or in place of, the recording of forensic interest are impartial and accurate. To achieve this, valid and reliable methods of transcription are required; this is the primary focus of the articles that make up this thesis.

2.8 References

- Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, 35(2), 520–529.
- Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken-word comprehension. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, 1925–1928.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.
- Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3), 383–398.
- Bruce, D. J. (1958). The Effect of Listeners' Anticipations on the Intelligibility of Heard Speech. *Language and Speech*, 1(2), 79–97.
- Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, 32(10), 1439–1465.
- Burda, A. N., Casey, A. M., Foster, T. R., Pilkington, A. K., & Reppe, E. A. (2006). Effects of Accent and Age on the Transcription of Medically Related Utterances: A Pilot Study. *Communication Disorders Quarterly*, 27(2), 110–116.
- Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: intelligibility and classification. *Language and Speech*, 51(Pt 3), 175–198.

- Collins, H., Leonard-Clarke, W., & O'Mahoney, H. (2019). "Um, er": how meaning varies between speech and its typed transcript. *Qualitative Research: QR*, 19(6), 653–668.
- Deamer, F., Richardson, E., Basu, N., & Haworth, K. (2022). For the Record: Exploring variability in interpretations of police investigative interviews. *Language and Law*, 9(1), 25–46.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from Four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16.
- DiChristofano, A., Shuster, H., Chandra, S., & Patwari, N. (2022). Performance Disparities Between Accents in Automatic Speech Recognition. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2208.01157>
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology. Human Perception and Performance*, 32(5), 1276–1293.
- Fong, J. Y., Borsky, M., Helgadóttir, I. R., & Gudnason, J. (2018). Manual Post-editing of Automatically Transcribed Speeches from the Icelandic Parliament - Althingi. In *arXiv [eess.AS]*. arXiv. <http://arxiv.org/abs/1807.11893>
- Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech Language and the Law*, 10(2), 203–226.
- Fraser, H. (2020). Forensic transcription: The case for transcription as a dedicated branch of linguistic science. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 416–431). Routledge.
- Fraser, H. (2021). The development of legal procedures for using a transcript to assist the jury in understanding indistinct covert recordings used as evidence in Australian criminal trials: A history in three key cases. *Language and law*, 8(1). <https://ojs.letras.up.pt/index.php/LLLD/article/view/10953>
- Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, 7. <https://doi.org/10.3389/fcomm.2022.898410>
- 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.
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2). <https://doi.org/10.1558/ijsl.v18i2.261>
- French, P., & Fraser, H. (2018). Why "Ad Hoc Experts" should not Provide Transcripts of Indistinct Audio, and a Better Approach. *Criminal Law Journal*, 298–302.

- Gibbons, J. (2003). *Forensic linguistics: An introduction to language in the justice system*. Blackwell Publishing.
- Giroux, M. (2022). *Confirmation bias for degraded forensic audio evidence* [Simon Fraser University]. <https://summit.sfu.ca/item/35863>
- Green, J., Franquiz, M., & Dixon, C. (1997). The Myth of the Objective Transcript: Transcribing as a Situated Act. *TESOL Quarterly*, 31(1), 172–176.
- Harrington, L., Love, R., & Wright, D. (2022, July). *Analysing the performance of automated transcription tools for covert audio recordings*. Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, 22(4), 428–450.
- Innes, B. (2011). R v David Bain--a unique case in New Zealand legal and linguistic history. *International Journal of Speech, Language & the Law*, 18(1).
<https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=17488885&asa=Y&AN=67122231&h=5rONQ3tYVrktv2aohXj5dFgp%2BjP%2B7qc6uQ8wflTsZa5M8HVhk9L18ZgLNJY0oy2rbX5xYfOunXeS73NKBGNouA%3D%3D&crl=c>
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. *Conversation analysis*, 13-31.
- Jenks, C. J. (2013). Working with transcripts: An abridged review of issues in transcription. *Language and Linguistics Compass*, 7(4), 251–261.
- Jones, T., Kalbfeld, J. R., Hancock, R., & Clark, R. (2019). Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2), e216–e252.
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7684–7689.
- Kowal, S., & O'Connell, D. C. (2004). The Transcription of Conversations. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A Companion to Qualitative Research* (pp. 248–252). SAGE.
- 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.
- Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104.
- Lima, L., Furtado, V., Furtado, E., & Almeida, V. (2019). Empirical Analysis of Bias in Voice-based Personal Assistants. *Companion Proceedings of The 2019 World Wide Web Conference*, 533–538.

- 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
<https://apps.dtic.mil/sti/citations/ADA414420>
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7.
<https://doi.org/10.3389/fcomm.2022.803452>
- Loakes, D. (2024). Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare? *Frontiers in Communication*, 9.
<https://doi.org/10.3389/fcomm.2024.1281407>
- MacLean, L. M., Meyer, M., & Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research*, 14(1), 113–123.
- Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 521–534.
- Meyer, J., Rauchenstein, L., Eisenberg, J. D., & Howell, N. (2020). Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6462–6468.
- Mimura, M., Sakai, S., & Kawahara, T. (2021). An End-To-End Model from Speech to Clean Transcript for Parliamentary Meetings. *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 465–470.
- Science & Technology in Policing. (2023). *Covenant for Using Artificial Intelligence (AI) in Policing*. Science & Technology in Policing.
<https://science.police.uk/delivery/resources/covenant-for-using-artificial-intelligence-ai-in-policing/>
- Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R.-A. (2014). Cross-accent intelligibility of speech in noise: long-term familiarity and short-term familiarisation. *Quarterly Journal of Experimental Psychology*, 67(3), 590–608.
- Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4), 487–501.
- Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59.
- Tompkinson, J., Haworth, K., Deamer, F., & Richardson, E. (2023). Perceptual instability in police interview records: Examining the effect of pauses and modality on people's perceptions of an interviewee. *International Journal of Speech, Language and the Law*, 30(1), 22–51.
- 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.*
Conference of the International Investigative Interviewing Research Group, Winchester.
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test.*
Conference of International Association for Forensic Phonetics and Acoustics,
Santander, Spain.
- Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment
completion are independent of recognition memory. *Journal of Experimental
Psychology. Learning, Memory, and Cognition*, 8(4), 336–342.
- Walker, A. (2018). The effect of long-term second dialect exposure on sentence transcription
in noise. *Journal of Phonetics*, 71, 162–176.
- Walker, A. G. (1990). Language at work in the law. In *Language in the Judicial Process* (pp.
203–244). Springer US.
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917),
392–393.

3. Thesis aims

The ultimate goal of the work carried out as part of this thesis is the provision of better transcripts within the criminal justice system. 'Better' here refers to aspects such as:

- More accurate content of transcripts
- The use of methods which have been validated through extensive empirical testing
- Clear presentation of transcripts which has been shown to be effective
- Proven transcriber competency
- The use of appropriate hardware and software

To work towards this goal, it is first necessary to understand how transcripts are currently being produced, which can then be followed by empirical research investigating ways to improve current practices. There is currently a distinct lack of research into the production of transcripts within the criminal justice system, and so the scope of this thesis is intentionally broad in order to cover a multitude of related issues. This thesis will contribute to the gap in the literature and act as a basis for further empirical research addressing transcription and the production of transcripts in the criminal justice system

More specifically, this thesis has three main research aims.

1. To gain a better understanding of current practices.

This will be achieved through a comprehensive overview of what is currently known about the transcripts presented to juries and their production, as well as through the results of a survey of international expert transcription practices. The survey (presented in Article 1) is the first of its kind to explore forensic transcription methods in depth, and will provide information about the practices employed by forensic practitioners across the world to produce transcripts of poor quality evidential audio. This will reveal parts of the methods that are carried out similarly across practitioners as well as those which are subject to disagreement. In exploring what is currently being done, areas which require further research can be identified, which will in turn contribute to the development of better, more robust methods based on empirical evidence.

2. To explore how factors relating to audio quality and regional accent can affect the content of transcripts.

Varying audio quality and regional accents are commonly encountered in evidential recordings and police-suspect interviews, and were the two factors rated by expert practitioners as the most influential on the perception and transcription of speech in poor quality evidential audio recordings in Article 1. Both the level of background noise and regional accent have been found to significantly impact transcription performance in a range of psycholinguistic studies, but the analysis in this thesis focuses more specifically on the content of transcripts (i.e. errors) rather than overall measures of success.

Article 2 and Article 3 present experiments which focus on the effects of these two factors on the content of transcripts, albeit via different approaches. Article 2 explores human transcription with three different levels of background noise, and the transcriber's degree of familiarity is manipulated such listeners are either a native speaker of the accent or are highly familiar with the accent through exposure despite being a speaker of a phonologically-different regional variety of British English. Article 3 explores automatic transcription of a standard and a non-standard regional variety of British English, as well as the effects of relatively minimal background noise.

3. To develop a method for analysing transcription performance for forensic purposes.


A new approach to analysing and evaluating the transcripts produced by humans or machines will be developed and used within this thesis. This approach will consider the forensic implications of the errors and guide the focus of the analysis within the two experimental studies presented in Article 2 and Article 3. In Article 2, the method of analysis will be applied to large-scale human transcription data, while in Article 3 it will be applied to small-scale machine transcription data. This novel evaluation method will begin to investigate which parts of transcripts are crucial to consider in forensic contexts, and ideally will act as a basis for an official framework which can later be used in the development of proficiency testing for experts (as discussed in Article 1).

University of York
York Graduate Research School
Research Degree Thesis Statement of Authorship

Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

Candidate name	Lauren Harrington
Department	Language and Linguistic Science
Thesis title	Towards improving transcripts of audio recordings in the criminal justice system


Title of the work (paper/chapter)	Forensic transcription practices: survey results from experts in Europe and North America	
Publication status	Published	
	Accepted for publication	
	Submitted for publication	x
	Unpublished and unsubmitted	
Citation details (if applicable)		

Description of the candidate's contribution to the work*	Conceptualisation (lead) Methodology Investigation Data curation Writing - original draft
Approximate percentage contribution of the candidate to the work	85%
Signature of the candidate	
Date (DD/MM/YY)	29/02/2024

Co-author contributions

By signing this Statement of Authorship, each co-author agrees that:

- (i) the candidate has accurately represented their contribution to the work;
- (ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	Richard Rhodes
Contact details of co-author	Forensic Voice Centre, York, UK richard.rhodes@york.ac.uk
Description of the co-author's contribution to the work*	Conceptualisation (supporting) Writing - review & editing
Approximate percentage contribution of the co-author to the work	15%
Signature of the co-author	
Date (DD/MM/YY)	08/02/2024

*The description of the candidate and co-authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co-authors may find it helpful to consider the [CRediT \(Contributor Roles Taxonomy\)](#) approach to recognising individual author contributions.

4. Article 1 - Forensic transcription practices: survey results from experts in Europe and North America

Lauren Harrington¹ & Richard Rhodes^{1,2}

¹Department of Language & Linguistic Science, University of York, UK

²Forensic Voice Centre, York, UK

Abstract

In this article we present the results of a survey about expert transcription practices, which received responses from 28 forensic practitioners from Europe and North America. We found a number of areas of convergence in transcription methods, such as the production of multiple drafts and widespread awareness of priming and cognitive bias. Responses also revealed some areas with less agreement among practitioners, particularly concerning drafting methods and how and when contextual information is consulted by transcribers. We discussed these issues at a workshop with practitioners and have explored some of their comments within this article. This study has highlighted the need for more empirical research to be carried out about forensic transcription methods, and so we have suggested some potential research questions which arise from the survey results and the discussions with practitioners.

1 Introduction

1.1 Transcription in forensic casework

Forensic phonetics experts are often instructed to transcribe the speech in audio or video recordings which arise in criminal matters or other types of court cases. The transcription of forensic audio materials may also be carried out by a range of people including police officers or police staff. In this paper, we are interested only in those practitioners who have specialist qualifications, training and/or experience in forensic speech science, and who would be considered expert witnesses in court (within this paper, unless otherwise specified, 'practitioner' refers to a forensic phonetic expert who carries out forensic transcription casework). Some research has been conducted on the transcription methods used by non-expert practitioners, such as police interview transcribers in the UK (Haworth, 2018), but the methods employed by experts have received very little attention, particularly in comparison with other types of forensic speech analysis such as voice comparison.

In the experience of the second author, who is a forensic expert with over 12 years of forensic casework experience in the UK, transcription is the second most frequently requested forensic service after forensic voice comparison. This seems to also be the experience of many practitioners working in mainland Europe (personal communication), though it may not be the case for all practitioners in all jurisdictions. Transcription differs from voice comparison and other types of forensic examinations like DNA, drug or fingerprint tests in that it does not produce a singular test result or conclusion; rather, transcription is a service which produces a product to assist an investigation or trial process. The purpose of a forensic transcript is to assist the user within the criminal justice system in interpreting speech in an audio or video recording (Fraser, 2022); the recording is the primary evidence, and the transcript is an aid to interpreting and following the speech, for the police officer, lawyer, defendant, judge or jury member or any other participant in a legal process.

In the second author's experience, typical situations where practitioners are instructed to prepare transcripts include, but are not limited to, the following:

- 1) **Telephone calls to the emergency services:** all emergency calls are recorded, and can contain key evidence. The speech of interest is often in the background of the call, and may be simultaneous to interactions between the caller and the emergency operator.
- 2) **Other types of telephone call:** for example, bank or insurance calls in fraud cases.
- 3) **Telephone intercepts:** telephone intercept recordings are not admissible as evidence in UK courts, but they are used commonly across Europe and the rest of the world.
- 4) **Covert or undercover officer recordings:** recording devices planted in locations or operated by undercover officers can capture relevant conversations from target speakers.
- 5) **Recording of conversations made (non-)deliberately by a participant:** some speakers accidentally or deliberately/covertly record themselves or others making admissions or discussing offences; this might be done using a smartphone voice recording application.
- 6) **CCTV recordings:** CCTV recordings might show a violent incident, a burglary or discussions about offending, for example. These can include fixed CCTV camera systems in houses or commercial properties, video doorbells (such as Ring or Nest), or portable devices such as body-worn cameras which are increasingly being used by police and security staff.

- 7) **Smartphone videos or recordings (sometimes via social media):** material showing, or relating to, offences may be posted online or found on seized digital devices, often smartphones. These can include videos from social platforms like Facebook, X or YouTube, voice messages from WhatsApp, Snapchat and similar applications.

Practitioners are generally instructed to provide forensic transcripts based on the following requirements:

- 1) **To provide an impartial transcript**, i.e., a transcript that is prepared by an independent and impartial party, rather than by members of the investigating police or prosecution team or defence team.
- 2) **To provide an accurate interpretation of what was said in the recording**; or a more accurate version”, if another transcript is in evidence and is being contested.
 - a) There may also be disputes about certain words or phrases in a longer recording; these can be analysed using a disputed or questioned utterance approach (see further in French et al., 2013; Innes, 2011; Morrison, Enzinger & Zhang, 2018).
- 3) **To attribute speech in the recordings to different speakers.** This can be done by attributing speech within a recording to the different speakers present, and also by using voice comparison methods to compare voices attributed in a transcript to different known speakers from reference recordings of their voice(s).
- 4) **To provide a clear, readable document with verbatim transcription** which can be followed while listening to an audio recording or viewing a video recording.

The technical method for transcribing is, in the experience of the authors, similar across most practitioners. This is to use high-quality headphones (these are typically closed-cup, the kind that fit over the ears, and are often closed-back, meaning they reduce interference from external sounds) and audio playback equipment in a controlled environment to listen to recordings within specialist software; this software allows the expert to closely control how the audio is played. Experts listen repeatedly to sections of different lengths - this might be a sentence, a phrase, a word, or even a part of a word or an individual sound - and write down what was heard orthographically, i.e., in words rather than using phonetic notation or other systems. Some of these software applications also allow the expert to make sections louder or apply enhancement processes, to play them at different speeds, and to view spectrographic displays of the sounds which can help identify and determine speech content.

However, there are different approaches that can be taken during the production of a transcript, such as the number of analysts involved and the way in which drafts are produced and analysed. It is these approach-level differences that can vary widely between practitioners and are therefore the subject of discussion in this article.

1.2 Motivation

The main aim of this article is to present basic factual information about current expert transcription practices, much like in Gold and French's (2011) paper about their survey on international forensic speaker comparison practices. This is to inform practitioners and researchers about forensic transcription methods and to help identify questions for further research. Previously, one survey (Rhodes, 2016) has asked a limited number of questions about forensic transcription methods. The present paper offers a more comprehensive and up-to-date view of how transcripts of forensic audio materials are produced by practitioners.

Another motivation for the study, from the authors' perspective, is the growing demand for validation of forensic examination methods by regulatory bodies such as the UK Forensic Science Regulator (FSR). The current situation in the UK is that forensic transcription is recognised by the FSR Code of Practice as a 'Forensic Science Activity' but does not require accreditation to standards such as ISO/IEC 17025; this is unlike the situation for other forensic processes such as DNA or cell-site analysis, or footwear, document or toolmark examinations, which require such accreditation. Experts carrying out forensic transcription are nonetheless expected to follow the Code, which requires experts to demonstrate the validity of their methods, among other requirements. UK audio and speech laboratories are expected to follow a standard procedure and demonstrate the validity of their methods.

Our main research focus, then, is to illustrate what methods forensic experts are currently using to transcribe evidential recordings. The secondary purpose is to identify further research routes to test, validate and improve the methods that are used by experts in criminal and other legal cases.

1.3 Article structure

In this paper, Section 2 ('Existing research') provides an overview of the existing research into methods of producing forensic transcripts. Our focus is necessarily limited to the orthographic transcription of speech in the practitioner's native language, though we acknowledge that activities like forensic translation may involve similar methodologies. In

section 3 ('Methods'), we outline the methods employed to gather the responses discussed in this paper; these were collected via both an online survey and a group discussion with forensic practitioners at a workshop organised by the authors. The responses from the first phase of data collection – the online survey – are presented in Section 4 ('Responses'), while details of the discussion that took place at the workshop are presented in Section 5 ('Discussion'). In Section 5, we also suggest multiple areas which require further research, including specific research questions that may be of interest to researchers.

1.4 Scope

The focus of this paper is limited to:

- Transcription carried out by experts in forensic speech science
- Forensic transcription of evidential material (not translation, interpreting, court transcripts, interview transcripts, etc.)
- Orthographic transcription of speech content rather than attribution of speech in multi-speaker dialogues

Much research has been conducted on the act of transcription (e.g. Biber, 1988; Bucholtz, 2000), academic applications of transcription (e.g. Jenks, 2013; MacLean, Meyer & Estable, 2004) and legal applications such as court reporting (e.g. Walker, 1990; Jones et al., 2019) and police interview transcription (e.g. Haworth, 2018). However, there is a distinct lack of literature concerning forensic transcription practices, and our aim is for this paper to provide a foundation for further discussion about methods of forensic transcription in the literature as well as further research into best practice.

2 Existing research

Forensic transcription is an under-researched area of forensic speech science. While the topic has received attention over the last two decades, mainly by Fraser and colleagues, the majority of research has focused on issues surrounding the presentation of non-expert transcripts to the court and the 'priming' power of transcripts (e.g. Fraser, 2003; Fraser, Stevenson & Marks, 2011; Fraser & Kinoshita, 2021). These articles explore the complexity of transcribing poor quality audio and the risks associated with unreliable transcripts of forensic audio materials. In more recent work there has been a focus on establishing an evidence-based method for producing reliable transcripts (e.g. Fraser, 2021), with an emphasis on competent transcribers with appropriate training in the style of transcription required (Fraser, 2022). However, this method is still in the process of being developed and

there is scant information available on the methods currently employed by practitioners when producing forensic transcripts.

An upcoming chapter in the *Oxford Handbook of Forensic Phonetics* (Harrison & Wormald, in press) provides a description of the production of forensic transcripts, though Harrison and Wormald acknowledge their focus on methods they are most familiar with, i.e., those employed by the private audio and speech laboratory in the United Kingdom where both authors formerly worked.

In terms of experimental work concerning practical methods of forensic transcription, researchers at the Bundelskriminalamt (BKA), the Federal Criminal Police Office of Germany, conducted a small-scale proficiency testing study to investigate whether experts perform better than lay people, and whether groups of experts perform better than individual experts (Tschäpe & Wagner, 2012). They found that transcripts produced by experts were significantly better than those produced by lay listeners (non-phonetician employees of the BKA) in terms of the number of errors made. It was also reported that transcripts produced by multiple transcribers were better than those produced by individual transcribers. However, it should be noted this was a small-scale study, with only two groups of experts (one group of 2, one group of 3) and seven individual experts.

Recent research has also investigated whether automatic speech recognition (ASR) software can be applied to forensic audio materials. Loakes (2022) tested the performance of two automatic systems (BAS Services and Descript) using a 44-second forensic-like audio recording of a band rehearsal. She found that the first system did not attempt to transcribe the audio, returning an error due to a bad quality signal, and the second system identified the words “yeah”, “yes” and “okay”, comprising 1.7% of the overall speech content of the recording. In a follow-up study carried out two years later and using the same audio recording, Loakes (2024) found that Open AI’s Whisper performed much better than the other commercial ASR systems in the study, though still only managed to successfully transcribe 50% of the recording. Loakes concludes that this technology is not accurate enough for use in forensic transcription cases. This conclusion is shared by Harrington, Love and Wright (2022), who tested 12 commercial ASR systems on a 4-minute recording of conversation in a busy restaurant; even in short, relatively clear sections of speech, the best performing system (Microsoft) achieved a successful transcription rate of around 70%.

In terms of other expert surveys, a survey on cognitive bias in forensic speech science was conducted by Rhodes (2016) and featured a limited number of questions on transcription

methods. Findings from 26 respondents from Europe and North America concluded that transcription was considered an area which could be highly susceptible to bias, particularly from priming, and that transcription methods are not standardised across labs and practitioners. Responses also showed that multiple analysts (most commonly two or three) are typically involved in the production of a transcript and that the most common number of drafts produced is two or three. Of the 26 respondents in Rhodes' survey, nearly all stated that transcripts provided by other agencies (such as the police) are consulted at some stage of the transcription process, with 20% of respondents consulting these from the beginning of the process.

3 Methods

We carried out two phases of data collection to elicit responses concerning forensic transcription practices from practitioners: an online survey and a group discussion with forensic practitioners at a workshop. The data collected is necessarily anecdotal and based on the personal experience of the respondents. Ethics approval was granted for both phases of data collection by the University of York.

3.1 Survey

The survey was conducted online, in English, using Google Forms; data collection took place between September 2021 and September 2022. The survey consisted of 30 questions divided into five topics:

- 1) Details of work & affiliations
- 2) Transcription methods
- 3) Content of transcripts
- 4) Cognitive bias
- 5) Concluding remarks

Questions drew inspiration from those asked in Gold and French (2011) and Rhodes (2016), with additional questions investigating specific aspects of the transcription process, e.g. how certain features are represented and what type of equipment is used. Most questions required an answer before the participant could progress, though the first (details of work & affiliations) and last (concluding remarks) sections were entirely optional¹⁶.

¹⁶ The possibility of complete anonymity was offered in order to encourage a larger number of responses, though we acknowledge that this methodology does not allow us to confirm the respondents' credentials. However, given the content of the responses received, we believe that each

The survey was distributed via direct email or mailing lists to relevant organisations, such as the International Association for Forensic Phonetics and Acoustics (IAFPA) and the European Network of Forensic Science Institutes (ENFSI)¹⁷. This phase of data collection was intended as an opportunistic survey of any relevant forensic practitioners. In total, 28 forensic practitioners completed the survey. For reference, 36 practitioners responded to Gold and French (2011) and 39 responded to Gold and French (2019) on forensic voice comparison methods, both of which had a slightly larger scope than the present study.

The survey contained a range of question types, such as multiple-choice responses and free text responses. Analysis of the responses involved a combination of quantitative and qualitative methods. For some of the multiple-choice questions, there was an ‘Other’ category where respondents could instead enter a free text response; for these questions, and the other free text questions, it was necessary to manually code the responses. For some questions, it was necessary to cross-reference responses to multiple questions to accurately categorise responses.

3.2 Workshop discussions

The second phase of data collection took place at an IAFPA webinar run by the authors in November 2023 entitled “Workshop: forensic transcription practices”. At the workshop, results of the online survey were presented and a number of specific points were highlighted for discussion. This group discussion allowed practitioners to provide further details on their methods and to highlight parts of their methods which could benefit from more research in the field of forensic transcription. Comments from the workshop are included in the discussion in Section 5.

4 Responses

The responses of the survey are presented in this section. The following subsections cover survey topics 1 to 4.

of the respondents do belong to the target community, i.e. practitioners carrying out (expert) forensic transcription as part of their work duties.

¹⁷ The survey was also disseminated on the FORENSIC-LINGUISTICS JISCMail mailing list. Members of the list include academic linguists, academic and practising lawyers and legal interpreters according to the [mailing list's description](#).

4.1 Details of work & affiliations

This section of the survey was optional to offer complete anonymity to respondents if desired. One respondent chose not to answer the questions regarding their background information.

4.1.1 Country of work

Participants were asked to specify the country or jurisdiction in which they primarily work (Figure 1). The majority of respondents (24 out of 28) primarily work within Europe, with a quarter (8 out of 28) practising in the UK.

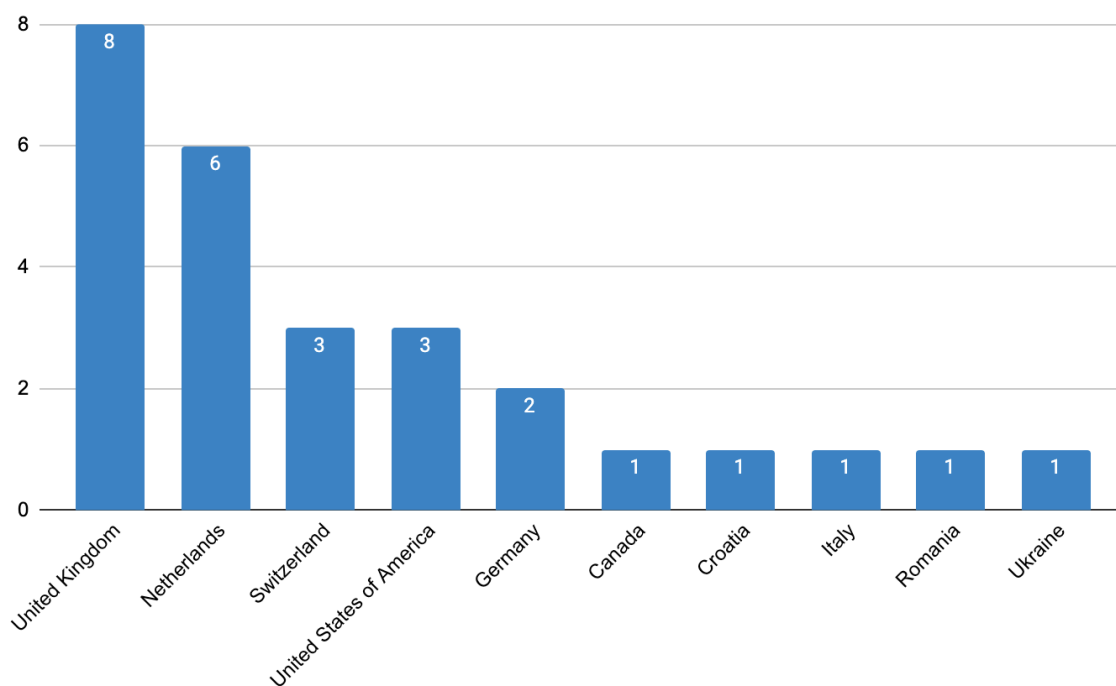


Figure 1: Countries or jurisdictions in which the respondents primarily work, ordered from most to least frequently answered.

4.1.2 Working affiliation(s)

Participants were asked to indicate their working arrangements or affiliation(s) (Table 1). The majority of respondents reported working in a governmental laboratory. It is worth noting that two of the fifteen respondents who work at a governmental laboratory and four of the seven respondents who work as an individual private practitioner indicated additional affiliations with a research institute or university department.

Response	N	% of respondents
Governmental laboratory	15	55%
Individual private practitioner	7	26%
Independent facility with multiple members of staff / private provider	4	15%
Research institute / university department	1	4%

Table 1: Working arrangements or affiliations of respondents.

4.2 Methods of transcription

4.2.1 Frequency of transcription work

Participants were asked how often they carry out forensic transcription as part of their work duties (Figure 2). Around half of the respondents frequently or very frequently carry out casework of this type, while the other half reported that forensic transcription is carried out occasionally or rarely. This aligns with our expectation, given that forensic transcription tends to be carried out much less regularly than forensic voice comparison.

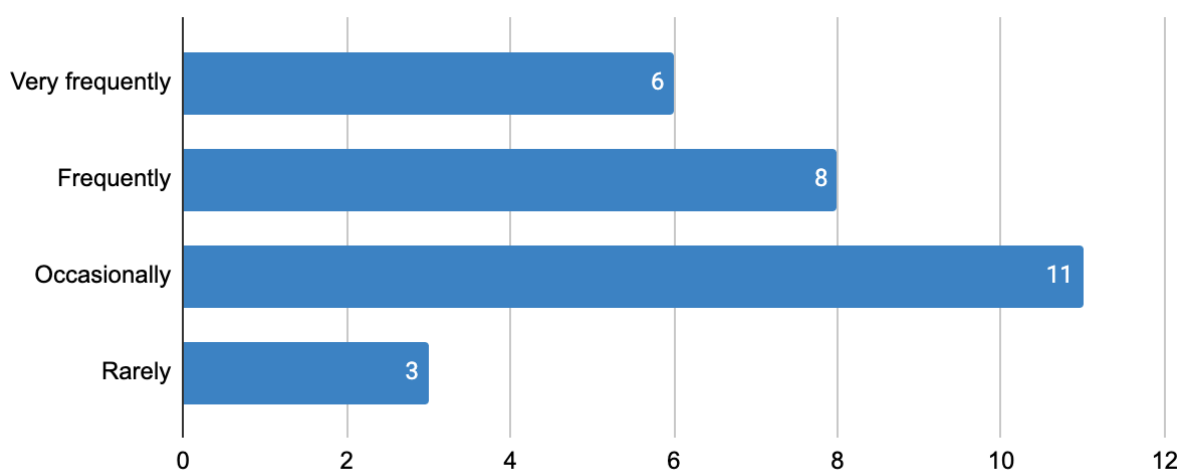


Figure 2: How frequently respondents carry out transcription as part of their work duties.

4.2.2 Number of transcribers

Participants were asked how many transcribers typically contribute to the production of a transcript (Table 2). The majority of respondents stated that multiple transcribers work on a transcript; most commonly, two or three transcribers work on a transcript, though one respondent indicated that four or more transcribers typically contribute. It is worth noting that of the eight respondents who indicated that one transcriber works on a transcript, five work as individual private practitioners (and therefore do not have the opportunity to involve multiple transcribers).

Response	N	% of respondents
1 transcriber	8	28%
2 transcribers	12	43%
3 transcribers	7	25%
4+ transcribers	1	4%

Table 2: Number of transcribers that typically work on a transcript.

4.2.3 Number of drafts

Participants were asked how many drafts are typically made during the production of a forensic transcript (Table 3). The vast majority of respondents reported the production of multiple drafts of a transcript, with most respondents producing between two and five drafts. It is worth noting that all four respondents who specified that one draft is typically produced also indicated that a single transcriber works on the transcript.

Response	N	% of respondents
1 draft	4	14%
2-3 drafts	12	43%
4-5 drafts	7	25%
6+ drafts	5	18%

Table 3: Number of drafts typically produced.

4.2.4 Drafting methods

Participants were asked whether they typically opt for a sequential drafting method, whereby each new draft builds on an existing draft, or a parallel drafting method, whereby independent drafts are produced by multiple transcribers and then merged together:

- Of the 20 respondents who indicated that multiple transcribers work on a transcript,
 - 7 respondents typically employ a sequential drafting method
 - 13 respondents typically employ a parallel drafting method
- The remaining 8 respondents, all of whom work alone, either produce only one draft or typically employ a sequential drafting method

It should be noted that many of the respondents specified that they would employ a parallel method initially, followed by a sequential method; such responses were categorised as 'parallel drafting'. Some responses which were categorised as 'sequential drafting' contained detail of parallel methods being used for shorter recordings. This issue is discussed in greater detail in Section 6.3.2.

4.2.5 Equipment

Participants were asked what equipment is used by the person(s) carrying out the transcription. This question was answered with varying amounts of detail. In terms of software, the mostly commonly mentioned computer program was Praat (14 respondents), followed by Adobe Audition (6 respondents) and Sound Forge (6 respondents). Other computer programs mentioned by between 1 and 4 respondents include Audacity, WaveLab, iZotope RX and ELAN. In terms of hardware, 22 respondents specified that they use high quality closed-cup headphones, with mentions of brands including Sennheiser (9

respondents), Sony (4 respondents), AKG (3 respondents) and Bose (2 respondents). Other pieces of hardware mentioned include amplifiers and sound cards.

4.2.6 Phonetic analysis

Participants were asked how often phonetic and/or acoustic analysis (e.g. acoustic measurements, spectrographic analysis, narrow transcription using the International Phonetic Alphabet) is employed when producing a forensic transcript (Figure 3). There were a range of responses, though most respondents indicated that this type of analysis occasionally or frequently takes place.

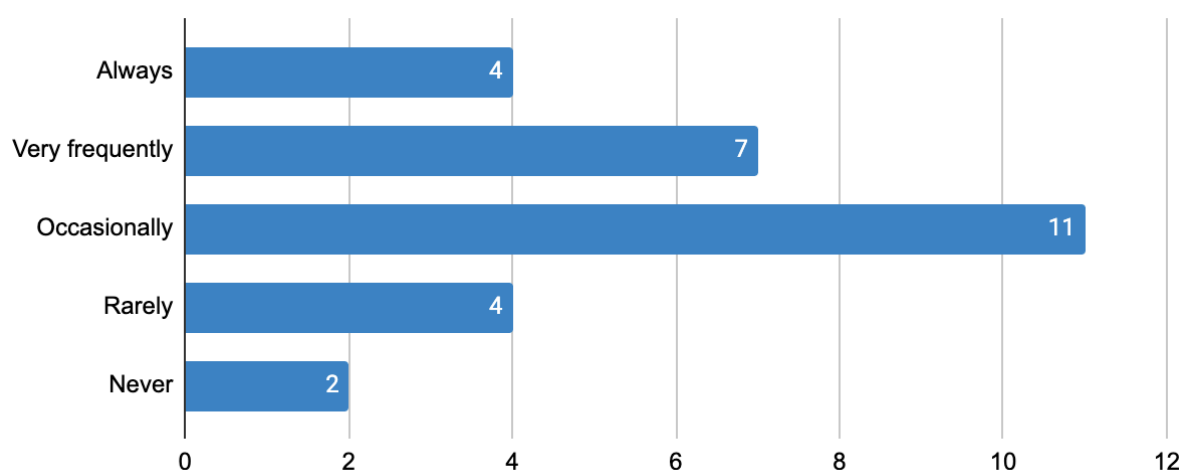


Figure 3: How often respondents employ phonetic and/or acoustic analysis during the production of a transcript.

4.2.7 Audio enhancement

Participants were asked whether recordings for transcription are usually enhanced beforehand (Figure 4). There were a range of responses with a relatively even split between those who do not typically enhance audio prior to transcription and those who sometimes or typically do. It is worth noting that this question presented respondents with three options ('Yes', 'No', 'Other'); some respondents opted for 'Other' in order to expand their answer, and these were categorised accordingly. However, multiple responses did not align with either 'Yes' or 'No', hence the creation of the 'Sometimes' category.

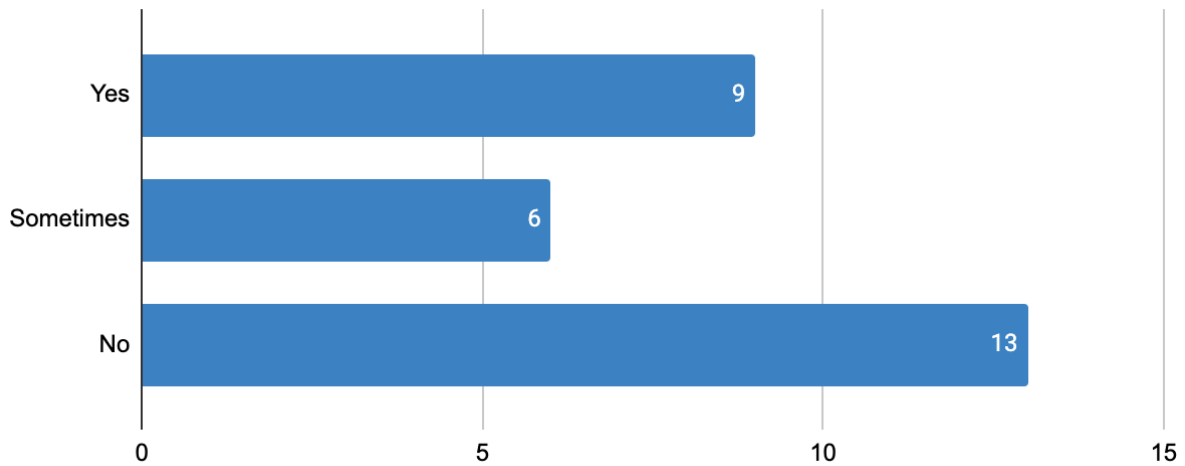


Figure 4: Responses concerning whether audio recordings are typically enhanced before transcription.

5 respondents that indicated that enhancement typically or sometimes takes place commented that while both versions of the recording are incorporated into the transcription process, they tend to primarily focus on the original version. Participants were additionally asked to specify which enhancement methods are commonly used:

- 11 respondents mentioned the use of noise reduction techniques
- 10 respondents mentioned the use of filtering techniques
- 6 respondents mentioned the use of amplification techniques
- 4 respondents mentioned the use of equalisation techniques

4.2.8 Automatic speech recognition

Participants were asked whether automatic speech recognition is used in the production of a transcript (Table 4). One respondent stated that they use software for the purpose of automatic speech-silence detection; however, none of the respondents employ automatic methods to transcribe speech within forensic recordings.

Response	N	% of respondents
No	27	96%
Yes: for automatic speech-silence detection only	1	4%

Table 4: Responses concerning the use of automatic speech recognition during the production of a transcript.

4.3 Content of transcripts

4.3.1 Levels of confidence

Participants were asked how many levels of confidence are represented in their transcripts:

- 2 respondents indicated that only speech that has the transcriber's full confidence is included in the transcript
- 26 respondents indicated that they include multiple levels of confidence within their transcripts, most commonly two levels (i.e. speech with the transcriber's full confidence and speech with a lower level of confidence)

Participants were asked to provide details of how these levels of confidence are represented. In all cases, plain, unmarked or unbracketed text represents transcription that has the transcriber's full confidence. The 26 respondents who indicated that they include multiple levels of confidence use the following representations:

Response	N	% of respondents
Brackets for speech with lower confidence e.g. "good (night)"	16	61%
Numerical or discursive system to score confidence, e.g. separate column with 0 (no confidence) to 4 (very confident)	2	8%
Question mark in brackets for speech with lower confidence, e.g. "good night(?)"	1	4%
No clear indication of how levels of confidence are represented	7	27%

Table 5: Ways in which sections of lower confidence are represented within transcripts.

Furthermore, 13 respondents stated that alternatives are given in cases where a word could be one of two options, and this is most commonly represented via the use of brackets to indicate uncertainty and a forward slash between the options, e.g. "good (night/might)".

4.3.2 Representation of unintelligible speech

Participants were asked how unintelligible speech is represented within their transcripts (Table 6). Half of respondents use an ellipsis to signal a section of unintelligibility, and most

of the rest use some kind of descriptive label. Furthermore, seven respondents within the group mentioned some form of representation of the number of syllables within the unintelligible section: four respondents use either an ellipsis or a number to represent the number of syllables (if possible to determine); two respondents use a descriptive label with the number of syllables included, e.g. “[approx. 3 syllables unintelligible]”; and one respondent uses one “x” per syllable (as per the final row of Table 6).

Response	N	% of respondents
Ellipsis, i.e. “...”	14	50%
Descriptive label in brackets, e.g. “[unintelligible]” or “{INDISTINCT}”	11	39%
Blank space or empty bracket, e.g. “[]”	2	7%
“x” per syllable in the unintelligible section	1	4%

Table 6: Ways in which unintelligible speech is represented within transcripts.

4.4 Cognitive bias

4.4.1 Factors affecting transcription

Participants were presented with the following list of factors that could potentially influence the perception and transcription of speech within forensic recordings:

- Channel/quality degradation (e.g. level of background noise)
- Your level of experience with the speaker's accent/dialect
- Information from instructing party (e.g. incriminating evidence linked to speaker)
- Information about the type of offence (e.g. terrorism, drug trafficking, theft)
- Expectations of instructing party (e.g. pressure to get a certain result)
- Content of recordings (e.g. highly emotive or distressing speech)

They were asked to rate each factor on a scale of 1 (no influence) to 6 (great influence) to show the extent to which they believe this factor could influence the perception and transcription of speech within forensic recordings. In Figure 5, ratings are grouped into three categories: 1 and 2 indicating little to no influence, 3 and 4 indicating some influence, and 5 and 6 indicating a great influence.

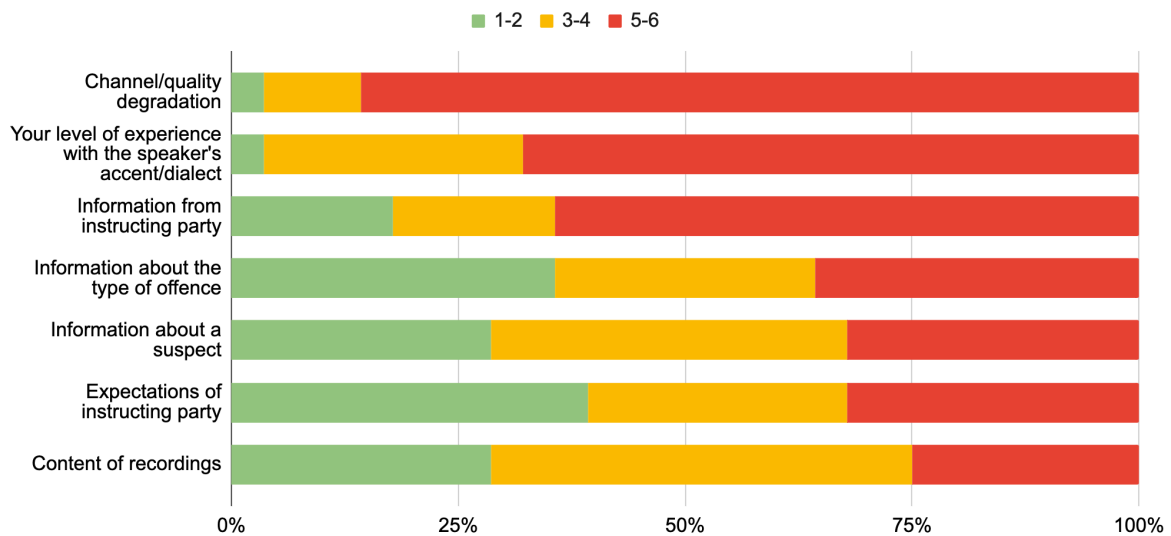


Figure 5: Distribution of respondents' ratings for each factor.

Channel/quality degradation was clearly viewed by respondents to be the factor with the most potential to influence the perception and transcription of speech in poor quality recordings (average rating = 5.29). Following this was the respondent's level of experience with the accent or dialect of the speaker (average rating = 4.82). This aligns with experimental research which has shown both factors to have a significant impact on transcription performance (e.g. Clopper & Bradlow, 2008; Smith et al., 2014).

4.4.2 Awareness of cognitive bias

Participants were asked whether they consider cognitive bias to play a "significant role" in forensic transcription (Table 7). The vast majority acknowledged that cognitive bias could affect forensic transcription; participants were later asked about the strategies employed to mitigate the risks of cognitive bias.

Response	N	% of respondents
Yes	24	86%
No	4	14%

Table 8: Responses concerning respondent's views on whether cognitive bias could play a significant role on the transcription of forensic audio materials.

4.4.3 Protocol

Participants were asked whether they have some form of protocol or any procedures in place to mitigate the effects of cognitive bias and priming (Figure 6). The vast majority of respondents stated that they do.

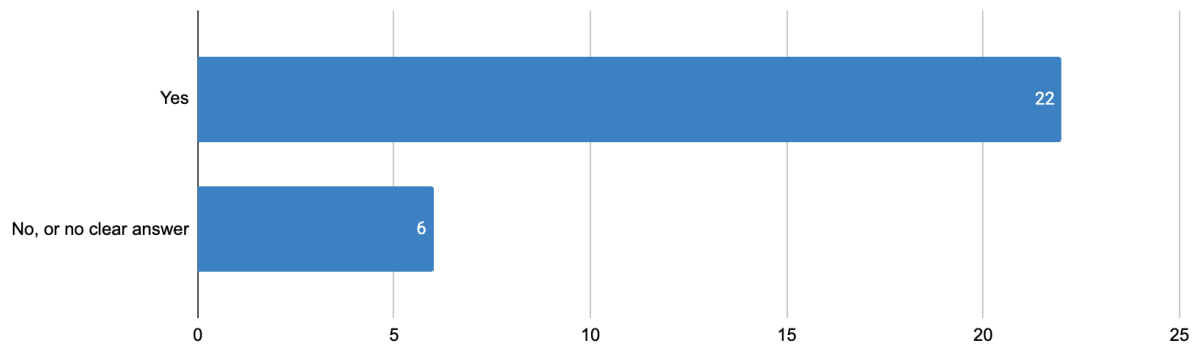


Figure 6: Responses concerning the existence of protocol or precures to mitigate the effects of cognitive bias and priming.

Participants were then asked to describe these procedures. Common procedures include:

- No background information or minimal background information given to transcribers (10 respondents)
- Revealing contextual information after blind drafts have been produced (10 respondents)
- Multiple transcribers working on independent drafts (9 respondents)
- An independent person acting as 'information manager' (7 respondents)

Information management strategies are discussed in Rhodes et al. (in press).

4.4.4 Transcripts from instructing party

Participants were asked if they refer to transcripts produced by the instructing party at any stage during their transcription process (Figure 7). Of those who receive such transcripts, there was a relatively even split between respondents who opt to refer to them after blind drafts have been produced and those who choose not to refer to them at all.

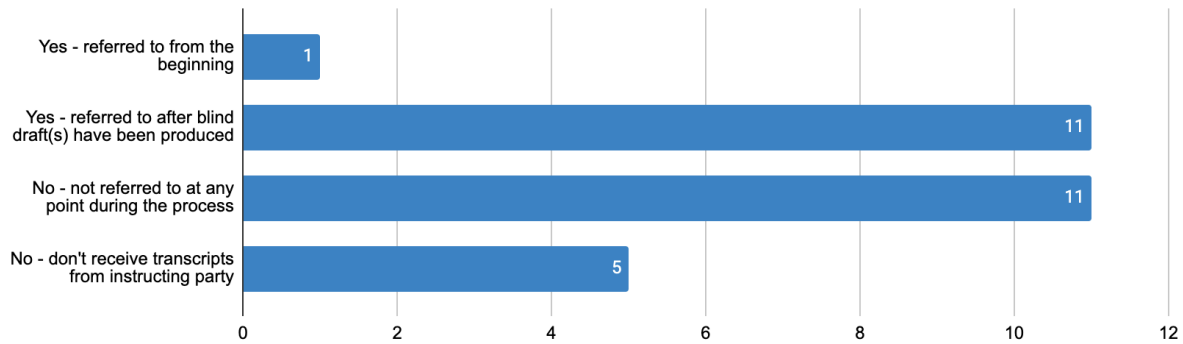


Figure 7: Responses concerning reference to transcripts produced by the instructing party.

Participants were also asked to comment on the quality of transcripts produced by the instructing party. Responses generally indicated that the quality of existing transcripts (from the instructing party, police, etc.) was variable. Some common issues that arose in participants' responses include:

- Incorrect speaker attribution
- No indication of confidence
- 'Cleaning up' of speech, e.g. repetitions and other disfluencies omitted
- Sections of speech summarised rather than verbatim transcription

5 Discussion

In this section, we first acknowledge some of the limitations of the present study. Then we highlight the main findings of the survey and relevant contributions from the workshop's discussion. Multiple future research topics and questions are also suggested.

5.1 Limitations

Before discussing the findings, we should acknowledge a few limitations of the present study. Firstly, there is a heavy skew in respondents towards practitioners working in Europe, in particular the United Kingdom and the Netherlands, with half of respondents working in these two jurisdictions. It is also likely that multiple responses have been collected from some organisations; however, participants were not required to state the name of their specific workplace due to anonymity requirements, and so it is not possible to verify this information.

Secondly, many of the survey questions, particularly in the section on transcription methods, ask participants to estimate numbers for certain aspects of their methods, such as the

number of drafts typically produced. We acknowledge that these numbers may vary depending on many factors, such as the requirements of the task or the quality, difficulty or type of recording. We therefore acknowledge that the responses, particularly those presented in Section 4.2, are estimated averages or typical approaches, and are not representative of every case worked on by practitioners.

5.2 Survey findings

The findings of this survey reveal that some areas of methods used by practitioners are similar across the majority of respondents; for example, most respondents reported that multiple transcribers are involved in the process and multiple drafts are produced. There are mixed responses with regard to the use of audio enhancement techniques and phonetic analysis, although the use of these procedures will depend on the quality of the recording (for example, whether audio enhancement could help with intelligibility and/or listenability) and the specific task (for example, disputed utterance analysis will require phonetic analysis but transcription of longer, clearer stretches of speech may not). Unsurprisingly, given recent findings on the ineffective performance of automatic systems on poor quality forensic-like audio recordings (Loakes, 2022; Harrington et al., 2022; Loakes, 2024), no respondents reported the use of automatic systems to aid transcription of speech within forensic audio recordings.

There seems to be general agreement on the factors that respondents believe to be the most influential when transcribing forensic audio recordings, with a clear consensus that audio quality, i.e. channel degradation and level of background noise, is the factor with the greatest influence on the perception and transcription of speech. This was followed by level of experience with the speaker's accent or dialect, and information from the instructing party, such as incriminating evidence linked to the speaker. These three factors in particular stood out from the rest, which received responses spread relatively evenly across the scale from 1 (no perceived influence) to 6 (great perceived influence).

There is widespread recognition of cognitive bias and priming by practitioners, with the majority of respondents reporting awareness of these phenomena and implementation of procedures to mitigate their effects. The findings in Rhodes (2016) demonstrate that around six years prior to our research, the majority of respondents consulted existing transcripts during their transcription process and 20% of respondents did so from the beginning of the process. In the present survey findings, only 1 respondent out of 28 reported consulting existing transcripts from the beginning of the process, and of the remaining respondents who

receive existing transcripts, there is an even divide within the group regarding whether these are consulted after blind drafts have been produced or not at all. This suggests a developing approach to bias issues that has taken place in the period between Rhodes' 2015-2016 survey and our 2021-2022 survey.

5.3 Areas of divergence

In this subsection, we will present two main issues and the surrounding discussions that took place during the survey and the workshop. The purpose of this section is to highlight areas of disagreement among experts concerning approaches, to explore common problems encountered by practitioners, and to make suggestions for further empirical research.

5.3.1 Contextual information

There seem to be mixed opinions within the forensic speech science community concerning the use of contextual information during the transcription process. One question in the survey asked "How is contextual information (e.g. case information, context of speech within recording, previous transcripts) managed?". Responses contained varying amounts of detail, but it is clear that some practitioners prefer to work with no information at all, while others introduce contextual information and existing transcripts after blind drafts have been produced.

Some survey respondents reported that they exclude all case information in order to avoid any priming that may influence the way in which they perceive the speech within a recording. This is an understandable approach, given that experts are often hired to provide an 'impartial' transcript and it could be argued, particularly by opposing lawyers, that knowledge of any context leads analysts to 'lose their impartiality'. However, it can also be argued that 'top-down' information, such as contextual clues, is required for speech perception and that forensic transcribers should not be expected to use a 'bottom-up' only approach in their work.

At the workshop, practitioners discussed an alternative view: that the goal of forensic transcription is to provide the end user with the "maximally useful" and "most objectively reliable" transcript. In order to achieve that, one practitioner argued that it is "absolutely crucial" to have context but that this must be done in a responsible way. This stance is also taken by the Research Hub for Language in Forensic Evidence at the University of Melbourne, whose recent findings (Fraser et al., 2023) confirm that it is "unrealistic to expect

individual transcribers with no contextual information to produce demonstrably reliable transcripts". The use of a 'linear sequential unmasking' process, as suggested by Dror and colleagues (Dror et al., 2015) and recommended by the UK Forensic Science Regulator's guidance on Cognitive Bias Effects Relevant to Forensic Science Examinations (FSR-G-217, 2020), seems to be a common approach among practitioners who do refer to contextual information during their transcription process. Practitioners at the workshop highlighted the need for transparency in this process, with clear documentation of the information that is introduced and at what stage, so that the user of the transcript can determine whether the content of the transcript was affected by the information. Details of how a UK speech and audio laboratory deal with information management can be found in an upcoming paper (Rhodes et al., in press).

Something important to consider is that there are different layers and levels of contextual information. Knowledge that the recording takes place in an aeroplane cockpit, for example, may help a transcriber with the recognition of domain-specific terminology. Other pieces of contextual information may not be relevant and could potentially bias the transcriber; the task then becomes differentiating between what is or is not relevant and subsequently introducing the contextual information that has been deemed as relevant in a managed and well-documented way.

One practitioner at the workshop then posed the question: what counts as relevant contextual information? There is currently no empirical research concerning this question, and it is up to individual practitioners to make those decisions. This highlights the need for more research on contextual information with regard to transcription:

- What sort of information is typically useful? Are analysts able to reliably transcribe more of the audio's speech content after the introduction of contextual information?
- What sort of information could have an adverse effect on the perception of speech in a forensic audio recording?
- When should the information be introduced? Gradually over multiple drafts or just before finalising the transcript?

Furthermore, as mentioned in section 5.2, there is disagreement among practitioners regarding the consultation of existing transcripts. Of the respondents who typically receive transcripts from the instructing party, roughly half opt to not consult the transcripts at any stage of their transcription process. The other half choose to refer to these transcripts once blind drafts have been produced; this may be midway through the drafting process or right before finalisation of the transcript to check, for example, place names.

The divergence in the approach to existing transcripts reveals the need for more empirical research on the topic of cognitive bias and priming:

- Is it necessary to avoid any reference to existing transcripts to remain 'impartial'?
- Can reference to existing transcripts be useful for accurate transcription of content such as place names and domain-specific terminology?
- At what point should analysts refer to existing transcripts, if at all?

5.3.2 Drafting methods

The survey findings demonstrate some differences in experts' practices with regard to the drafting methods typically implemented. Of the respondents who have multiple transcribers working on a transcript, nearly two thirds typically implement a parallel drafting method, whereby at least two transcribers produce independent transcript drafts and these are then 'merged' together. The remaining third of respondents reported that they typically opt for a sequential method, in which each new draft builds on a previous version of the transcript.

85% of those who typically choose a parallel method hold affiliations with a European government lab, while around 70% of the respondents who typically choose a sequential method hold affiliations with independent facilities. One practitioner at the workshop suggested that this could play a role in the choice of drafting methods, as there may be a 'marketplace' difference between public organisations and private companies; as parallel drafting may take longer, it therefore may be more difficult to secure funding for this approach in a competitive environment where each provider must provide a quote for their work. Self-evidently, some methods like parallel drafting or sequential drafting using multiple transcribers are not available to sole practitioners or those working alone in a larger organisation.

From discussions that took place at the workshop, two main approaches seem to be taken to transcript drafting in actual practice. The first approach would always involve parallel drafts, but this work would be targeted at shorter sections of a recording as it may be too expensive or time-consuming to parallel-draft longer recordings; in cases of longer recordings, they will ask the client to pinpoint these sections. However, some cases may involve documenting the content of longer stretches of relatively clear speech across multiple recordings. Some practitioners at the workshop argued that parallel drafting in these circumstances would duplicate the amount of effort, the majority of which would be wasted, while others argued that a parallel approach should still be taken to mitigate priming effects. In cases of

‘questioned utterances’ or short sections of poor quality, there seemed to be a consensus at the workshop that these should be addressed with parallel transcription, if possible.

The second approach involves sequential drafting, or a combination of sequential and parallel drafting depending on the type of recording. Discussions at the workshop around this method focussed on a risk-assessment approach which balanced the need to avoid priming in poorer sections with concerns about time- and cost-effectiveness. In this method, longer stretches of clearer speech could be transcribed by one analyst and checked sequentially by another, while areas of ambiguity or sections of poor audio quality could be transcribed in parallel by independent analysts and then discussed.

In both approaches, there would be some cross-checking and further development of either sequential or parallel drafts as they are finalised into the produced transcript.

This area of disagreement highlights the need for further empirical research on the topic of drafting methods and priming, which could address questions such as:

- Does one drafting approach produce reliably better transcripts than the other?
- In a sequential approach, are second transcribers able to reliably identify errors or potential alternatives within a transcript? To what extent are sequential drafters ‘primed’ by other expert transcripts?
- Does parallel drafting take a significantly longer amount of time than a sequential (or hybrid) drafting approach? Does this approach produce fewer ‘primed’ errors compared to a sequential approach?

5.4 Other areas of interest

In this section we will report some other issues that have been discussed as part of this study and suggest directions for future research. This is by no means an exhaustive list of issues, but we hope that it serves as an inspiration for future research projects about forensic transcription practices.

5.4.1 Number of transcribers

Many practitioners at the workshop perceived that their transcripts are better when multiple analysts have contributed to their production. Some practitioners who employ a parallel drafting method highlighted that having multiple transcribers is a useful way to locate sections of ambiguity within a recording, which can then be discussed. Some practitioners

who employ a sequential drafting method noted that having a second analyst check the transcript is a useful way to reduce the number of errors. Generally, then, there seems to be agreement that it is ideal to have multiple analysts working on a transcript.

This sentiment is supported by the findings of Tschäpe and Wagner (2012), which reported that transcripts produced by multiple transcribers were better than those produced by individual transcribers.

- Further empirical research on this topic could demonstrate whether these findings are replicable, i.e. are multiple experts working together better than experts working alone?

5.4.2 Levels of confidence

In the survey results, 93% of respondents reported that they represent multiple levels of confidence within their transcripts. This is most commonly done with the use of brackets, whereby speech within brackets is designated a lower level of confidence than speech outside of brackets. However, one respondent noted in their survey response that “a non-expert listener who reads the transcript while listening to the audio will be primed in the same way irrespective of whether some of the speech is designated as carrying full confidence and other speech being designated as carrying a lower confidence rating”. This sentiment was shared by an attendee at the workshop, who confirmed from experience that lay listeners do not look at the brackets when reading a transcript and that, in cases where alternative options for a particular word are given, readers will choose “whichever option suits their understanding of what the case is all about”. Furthermore, findings by Tompkinson et al. (2023) suggest that, even when presented with a transcript key, readers are not always able to correctly identify the meaning of certain conventions. Given that the end user should be at the forefront of considerations, this raises questions regarding whether the inclusion of multiple possibilities for a particular word (as reported in section 4.3.1) is counterproductive and furthermore, whether transcripts should contain content that does not have the transcriber’s full confidence.

When this issue was discussed at the workshop, additional questions were raised concerning the meaning of ‘level of confidence’. One practitioner asked how to decide which parts are designated a lower level of confidence; for example, if something was not heard during the production of the first draft, is the analyst obliged to mark this section as having a lower level of confidence? Another practitioner raised the issue that levels of confidence can vary within recordings as well as across recordings, and it can vary across practitioners as

well. They gave the example that non-bracketed speech in a transcript for a clearer recording could have a much higher level of confidence than non-bracketed speech for a recording of poor audio quality, despite both being represented as having the transcriber's full confidence.

Rather than a subjective judgement of confidence on the part of the transcriber, one practitioner suggested that this concept could be framed differently: as a measure of the extent to which the available phonetic evidence supports the inclusion of a particular word in a transcript. This sentiment was shared by others at the workshop; it was also suggested that the idea of 'confidence' could be replaced by 'certainty' or 'evaluation of degree of support for the interpretation, based on the clarity of the recording'.

- Further research on lay-people's perceptions of transcript content would be welcomed in order to better understand how these confidence-related features are perceived by end users. It could establish what information lay-people and legally trained people receive from transcripts, and what briefing information helps them to understand different levels of certainty or support.

5.4.3 Fatigue or the 'fresh ears' method

In the 'concluding remarks' section of the survey, one respondent commented that they employ a 'fresh ears' method, whereby they spread the transcription process over a few days. This sentiment was shared by others at the workshop; one practitioner reported their experience of returning to a transcript draft the next day and, having relistened to the audio, questioning how they heard the content that they had previously transcribed. Another practitioner highlighted the importance of working in manageable periods of time in order to avoid cognitive fatigue.

These suggested methods are based on experience rather than grounded in empirical findings; it would therefore be useful to carry out research on time-related features of transcription procedures, for example:

- How long should be spent on a transcript in a single session?
- What is the ideal time period between sessions?
- How many sessions are optimal for the production of the most reliable transcript?

5.4.4 Competence

Another comment that was made by a respondent in the ‘concluding remarks’ section of the survey was that it is necessary to have a lot of experience to become a good transcriber but “you never know when you are good in a certain case”. One of the issues with forensic recordings is that the ground truth of what was said is seldom known (Fraser, 2021), and therefore lots of experience does not necessarily mean that a transcriber is accurately and reliably transcribing the speech content of forensic audio recordings.

An approach to address this issue would be to carry out proficiency testing with recordings where the ground truth of what was said is known; experts would then receive useful feedback on transcription training exercises and be able to monitor their competence in producing accurate transcript drafts. This would also go towards providing empirical validation of the overall methods used by experts for transcribing forensic recordings.

In order to produce proficiency tests, forensically realistic audio materials where the ground truth is known should be obtained; this could be done via a dual-microphone set up whereby forensic-quality and high-quality recordings of an interaction are simultaneously collected, similar to the set-up used by Tschäpe and Wagner (2012). A ground truth transcript can be produced from the high-quality feed, to which transcripts of the forensic-quality material can be compared. Another consideration when producing proficiency tests is how to compare the transcripts; this could be done in a number of ways, such as comparisons carried out at a syllable-level (e.g. Tschäpe & Wagner, 2012), at a word-level (e.g. Harrington, 2023) or at a phrasal-level (e.g. Fraser et al., 2023).

Further research on how to carry out proficiency testing in terms of the materials used and the methods of analysis would be welcomed, including:

- How many tests each practitioner should carry out?
- What range of forensic situations should be represented?
- What threshold or score does a practitioner need to achieve to be declared competent? Is it enough to outperform lay-listeners who we can assume represent triers-of-fact? Should this be reported, and if so, how?

6 Conclusion

In this article we have presented the results of a survey about expert transcription practices. While we found a number of areas of convergence in expert methods, the results of the

survey also revealed some areas with less agreement among practitioners, particularly concerning drafting methods and the use of contextual information. We discussed these issues with practitioners at a workshop and have explored some of their comments in the discussion above. This paper has highlighted that there are many unanswered questions on the topic of forensic transcription, and that more empirical research should be carried out on the methods employed to produce transcripts in criminal and legal matters. We hope that this article, along with the suggested research questions in the discussion, encourages further critical exploration of forensic transcription practices.

7 References

- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.
- Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, 32(10), 1439–1465.
- Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: intelligibility and classification. *Language and Speech*, 51(Pt 3), 175–198.
- Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Letter to the Editor- Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making. *Journal of Forensic Sciences*, 60(4), 1111–1112.
- Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech Language and the Law*, 10(2), 203–226.
- Fraser, H. (2021). The development of legal procedures for using a transcript to assist the jury in understanding indistinct covert recordings used as evidence in Australian criminal trials: A history in three key cases. *Language and law*, 8(1).
<https://ojs.letras.up.pt/index.php/LLLD/article/view/10953>
- Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, 7.
<https://doi.org/10.3389/fcomm.2022.898410>
- 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.
- Fraser, H., Loakes, D., Knoch, U., & Harrington, L. (2023). *Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio*. Conference of the International Association for Forensic Phonetics and Acoustics, Zurich.
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language*

- and the Law*, 18(2). <https://doi.org/10.1558/ijssl.v18i2.261>
- French, P., Stevens, L., Jones, M., & Knight, R. A. (2013). Forensic speech science. Bloomsbury companion to phonetics, 183-197.
- Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, 18(2), 293–307.
- Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication*, 8. <https://doi.org/10.3389/fcomm.2023.1165233>
- Harrington, L., Love, R., & Wright, D. (2022, July). *Analysing the performance of automated transcription tools for covert audio recordings*. Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.
- Harrington, L., & Rhodes, R. (2023). *Forensic transcription: a survey of expert transcription practices in Europe and North America*. Conference of the International Association for Forensic Phonetics and Acoustics, Zurich.
- Harrison, P., & Wormald, J. (in press). Forensic transcription and questioned utterance analysis. In F. Nolan, T. Hudson, & K. McDougall (Eds.), *Oxford Handbook of Forensic Phonetics*. Oxford: OUP.
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, 22(4), 428–450.
- Innes, B. (2011). R v David Bain--a unique case in New Zealand legal and linguistic history. *International Journal of Speech, Language & the Law*, 18(1).
- ISO/IEC 17025:2017 - International Standard: General requirements for the competence of testing and calibration laboratories. [URL - <https://www.iso.org/ISO-IEC-17025-testing-and-calibration-laboratories.html>]
- Jenks, C. J. (2013). Working with transcripts: An abridged review of issues in transcription. *Language and Linguistics Compass*, 7(4), 251–261.
- Jones, T., Kalbfeld, J. R., Hancock, R., & Clark, R. (2019). Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2), e216-e252.
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7. <https://doi.org/10.3389/fcomm.2022.803452>
- Loakes, D. (2024). Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?. *Frontiers in Communication*, 9, 1281407. <https://doi.org/10.3389/fcomm.2024.1281407>
- MacLean, L. M., Meyer, M., & Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research*, 14(1), 113–123.

- Morrison G.S., Enzinger E., Zhang C., 2018. Forensic speech science. In Freckelton I., Selby H. (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters.
- Rhodes, R. (2016). *Cognitive bias in forensic speech science: a survey on risks and proposed safeguards*. Conference of the International Association for Forensic Phonetics and Acoustics, York.
- Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R.-A. (2014). Cross-accent intelligibility of speech in noise: long-term familiarity and short-term familiarisation. *Quarterly Journal of Experimental Psychology* , 67(3), 590–608.
- Tompkinson, J. & Haworth, K. (2023). The perception and interpretation of additional information in legally relevant transcripts. Conference of International Association for Forensic Phonetics and Acoustics, July (Zurich).
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test*. Conference of International Association for Forensic Phonetics and Acoustics, Santander, Spain.
- UK Forensic Science Regulator. (2020). FSR-G-217: Cognitive Bias Effects Relevant to Forensic Science Examinations (Issue 2) [URL: <https://www.gov.uk/government/publications/cognitive-bias-effects-relevant-to-forensic-science-examinations>]
- UK Forensic Science Regulator. (2023). Forensic science activities: Statutory Code of Practice (version 1 - March 2023) [URL: <https://www.gov.uk/government/publications/statutory-code-of-practice-for-forensic-science-activities>]
- Walker, A. G. (1990). Language at work in the law. In *Language in the Judicial Process* (pp. 203–244). Springer US.

Article 1 - Appendix A

Survey questions

1. Please indicate the country/jurisdiction in which you primarily work.
2. If you work in multiple countries/jurisdictions, please indicate any additional countries/jurisdictions in which you work.
3. Which language(s) do you most commonly work with?
4. Please specify your working arrangements or affiliations. Tick all that apply.
 - ☐ Government laboratory
 - ☐ University department
 - ☐ Individual private practitioner
 - ☐ Independent facility with other staff
 - ☐ Research institute
 - ☐ Other
5. How often do you carry out forensic transcription as part of your work duties?
 - ☐ Very frequently
 - ☐ Frequently
 - ☐ Occasionally
 - ☐ Rarely
6. How many people typically work on a transcript?
 - ☐ 1
 - ☐ 2
 - ☐ 3
 - ☐ 4+
7. How many drafts are typically made?
 - ☐ 1
 - ☐ 2
 - ☐ 3
 - ☐ 4
 - ☐ 5
 - ☐ 6
 - ☐ 7
 - ☐ 8+
8. If multiple drafts are produced, are these typically made in parallel or sequentially?
 - ☐ In parallel (i.e. analysts working independently on separate transcripts)
 - ☐ Sequentially (i.e. analysts building on previous versions of transcript)
 - ☐ Other

9. Roughly how long would a transcript typically take to produce for a 10-minute, slightly noisy recording?
10. What equipment is used by the person(s) carrying out the transcription? Please provide details of both the hardware and software used.
11. How is contextual information (e.g. case information, context of speech within recording, previous transcripts) managed?
12. How often is phonetic and/or acoustic analysis employed when producing a transcript?
 - Always
 - Very frequently
 - Occasionally
 - Rarely
 - Never
13. Are recordings for transcription usually enhanced beforehand?
 - Yes
 - No
 - Other
14. If so, which enhancement methods are used?
15. Do you use automatic speech recognition systems in the production of a transcript?
 - Yes
 - No
 - Other
16. If so, how are these systems used?
17. How are levels of confidence represented?
18. How many levels of confidence are represented?
 - 1
 - 2
 - 3
 - 4+
19. How is unintelligible speech represented?
20. How is overlapping speech represented?
21. How are non-speech sounds (e.g. coughs) represented?
22. How are disfluency phenomena (e.g. self-interruptions) represented?
23. To what extent do you believe the following factors could influence the perception and transcription of speech within forensic recordings? Please rate the factors on a scale of 1 to 6, where 1 represents no effect at all on the transcription and 6 represents a great effect on the transcription.

- Information from instructing party (e.g. incriminating evidence linked to speaker)
 - Expectations of instructing party (e.g. pressure to get a certain result)
 - Content of recordings (e.g. highly emotive or distressing speech)
 - Information about a suspect (e.g. age, profession, criminal record)
 - Information about the type of offence (e.g. terrorism, drug trafficking, theft)
 - Channel/quality degradation (e.g. level of background noise)
 - Your level of experience with the speaker's accent/dialect
24. Do you consider cognitive bias to play a significant role in transcription?
- Yes
 - No
25. Please explain your reasoning regarding the role of cognitive bias.
26. Do you have some form of protocol or any procedures in place to protect against the effects of cognitive bias?
- Yes
 - No
 - Other
27. If so, please provide details.
28. If you receive transcripts produced by the instructing party, do you refer to these during the transcription process?
- Yes - referred to from the beginning
 - Yes - referred to after blind draft(s) have been produced
 - No - not referred to at any point during the process
 - No - don't receive transcripts from instructing party
 - Other
29. How would you rate the quality of transcripts produced by the instructing party?
- Please give details of any common errors or concerns if applicable.
30. If you have any additional comments about your forensic transcription methodology which have not been addressed in this survey, please type them here.

5. Additional resource - Focus interview: non-expert transcripts in England and Wales

In the survey of international expert transcription practices carried out as part of this thesis, respondents were asked to comment on the quality of transcripts produced by non-experts. It is often the case that the instructing party, which may be the police, the prosecutor or the defence lawyer, provide transcripts that they have produced or that have been produced by the opposing side. The transcript may be for the purpose of providing forensic practitioners with a record of the speech content to aid their speaker comparison or audio authentication casework, or it may be that the speech content is of interest and they are (a) offering their own interpretation or (b) questioning an interpretation offered by the opposing side.

Responses to this particular survey question indicate a number of issues with police transcripts, such as:

- Wrongful speaker attribution, often guessed or based on knowledge of the case
- Wrongful representation of turn-taking and overlapping speech
- Objectively incorrect transcription of the speech content
- No indication of uncertainty, such that everything seems to be transcribed with confidence
- Speech often 'tidied up' and standardised, or summarised rather than verbatim
- Lack of timing details
- Disfluencies omitted

Overall, the most common response concerned the huge variability in the quality of these transcripts; at times a non-expert transcriber can do a relatively good job at accurately and clearly portraying the speech, while at other times the quality can be extremely poor.

In order to contextualise the research in this thesis concerning non-expert transcription (i.e. Article 2), a focus interview was conducted with two forensic practitioners, who have a combined total of over 30 years' experience with forensic speech science casework. The aim of the interview was to provide a more comprehensive description of issues that practitioners have encountered with transcripts of poor quality evidential recordings produced by non-expert transcribers, with a specific focus on the situation in England and Wales (the jurisdiction in which the doctoral research has taken place). The resource presented in this section was created for the purpose of providing a reference for a number of issues

discussed in other parts of the thesis; there is no analysis or further discussion of the interviewees' responses within this section.

The focus interview took place on Zoom on 19th January 2024. The interviewees were sent the following list of questions prior to the Zoom meeting to act as a starting point for the discussion:

- What are some common issues with the layout of non-expert transcripts?
- What are some common issues with the content of non-expert transcripts?
- Are keys provided to help the reader understand what certain symbols mean?
- Are time points provided within transcripts? Are these transcripts generally easy to follow alongside the audio recording?
- Are these transcripts accompanied by a methodological report?

The following section recounts the discussion that took place within the interview, which addressed a range of issues including, and additional to, those listed above. The discussion is not reported in a verbatim manner; instead, the interviewees' responses have been summarised by topic and 'cleaned up' to create a clear, readable text. Text enclosed within brackets provides clarification or contextual information inserted by the author. The content below has been approved by the interviewees, who will henceforth be referred to as Practitioner 1 and Practitioner 2.

Content of the interview

Practitioner 1 started the discussion by providing some context regarding transcripts produced by non-experts:



The police will prepare many more transcripts than experts because those transcripts will be used in both the investigation and in charging decisions. Experts tend to be brought in towards the end of the process, and are involved mostly in preparing evidence for a trial, so the function of non-expert and expert transcripts is often different. The decision to engage expert forensic practitioners is, however, less clear cut for the task of transcription compared with voice comparison casework; there is a clear legal boundary that police officers cannot really give voice evidence (e.g. speaker identification or voice comparison) because you must be an expert in the relevant field to do so. There seems to be an expectation that as a transcript becomes more difficult or more central to the case, that is when an expert is required, but it is definitely less concrete than, for example, voice comparison.

Practitioner 2 echoed some of the points made by Practitioner 1:

The initial usage or requirement for having a transcript in the first place is very different for non-experts compared with the expert context. It seems that many of the transcripts almost end up in court by accident, having a bigger role or status than was probably ever imagined by the transcriber. This is particularly true of the police interview context. The legal status of transcripts is different to voice comparison evidence, given that the transcript is just there as a guide and it is ultimately for the jury to decide what they hear. And so there is nothing to stop a transcript produced by, for example, an investigating officer, being entered into evidence.



Moving onto some more specific issues with non-expert transcripts, Practitioner 1 discussed a lack of references to time points:

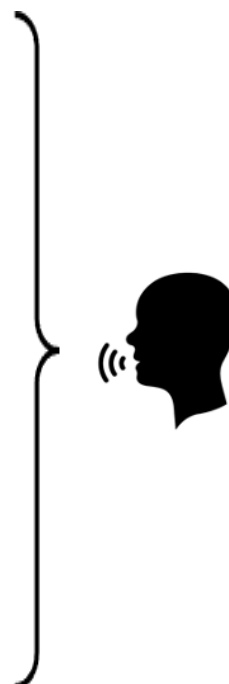


A very common issue would be not being able to locate utterances within the transcript in time. Some non-expert transcripts will contain only the speech content, with no indication of when in the recording the speech content takes place or who is saying it. Often with transcripts of very long recordings, e.g. those captured with covert recording devices planted in a location frequented by the suspect(s), experts will be presented with transcripts containing 30 pages of speech, with no way to link any of that speech with particular time points in the recording. Sometimes the transcripts may contain a reference to the time point in the recording once per page; the lack of timing indications makes the transcripts really hard to follow and it is not clear who is speaking when¹⁸. In over ten years of experience, I have only seen one or two transcripts that have a time point for every utterance (as is standard for practitioners). For the rest of the transcripts, there seems to be a fairly even divide between those which contain relatively regular indications of time and those which do not. Some sort of standardisation could help with this problem.

¹⁸ Practitioner 1 also acknowledged that a lack of clarity in the transcripts can also be applied to those produced by experts. These issues will not be discussed in depth here, but Practitioner 1 highlighted that they have observed an overcomplication of conventions or veneer of high technicality in expert transcripts, which can undermine the main purpose of the transcript: to help the jury follow the speech content of a poor quality evidential recording.

On the topic of time points within transcripts, Practitioner 2 added:

The biggest issue with the lack of time points within non-expert transcripts is trying to find things in the recording, and being certain that the part that you are reading in the transcript actually matches with the audio recording. When carrying out tasks such as voice comparison or audio enhancement, the transcript is often used as a kind of map of the content of the recording, but they often contain a lot of ambiguity. And the practitioner has to work out whether the ambiguity is because the content is wrong, or because a phrase was said multiple times but has only been transcribed once, or because the speech has been summarised rather than transcribed verbatim. Sometimes experts are working with days of covert recordings and trying to use the transcript as a guide can be really hard work. Often significant time is wasted attempting to locate sections that don't have adequate timings.



Practitioner 1 then discussed some issues with the layout of non-expert transcripts:

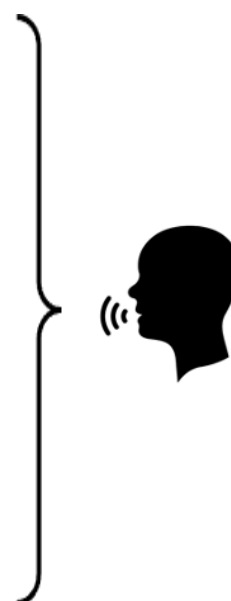


Sometimes transcripts are presented as Microsoft Word documents, and we have also seen transcripts created in Microsoft Excel (which tend to be more technical, e.g. they might include other information from visual surveillance or call records). They can be laid out in a really sensible way, with one line per speaker, but I have also seen transcripts where the speech content is presented in a block of text which switches between verbatim transcription and summarising. There is often no indication of when speech is being transcribed verbatim and when it is being summarised; for example, a transcript may say “conversation about football for 10 minutes” and then return to verbatim transcription with no indication of doing so, which is further compounded by a lack of time points, making it very difficult to find the relevant speech within the recording. The representation of overlapping speech can also be problematic within non-expert transcripts as this is very rarely done in a clear or systematic way.

On the topic of summarising sections of the speech content, which is common within non-expert transcripts, Practitioner 2 added:

Practitioners avoid summarising parts of the speech content as a matter of principle. It is not for practitioners to judge or interpret the evidential relevance of something because they don't know all the details. Something that might appear on the surface as a throwaway comment may turn out to be crucial.

Non-experts are often not actually producing verbatim transcripts, even when that seems to be the goal. There is inherently some level of summary happening, given that they often do not include all the full stops, hesitations, repetitions, interruptions, etc. Transcribers may completely miss out interjections, which sometimes doesn't matter but at other times it does.



Practitioner 1 recounted a recent experience in court with a police officer regarding the layout of transcripts:



A police officer who specialises in guns and slang usage was asked to prepare transcripts of evidential recordings given his specialist knowledge of some of the words being used. To some extent, he was independent from the investigation team, which is good. This police officer does a lot of transcription because he is brought in to consider the (code) words used and what they mean. We (a team of forensic practitioners) had also prepared transcripts of the recordings, and the officer was impressed by the layout, which is essentially just a table with columns for the time point, attributed speaker and speech content, and with background sounds represented within square brackets. He showed enthusiasm for having a standard format that you can fill in and a set way of doing things, which suggests that police officers are afforded no guidance and it is left to individual officers to decide what is best.

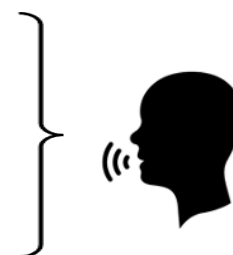
Practitioner 1 also highlighted some of the issues concerning speaker attribution within non-expert transcripts:



Sometimes non-expert transcribers attribute speech to named people even if their presence in the recorded interaction is questioned; or utterances will be attributed to, for example, “male” for a male speaker or “u/k” for an unknown speaker, but even within cases this isn’t always consistent. Just having a method, such as using tables and consistent terms like “M1”, “M2”, etc.¹⁹ to represent different speakers, would be a simple but effective way to improve transcripts.

On the topic of speaker attribution, Practitioner 2 added briefly:

There have been occasions where speech has been attributed correctly to the speakers at the beginning of a recording but at a certain point those speakers are flipped, and then the rest of the transcript contains utterances which have been attributed to the wrong person. That can have quite significant consequences.



¹⁹ This is how unidentified (male) speakers are represented in the transcripts produced by Practitioner 1.

Practitioner 1 discussed the variability in quality of non-expert transcripts:



One of the key problems is that the transcripts are all really different. Different officers are good and bad at doing it, and from the quality I would assume that some have obviously spent more (or less) time working on the transcripts. One issue that is quite common is that non-expert transcribers will try and transcribe everything, without having an appreciation that you can't always get everything that someone says. I have seen transcripts that are complete of very difficult recordings (i.e. there are no words omitted or marked as unintelligible), and when that happens it is most likely the case that a lot of it is wrong. Quite often, non-expert transcripts will have more content than those produced by experts, because they expect to be able to transcribe everything.

Practitioner 2 agreed with the variable quality of non-expert transcripts:

You have some people doing really good things and, at the other extreme, some people doing really bad things. It often stems from a lack of understanding. Transcription has not been given the level of oversight and scrutiny that other forensic practices have and most people are not aware of the psychological processes and acoustic effects that can impact what they hear. Maybe the problem is that these procedures are not standardised in any way²⁰.

There are also cases where transcripts have been submitted by non-experts for other purposes, e.g. audio enhancement or authentication cases, and the quality of these can be extremely poor. Sometimes they have missed out large portions of the recordings or the content is fundamentally wrong, yet to some extent the case is being based on this evidence. And it is by chance that the mistakes have been picked up on by experts. The issue is that this could be happening frequently but experts are not being asked to do this work, and therefore the problems are not being identified at an early stage (or at all). This highlights a common misunderstanding about how good enhancement can be. Very often a case will come in for audio enhancement, but it ends up becoming a transcription case because, contrary to their prior belief, enhancement cannot lead to the officers suddenly being able to hear and transcribe everything.



²⁰ Practitioner 1 then asked whether it is known that there aren't guidelines or standard procedures. Practitioner 2 stated that if these guidelines exist then they are not as good as they might be. Following this discussion, I searched the College of Policing website (<https://www.college.police.uk/>), the Health and Safety Executive Enforcement Guide (England & Wales) (<https://www.hse.gov.uk/enforce/enforcementguide/index.htm>), and the Crown Prosecution Service's website (<https://www.cps.gov.uk/>) for guidelines concerning the production of transcripts of evidential recordings, with no success.

Practitioner 1 highlighted that the variety spoken by the speaker can have an impact on transcripts:



Sometimes there is variation in what transcribers can get because they are clearly unfamiliar with the variety that is being spoken; there will often be clearly uttered words but because they are non-standard, they will be marked as 'inaudible'. Non-standard varieties, L2-accented speech and slang can all cause things to be missed or mistranscribed.

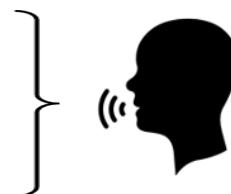
Practitioner 1 then discussed the lack of clear transcription conventions in non-expert transcripts:



There will often be multiple ways of doing the same thing and it isn't always clear what some of the conventions used actually mean. For example, does an ellipsis mean a pause or is it actually some speech that's been left out? Does "(inaudible)" mean you cannot hear anything or does it actually mean unintelligible, i.e. that you cannot understand any of what is being said? Consistency is key for that. There is very rarely a key of transcription conventions at the beginning of non-expert transcripts. Standardised guidelines could help to make these features more consistent across transcripts.

Practitioner 2 added that levels of confidence are rarely, if ever, included within transcripts produced by non-experts:

There are very often no levels of confidence; in fact, I cannot immediately recall any non-expert transcripts that used some kind of symbol, e.g. brackets (as is fairly standard in expert transcripts), to represent a level of uncertainty.



Practitioner 1 highlighted that little is known about the equipment or other methods employed by non-expert transcribers:

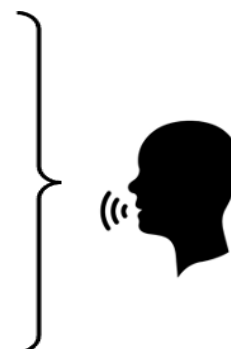


Very rarely will there be any indication of the equipment used by non-experts to produce the transcripts. Often, the only way for the inclusion of such information is when police officers need to justify their status as an 'ad hoc' expert, and they will produce a longer statement (than is usually included with a transcript) detailing that they used 'these headphones' and 'that software' and they listened to the audio for 'hundreds of hours'.

Another thing that we may be able to assume about non-expert transcription is that it is likely quite rare that more than one person will work on a transcript. The police officer specialising in slang usage (mentioned earlier) seemed almost confused when he learned that myself and colleagues had produced our transcript as a team. It is often the case that a transcript is one person's exhibit. Sometimes the transcript is contained within a sort of witness form and someone else has signed that they have 'checked it' at the bottom, although what that means is unclear.

On the topic of equipment and other methods employed by non-experts, Practitioner 2 added:

In the majority of cases, all that will be included is a witness statement saying that they are producing the transcript as their exhibit, or that they did it between these particular hours on these particular days. Maybe one or two transcribers over the years might have included the equipment they used, but this is not the norm. There certainly isn't a methodological report detailing the equipment and, e.g. the number of drafts made, as is produced by experts.



Moving on from non-expert transcripts and onto some general issues with the use of transcripts and audio recordings in court, Practitioner 1 discussed problems with audio playback in England and Wales:



The court speaker systems are terrible; often there will be two small speakers high in the corner of a large courtroom which is covered in wood with glass panelling and, in recent years, covid screens. Without fail, the audio sounds ten times worse than it actually is. There have also been times when they can't get the audio to play and so the barrister will play the audio recording on their laptop and hold it next to the court microphone. Despite instructions to listen carefully to the audio recordings, the jury isn't really hearing the audio because of the poor quality of the audio playback procedures, and this undermines the fact that the audio itself is the main evidence and the transcript is just an aid.

There has been one case where the audio was central and everyone in the courtroom had their own pair of headphones, but this is very rare. A police officer has even contacted the forensic lab a few weeks before a trial to ask what equipment they should buy for the jury to listen to the audio; why does the court not have a service for this sort of thing? Good audio equipment would be an easy win for the court system.

On the topic of audio playback, Practitioner 2 added:

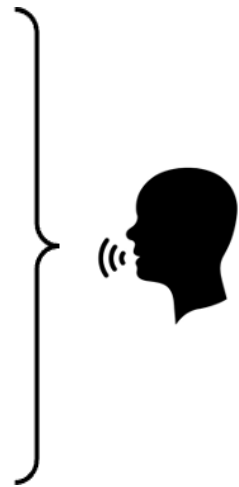
It stems from a lack of understanding on lots of levels about speech and audio and everything surrounding it. Historically, speech evidence has maybe been a less common kind of evidence, i.e. it hasn't been included in every case, so people haven't thought that there was a need to consider how audio is presented and therefore haven't invested the money. There's also a lack of money within the court system so it is not possible to furnish every courtroom with high quality audio playback equipment. Then it comes down to the fact that the Crown Prosecution Service has to pay for equipment hire for the length of a trial, which is expensive, but we would always advise them to hire headphones as a bare minimum. It's often an issue of oversight too; even in extremely modern courtrooms with brand new facilities, audio playback can be completely overlooked and lawyers can end up playing audio from their laptop speakers. Ultimately it is about putting money in the right place and in the hands of people who actually know what they're doing.

The jury is not given the best chance when it comes to audio evidence. Not only are they disadvantaged by the poor quality of the audio being played through small speakers located far away from their bench, but the listening experience is out of their control. Someone else is deciding how many times they can listen, and whether they hear it all in one block or in shorter sections. That is completely different to how the experts or non-experts transcribing the audio listen to it. Furthermore, the jury often won't have any means of seeing the time of the audio playback, so even if there are time points within the transcript, it can be very challenging to follow and very difficult to get back on track if they lose their focus or encounter slang or language use that they are not familiar with.



As a result of poor audio playback procedures, Practitioner 2 highlighted that juries tend to more heavily rely on transcripts:

The jury doesn't really get the opportunity to think about listening to the audio independently of the transcript or scrutinising the transcript's contents. The audio is played once or twice and in order to be able to immediately follow the speech content, it is necessary to read the transcript; and then the transcript becomes what they hear. It is not clear to what extent the jury is told that the transcript is simply a guide and that what they hear is the important thing, but it is unreasonable to expect them to be able to critically listen and review a transcript alongside a poor quality audio recording that they listen to a few times at most.




University of York
York Graduate Research School
Research Degree Thesis Statement of Authorship

Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

Candidate name	Lauren Harrington
Department	Language and Linguistic Science
Thesis title	Towards improving transcripts of audio recordings in the criminal justice system

Title of the work (paper/chapter)	A forensic approach to transcription errors: factors affecting human performance	
Publication status	Published	
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	x
Citation details (if applicable)		

Description of the candidate's contribution to the work*	Conceptualisation Methodology Formal analysis Investigation Writing - original draft
Approximate percentage contribution of the candidate to the work	100%
Signature of the candidate	
Date (DD/MM/YY)	29/02/2024

*The description of the candidate and co-authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates

and co-authors may find it helpful to consider the [CRediT \(Contributor Roles Taxonomy\)](#) approach to recognising individual author contributions.

6. Article 2 - A forensic approach to transcription errors: factors affecting human performance

Lauren Harrington

Department of Language & Linguistic Science, University of York, UK

Abstract

Transcripts are used for a number of purposes within the criminal justice system in England and Wales, one of which is to aid the jury in understanding the speech within poor quality evidential recordings. In these recordings, the speech content often suffers from poor intelligibility such that it is extremely difficult to decipher what is being said, hence the need for a transcript. This study explores the way in which two common factors in evidential recordings - the level of background noise and the accent background of the transcriber - can affect the content of transcripts. Participants transcribed speech in a Standard Southern British English (SSBE) accent at three different signal-to-noise ratios (SNR), and results showed no significant differences in performance between speakers of SSBE and speakers of other varieties of British English, but significant differences across SNRs; as the level of background noise increased, the likelihood of achieving a correct transcription significantly decreased and the distribution of errors changed such that deletion errors made up a significantly larger proportion of the overall error count. This study presents a novel method for the analysis of transcription errors, and serves as a basis for further development of a framework for forensic transcript analysis.

1 Introduction

When speech evidence is presented in courts of law in England and Wales, orthographic transcripts are often provided alongside the evidence for a number of reasons; most simply, a written record is much easier to refer to than an audio recording and serves as a reminder of the speech content after the presentation of the evidence. Most importantly, though, these recordings often suffer from poor intelligibility given the nature of data collection (e.g. covert recording devices planted in a location, simultaneous interactions during a phone call to the emergency services, etc.), and a transcript is therefore necessary to assist the court in hearing the speech content and following the discourse. The audio recording remains the evidence while the transcript is viewed as an aid to the user.

Such transcripts may be produced by experts in forensic speech science, who implement systematic and considered procedures when producing transcripts of forensic materials and are aware of psychological phenomena such as priming and cognitive bias that may influence their perception of the speech content (Article 1 of thesis). However, in many cases, these transcripts will be produced by non-experts, often police detectives involved in the case or other individuals employed by the police (such as police interview transcribers; Tompkinson et al., 2022). French & Fraser (2018) argue that this approach is less than ideal; police have access to contextual information which can substantially affect the way in which speech is perceived, leading to expectations about the speech content that are far from neutral (French & Fraser, 2018).

Most experts in forensic speech science would agree that transcripts of poor quality recordings to be used in court are most appropriately made by expert practitioners who are independent from the investigation (French & Fraser, 2018). However, most transcripts presented alongside evidential recordings will inevitably be produced by non-experts given, firstly, misconceptions about the challenges of transcription; many lay people assume that the task is as simple as putting some headphones on and typing out what the speakers are saying, with no consideration of how their perception may be affected by their external knowledge of the situation, or personal biases and expectations. There are also practical challenges as a result of a lack of funding within English and Welsh police forces, as well as a lack of access to expert practitioners. It is therefore necessary to investigate transcription carried out by lay people and the factors that can affect the transcripts that they produce.

Current work on the transcription of poor quality evidential audio recordings tends to focus on transcription procedures and guidelines (e.g. work carried out by the Research Hub for Language in Forensic Evidence at the University of Melbourne), with a particular focus on transcription carried out by experts or highly proficient transcribers. There is very little work that has considered transcription *performance* within the forensic domain, and the research that has been conducted has been small-scale proficiency testing comparing experts and non-experts (Tschäpe & Wagner, 2012).

Transcription can be affected by many factors related to (a) the speaker in the recording, (b) the audio quality and (c) the transcriber (Fraser, 2022). In the present study, the focus lies on an audio factor - the level of background noise - and a transcriber factor - the accent background of the transcriber (or, alternatively put, the transcriber's familiarity with the speaker's accent). Both factors are often involved in the transcription of evidential audio recordings, which frequently contain sections of indistinct audio and are transcribed by

people who do not share an accent with the speaker in the recording. These factors have both been shown to have a significant impact on transcription; the accuracy of transcripts tends to decrease with increased background noise (e.g. Clopper & Bradlow, 2008) and a lack of familiarity with an accent (Clopper & Bradlow, 2008; Smith et al., 2014).

The present study aims to investigate how these factors - level of background noise and listener's accent background - affect transcription performance when transcribing a 'standard' variety. Though similar studies have been done in the field of psycholinguistics, this paper takes a forensic approach with a much larger focus on the types of errors that transcribers produce. A standard measurement for performance in previous studies is the percentage of words correctly transcribed, and while this measure does reveal some information about general patterns in performance, it does not address issues produced by the words that were not correctly transcribed. In forensic contexts, the types and magnitude of the errors are of particular interest, given that the mistranscription of a single word could entirely change the meaning of an utterance. The novelty of this study lies in the large-scale evaluation of transcription errors and their implications in forensic transcripts.

2 Background

2.1 Transcription of noisy audio

In order to understand why certain factors may have an effect on transcription performance, it is necessary to consider how speech perception works. There is a common lay misconception that understanding what a speaker says relies purely on the sounds which they utter (Fraser, 2003). While 'bottom-up' processing, which relies on auditory information from the signal, does play a substantial role in speech perception, 'top-down' processing is also a key component and one which non-linguists are often unaware of. Top-down information, such as the listener's knowledge and expectations about "the language the speaker is using and the situation in which they are speaking" (Fraser, 2003, p. 204) can help the listener to predict what is likely to be said.

There are many types of forensic recordings that may be relevant in an investigation or for evidential purposes, such as covert or undercover officer recordings, CCTV recordings, audio or video recordings uploaded to social media, and telephone calls to the emergency services. Factors such as background noise, overlapping speech, (varying) distance from the microphone and poor transmission quality frequently feature in these types of audio materials given the ways in which the data is collected. Furthermore, the content is usually

highly contextualised and therefore difficult to understand without knowledge of the context, and combined with the poor technical quality discussed above, this can lead to speech content which is almost unintelligible to general listeners (Fraser, 2022).

Degradation of the acoustic signal leads to a substantial reduction in the bottom-up information that a listener can use to decode and understand the speech. This, in turn, leads to a greater reliance on top-down information, such that the listener's expectations of what the speaker is likely to have said play a stronger role (Assmann & Summerfield, 2004). A transcriber is abstracted from the reality of the situation; they are not present during the interaction which means that they cannot use facial expressions or gestures to help guide their understanding of the speech, and they do not share the contextual knowledge of the speakers. This means that the amount of top-down information is reduced substantially when transcribing audio, yet perception must rely more heavily on such information, i.e. the listener's knowledge, expectations and assumptions about the circumstances in which the recording was made, in poor listening conditions (Fraser, 2020).

Background noise is well-documented as creating a challenge for transcribers, with much lower word recognition accuracy for poorer quality audio recordings (e.g. Clopper & Bradlow, 2008; Lange et al., 2011). Most studies that have explored transcription performance of audio with different levels of noise have had a non-forensic focus and have therefore typically looked solely at overall accuracy and how higher levels of background noise can affect this measure. However, no work on the transcription of poor quality audio has yet explored the types of errors produced as a function of different levels of background noise. In the present study, interest lies in the errors that transcribers produce; two transcribers may both achieve a word recognition accuracy rate of 90%, where one has omitted the remaining 10% of words and the other has mistranscribed them, leading to substantial changes in meaning. In forensic contexts, the differences in these transcripts can be critical.

2.2 Familiarity effects

'Familiarity' with an accent has been shown to affect performance in a range of lexical processing tasks (Sumner & Samuel, 2009; Floccia et al., 2006; Adank & McQueen, 2007). A significant drop in performance tends to be observed for 'unfamiliar' accents while accents which are judged to be familiar, such as the speaker's home accent and their country's standard variety, both generally elicit a higher and similar level of performance. For sentence processing and transcription tasks, this effect is particularly prevalent in poorer listening conditions (Adank et al., 2009; Smith et al., 2014).

In a sentence verification task carried out by Adank et al. (2009), a group of speakers from Greater London and a group of speakers from Glasgow were presented with sentences spoken in Standard Southern British English (SSBE) or Glaswegian English, and asked to indicate whether the sentence was true or false. When the sentences were presented in quiet audio (i.e. with no background noise), no significant difference was found in performance across the accents within each listener group. However, when the audio had been mixed with speech-shaped noise at signal-to-noise ratios of +3 dB and 0 dB, speakers of Standard Southern British English showed a significant drop in performance when listening to Glaswegian English as compared with their own accent. Speakers of Glaswegian English showed no such decline in performance at these noise levels, performing equally for both accents, and both groups showed no significant differences in performance across the accents when the signal-to-noise ratio was -3 dB, i.e. adverse listening conditions in which the background noise was louder than the speech itself.

Smith et al. (2014) carried out a similar study with speakers of SSBE and speakers of Glaswegian English where participants were asked to orthographically transcribe a series of unpredictable and ambiguous sentences. The sentences were uttered in either SSBE or Glaswegian English and, continuing the theme of speech perception in noise, were mixed with randomly varying cafeteria noise at an average signal-to-noise ratio of +2 dB. Similar to the effects observed by Adank et al. (2009), a drop in performance was observed for the Standard Southern British English group when transcribing Glaswegian English as compared with transcribing their own accent, while the Glaswegian English group showed no difference in performance across the two speaker accents. These results, along with those presented by Adank et al., demonstrate that long-term familiarity with an accent such as the “media-standard” (i.e. SSBE) can lead to very similar performance across both the home and standard accents for a speaker of a non-standard regional variety; and no familiarity with an accent (as assumed for the SSBE listeners with Glaswegian English) leads to worse performance than with a familiar accent.

Smith et al. (2014) highlight that the analysis within the study carried out by Adank et al. (2009) focused solely on within-group behaviour, such that the performance for each accent was not compared across the two listener groups. Results of the study carried out by Smith et al. showed that words were more accurately identified by listeners who shared an accent with the speaker in the recording, i.e. SSBE listeners responded less accurately to Glaswegian speech than Glaswegian English listeners did, and Glaswegian English listeners responded less accurately to SSBE speech than SSBE listeners did. Closer inspection of the

results presented by Adank et al. reveals a similar pattern wherein the Glaswegian English listeners do not perform as well for SSBE as the SSBE listeners did. These results demonstrate that long-term familiarity with an accent other than one's own does not lead to native-like perception as straightforwardly as had been previously assumed. Smith et al. conclude by stating that long-term familiarity with an accent other than one's own can *mitigate* the associated processing costs but cannot *compensate* for them entirely.

A similar study on dialects of American English by Clopper and Bradlow (2008) found that the standard national variety (General American) was more intelligible in noise than non-standard varieties for listeners from all regional dialect backgrounds. Listeners were divided into three groups according to their dialect background: speakers of General American, speakers of Northern American English and 'Mobile' listeners, categorised as those who had lived in at least two different dialect communities before the age of 18. Participants transcribed short sentences that had been mixed with speech-shaped noise at SNRs of +2 dB, -2 dB and -6 dB and were spoken in four North American dialects: General American, Northern, Southern and Mid-Atlantic. Performance was found to be worst at the hardest SNR of -6 dB and best at the easiest SNR of +2 dB, though at all SNRs, General American was found to be the most intelligible variety and Mid-Atlantic (judged to be unfamiliar to all listeners except some of those in the 'Mobile' group) was the least intelligible for all listeners. Listener dialect was not found to have a significant effect on performance, and Northern listeners actually performed better for General American than their own accent. These findings suggest transcribers from a range of accent backgrounds will perform very similarly when transcribing the standard variety, even at moderate SNRs around 0 dB, with no significant advantage for native speakers of the standard variety. It should be noted that this particular study investigated transcription of different accents and by a different population to those in Smith et al. (2014) and the present study, i.e. American listeners rather than British listeners and with General American as the standard variety rather than SSBE.

An issue with the studies discussed above is the highly-controlled and unrealistic speech content of the stimuli. The stimuli for transcription in Smith et al. (2014) are pairs of sentences which are phonemically identical in SSBE but differ in a critical (usually grammatical) portion, e.g. "but Ralph surpasses" and "but Ralph's are passes"; these can only be disambiguated by their preceding context, "I agree that Simon excels" for the former and "Geoff's grades are all distinctions" for the latter, though this was not presented alongside the stimuli in the transcription task. The stimuli in Clopper and Bradlow (2008) are taken from the Speech-In-Noise (SPIN) test (Kalikow et al., 1977) and comprise short

sentences containing 5-8 words and ending with a highly predictable word. These stimuli are contrived for the purposes of a controlled experiment and their content, understandably, does not resemble the type of speech content encountered in a forensic recording, particularly in terms of length; contextual clues from surrounding utterances are very helpful in disambiguating words or phrases but this is lacking when the stimuli for transcription contain so few words.

This poses the question of whether a significant difference between native speakers of SSBE and speakers of non-standard regional varieties of British English would be observed when transcribing forensically-realistic recordings in terms of speech style, linguistic content and audio quality. The results of Smith et al. (2014) suggest that SSBE speakers would have an advantage, while the results of Clopper and Bradlow (2008) suggest that all listeners would perform at a similar level for the standard variety.

2.3 Performance evaluation

Note: this subsection contains some repeated content from Section 2.6 (in the ‘Research background’ section) of this thesis, concerning different existing transcription accuracy metrics. The final paragraph of this subsection details the plan for the present study.

There are many ways in which transcription performance can be measured and these often vary according to the specific application of the analysis. Previous work on the effects of ‘familiarity’ tends to measure performance by calculating the number or percentage of words correctly transcribed (e.g. Burda et al., 2006; Derwing & Munro, 1997; Jones et al., 2019; Smith et al., 2014), or more specifically *content words* (Clopper & Bradlow, 2008) or selected *keywords* (Walker, 2018) correctly transcribed. Linguistic research tends to focus on success rates, i.e. words correctly transcribed, rather than error rates. Conversely, the industry standard used for measuring transcription accuracy in automatic transcription systems is word error rate (WER). This measure is calculated by counting the number of errors (insertions, deletions and substitutions) within a transcript and dividing by the number of words uttered. This metric is usually presented as a percentage, where 0% demonstrates a perfect match between the content of the transcript and the speech uttered, and it can surpass 100% in scenarios where there are more errors in the transcript than words contained within the speech content (i.e. the reference transcript).

From a forensic perspective, it is much more important to consider transcription errors than the percentage of words transcribed correctly. However, quantified overall error rates employed in automatic transcription assessment, such as word error rate, can be distracting in forensic contexts and obscure details about performance. For example, two systems could achieve the same WER where one has deleted most of an utterance but the other has substituted key words, significantly changing the meaning. Consider the example in Table 1, where two possible transcriptions of the utterance “he was having the dragon themed curry” are presented. Both transcripts contain five errors and achieve a WER of 71%, despite huge differences in their length, meaning and level of incrimination; the substitutions in Transcript 1 change a completely innocent utterance into an incriminating one, and if combined with the court’s expectations about the crime (e.g. if the speaker was suspected of drug-related crimes) and an extremely poor quality audio recording, such mistranscriptions could be accepted by the court and the content of the transcript could falsely act as incriminating evidence against the speaker. The deletions in Transcript 2 lead to a transcript that is noticeably incomplete but the central theme of the utterance is retained (i.e. it is about a curry) and, crucially, these errors do not create a falsely incriminating interpretation of the speech content. In a forensic context, therefore, it is crucial to consider not only the number of errors but also the type and magnitude of those errors.

Reference	he	was	having	the	dragon		themed	curry
Transcript 1	he	was	hiding		drugs	in	the	curry
Transcript 2				the				curry

Table 1: Comparison of two transcripts with a reference transcript, with errors highlighted. A shaded red cell shows a deletion, bold red text shows a substitution, and bold blue text shows an insertion.

There is very little work concerning transcription performance within the field of forensic speech science. One attempt at analysing the accuracy of forensic transcripts was carried out by researchers (Tschäpe & Wagner, 2012) from the Department of Speaker Identification and Audio Analysis at the German Federal Criminal Police Office (Bundeskriminalamt; BKA). A small-scale proficiency test was devised to investigate the validity of their transcription methods and to compare the performance of multiple experts, individual experts and non-linguists. The test material used in this experiment consisted of a spontaneous conversation between two German adult males during a car journey that had been recorded on an undercover microphone located between the car’s speakers and transmitted via GSM (Global System for Mobile Communications, i.e. by mobile telephone). This style of (covert)

recording replicates a common type of sample used by forensic phonetic practitioners in casework. In order to create a reference transcript against which the participant responses could be compared, a high-quality reference recording of the conversation was made simultaneously with microphones attached to each speaker's clothing.

Participants in the proficiency test were 12 German-speaking forensic phonetic experts (including one group of 2 and one group of 3) and 8 non-phonetician BKA employees. The test involved transcribing 211 seconds of speech which was then compared with the reference transcript on a syllabic level and marked for substitutions, insertions or omissions (deletions). When presenting the experts' results, Tschäpe and Wagner offer three measurements: the percentage of correct identifications, the percentage of omissions and the percentage of substitutions and insertions combined. The authors offer an assumption that within forensic casework, reduced information causes less damage than false information. This view coincides with the work on the 'priming' power of transcripts; once an interpretation has been put forth, it can be almost impossible to dismiss (e.g. Fraser et al., 2011; Fraser & Kinoshita, 2021). Taking this assumption into account, Tschäpe and Wagner offer a different measurement when comparing the performance of experts and non-linguists, which assigns two error points to each substitution or insertion and only one error point to each deletion. The average number of error points within each of the experts' transcripts and the non-experts' transcripts is then compared.

The approach taken by Tschäpe and Wagner (2012), where the focus lies on the errors within the transcripts, is a much more appropriate way to analyse transcripts from a forensic perspective. The differentiation made between the types of errors where information is added (insertions and substitutions) and those where information is omitted (deletions) is a crucial part of interpreting the quality of transcripts. However, one of the limitations of the above study is the small number of participants. The plan in the present study is to more systematically assess transcription errors with a larger dataset, essentially combining the 'experimental' approach of the psycholinguistic studies discussed in section 2.2 with the more 'content-specific' approach taken by Tschäpe and Wagner (2012). The present study also carries out a more detailed analysis of the errors that are occurring, particularly in the case of substitutions. In many cases, substitutions may be entirely inconsequential in terms of their effect on the meaning and interpretation of the utterance, e.g. "try *and* help" versus "try *to* help". In other cases, though, substitutions could substantially change the meaning of an utterance from completely innocent to incriminating (see Table 1).

3 Aims

The main aim of the present research is to assess transcription performance across accent groups and audio qualities, with a particular focus on the types and magnitudes of errors within the transcripts. Such factors are often relevant within the task of forensic transcription, given the poor quality of evidential audio materials and the range of speaker accents. This study offers a novel approach to analysing transcription performance in forensic contexts, with consideration of different types of errors, different types of substitution errors, and their potential impact on a transcript.

The specific research questions are:

1. What impact does increased background noise (decreased signal-to-noise ratio) have on the transcripts produced?
2. Do native speakers of Standard Southern British English have an advantage over speakers of non-standard regional varieties of British English in terms of the quality of transcripts produced?
3. What types of errors are often produced, and what are the implications of these errors?

4 Methods

4.1 Participants

A total of 174 responses were collected from participants recruited online using Prolific. However, 21 participants did not complete the experiment and 18 participants were removed from the analysis due to answers to the phonology-based questions in the accent background survey that did not align with the chosen constraints (see below). There were therefore a total of 135 participants whose responses were analysed: 70 were speakers of SSBE and 65 were speakers of other regional accents of British English. All participants reported no issues with hearing. Participants were divided into two listener groups: those who speak Standard Southern British English and those who speak a non-standard regional variety of British English. The experiment was therefore targeted at (a) participants who had been born and raised in the South of England and (b) participants from the North of England, Scotland, Wales or Northern Ireland. These locations were targeted in order to recruit participants with an accent that is phonologically divergent from SSBE.

To verify the accent background of participants, a survey was included after the transcription task requesting details such as where they were born and raised as well as a self-identification of their accent (Article 2 - Appendix A). Some phonology-based questions regarding diagnostic vocalic differences in British accents were also included; for example, participants were asked to indicate via the use of short audio clips (recorded by the author) whether they pronounce “bath” as [bɑ:θ] (typical SSBE pronunciation with a long back vowel) or [baθ] (typical Northern English pronunciation with a short front vowel). In order for participants to be included in the SSBE group, they had to indicate typical southern pronunciations of “bath” (i.e. with a long back vowel) and “strut” (i.e. with an unrounded low vowel). Their responses to the accent identification and location-based responses were also considered. Any participants in the non-SSBE group who indicated typical southern pronunciations for both BATH and TRAP were excluded to ensure that those in the non-SSBE group speak with an accent that is substantially divergent from SSBE.

4.2 Materials

Stimuli were extracted from the Dynamic Variability in Speech (DyViS; Nolan et al., 2009) database which contains forensically-relevant recordings of 100 adult male speakers of SSBE. Short utterances were extracted from the ‘mock police interview’ task (Task 1), in which participants were interviewed by a ‘police officer’ (played by a Research Assistant) about a mock crime. The speech was semi-spontaneous since participants had visual prompts containing information about the events in question, including certain incriminating facts to be denied or avoided. A total of 24 utterances were collected from four speakers (six utterances per speaker; Article 2 - Appendix B), ranging between 15-22 words and 3-6 seconds in length. The mock police interview task in DyViS is an adapted map task, therefore much of the content involves proper nouns such as place names or surnames. To ensure that transcription difficulties did not occur as a result of infrequent lexical content, the extracted speech contained no proper nouns or names, with the exception of one occurrence of the popular telecommunication application “Skype”.

Utterances were extracted using Praat (version 6.1.30, Boersma & Weenink, 2020) and trimmed such that the beginning and end of the file aligned with the onset and offset of speech. 500 milliseconds of silence was added to the beginning and end of each file, and dynamic compression was applied to the recordings to reduce the difference in amplitude between the loudest and quietest sections of each utterance. In order to manipulate audio quality, the files were mixed with speech-shaped noise derived from the stimuli and the signal-to-noise ratio (SNR) was manipulated to create three listening conditions representing

fair, moderate and poor intelligibility. This was achieved by adjusting the speech levels and keeping the noise level constant across samples. The SNRs used for this experiment were loosely based on Adank et al. (2009): studio quality, +3 dB SNR, 0 dB SNR, and -3 dB SNR. Initial piloting revealed a similar (and relatively poor) level of performance between SNRs of +3 dB and 0 dB, but a relatively high intelligibility for +6 dB SNR; +6 dB SNR was therefore chosen as the baseline for the present study (i.e. “fair” audio quality), 0 dB SNR represented “moderate” audio quality and -3 dB SNR represented “poor” audio quality.

Since each utterance was mixed with noise at three SNRs, a total of 72 stimuli were produced. To ensure that each version of each utterance was transcribed roughly an equal number of times across the experiment, the stimuli were divided into three sets of recordings; each set contained the 24 unique utterances as well as a mixture of the three chosen SNRs (eight stimuli at each SNR). Participants within each listener group were assigned to each condition in roughly similar numbers.

4.3 Procedure

The experiment was carried out using Qualtrics and the survey contained four parts: a consent page, demographic questions, a transcription task, and an accent background survey. For the transcription task, participants were instructed to transcribe every word that they could hear and to try to make their transcription as accurate as possible. Participants were also instructed to include hesitation markers such as “er” and “erm”, and to not ‘clean up’ the text (e.g. correcting grammar or replacing slang/contractions such as “dunno” with “I don’t know”).

The transcription task contained 24 questions each containing an audio file, a text box for transcription, and a scale from 1 to 5 for participants to rate how challenging they found the recording to transcribe. Participants were assigned to a set of the recordings such that they heard the 24 unique utterances and a mixture of SNRs throughout. Stimuli were presented in a random order to ensure that familiarity with a speaker or a noise level did not occur. Participants were allowed to listen to the audio files as many times as they wished.

4.4 Transcript alignment

Reference transcripts were produced by the author for each utterance by transcribing the studio quality version of each file. Participants’ transcripts were aligned with the corresponding reference transcript on a word-level basis, using bespoke software developed

specifically for this project. The software uses a JavaScript implementation of the Hunt-McIlroy algorithm (Hunt & McIlroy, 1976) to align the reference transcript (Truth) with the participant transcript (Hypothesis) such that any word pairings are automatically aligned. The software automatically highlights any non-matches, i.e. any errors, and includes tools for the user to manipulate the alignment (e.g. insert rows above or below, move cells up or down) so that the manual alignment of substitutions can take into account phonetic similarities between the words. An example of transcript alignment using this software is shown in Figure 1.

Truth	Hypothesis
uh	uh
did	did
	i
have	have
a	a
sack	sack
of	of
potatoes	debt
	is
in	in
front	front
could	could
have	have
been	been
that	not
but	boats
um	

Figure 1: Alignment of reference transcript (Truth) and participant transcript (Hypothesis) using a custom-built tool. Cells highlighted in yellow indicate a non-match, i.e. an error.

4.5 Transcript marking

Once the reference and participant transcripts had been aligned, word pairs were either marked as a match or as an error. Errors were classified as either deletions (whereby a word appears in the reference transcript but not the participant transcript), insertions (whereby an additional word features in the participant transcript), or substitutions (whereby the word pair was not a match). Examples of each error type can be observed in Figure 1: insertion of “I”, substitution of “but” with “boats”, and deletion of “um”.

The software looks for exact matches, therefore some ‘errors’ were actually deemed by the author to be correct transcriptions of the content, even if not a perfect match to the reference

transcript (see Article 2 - Appendix C for an outline of corrections made). For example, spelling errors were corrected and compound words were separated (or vice versa) to match the reference transcript, e.g. “steakhouse” was corrected to “steak house” so that two matches were registered rather than a substitution and a deletion. If a filled pause was correctly recognised but represented in a different way within the participant transcript, it was adjusted to match the reference transcript, e.g. “er” and “erm” were considered a match. Similarly, contractions were expanded (or vice versa) such that “it’s” and “it is” were considered a match. Any transcripts which seemed to contain reference to the experiment such as “I can’t hear this bit” were removed, and any participant transcripts which contained no text were removed from analysis to ensure that the number of deletions across the experiment was not artificially inflated by non-attempts.

4.6 Substitution classification

A word pairing that involves two different words is categorised as a substitution, and this error type is considered to have the most potential to cause damage in forensic contexts since it introduces information that is not contained within the acoustic signal (Tschäpe & Wagner, 2012) and readers of the transcript may be subconsciously influenced to hear the substituted term (e.g. Fraser & Kinoshita, 2021). In many cases, though, the effects of a substitution can be minimal; for example, “one guy **who** works there” being transcribed as “one guy **that** works there” will have very little effect on the meaning of the utterance or the listener’s perception of the speaker. For these reasons, substitutions were explored in greater detail, with further error classification taking place. Each substitution error was classified as belonging to the following categories (Table 2):

Classification	Explanation	Example
Form	Morphologically-related words, including verb conjugations (with the exception of tense changes)	Went forward → went forwards
Grammar	Grammatically similar words with very little change in meaning	Spoken with him → spoken to him
Synonym	Retention of meaning but not morphologically-related	Work's pretty hard → work's really hard
Tense	Change of tense	Messes up → messed up
Pronoun	Change of pronoun leading to potential change in meaning	I may have seen him → I may have seen her
Antonym	Words with contradictory meanings	Couldn't put → could put
Involving filled pause	Filled pause transcribed as a word, or vice versa	And er → to her
Unmarked	None of the above	To pick people up → to beat people up

Table 2: Classifications of different types of substitution errors.

The above classification scheme was created as a way of trying to broadly classify the effect of substitution errors in a way that might produce generalisable results for forensic contexts, where the impact of the error is much more important than the overall number of errors. Substitutions belonging to the top three categories (form, grammar, synonym) were categorised as relatively 'minor' with regard to their potential impact on the meaning of an utterance. Substitutions in the final five categories (tense, pronoun, antonym, involving filled pause, unmarked) were judged to have a higher potential of substantially impacting a reader's understanding of an utterance, and were therefore categorised as errors with 'major' potential to change meaning.

Another way of categorising the substitutions errors is examining the type of word that was substituted, i.e. function word or content word. In general, the substitution of content words

(nouns, verbs, adjectives, adverbs, etc.) could have a larger impact on the transcript reader's interpretation of the utterance since these words convey more meaning than function words. In forensic contexts, the substitution of a function word could lead to a substantial change in the meaning in a specific context (e.g. substitution of pronouns), though in many cases it would likely make little difference (e.g. "the" in place of "a"). Each word in the stimuli was categorised as a content word, function word or hesitation; a list of 264 function words developed for research in human-machine communication (O'Shea et al., 2012) was used to classify function words, hesitations were marked as such, and everything else was classified as a content word.

4.7 Evaluation of results

The percentage of words correctly transcribed was initially calculated for each participant transcript in this study. This metric was then compared across listening conditions and listener groups to provide an overview of overall performance. Then, in order to assess the effects of listener accent background and level of background noise on the likelihood of a word pair being a match, i.e. achieving a correct transcription, a binomial logistic mixed effects regression model was fitted to the data, using the `glmer()` function in R (Bates et al., 2015). The binary variable of Match (versus Non-Match, i.e. error) was included as the dependent variable, and accent background (SSBE or non-SSBE) and level of background noise (+6 dB, 0 dB or -3 dB SNR) (as well as the interaction between those two variables) were included as independent variables. The speaker within the stimulus was included as a random intercept and random slopes modelling SNR by listener were also included. Two reduced models were also fitted, identical to the full model except for the exclusion of one of the fixed effects (listener accent background or level of background noise). The full and reduced models were then compared using the `anova()` function in R to examine the effect of each variable on the likelihood of achieving a correct transcription.

Following this, the non-matched word pairs (i.e. errors) were separately analysed to examine the different types of errors produced within each listening condition and by each listener group. To assess the effect of the two independent variables (listener accent background and level of background noise) on the distribution of error types, a multinomial logistic mixed effects regression model was fitted with error type (DEL, SUB or INS) as the dependent variable and speaker and participant as random effects. This approach, essentially, involves a series of pairwise binomial regression comparisons between each pair of error types, and was carried out using the `mblogit()` function in R (Elff, 2022). A more complex random effects

structure, with random slopes, was not feasible due to the small amount of data (models with random slopes did not converge).

5 Results

5.1 Correct transcriptions

Figure 2 displays the percentage of words correctly transcribed by each accent group (speakers of SSBE and speakers of other varieties of British English) in each of the noise conditions (signal-to-noise ratios of +6 dB, 0 dB and -3 dB). A substantial decrease in the number of words correctly transcribed is observed as the signal-to-noise ratio decreases (i.e. as the level of background noise increases), with average medians of 91.7% in the 6 dB SNR condition, 75.6% in the 0 dB SNR condition and 48.4% in the -3 dB SNR condition. The mean accuracy is consistently higher for the SSBE group compared with the non-SSBE group, but the absolute difference is very small. Both groups show an increasing amount of variability as the noise level increases, with standard deviations of 19.1 and 17.5 for the SSBE group and the non-SSBE group respectively in the 6 dB SNR condition, 24.5 and 26.1 in the 0 dB condition, and 29.3 and 28.5 in the -3 dB condition.

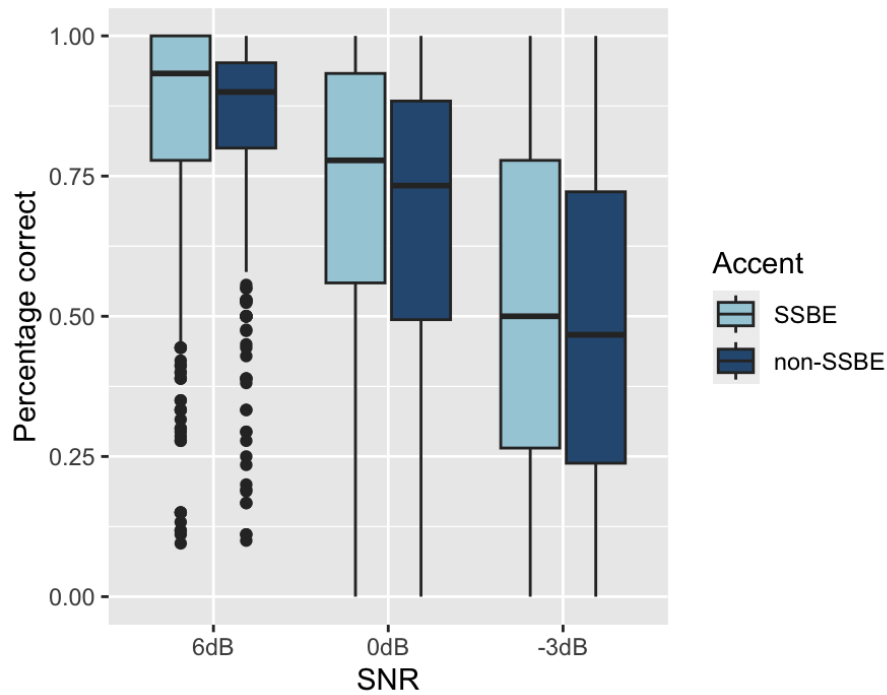


Figure 2: Percentage of words correctly transcribed across accent groups (SSBE speakers, non-SSBE speakers) and different signal-to-noise ratios (SNR; 6 dB, 0 dB and -3 dB). The level of background noise increases from left to right.

Table 3 provides an overview of the model comparisons between the full and null models for each of the independent variables. Results showed that the model fit was significantly improved when including the level of background noise (i.e. SNR) as a fixed effect, but that including accent background (i.e. SSBE versus non-SSBE) did not improve the model.

Variable	χ^2	Df	<i>p</i>
Accent background	2.97	3	0.397
Signal-to-noise ratio	177.72	4	< .001 (***)

Table 3: Output of model comparisons between the full model (including accent background and level of background noise/SNR) and the null models which excluded each of the independent variables in turn.

Inspection of the summary output of the model including both independent variables (and their interaction) revealed that accent background did not have a significant impact on the probability of achieving a correct transcription, although listeners in the SSBE group were 19% (95% CI [.93, 1.53]) more likely to achieve a correct transcription than those in the

non-SSBE group. However, the level of background noise was shown to have a significant impact on the likelihood of a correct transcription, with statistically significant differences in performance observed across all signal-to-noise ratios. Closer inspection of the model reveals that the odds of a word being transcribed correctly increased by 203% (95% CI [2.45, 3.75], $p < .001$) when the signal-to-noise ratio was 6 dB compared with 0 dB, and by 158% (95% CI [2.12, 3.15], $p < .001$) when the signal-to-noise ratio was 0 dB compared with -3 dB. The interaction between accent background and signal-to-noise ratio was not found to be significant in any of the comparisons.

Participants were also asked to rate each recording on a scale of 1 to 5 according to how challenging they found it to transcribe, where 1 represents no difficulty and 5 represents great difficulty. These ratings were compared against the percentage of words that participants correctly transcribed to investigate whether perceived difficulty correlates with transcription accuracy. Figure 3 shows a clear pattern whereby the recordings rated the easiest to transcribe achieved accuracy rates surpassing 90%, and the average accuracy decreases as the perceived difficulty increases.

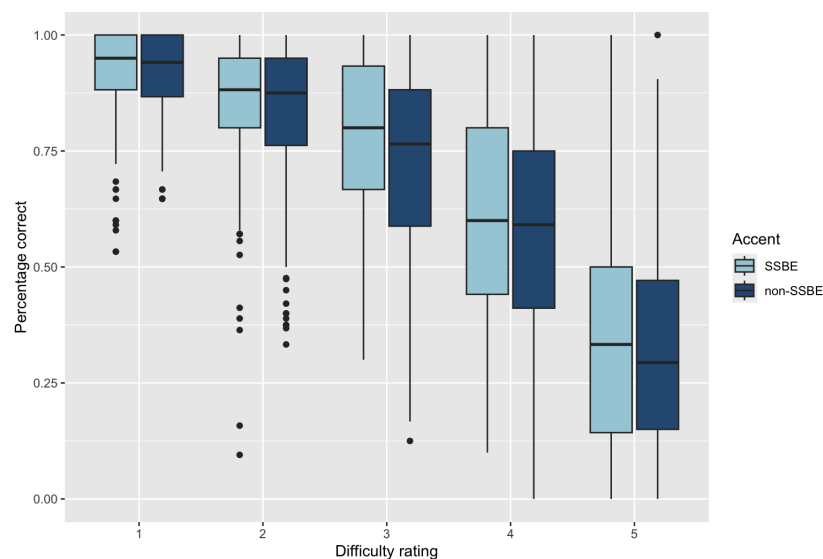


Figure 3: Percentage of words correctly transcribed across accent groups (SSBE speakers, non-SSBE speakers) and different difficulty ratings (where 1 represents least difficult and 5 represents most difficult).

5.2 Transcription errors

Figure 4 shows the total number of errors and the distribution of the three error types produced in each noise condition. A clear relationship is observed between the number of errors and the level of background noise, whereby an increase in background noise leads to a substantially higher number of errors. This effect is consistent across the accent groups, whereby for both groups there was around a 93% increase in the number of errors from the 6 dB SNR condition to the 0 dB SNR condition, and a 60% increase in the number of errors from the 0 dB SNR condition to the -3 dB SNR condition.

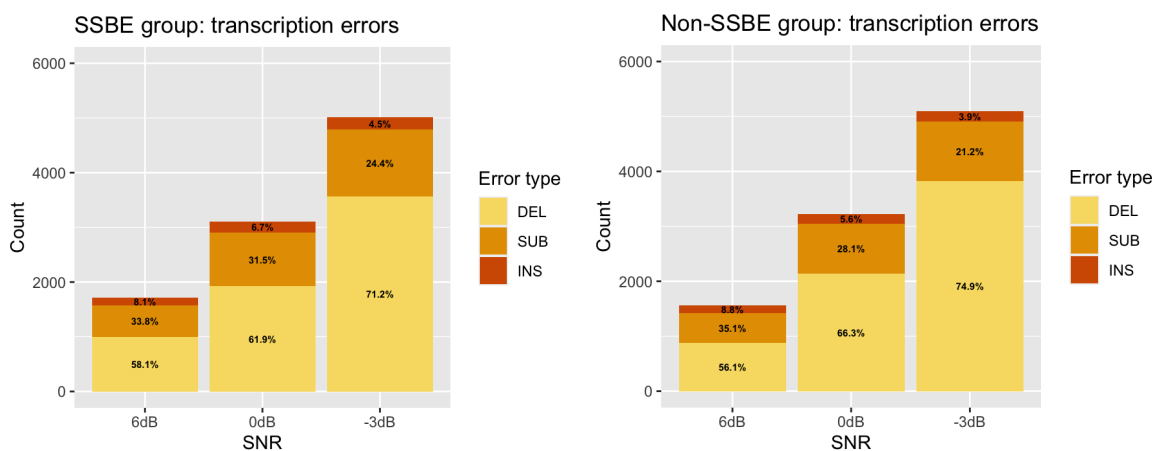


Figure 4: Raw number of errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of deletions (yellow), substitutions (orange) and insertions (red), and percentages within each noise condition are included. The level of background noise increases from left to right on each plot.

The increase in the total number of errors is largely driven by the number of deletions, which increases substantially with the level of background noise. Deletions are the most common error type across all conditions, and account for roughly 57% of all errors in the 6 dB SNR condition, 64% in the 0 dB SNR condition, and 73% in the -3 dB SNR condition. The number of errors that involve word substitutions also increases with the level of background noise, though to a lesser extent than deletions. Despite the increase in the number of substitution errors, the proportion of overall errors that involve substitutions decreases with higher levels of background noise. Substitutions account for around 34% of all errors in the 6 dB SNR condition, 30% in the 0 dB SNR condition, and 23% in the -3 dB SNR condition. Insertions are produced infrequently and at a relatively consistent rate across all conditions, constituting only 4-9% of all errors.

The output of a multinomial logistic regression model (specifically the odds ratio: OR) indicates the odds of producing one type of error in relation to another type of error across the different accent and SNR conditions. For example, the odds ratio of 1.15 on row 1 of Table 4 reveals that a transcriber is 1.15 times more likely to produce a substitution relative to a deletion when they are a native speaker of SSBE compared with a speaker of another variety of British English. Alternatively put, the odds of producing a substitution relative to a deletion increase by 15% when the transcriber belongs to the SSBE group compared with the non-SSBE group. The p value reveals whether the comparison is found to be statistically significant; in the example given above the p value exceeds the alpha level of 0.05 and therefore the difference between accent groups is not found to be statistically significant.

In all pairwise comparisons of the error types, accent background was not found to have a significant effect on the odds of producing one error type in comparison with another. Therefore, there is no statistically significant difference between the distribution of error types across the SSBE group and non-SSBE group, which can also be clearly visually observed in Figure 4. The level of background noise, however, was found to significantly affect the distribution of error types in almost every pairwise comparison (see Table 3); only the comparisons involving insertions and substitutions were not found to be statistically significant, which is likely a result of the small number of insertions. Simply put, as the noise level increases, the proportion of errors that involve deletions increases and the proportion of errors that involve substitutions decreases.

Model	Baseline	Comparison	OR	CI lower	CI upper	<i>p</i>
SUB vs DEL	non-SSBE	SSBE	1.15	0.89	1.53	0.27
	-3 dB	0 dB	1.06	1.65	2.06	< .001 (***)
	0 dB	6 dB	1.08	1.34	1.78	< .001 (***)
INS vs DEL	non-SSBE	SSBE	1.19	0.85	1.70	0.29
	-3 dB	0 dB	1.12	1.69	2.61	< .001 (***)
	0 dB	6 dB	1.14	1.48	2.43	< .001 (***)
INS vs SUB	non-SSBE	SSBE	1.13	0.82	1.31	0.78
	-3 dB	0 dB	1.12	0.92	1.44	0.22
	0 dB	6 dB	1.14	0.95	1.56	0.12

Table 4: Multinomial logistic regression analysis of the distribution of error types across conditions. Results from the model output (beta and Standard Error) were used to calculate the odds ratio (OR) and upper and lower confidence intervals (CI). Statistical significance is demonstrated with asterisks according to the degree of significance.

5.3 Substitution errors

The number of substitution errors produced is not affected by the accent background of the transcriber, but a higher level of background noise leads to a higher number of word substitutions. Closer examination of this type of error reveals that a number of substitutions are relatively ‘minor’ in their impact. For example, transcribing “it’s” in place of “that’s” in the utterance “that’s quite hard to see from there” is unlikely to have an effect on the perception of the speech or the speaker. There was very little difference in the proportions of ‘major’ and ‘minor’ substitutions between the two listener groups (see Figure 5). It was found that ‘minor’ substitution errors, such as those involving a small grammatical or morphological change that does not affect the meaning, account for between 31-33% of substitutions in the 6dB SNR condition, 23-27% in the 0dB condition and around 20% in the -3dB SNR condition.

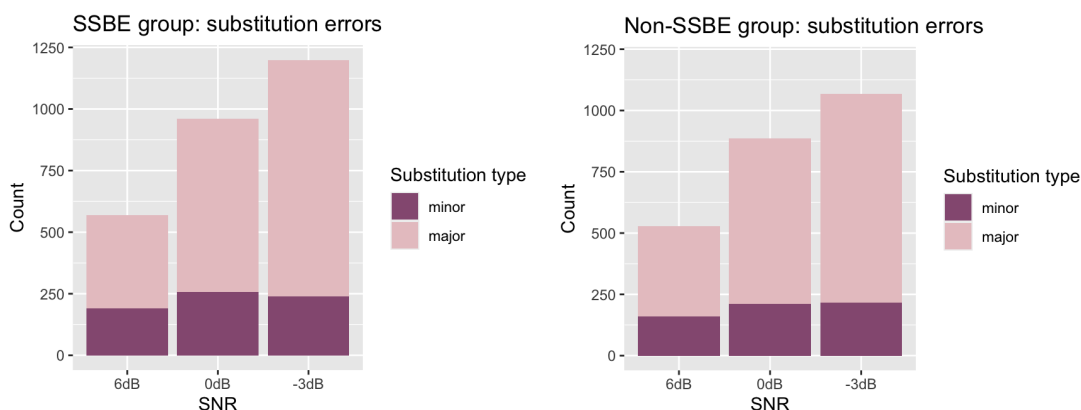


Figure 5: Raw number of substitution errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of minor errors (purple) and major errors (pink). The level of background noise increases from left to right on each plot.

The majority of errors made by participants in every condition were therefore categorised as ‘major’, with potentially significant changes in meaning arising from the error. For example, a few participants transcribed “it’s usually just me **and the girl**” in place of “it’s usually just me **in the car**”; the two word substitutions (“and” in place of “in”, and “girl” in place of “car”) have a substantial effect on a reader’s interpretation, with the original utterance reporting that the speaker is alone and the mistranscription conveying the opposite (that the speaker was with another person).

In terms of the type of word that was substituted (function words vs content words), there is very little difference in the proportions between the accent groups, as well as very little difference in proportions across the noise conditions (see Figure 6). There was an almost even divide in every condition, with function words accounting for around 55% of substituted words in the 6 dB SNR condition, 54% in the 0 dB SNR condition and 52% in the -3 dB SNR condition.

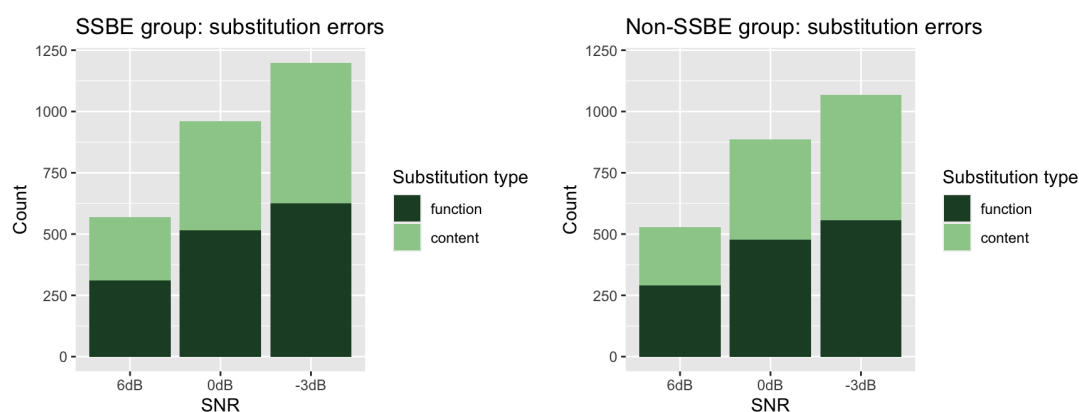


Figure 6: Raw number of substitution errors in each noise condition for the SSBE group (left) and the non-SSBE group (right). The bars are composed of function words that were substituted (dark green) and content words that were substituted (light green). The level of background noise increases from left to right on each plot.

6 Discussion

The present study set out to investigate how the quality of orthographic transcripts produced by lay listeners varies as a function of the level of background noise and the speaker's accent background. Unsurprisingly, results showed that transcripts were of much poorer quality as the level of background noise increases, in terms of the number of words correctly transcribed and the number and types of errors produced. However, accent background has very little effect on the transcripts produced, which will be further explored later in this discussion.

6.1 Level of background noise

The percentage of words correctly transcribed was relatively high in the least noisy condition (+6 dB SNR), with both listener groups averaging scores of over 90%. A significant decline in performance was observed as the SNR decreased, with averages of around 75% in the 0 dB SNR condition and of less than 50% in the noisiest condition at -3 dB SNR. The noisiest condition proved to be extremely challenging for transcribers, and there was a substantial increase in the amount of variability in performance across listeners at this noise level (standard deviation of roughly 0.29 in this condition, compared with around 0.25 in 0 dB SNR and 0.18 in +6 dB SNR). Given that increased noise levels reduce the amount of bottom-up information that transcribers are able to work with, it is unsurprising that the percentage of words that are correctly transcribed decreases with the SNR, and this reflects findings in

previous transcription studies where SNR has been manipulated (e.g. Clopper & Bradlow, 2008).

A substantial increase was observed in the number of errors produced as the SNR decreased, with almost triple the number of errors in the -3 dB SNR condition compared to the +6 dB SNR condition. The most frequent type of error produced across all conditions was deletions, and the number of deletions increased substantially as the noise level increased, roughly doubling with each SNR decrease. The distribution of errors was also significantly impacted by the varying levels of background noise; unsurprisingly, given the aforementioned huge increase in deletions, the proportion of errors that involve deletions significantly increased as the recordings became noisier. Despite the number of substitutions increasing with more noise, this type of error made up a smaller proportion of the total errors as the SNR decreased.

These results suggest that transcribers generally opt for omitting material they cannot confidently transcribe rather than guessing what may have been said (i.e. producing significantly more substitutions as noise increases). This is somewhat reassuring given that participants were asked to transcribe as much as they could and yet did not produce increasing rates of substitution errors as the intelligibility of the speech decreased. It should be noted that the transcribers in this experiment were given very basic instructions and, crucially, no contextual information regarding the content of the audio recordings. If they had received some indication of the conversational topics, they may have been more confident in offering (inaccurate) interpretations of the more challenging sections given that they would have had an increased amount of top-down information to guide their comprehension.

6.2 Accent background

No significant differences were found between the SSBE group and the non-SSBE group in this experiment, with regard to the percentage of words correctly transcribed, the types of errors produced and, more specifically, the different types of substitutions produced. This is somewhat unsurprising given that both groups are judged to be familiar with SSBE, though the results of Smith et al. (2014) suggested an advantage for the native SSBE speakers in moderate listening conditions (e.g. the 0 dB SNR condition in this study). However, the study conducted by Smith et al. (2014) presented stimuli containing ambiguous linguistic content without the disambiguating context. In this study, the stimuli were much longer, providing more information for transcribers to use when interpreting the signal, and the stimuli chosen were not intended to be ambiguous. This experiment attempted to conduct a more realistic

transcription task, though it should be acknowledged that even the stimuli in this study were much shorter than would usually be transcribed in forensic contexts.

In a transcription study on American dialects, Clopper and Bradlow (2008) demonstrated that transcribers from all accent groups (General American, Northern and 'Mobile' listeners) performed best and at a relatively similar level for stimuli in a General American accent. The authors suggest that listeners rely on 'standard' representations of words for acoustic-phonetic mappings in situations where less auditory information is available. The findings of the present study seem to support that hypothesis; despite the non-SSBE listeners likely having differing productions of many of the words contained within the stimuli, they were able to identify and transcribe such words with no less success than the SSBE listeners.

Two common phonological variables that separate Southern, more 'standard' dialects of British English from Northern dialects are the vowels in the BATH and STRUT lexical sets; in the accent background survey at the end of the experiments, participants had to indicate typical Southern pronunciations of these two vowels to be assigned to the SSBE group, and non-Southern pronunciations for at least one of the vowels in order to be categorised as non-SSBE listeners. Within the stimuli, there was only one example of a word belonging to the BATH lexical set: 'passed'. All participants either correctly transcribed this word or transcribed the homophone 'past'. There were numerous examples of the STRUT vowel within the stimuli, e.g. 'cut' and 'running'. Inspection of the participants' responses showed no obvious mistranscriptions as a result of a different production of this vowel for many of the non-SSBE listeners.

6.3 Substitutions

One of the ways in which this study aims to present a novel analysis of transcription performance is the focus on substitution errors, given the potential consequences of this error type in forensic contexts. Inspection of the substitution errors revealed a wide range of degrees of inaccuracy. At one end of the scale there was obvious miscomprehension of the speech content, such as "bartender" in place of "barber", but at the other end were much more insignificant substitutions, such as "spoken *to* him a few times" in place of "spoken *with* him a few times".

Given that all substitutions would otherwise be treated equally, a qualitative classification scheme was devised to categorise errors such that a differentiation could be made between

those which could potentially be harmful in a transcript and those which were very unlikely to have any substantial impact on the meaning of the utterance or the listener's perception of the speech. Substitutions categorised as relatively 'minor' in their potential impact (such as morphological and grammatical changes and synonyms) accounted for between one fifth and one third of all substitution errors across the six conditions (see Figure 3). This means that the majority of substitutions were deemed as potentially 'major' in their impact on the transcript. At first glance, this may sound like a worrying statistic; however, closer inspection of the 'major' substitutions once again showed a range of degrees of potential damage.

Some transcriptions involving 'major' substitutions retained most of the meaning, such as:

- "Pop in for a *quiet one* after work" → "pop in to a *bar* after work" or "go for a *pint* after work"
- "Internet *phone* thing" → "internet *web* thing"

Some transcriptions involving 'major' substitutions were nonsensical as a result, such as:

- "The *clocks* that weekend also went forward" → "the *crops* that weekend also went forward"
- "I *quite* generally *go*" → "I've *got* generally *goats*"

And some transcriptions involving 'major' substitutions could have a significant (potentially incriminating) impact on the listener's interpretation of the speech content, such as:

- "It's usually just me in the *car*" → "it's usually just me and the *guys*" or "it's usually just me and the *girl*" or "it's usually just me and the *gun*"
- "I was just *giving* a lift back maybe" → "I was just *getting* a lift back really"

This goal of the classification system and differentiating between potentially 'minor' and potentially 'major' substitution errors is an attempt to generalise for forensic purposes; however, it should be noted that there is so much nuance with regard to substitutions and the realistic impact that they could have in a forensic context. 'Disputed utterances', whereby different interpretations of a word or phrase are put forth (usually one of which is incriminating), arise in cases where poor audio quality is combined with badly used top-down information. Very small differences in the transcript, such as one mistranscribed word, could have a substantial effect on the court's judgements of guilt regarding the speaker (see Fraser (2020) for real examples of disputed utterance cases). These situations are relatively infrequent, given that the acoustic-phonetic information and the contextual information have to allow for an incriminating misinterpretation, but can be incredibly damaging when such occurrences do take place.

In many cases, the ‘major’ substitutions examined in this study would likely be picked up as a result of the listener’s knowledge of the context, such as the examples of ‘nonsensical transcriptions’ above. Issues generally arise in situations where there is great phonetic similarity between the target word and the mistranscription, and the mistranscription makes sense within the context; the examples in the ‘potentially incriminating transcriptions’ section above are a good illustration of this. The speaker in the first example claims that he was driving alone in his car, but the first two mistranscriptions demonstrate the opposite: that he was in company. The third mistranscription (involving “gun”) introduces a weapon, which could be construed as substantially incriminating in the right context, e.g. if the speaker was accused of shooting someone, though in other contexts would likely be extremely nonsensical.

The task of analysing substitution errors in terms of their impact on the transcript is extremely challenging on a large scale, and it is hoped that this paper will encourage further attempts to classify errors in a forensically-relevant manner. The current study has approached error analysis on a word-level basis, which allowed for a degree of automation within the analysis. It would be interesting to analyse transcription errors at a phrasal-level, given that many of the word substitutions in this study were analysed as relatively ‘minor’ in their impact and this would likely be addressed by a phrasal-level approach; however, this method would likely require a lot more manual data analysis (and therefore be more challenging on a large scale) as well as introduce a greater level of subjectivity in accuracy judgements.

The development of a framework for assessing the severity of transcription errors is essential for proficiency testing, which is a potential area of interest for forensic practitioners given a recent focus on validation of methods within the forensic sciences. Such a framework would not only allow for the production of proficiency tests as a resource for practitioners to prove their competency, it would also allow for further studies to explore transcription performance of both experts and non-experts, with research questions such as:

- What types of errors do experts tend to make, and do these differ from those made by non-experts?
- Do experts respond to noise levels in a different way to non-experts in terms of the content of the transcripts produced?

- How do transcripts produced by individual forensic experts compare to those produced by a team of forensic experts?²¹

7 Conclusions

In many cases, non-experts will carry out transcription of poor quality evidential audio recordings, and so this study aimed to explore the effects of two common factors that arise in forensic transcription: background noise and accent background of the transcriber. Unsurprisingly, an increase in background noise led to a significant decline in performance; however, the extent of this decline may be shocking to non-experts. It is crucial that the quality of the recording is taken into consideration when transcribing evidential audio materials. Expert practitioners in forensic speech science are acutely aware of the limitations of transcribing poor quality audio; there are often situations where the audio cannot be transcribed due to the speech being completely unintelligible as a result of, for example, background noise or overlapping speech/other noises. The approach taken by non-experts with regard to unintelligible speech seems to vary; some transcripts contain appropriate indications of unintelligible speech while others contain full transcriptions of sections which are judged by experts to be unintelligible and may not even contain speech (see section 5 of thesis - “Additional Resource”). Those unfamiliar with the complexities of transcription may lack an appreciation that it is not often possible to transcribe every word within forensic recordings as a result of the poor audio quality. In situations where such evidence plays a major role in a court case, it is necessary for transcripts to be as reliable as possible, and this may not be achieved by transcribers without specialist knowledge.

The results of this experiment demonstrate that, for the most part, listeners tend to omit content that they are not sure of, though substitutions made up over 20% of errors in all conditions. Even one mistranscribed word could lead to an innocuous utterance being interpreted as incriminating (i.e. in ‘disputed utterance’ cases), therefore it is crucial that transcribers in all domains of the criminal justice system are aware of the potential consequences of inaccurate transcripts. At the very least, training should be provided for non-expert transcribers on the dangers of poor quality audio, as well as the seriousness of some types of transcription errors.

²¹ This is a similar research question to that posed in a small-scale study conducted by Tschäpe and Wagner (2012), which found that teams of experts tended to perform better than individual experts, though what that means in terms of the content of the transcripts is unclear.

Accent background was not found to have an effect on transcription performance in this study, though it should be noted that all participants were judged to be ‘familiar’ with SSBE, either through their status as a native speaker of this variety or through long-term familiarity with the ‘standard’ variety via the media, education, etc. If this experiment had manipulated familiarity through the use of non-familiar accents, such as SSBE speakers transcribing Glaswegian English, it is expected that differences would emerge between accent groups in poor listening conditions, according to the findings of Smith et al. (2014). Clopper and Bradlow (2008) suggest that in poor listening conditions speakers rely on standard acoustic-phonetic representations, which would explain why the speakers of other varieties of British English performed at the same level as the native SSBE speakers in this study. Introducing an unfamiliar accent would lead to phonetic productions of words that do not easily map to the listener’s lexical representations, and it is therefore likely that more errors would occur. It is hypothesised that more substitutions would be produced for unfamiliar accents, given that transcribers would likely misinterpret words containing unfamiliar sounds; it would be interesting for future research to explore the transcription of unfamiliar accents in forensic-like conditions rather than in typical highly-controlled speech-in-noise tests.

Something that hasn’t been addressed in the present study is how external contextual information can affect the content of transcripts. In this experiment, participants were given relatively short extracts from a mock police interview with no common theme throughout the stimuli, and the only guidance participants were given was the relatively vague instructions to “transcribe every word that you can hear”. The lack of contextual information in this experiment was deliberate in order to isolate the effects of the two variables of interest. However, the knowledge of external contextual information is a key concern in the transcription of poor quality evidential audio materials, and in many cases of non-expert transcription (e.g. by police detectives), transcribers will be aware of details of the case and will therefore have certain expectations about what was said. Lange et al. (2011) found that simply being aware that the speech content was taken from criminal suspects’ interviews can lead to significantly higher numbers of incriminating misinterpretations. Expert practitioners generally follow procedures whereby the first draft of a transcript is produced ‘blind’ (without any information at all), and relevant contextual information is revealed to the transcribers in a gradual and controlled manner; such methods attempt to mitigate the effects of priming produced by knowledge of contextual information in order to facilitate the production of impartial transcripts. In the case of non-expert transcription, the implementation of similar methods could help to ensure the impartiality of transcripts for use alongside speech evidence.

8 References

- Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, 35(2), 520–529.
- Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken-word comprehension. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, 1925–1928.
- Assmann, P., & Summerfield, Q. (2004). The Perception of Speech Under Adverse Conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), *Speech Processing in the Auditory System* (pp. 231–308). Springer New York.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using “lme4.” *Journal of Statistical Software*, 67(1), 1–48.
- Boersma, P., & Weenink, D. (2020). *Doing phonetics by computer [Computer program]* (Version 6.1.30). <http://www.praat.org/>
- Burda, A. N., Casey, A. M., Foster, T. R., Pilkington, A. K., & Reppe, E. A. (2006). Effects of Accent and Age on the Transcription of Medically Related Utterances: A Pilot Study. *Communication Disorders Quarterly*, 27(2), 110–116.
- Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: intelligibility and classification. *Language and Speech*, 51(Pt 3), 175–198.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from Four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16.
- Elff, M. (2022). *Multinomial Logit Models, with or without Random Effects or Overdispersion*. <https://cloud.r-project.org/web/packages/mclogit/mclogit.pdf>
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology. Human Perception and Performance*, 32(5), 1276–1293.
- Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech Language and the Law*, 10(2), 203–226.
- Fraser, H. (2020). Forensic transcription: The case for transcription as a dedicated branch of linguistic science. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 416–431). Routledge.
- Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, 7. <https://doi.org/10.3389/fcomm.2022.898410>
- 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.
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2). <https://doi.org/10.1558/ijssl.v18i2.261>
- French, P., & Fraser, H. (2018). Why “Ad Hoc Experts” should not Provide Transcripts of Indistinct Audio, and a Better Approach. *Criminal Law Journal*, 298–302.
- Hunt, J., & McIlroy, M. (1976). An Algorithm for Differential File Comparison. *Murray Hill: Bell Laboratories*, 9.
- Jones, T., Kalbfeld, J. R., Hancock, R., & Clark, R. (2019). Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2), e216–e252.
- Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337–1351.
- 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.
- MacLean, L. M., Meyer, M., & Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research*, 14(1), 113–123.
- Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*. <https://journal.equinoxpub.com/IJSSL/article/view/10005>
- O’Shea, J., Bandar, Z., & Crockett, K. (2012). A Multi-classifier Approach to Dialogue Act Classification Using Function Words. In N. T. Nguyen (Ed.), *Transactions on Computational Collective Intelligence VII* (pp. 119–143). Springer Berlin Heidelberg.
- Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R.-A. (2014). Cross-accent intelligibility of speech in noise: long-term familiarity and short-term familiarisation. *Quarterly Journal of Experimental Psychology*, 67(3), 590–608.
- Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4), 487–501.
- 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*. Conference of the International Investigative Interviewing Research Group, Winchester.
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test*. Conference of International Association for Forensic Phonetics and Acoustics, Santander, Spain.

Walker, A. (2018). The effect of long-term second dialect exposure on sentence transcription in noise. *Journal of Phonetics*, 71, 162–176.

Article 2 - Appendix A

Accent background survey questions

Free text responses were collected for questions 1-3 and question 5. Participants had to select one of the options provided for question 4 and questions 6-9. Responses were required for all questions except question 3, which was optional for those who had moved around before the age of 18.

1. What accent do you have?
2. Where were you born and raised?
3. If you moved around before the age of 18, please list all the places where you have lived.
4. If you are from England, would you say your accent is associated with the North, the South or the Midlands?
 - a. The South
 - b. The Midlands
 - c. Not from England (from Wales, Scotland or Northern Ireland)
 - d. Not sure
5. Is your accent associated with a certain city or region? E.g. Manchester, Derbyshire, South West. Please provide details.
6. Paying particular attention to the vowel, does the word "bath" sound more like option A or option B when you say it out loud?
 - a. Audio file - [bɑ:θ]
 - b. Audio file - [bæθ]
7. Do the words "trap" and "bath" contain the same vowel when you say them out loud?
 - a. Yes
 - b. No
 - c. Not sure
8. Paying particular attention to the vowel, does the word "strut" sound more like option A or option B when you say it out loud?
 - a. Audio file - [stɹʌt]
 - b. Audio file - [stʊt]
9. Do the words "foot" and "strut" contain the same vowel when you say them out loud?
 - a. Yes
 - b. No
 - c. Not sure

Article 2 - Appendix B

Linguistic content of stimuli

Speakers are identified by the database from which they were extracted, whereby 'D' represents the DyViS database, and the speaker number given to them within that database.

Speaker	Extract no.	Utterance
D002	1	er no I don't think so actually my phone line has er been cut off
D002	2	we sometimes go and eat together at the steak house he's also a barber there
D002	3	right well it could have been somebody from work that I was just giving a lift back maybe
D002	4	yeah I may have seen him around but I couldn't put a name to a face
D002	5	and erm there's also a boat house but that's obviously that's quite hard to see from there
D002	6	well she hasn't actually passed her test yet actually to be honest but erm she's quite environmentally conscious as well
D023	7	yeah I dunno er I didn't look at my watch it was running a bit slow at the time
D023	8	I think the clocks that weekend also went forward and that just messes up my entire concept of time
D023	9	and er was just flicking on and then saw the end of erm a film I like
D023	10	I don't see that much of him but I chat quite a bit on an internet phone thing called Skype
D023	11	I don't always drive on my own I occasionally give my sister lifts places but it would have been a girl
D023	12	I know one guy who works there erm not very well spoken with him a few times
D073	13	yeah couple of soap operas think I caught the end of a film as well that was about it
D073	14	yeah we play a bit of football erm together erm occasionally like the odd round of golf when we can afford it

D073	15	no I don't usually tend to pick people up to be honest erm yeah it's usually just me in the car
D073	16	work's pretty hard at the moment I've been er pretty shattered so er not up to too much
D073	17	she's alright a bit erm a bit flouncy she kind of keeps poodles and er drives a scooter you know
D073	18	no it's not that often that I go I just kind of pop in for a quiet one after work
D060	19	I've been in there a couple of times but I've never got to know any of the people who work there
D060	20	absolutely not I haven't committed any crimes certainly not of the kind that you're suggesting
D060	21	sometimes I go for a walk there with my sister but I certainly wasn't there Wednesday
D060	22	it's reasonably new I think it's er it's a racer rather than a mountain bike
D060	23	well I quite generally go it's you know it's cheap food it's nothing special but i like it
D060	24	I've told you everything that I did Thursday night on Wednesday night on Friday morning

Article 2 - Appendix C

Data cleaning prior to analysis

Many of the non-matches identified by the transcript alignment software in this project did not actually constitute errors; for example, spelling mistakes were initially marked as substitution errors. There were many participant responses that needed to be amended in order to register as a match and therefore a correct transcription.

A number of decisions had to be made regarding what constitutes a match in particular circumstances, e.g. “it is” versus “it’s”. Taken into consideration throughout this process were (a) whether the listener’s perception had been affected, (b) whether a transcript reader’s perception would be affected, and (c) whether this type of error would feature in a finalised transcript (i.e. spelling mistakes).

An outline of the types of responses corrected (or left uncorrected) is included below.

Filled pauses

Any indication of a filled pause was marked as a match, regardless of the inclusion/exclusion of a nasal portion.

Reference transcript	Participant transcript	Marked as a match?
Er	Uh, err, ah, ur, um, uhm, ermm	✓
Erm	Uh, err, ah, ur, um, uhm, ermm	✓

Compound words

Terms that had been correctly recognised but transcribed with an additional space (e.g. separated compound words) or without a space (e.g. joined words) were marked as a match.

Reference transcript	Participant transcript	Marked as a match?
Steak house	Steakhouse	✓
Boat house	Boathouse	✓
Phone line	Phoneline	✓
Mountain bike	Mountainbike	✓
A bit	Abit	✓
As well	Aswell	✓
Think so	Thinkso	✓
Kind of	Kindof	✓
House but	Housebut	✓
Alright	All right	✓
Everything	Every thing	✓
Forward	For ward	✓
Somebody	Some body	✓

Contractions & expansions

Grammatical contractions and expansions were marked as a match. This included some colloquial language.

Reference transcript	Participant transcript	Marked as a match?
You're	You are	✓
It's	It was	✓
He's	He is	✓
That's	That was	✓
Would have	Would've	✓
That was	That's	✓
Could have	Could've	✓
Kind of	Kinda	✓
Dunno	Don't know	✓

Representational spelling changes

Instances where the participant's response was clearly the same as the reference transcript but represented in a different manner were marked as correct. This included some common grammatical errors/misconceptions.

Reference transcript	Participant transcript	Marked as a match?
Yeah	Yeh, yah, ye, yeahh, yea	✓
One	1	✓
No	Nah, nar	✓
(Could) have	(Could) of	✓
To (see)	Too (see)	✓
(Seen her) around	(Seen her) round	✓

Spelling mistakes

Very obvious spelling mistakes were marked as a match.

Reference transcript	Participant transcript	Marked as a match?
Football	Ffootball	✓
Conscious	Concious, concouis	✓
Committed	Committed	✓
Occasionally	Occasionally, occassionally	✓
Thursday	Thursdxay, thursday	✓
Haven't	Havnt	✓
Flouncy	Flounsy	✓
Usually	Usally	✓
Scooter	Scotter	✓
It	Ot	✓
Guy	Gy	✓
Morning	Mornin	✓
Film	Fim	✓
Caught	Cought	✓
Couldn't	Coudnt, couldny	✓
Around	Aroun	✓
A	An	✓
Certainly	Ceertainly, certainy	✓
Sometimes	Sometime	✓

Morphological changes

Participant responses that contained the same root word but with some kind of morphological alteration were marked as a non-match. This included changes in grammatical number and tense.

Reference transcript	Participant transcript	Marked as a match?
Crimes	Crime	✗
Go	Goes	✗
Been	Being	✗
Work	Worked	✗
Cheap	Cheapest	✗
Special	Especially	✗
Don't	Didn't	✗
Giving	Give	✗
Food	Foods	✗
Reasonably	Reasonable	✗
Like	Liking	✗
Racer	Race, racing	✗
Seen	See	✗
Certainly	Certainty	✗
Somebody	Someone	✗

Homophones & synonyms

Homophones and synonyms were marked as a non-match.


Reference transcript	Participant transcript	Marked as a match?
Passed	Past	✗
New	Knew	✗
Drive	Travel	✗
Pretty	Really	✗
Couple	Few	✗
Kind	Type	✗
Occasionally	Sometimes	✗
Own	Alone	✗
Round (of golf)	Game, around	✗
Entire	Whole	✗
Chat	Talk	✗

University of York
York Graduate Research School
Research Degree Thesis Statement of Authorship

Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

Candidate name	Lauren Harrington
Department	Language and Linguistic Science
Thesis title	Towards improving transcripts of audio recordings in the criminal justice system

Title of the work (paper/chapter)	Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance	
Publication status	Published	x
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	
Citation details (if applicable)	Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. <i>Frontiers in Communication</i> , 8, 1165233.	

Description of the candidate's contribution to the work*	Conceptualisation Methodology Formal analysis Investigation Writing - original draft
Approximate percentage contribution of the candidate to the work	100%
Signature of the candidate	
Date (DD/MM/YY)	29/02/2024

*The description of the candidate and co-authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co-authors may find it helpful to consider the [CRediT \(Contributor Roles Taxonomy\)](#) approach to recognising individual author contributions.

7. Article 3 - Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance

Lauren Harrington

Department of Language & Linguistic Science, University of York, UK

Abstract²²

Introduction: In England and Wales, transcripts of police-suspect interviews are often admitted as evidence in courts of law. Orthographic transcription is a time-consuming process and is usually carried out by untrained transcribers, resulting in records that contain summaries of large sections of the interview and paraphrased speech. The omission or inaccurate representation of important speech content could have serious consequences in a court of law. It is therefore clear that investigation into better solutions for police-interview transcription is required. This paper explores the possibility of incorporating automatic speech recognition (ASR) methods into the transcription process, with the goal of producing verbatim transcripts without sacrificing police time and money. We consider the potential viability of automatic transcripts as a “first” draft that would be manually corrected by police transcribers. The study additionally investigates the effects of audio quality, regional accent, and the ASR system used, as well as the types and magnitude of errors produced and their implications in the context of police-suspect interview transcripts.

Methods: Speech data was extracted from two forensically-relevant corpora, with speakers of two accents of British English: Standard Southern British English and West Yorkshire English (a non-standard regional variety). Both a high quality and degraded version of each file was transcribed using three commercially available ASR systems: Amazon, Google, and Rev.

Results: System performance varied depending on the ASR system and the audio quality, and while regional accent was not found to significantly predict word error rate, the distribution of errors varied substantially across the accents, with more potentially damaging errors produced for speakers of West Yorkshire English.

Discussion: The low word error rates and easily identifiable errors produced by Amazon suggest that the incorporation of ASR into the transcription of police-suspect interviews

²² The formatting of this abstract conforms to the requirements set by *Frontiers in Communication*, the journal in which this article has been published.

could be viable, though more work is required to investigate the effects of other contextual factors, such as multiple speakers and different types of background noise.

1 Introduction

Orthographic transcripts of spoken language can be admitted as evidence in courts of law in England and Wales in a number of scenarios. When the speech content of an audio or video recording is used as evidence, e.g. a threatening voicemail message, the recording is often accompanied by a transcript to assist the court in “making out what was said and who said it” (Fraser, 2020, p. 416). These recordings tend to be of very poor quality such that the speech is often close to unintelligible without the aid of a transcript. However, this means that the transcript can be highly influential on what members of the court believe they hear in the recording, as highlighted by Fraser and Kinoshita (2021; see also Fraser et al., 2011). It is therefore crucial that transcripts presented alongside speech evidence are as accurate as possible since they can play an important role in listeners' perception of speech and speakers, potentially leading to miscarriages of justice in cases where an utterance is inaccurately interpreted as incriminating (Harrison & Wormald, in press).

Another use of orthographic transcripts in the legal system is transcripts of police-suspect interviews, which play an important role in the investigative process and are often admitted as evidence in court (Haworth, 2018). While the audio recording of the police-suspect interview is technically the “real” evidence in this context, the transcript is admissible as a “copy” and is often the only version of the police-suspect interview that is referred to in the courtroom (Haworth, 2018). Given that the court often does not hear the original audio recording, it is important that the transcripts are an accurate representation of the interview's contents. However, Haworth (2018; 2020) has identified issues with the transcripts created by police transcribers, including summarising large sections of the interview, paraphrasing the speech content and inconsistent representation across transcribers. A verbatim record of the speech would be ideal, but this is a time-consuming and laborious task.

Automatic speech recognition (ASR) technology is rapidly improving and can produce transcripts in a fraction of the time it would take a human to complete the same task. Transcripts produced by an ASR system would require manual checking and correction, but the output would be a verbatim record of the full interview, eliminating the issue of potentially important information being inaccurately paraphrased or omitted. A computer-assisted transcription method could lead to more reliable evidence being presented to courts without a significant increase in the time spent producing the records.

When considering the incorporation of ASR into the transcription process, it is important to take into account factors that have a significant impact on ASR performance, such as audio quality and regional accents. Background noise has been shown to decrease the accuracy of ASR systems in a number of contexts (Lippmann, 1997; Littlefield & Hashemi-Sakhtsari, 2002) including for forensic-like audio recordings (Harrington et al., 2022; Loakes, 2022). In recent years, a growing body of research has focused on systematic bias within automatic systems, i.e. underperformance for certain demographic groups, and significant disparities in performance have been demonstrated across accents. Transcripts tend to be significantly less accurate for non-native speakers (DiChristofano et al., 2022) or speakers of non-standard regional varieties (Markl, 2022). However, a limitation of work in this area is the use of word error rate (WER) for evaluating performance. WER is the ratio of errors in a transcript to the total number of words spoken and can be useful to highlight differences in performance across groups. However, this metric does not provide insights into where and why systems produce errors, or how evidentially significant those errors could be.

This paper presents work on the topic of automatic speech recognition in the context of police-suspect interview transcription, employing a novel method of analysis that combines industry-standard measures alongside detailed phonetic and phonological analysis. While WER is useful for an overview of performance, incorporating fine-grained linguistic analysis into the method permits a deeper understanding of the aspects of speech that prove to be problematic for automatic systems. The performance of three commercial ASR systems is assessed with two regional accents, across different audio qualities; the purpose of this assessment is to evaluate how practical it would be for ASR systems to play a role in the transcription of police-suspect interviews.

2 Background

This section covers a range of topics relevant to the present study. Firstly, Section 2.1 outlines the current situation regarding police-suspect interview transcription in England and Wales, and highlights the issues. Automatic speech recognition (ASR) is offered as part of a potential solution, and Section 2.2 covers a brief history of ASR and its rapid improvement in recent years. Section 2.3 describes research on the use of ASR for transcribing forensic audio recordings, which leads into the potential incorporation of ASR in the transcription of comparatively better quality audio recordings, i.e. police-suspect interviews, in Section 2.4. Section 2.5 addresses potential speaker-related factors that may affect ASR performance, such as regional accent. Finally, Section 2.6 outlines the research aims of the present study.

2.1 Transcription of police-suspect interviews in England and Wales

In England and Wales, police-suspect interviews are recorded according to requirements of the Police and Criminal Evidence Act 1984. The audio recording is subsequently used to produce a Record of Taped Interview (ROTI), and if the case ends up going to trial, the ROTI is often admitted as evidence alongside the original audio recording. However, the transcript itself often becomes effectively “interchangeable [with] and (in essence) identical” (Haworth, 2018, p. 434) to the audio evidence in the eyes of the court, and is often used as a substitute for the original audio recording. Relying on the transcript as the primary source of the interview's contents could be problematic in cases where speech has been omitted or inaccurately represented.

The police interview transcribers, also known as ROTI clerks, tend to be employed as administrative staff, and the job-specific skills required often include proficiency in audio and copy typing and a specific typing speed (Tompkinson et al., 2022). ROTI clerks receive little to no training or guidance on the transcription process (Haworth, 2018), which has the potential to create a systematic lack of consistency in transcription production, even within police forces. This is highlighted by an example provided in Haworth (2018) in which three ROTI clerks transcribe an unanswered question in three unique ways: “no response”, “no audible reply” and “defendant remained silent”. Each representation could potentially generate varying interpretations of the interviewee's character. It is also worth noting that the 43 territorial police forces in England and Wales operate individually, which contributes to the issue of inconsistency in transcription and transcript production across forces.

Another issue with ROTIs is that much of the interview is summarised and the transcriber, untrained in legal issues and protocol, will ultimately decide what is deemed as important and worthy of full transcription. This decision-making process could lead to serious consequences given Section 34 of the Criminal Justice and Public Order Act 1994, which states that the court may draw inferences if something later relied upon as evidence is not mentioned during the initial interview stage.

In accordance with Haworth (2018), this assessment of problematic issues surrounding ROTIs does not serve as a critique of the clerks hired to produce the transcripts, but of the wider process. Transcription, particularly of long stretches of speech, is a time-consuming and labour-intensive task that can take four to five times the length of the audio recording to transcribe for research purposes (Walford, 2001; Punch & Oancea, 2014), and a time factor

of 40 to 100 for difficult forensic recordings (Richard Rhodes, personal communication). It is also prone to human error, for example spelling and punctuation mistakes (Johnson et al., 2014) and omission or misrepresentation of short function words, discourse markers and filled pauses (Stolcke & Droppo, 2017; Zayats et al., 2019). Transcribing spoken language, even when producing a verbatim transcript, is a complex and inherently selective process which carries the inevitable risk of systematic and methodological bias (Jenks, 2013; Kowal & O'Connell, 2014). Transcripts carry social and linguistic information, therefore transcribers have an enormous amount of power over the way in which people are portrayed (Jenks, 2013).

Discrepancies concerning the portrayal of speakers have been reported within legal transcripts (e.g. US court reports, UK police interviews), with standardised language and “polished” grammar for professionals such as lawyers, expert witnesses and police interviewers but verbatim transcription or inconsistently-maintained dialect choices for lay witnesses or suspects (Walker, 1990; Coulthard, 2013). Similar inconsistencies were observed in ROTIs (Haworth, 2018), as well as an assumption revealed in focus group discussions with ROTI clerks that the interviewee will be charged with or convicted of an offence, as demonstrated through the use of terms such as “defendant” or “offender” to refer to interviewees (89% of references; Haworth, 2018, p. 440).

The use of ASR could address a number of the concerns regarding the production of police interview transcripts. Automatic systems can process a large amount of data in a fraction of the time it would take a human to do the same task. This could allow for interviews to be transcribed in full, rather than mostly summarised, while saving time, effort and money on behalf of the police. An automatic system would not apply social judgements to the role of interviewer and interviewee, and would therefore likely remain consistent in its treatment of speakers in this regard, given that only the speech content would be transcribed. Furthermore, an ASR system would likely be consistent in its representation of phenomena such as silences; for example, unanswered questions simply would not be transcribed, and therefore the system would not inject potentially subjective statements such as “defendant remained silent”.

2.2 Automatic speech recognition

The field of automatic speech recognition (ASR) has received growing interest over the last decade given its expanding applications and rapid improvements in performance, though this technology has existed in different forms for over 70 years. The first speech recogniser

was developed in 1952 at Bell Telephone Laboratories (now Bell Labs) in the United States and was capable of recognizing 10 unique numerical digits. By the 1960's systems were able to recognize individual phonemes and words, and the introduction of linear predictive coding (LPC) in the 1970's led to rapid development of speaker-specific speech recognition for isolated words and small vocabulary tasks (Wang et al., 2019). The 1980's saw the creation of large databases (O'Shaughnessy, 2008) and the implementation of a statistical method called the "Hidden Markov Model" (HMM) which allowed systems to recognize several thousand words and led to substantial progress in the recognition of continuous speech (Wang et al., 2019). Combining HMM with a probabilistic Gaussian Mixture Model (HMM-GMM) created a framework that was thoroughly and extensively researched throughout the 1990's and 2000's, and remained the dominant framework until the last decade when deep learning techniques have become prevalent (Wang et al., 2019). In recent years deep neural networks (DNN) have been implemented to create the HMM-DNN model, achieving performance well beyond its predecessor.

Modern state-of-the-art ASR systems are typically made up of two main components, an acoustic model and a language model, both of which are concerned with calculating probabilities. As a basic summary according to Siniscalchi and Lee (2021), the acoustic model recognizes speech as a set of sub-word units (i.e. phonemes or syllables) or whole word units. It is then tasked with calculating the probability that the observed speech signal corresponds to a possible string of words. The language model then calculates the probability that this string of words would occur in natural speech. This is often evaluated using n-grams, which calculate the probability of the next word in a sequence given the n previous words, based on extensive training on large text corpora. Both models contribute to the estimated orthographic transcription produced by the ASR system.

Adaptations to the architecture of ASR systems have led to huge improvements in accuracy, which can be illustrated by observing the reported word error rates (WER) on a commonly-used dataset for measuring ASR performance, such as the Switchboard corpus (Godfrey & Holliman, 1993). This is a dataset of American English conversational telephone speech that is commonly used to benchmark ASR performance. The first reported assessment of speech recognition performance had a WER of around 78% (Gillick et al., 1993) and by 2005 state-of-the-art systems were yielding WER measures between 20 and 30% (Hain et al., 2005). Thanks to large amounts of training data and the application of machine learning algorithms, huge improvements in speech technology have been demonstrated in recent years. In 2016, Microsoft reported that their automatic system had achieved human parity, with a WER of 5.8% compared with a human error rate of 5.9% on a

subset of the Switchboard data (Xiong et al., 2016). In 2021, IBM reported an even lower WER of 5.0% on a subset of the Switchboard data, reaching a new milestone for automatic speech recognition performance (Tüske et al., 2021).

It is crucial to acknowledge, though, that performance is relative to the materials being transcribed. Though trying to mimic spontaneous conversations, the Switchboard corpus contains “inherently artificial” (Szymański et al., 2020, p. 2) spoken data due to factors such as the predefined list of topics, the localised vocabulary and the relatively non-spontaneous form of the conversations. These factors, paired with the relatively good audio quality, create conditions which are favourable to ASR systems, and while ASR may outperform human transcribers in some cases, there will be circumstances in which the reverse is true, especially in more challenging conditions such as forensic audio.

2.3 Automatic transcription of forensic audio recordings

Some work within the field of forensic transcription has considered whether automatic methods could be incorporated into the transcription of forensic audio samples, such as covert recordings. The audio quality of such recordings is generally poor given the real-world environments in which the recordings are made, and as a result of the equipment being deployed in a covert manner, rather than one designed to capture good-quality audio. They can also be very long, containing only a few sections of interest; it is often necessary to transcribe the recording in full to identify such sections, which is a time-consuming and arduous task for forensic practitioners.

Two studies in particular have explored automatic transcription in forensic-like contexts, the first of which uses an audio recording of a band rehearsal (Loakes, 2022), comparable to a covert recording. Two automatic transcription services (BAS Web Services and Descript) were employed to transcribe the 44 s recording containing the sounds of musical instruments and multiple speakers from a distance. BAS Web Services returned a system error when an orthographic transcription was requested, and when the in-built WebMINNI service was employed to segment the speech into phonemes, many sections of speech were identified as “non-human noise” and instrument noises were labelled as speech. Descript was also unsuccessful in its attempt to transcribe the speech, with the output containing only three distinct words (“yes”, “yeah”, and “okay”), a fraction of the total number of words uttered.

A second study on the topic of forensic transcription compared the performance of 12 commercial automatic transcription services using a 4-min telephone recording of a conversation between five people in a busy restaurant (Harrington et al., 2022). Talkers were positioned around a table upon which a mobile device was placed to record the audio, and all were aware of its presence. The transcripts produced by the automatic systems were of poor quality, making little sense and omitting large portions of speech, although this is not surprising given the high levels of background noise and numerous sections of overlapping speech.

A number of relatively clear single-speaker utterances were selected for further analysis, and results showed that even in cases of slightly better audio quality and more favourable speaking conditions, transcripts were far from accurate. The best performing system (Microsoft) produced transcripts in which 70% of words on average matched the ground truth transcript, though there was a high level of variability across utterances. Microsoft transcribed seven of the 19 utterances with over 85% accuracy, but many of the other transcriptions contained errors that could cause confusion over the meaning, or even mislead readers. For example, “that would have to be huge” was transcribed as “that was absolutely huge”, changing the tense from conditional (something that could happen) to past (something that has happened). In many cases, the automatic transcript would need substantial editing to achieve an accurate portrayal of the speech content.

The findings of such research, though valuable, are unsurprising given that commercial ASR systems are not designed to deal with poor quality audio; they are often trained on relatively good quality materials more representative of general commercial applications. Following recent advances in learning techniques to improve ASR performance under multimedia noise, Mošner et al. (2019) demonstrated that a system trained on clean and noisy data achieved better performance (i.e. higher reductions in WER) than a system trained only on clean data. It seems that training data has a direct effect on the capabilities of ASR systems. There could potentially be a place for automatic systems within the field of forensic transcription if the training data used is comparable to the audio recordings that would ultimately be transcribed. However, it is impractical to expect commercial ASR systems to perform at an appropriate level for the type of data that forensic practitioners handle.

Given the current state of the technology, ASR should therefore not be employed for the transcription of poor quality audio such as covert recordings, though the question remains as to whether it could be incorporated for comparatively better quality audio samples, such as police interviews. This type of audio recording is much better suited to automatic

transcription for many reasons. The quality of police-suspect interview recordings tends to be much better since the equipment utilised is built specifically for the purposes of recording audio, and all members present are aware of the recording process. The number of speakers is limited and known, and the question-and-answer format of the interview will most likely result in speech that is easier to transcribe, i.e. less overlapping speech. The level of background noise will also likely be much lower than a busy restaurant or a music practice room, although it must be noted that the audio quality of these interviews is not always ideal or comparable to studio quality audio. Reverberation, broadband noise or interference, the rustling of papers and the whirring of laptop fans (Richard Rhodes, personal communication) are examples of frequently occurring issues encountered within police interview recordings which can make some sections difficult to transcribe.

2.4 Incorporating automatic methods into police transcription

One approach to the use of automatic methods would be the use of an automatically-produced transcript as a starting point to which human judgements could be added i.e. “post-editing” an ASR output. Bokhove and Downey (2018) suggest that using automatic transcription services to create a “first draft” could be worthwhile in an effort to reduce the time and costs involved in human transcription. In their study, many of the errors made by the ASR system for interview data were relatively small and easily rectifiable, while recordings of a classroom study and a public hearing (with many speakers and microphones positioned far away from speakers) resulted in automatic transcriptions that deviated more substantially from the audio content. Nonetheless, Bokhove and Downey (2018) argue that, with little effort, reasonable “first versions” can be obtained through the use of freely available web services, and that these may serve as a useful first draft in a process which would involve multiple “cycles” or “rounds” of transcription (Paulus et al., 2013) regardless of the inclusion of automatic methods.

However, the baseline performance of the ASR system is a key issue in whether combining ASR and human transcription is viable. By artificially manipulating the accuracy of transcripts, Gaur et al. (2016) demonstrated that the time spent correcting an ASR output can exceed the time spent creating a transcript from scratch if the automatically-produced transcript is insufficiently accurate. By manipulating the WER of transcripts at rough intervals of 5% ranging between 15 and 55%, it was found that by the time the WER had reached 30% participants were able to complete the post-editing phase more quickly by typing out the content from scratch. However, participants only realised that the quality of the original transcript was a challenge when the WER reached around 45%. These findings suggest that

post-editing an ASR output could reduce the time taken to produce a verbatim transcript provided that the WER does not exceed a certain level; however, if the WER consistently approaches 30% then the incorporation of automatic methods into the transcription process fails to be a worthwhile avenue of research.

There are, however, some issues with using WER as the defining metric of system performance, as highlighted by Papadopoulou et al. (2021). Firstly, WER can be expensive and time-consuming to calculate due to the requirement of manual transcriptions to use as a reference. Secondly, quantified error metrics do not take into account the cognitive effort necessary to revise the ASR transcripts into a “publishable” quality. A more useful metric for analysing ASR outputs is the post-editing effort required. In their study, a single post-editor with intermediate experience in the field was tasked with post-editing transcripts produced by four commercial ASR systems (Amazon, Microsoft, Trint, and Otter). Both the time taken to edit the ASR output and the character-based Levenshtein distance between the automatic and post-edited transcripts were measured.

An interesting finding by Papadopoulou et al. (2021) is that the number of errors within a transcript does not always correlate with the amount of effort required for post-editing. Systems with the lowest error rates do not necessarily achieve the best scores in terms of the post-editing time and distance. Certain types of errors were shown to take longer to edit, such as those related to fluency, i.e. filler words, punctuation and segmentation. The authors also suggest that deletion and insertion errors are easily detectable and require less cognitive effort to edit than substitution errors. Although little justification for this claim is put forth in the paper, it does seem likely that deletions and insertions could be easier to identify given that the number of syllables will not match up between the speech content and the transcript. The post-editor may find substitutions more challenging to detect, especially if the phonetic content of the target word and transcribed word is similar. It is therefore crucial to consider the types of errors made, not just overall error rates, when assessing the viability of an automatic transcript as a first draft.

The study carried out by Papadopoulou et al. (2021) claims to be one of the first papers to evaluate the post-editing effort required to transform ASR outputs into usable transcripts and to conduct qualitative analysis on ASR transcription errors. Given that WER does not reveal sufficient information regarding the types of errors made and the difficulty of correcting those errors, there is a clear need for additional research on the topic of post-editing and alternative methods of analysis. This is particularly true when evaluating the practicality of incorporating ASR into the transcription process, as the effort required to transform an ASR

output into a fit-for-purpose verbatim transcript is the main consideration in whether this approach is advantageous, rather than the number of errors in the initial transcript.

2.5 Automatic systems and speaker factors

Given that the speakers taking part in police-suspect interviews will come from a range of demographics, it is important to consider how this may affect the performance of automatic speech recognition systems. Factors relating to a speaker's linguistic background, such as accent, can prove challenging for an automatic transcription system. Previous work has demonstrated that the performance of ASR systems declines significantly when confronted with speech that diverges from the “standard” variety; this has been found for non-native-accented speech in English (Meyer et al., 2020; DiChristofano et al., 2022; Markl, 2022) and Dutch (Feng et al., 2021), as well as for non-standard regionally-accented speech in Brazilian Portuguese (Lima et al., 2019) and British English (Markl, 2022).

Markl (2022) compared the performance of Google and Amazon transcription services across multiple accents of British English. One hundred and two teenagers from London or Cambridge (South of England), Liverpool, Bradford, Leeds, or Newcastle (North of England), Cardiff (Wales), Belfast (Northern Ireland), or Dublin (Republic of Ireland) were recorded reading a passage from a short story. Both systems demonstrated significantly worse performance, based on WER, for some of the non-standard regional accents compared with the more “standard” Southern English accents. Amazon performed best for speakers from Cambridge and showed a significant decline in performance for those from parts of Northern England (Newcastle, Bradford, and Liverpool) and Northern Ireland (Belfast). Much higher error rates were reported for Google for every variety of British English, likely as a result of much higher rates of deletion errors. Google performed best for speakers of London English and saw a significant drop in performance only for speakers from Belfast.

Many researchers have suggested that the composition of training datasets can cause bias within automatic systems (Tatman, 2017; Koenecke et al., 2020; Meyer et al., 2020; Feng et al., 2021) and that the underrepresentation of certain accents leads to a decline in performance for those varieties. Markl (2022) reports that certain substitution errors identified for speakers of non-standard regional accents of British English suggest that there is an overrepresentation of Southern accents in the training data or that acoustic models are being trained only on more prestigious Southern varieties, such as Received Pronunciation. Similarly, Wassink et al. (2022) claim that 20% of the errors within their data would be addressed by incorporating dialectal forms of ethnic varieties of American English (African

American, ChicanX, and Native American) into the training of the automatic systems. The implementation of accent-dependent (or dialect-specific) acoustic models has been found to improve performance, particularly for varieties deviating more substantially from the standard variety, such as Indian English and African American Vernacular English (Vergyri et al., 2010; Dorn, 2019).

2.6 Research aims

The main aim of the present research is to assess ASR transcription errors across accents and audio qualities. The implications of such errors being retained in a transcript presented to the court will be considered, and methods of analysis that are appropriate for this particular context will be employed. This work is centred on the transcription of recordings resembling police interview data, and a further aim of this work is to consider the practicality of incorporating ASR into the transcription of police-suspect interviews.

The present study will explore in much greater detail the types of errors produced across two different accents of British English, and will focus not only on the distribution of three main error categories (deletions, substitutions, and insertions), but also on the distribution of word types that feature in the errors. For example, some substitutions may be more damaging than others, or more difficult to identify in the post-editing of a transcript. Errors will also be assessed from a phonological perspective in order to identify errors resulting from phonological differences across the accents and highlight particularly challenging phonetic variables for the automatic systems. Although both the acoustic and language model will affect ASR performance, the analysis and interpretation of errors in this study will focus on those which are most likely a reflection of the acoustic model.

In this study, recordings that are representative of police interviews in the UK (in terms of speech style and audio quality) are used, which are expected to degrade ASR performance compared with previous studies that have typically used high quality materials. The present study considers, from a practical perspective, whether this technology could be incorporated into the transcription process for police-suspect interviews.

The specific research questions are:

1. How do regional accent and audio quality affect the performance of different ASR systems?
2. What types of errors are produced by the ASR systems, and what are the implications of these errors?

3. To what extent could ASR systems produce a viable first draft for transcripts of police-suspect interviews?

3 Materials and methods

3.1 Stimuli

In order to explore differences in ASR performance across different regional accents, two varieties of British English were chosen for analysis: Standard Southern British English (SSBE) and West Yorkshire English (WYE). SSBE is a non-localized variety of British English spoken mostly in Southern parts of England, and although linguistic diversity is celebrated in contemporary Britain, SSBE is heard more frequently than other accents in public life (e.g. TV programmes and films), especially in media that is seen on an international scale, and acts as a teaching standard for British English (Lindsey, 2019). SSBE is referred to in this study as a “standard” variety. WYE is a non-standard regional variety of British English which shares characteristics with many other Northern English accents²³ and whose phonology diverges substantially from SSBE (Hickey, 2015).

Stimuli were extracted from two forensically-relevant corpora of British English: the Dynamic Variability in Speech database (DyViS; Nolan et al., 2009) and the West Yorkshire Regional English Database (WYRED; Gold et al., 2018). DyViS contains the speech of 100 young adult males from the South of England (the majority of whom had studied at the University of Cambridge) taking part in a number of simulated forensic tasks, such as a telephone call with an “accomplice” and a mock police interview. WYRED contains the speech of 180 young adult males from three parts of West Yorkshire (Kirklees, Bradford, and Wakefield) and was created to address the lack of forensically-relevant population data for varieties of British English other than SSBE. The collection procedures employed in the production of the DyViS database were closely followed for WYRED, resulting in very closely matched simulated forensic conditions.

The mock police interview contained a map task in which specific speech sounds were elicited through the use of visual stimuli. Participants assumed the role of a suspected drug trafficker and had to answer a series of questions regarding their whereabouts at the time of the crime, their daily routine and their work colleagues, among other things. Visual prompts

²³ West Yorkshire English shares some features (e.g. lack of TRAP-BATH and FOOT-STRUT splits) with General Northern English (GNE), an emerging variety of Northern British English which is the result of dialect levelling (Strycharczuk et al., 2020). However, there are some features that make WYE distinct from GNE, such as the monophthongisation of vowels in words like “face” and “goat”.

were provided during the task, containing information about the events in question and incriminating facts shown in red text. Participants were advised to be cooperative but to deny or avoid mentioning any incriminating information. The speech was conversational and semi-spontaneous, and the question-and-answer format of the task was designed to replicate a police-suspect interview. On account of the focus on police-suspect interview transcription in this paper, the mock police interview task was selected for this study.

Two speakers of each accent were selected and eight short utterances were extracted per speaker, resulting in a total of 32 utterances. Much of the speech content in this task contained proper nouns such as the surnames of colleagues and place names. With the exception of two well-known brands, “Skype” and “Doritos”, proper nouns were not included in the extracted utterances in order to avoid inflated error rates as a result of misspellings or due to the name not featuring in the ASR system’s vocabulary. Other than filled pauses, which were extremely common in the spoken data, effort was also made to exclude disfluent sections. Disfluencies have been shown to be problematic for ASR systems (Zayats et al., 2019), therefore sections containing false starts or multiple repetitions were excluded in order to isolate differences in performance due to regional accent. Utterances ranged between 14 and 20 words in length and 3–6 s in duration, each containing a single speaker and unique linguistic content. Some examples of the utterances are included in Table 1 (and reference transcripts for all utterances can be found in Article 3 - Appendix A).

Speaker	Utterance
SSBE-1	And um there's also a boat house but that's obviously that's quite hard to see from there
SSBE-2	Not exactly I can't really remember their surnames but I might have known them i don't know
WYE-1	Uh can get a bit inebriated sometimes so not all the time no can't say
WYE-2	Yeah quarter of an hour half an hour something like that depending on traffic

Table 1: Examples of linguistic content of stimuli from each speaker.

To investigate the effects of low levels of background noise, such as that commonly found in real police interviews, the studio quality recordings were mixed with speech-shaped noise, derived from the HARVARD speech corpus. This was carried out using Praat (Boersma & Weenink, 2022), and the resulting files had an average signal-to-noise ratio (SNR) of 10 dB,

such that intelligibility was not hugely impacted but the background noise was noticeable. The studio quality files had a much higher average SNR of 22 dB, reflecting the lack of background noise in these recordings. To summarise, a studio quality version and a poorer quality version (with added background noise) of each file was created, resulting in a total of 64 stimuli for automatic transcription.

3.2 Automatic transcription

Three commercially-available automatic transcription services were used to transcribe the audio files: Rev AI²⁴, Amazon Transcribe²⁵, and Google Cloud Speech-to-Text²⁶. Many automatic transcription systems acknowledge that background noise and strongly accented speech can decrease transcription accuracy. Rev AI was chosen due to its claims of resilience against noisy audio and its Global English language model which is trained on “a multitude of... accents/dialects from all over the world” (Mishra, 2021). Services from Amazon and Google were chosen due to their frequent use in other studies involving ASR and the prevalent use of products from these technology companies in daily life. When uploading the files for automatic transcription, “British English” was selected as the language for Amazon and Google, and, since this option was not available for the third service, “Global English” was selected for Rev AI.

Reference (i.e. ground truth) transcripts were manually produced by the author for each utterance, using the studio quality recordings. The automatic transcripts produced by Amazon, Google, and Rev were compiled in a CSV file. Amazon and Google offer confidence levels for each word within the transcription but for the purpose of this research, only the final output (i.e. the highest probability word) was extracted.

3.3 Error analysis

Custom-built software was written to align the reference and automatic transcripts on a word-level basis, and each word pairing was assessed as a match or an error. Errors fall into three categories as outlined below:

- Deletion: the reference transcript contains a word but the automatic transcript does not.

²⁴ Rev AI accessed 12th November 2021.

²⁵ Amazon Transcribe accessed 17th October 2022.

²⁶ Google Cloud Speech-to-Text accessed 13th October 2022.

- Insertion: the reference transcript does not contain a word but the automatic transcript does.
- Substitution: the words in the reference transcript and automatic transcript do not match.

From a forensic perspective, insertions and substitutions are potentially more harmful than deletions, on the assumption that reduced information causes less damage than false information in case work (Tschäpe & Wagner, 2012). Table 2 shows an example of two potential transcriptions of the utterance “packet of gum in the car”, and demonstrates the different effect that substitutions can have in comparison with deletions. Both transcripts contain three errors, but the substitutions in transcript 2 could be much more damaging given the change in content and the new potentially incriminating interpretation of the utterance.

Reference	packet	of	gum	in	the	car
Transcript 1			gum	in		car
Transcript 2	pack	the	gun	in	the	car

Table 2: Two potential transcriptions of the utterance “packet of gum in the car”. Deletions are represented by a shaded red cell and substitutions are represented by bolded red text.

Some minor representational errors were observed, such as “steak house” transcribed as a compound noun “steakhouse” and numbers transcribed as digits. Since these substitutions do not constitute inaccuracies, rather slight changes in representation, the word pairing was marked as a match and these were not included as errors in the subsequent analysis. With regards to substitutions spanning multiple words, it was decided that the collective error would be marked as one substitution. For example, “cut and” transcribed as “cutting” was marked as a substitution rather than a combination of a substitution and a deletion, in an attempt to avoid inflated insertion and deletion rates.

Despite the limitations of WER, particularly in a forensic context, this metric can provide a brief overview of system performance across groups that can be used as a starting point for analysis. WER was therefore calculated for each utterance, by dividing the total number of errors (deletions, insertions, and substitutions) by the number of words in the reference transcript. The total number of each type of error in each condition was also calculated and compared to explore the differences across the ASR systems as well as the effects of

regional accent and level of background noise. In order to explore in greater detail the types of words involved in errors, each error pairing was manually evaluated as involving content words, function words, filled pauses or a combination of these. Substitutions involving morphological alterations were also highlighted, and transcripts were assessed in terms of the effort required to transform the ASR output into a more accurate, verbatim transcript.

Errors were also assessed on a phonological level in order to explore whether varying phonetic realisations of features across accents could be responsible for transcription errors, with a particular focus on marked vocalic differences across SSBE and WYE. Substitutions involving content words in the Yorkshire English transcripts were analysed by identifying which of Wells' lexical sets (i.e. groups of words all sharing the same vowel phoneme; Wells, 1982) the words in the reference and automatic transcripts belong to as well as transcribing the speaker's production of the word, with the goal of better understanding why the error may have been made.

Four vocalic variables in particular were analysed due to differences between the SSBE and WYE phonetic realisations (Wells, 1982; Hughes et al., 2005). These are outlined in Table 3, using Wells' (1982) lexical sets as a way of grouping words that share the same phoneme. Words in the BATH lexical set contain a long back vowel in SSBE, but typically contain a short front vowel in WYE, which overlaps with the production of the TRAP vowel [a] in both varieties. Words in the STRUT lexical set contain an unrounded low vowel in SSBE, but a rounded high vowel in WYE; the rounded high vowel [ʊ] is also produced in words belonging to the FOOT lexical set in both varieties. Words belonging to the FACE and GOAT lexical sets contain diphthongs in SSBE, but typically contain monophthongs in WYE.

Lexical set	SSBE	WYE
BATH	[ɑ:]	[a]
STRUT	[ʌ]	[ʊ]
FACE	[eɪ]	[e: ~ ɛ:]
GOAT	[əʊ]	[o:]

Table 3: Phonetic realisations of four vocalic variables across the two varieties of British English analysed in this study, Standard Southern British English (SSBE) and West Yorkshire English (WYE). Variables are defined using Wells' (1982) lexical sets.

3.4 Statistical analysis

In order to evaluate which factors had a significant effect on word error rate, three linear mixed effects models were fitted using the lme4 package (Bates et al., 2015) in R. In each model, regional accent, audio quality or ASR system was included as a fixed effect, and all models included Speaker and Sentence as random effects to account for variation across speakers within accent groups and the unique linguistic content of each utterance. A separate “null” model was fitted including only the random effects, and the ANOVA function in R was used to compare each full model with the null model. Results of the model comparisons indicate whether the full model is better at accounting for the variability in the data, and therefore whether the fixed effect has a significant impact on word error rate. Results of the model outputs, containing an Estimate, Standard Error rate and a p-value, were then examined to evaluate the relationship between variables. A threshold of $\alpha = 0.05$ was used to determine statistical significance.

A three-way comparison was carried out for the ASR system and in the first three models Amazon was used as a baseline, meaning that a comparison between Rev and Google had not been carried out. The “ASR system” variable was relevelled such that Rev became the baseline, and a fourth model was then fitted with ASR system as a fixed effect and Speaker and Sentence as random effects.

4 Results

4.1 ASR systems

The three automatic systems tested in this study performed with varying levels of success and were all clearly affected to some degree by the regional accent of the speaker and the level of background noise. Figure 1 shows WER in each condition for the three ASR systems. The four conditions are SSBE speech in studio quality audio, SSBE speech in audio with added speech-shaped noise, WYE speech in studio quality audio and WYE speech in audio with added speech-shaped noise; these will henceforth be referred to as SSBE studio, SSBE SSN, WYE studio and WYE SSN, respectively. Amazon was the best performing system with the lowest word error rate (WER) in each of the four conditions compared with Rev and Google, and achieved its lowest WER (13.9%) in the SSBE studio condition and highest WER (26.4%) in the WYE SSN condition. Google was the worst performing system, achieving the highest WER in every condition except for WYE speech in studio quality, for which Rev performed worst with a WER of 34.1%.

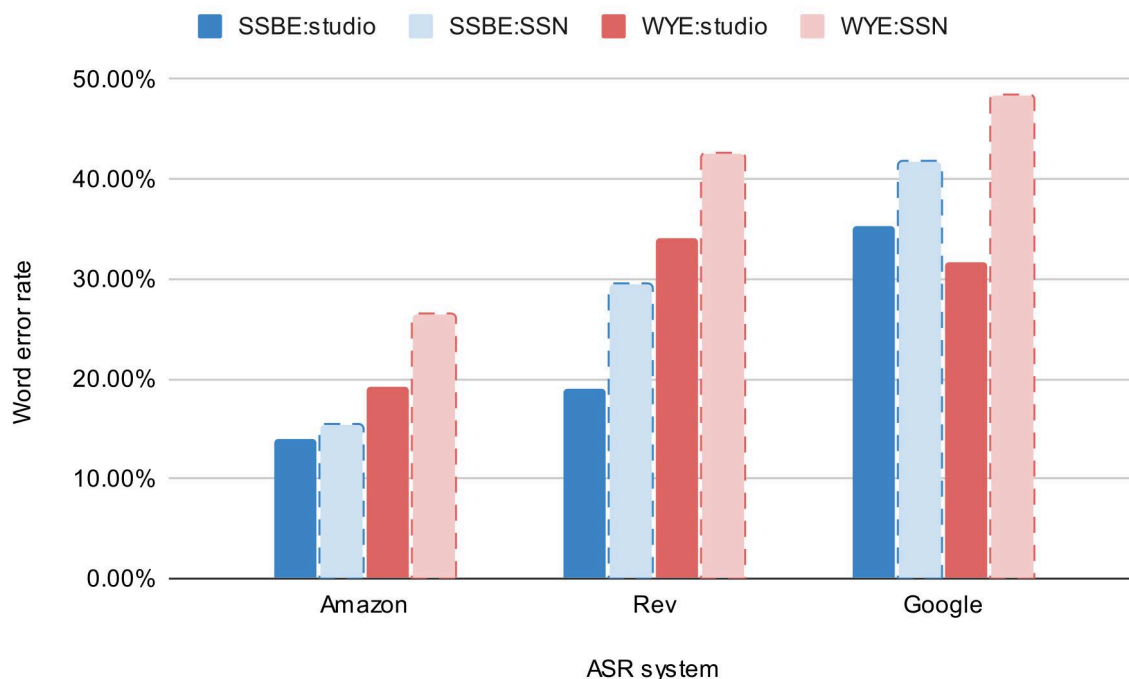


Figure 1: Average word error rate in each of the four conditions (SSBE studio, SSBE SSN, WYE studio, and WYE SSN) for all three ASR systems (Amazon, Rev, and Google). ASR systems are ordered from left to right according to lowest to highest average WER.

Results of a model comparison between the null model and the model with ASR system as a fixed effect revealed that ASR system has a significant impact on WER [$\chi^2(2) = 50.35, p < 0.0001$]. The summary output of the linear mixed effects model revealed that there was a significant difference in error rates between Amazon and both Rev ($\beta = 0.13, SE = 0.26, p < 0.001$) and Google ($\beta = 0.20, SE = 0.26, p < 0.001$). Rev achieved WERs that were on average 13% higher than those produced by Amazon, while Google produced WERs on average 20% higher than Amazon. When comparing the two worst performing systems, Google was found to produce significantly higher WERs than Rev ($\beta = 0.08, SE = 0.03, p < 0.005$).

A notable trend in the data was Google's high tendency toward deletion errors, with over double (and in some cases quadruple) the number of deletions that Amazon produced in the same condition. An example of this is the utterance “not exactly I can't really remember their surnames but I might have known them I don't know” which was transcribed in studio quality by Amazon as “not exactly I can't remember their names but I might have known him I don't you” (with one deletion and three substitutions) and by Google as “not exactly I can't remember this sentence I don't know” (with seven deletions and two substitutions).

4.2 Regional accent

There are some clear differences in performance between the two accents in this study. Word error rate is lower for SSBE than for WYE in all conditions except for Google in the WYE studio condition; however, the results of a model comparison between the null model and the model with regional accent as a fixed effect showed that the difference in performance across accents was not statistically significant [$\chi^2(1) = 1.28$, $p = 0.26$]. This is likely due to the extremely small sample size and variation in system performance across the speakers of each accent. All ASR systems produced higher WERs for one of the SSBE speakers, which were on average 13 and 20% higher than for the other SSBE speaker in studio quality audio and speech-shaped noise audio, respectively. One of the WYE speakers also proved more challenging for the ASR systems, though the difference was most pronounced in studio quality where WERs were on average 10% higher than for the other WYE speaker. An average difference of 4% was observed between the WYE speakers in speech-shaped noise audio, which is likely a result of the highest WERs in the study being observed in this condition.

The most common type of error also varied across accents, with deletions featuring most frequently for SSBE speech (see Table 4) and substitutions featuring most frequently for WYE speech (see Table 5). As discussed earlier in this paper, substitution errors can be viewed as more harmful than deletion errors in forensic contexts given that incorrect information has the potential to be much more damaging than reduced information. Substitutions may also have a stronger priming effect than other types of errors on the post-editors who are correcting an ASR transcript.

System	Audio Quality	INS	DEL	SUB	Total Errors
Amazon	Studio	0	20	16	36
Amazon	SSN	0	26	14	40
Rev	Studio	0	25	25	50
Rev	SSN	2	43	30	75
Google	Studio	0	57	36	93
Google	SSN	1	73	37	111

Table 4: Counts of each error type (insertions, deletions, and substitutions) produced by each system for Standard Southern British English speech. SSN refers to the audio quality with added speech-shaped noise.

55.6% of SSBE errors in studio quality audio and 62.7% of SSBE errors in speech-shaped noise audio were deletions. The number of deletions in SSBE was consistently higher than in WYE, though occasionally only by a relatively small margin. The majority of deletion errors involved short function words, such as “a” and “to”, which made up between 61.5 and 80% of all deletion errors for Rev and Google. Amazon made the fewest deletion errors out of all the ASR systems, and the majority of the deletions for SSBE speech involved the omission of filled pauses. The deletion of content words was much less frequent, accounting for 17.9% of all deletion errors for Rev and 16.3% of all deletion errors for Google. Amazon was the only system for which content words were never deleted.

System	Audio Quality	INS	DEL	SUB	Total Errors
Amazon	Studio	1	13	33	47
Amazon	SSN	3	16	46	65
Rev	Studio	3	22	53	78
Rev	SSN	5	30	64	99
Google	Studio	4	39	42	85
Google	SSN	4	71	50	125

Table 5: Counts of each error type (insertions, deletions, and substitutions) produced by each system for West Yorkshire English speech. SSN refers to the audio quality with added speech-shaped noise.

Substitutions accounted for the most frequently occurring type of error for West Yorkshire English speech, with an average of 62.5% of all errors in studio condition and 58.5% of all errors in the speech-shaped noise condition involving the substitution of words or phrases. The only condition in which substitutions were not the most frequently occurring type of error for WYE speakers was Google in the speech-shaped noise condition where deletions constituted 71 of the 125 errors. The distribution of word types involved in substitution errors also differed across accents. The majority of substitutions for WYE speech involved content words while most substitutions for SSBE speech involved function words (Figure 2).

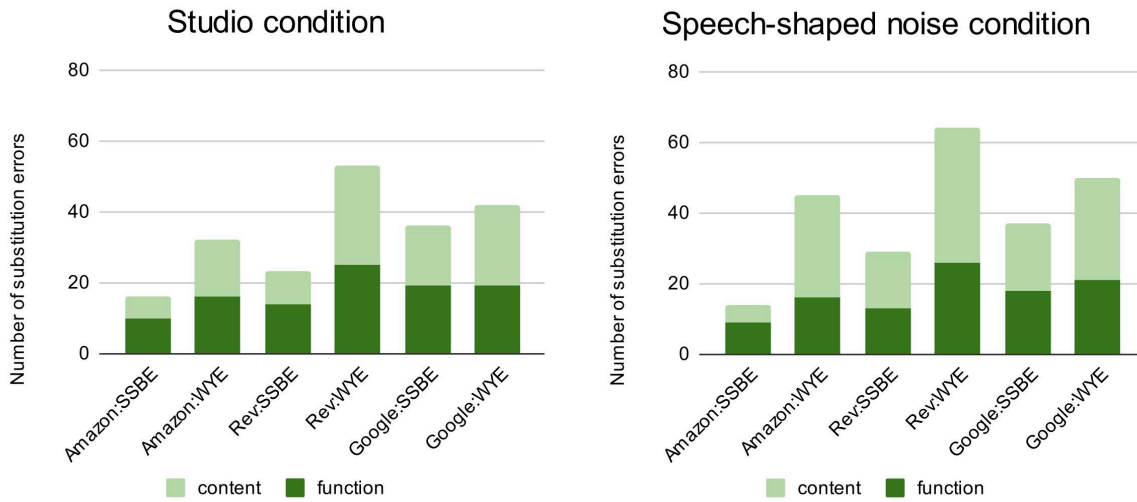


Figure 2: Number of substitution errors produced by each ASR system in each accent, in studio condition (left) and speech-shaped noise condition (right). ASR systems are ordered from left to right according to lowest to highest average WER.

Despite substitutions relating to function words accounting for a minority of substitution errors in WYE, there were more of this type of error in WYE than in SSBE. For Amazon and Rev, the number of content word-related substitutions was between 2 and 5 times higher for Yorkshire English than for SSBE, and the smaller increase for Google was likely a result of higher numbers of substitutions for SSBE speakers.

4.3 Audio quality

Higher error rates (by an average of 8%) were observed in speech-shaped noise audio compared with studio quality audio for all systems and for both accents. The results of a model comparison between the null model and the model with audio quality as a fixed effect showed that this difference was statistically significant [$\chi^2(1) = 11.42$, $p < 0.001$], and examination of the model output confirmed that WER was significantly higher in the degraded audio condition ($\beta = 0.08$, $SE = 0.02$, $p < 0.001$). An increase was observed in the number of insertions and deletions in all conditions when comparing the transcripts of the studio quality recordings to the recordings with added speech-shaped noise. Rev and Google in particular show large increases in the number of deletions from studio condition to the speech-shaped noise condition. A very similar number of substitutions was observed across the audio qualities in SSBE, but the number of substitutions in WYE was 19–40% higher in the speech-shaped noise condition. The change in audio quality also affected the

distribution of word types involved in substitutions. While the majority of substitution errors in SSBE were related to function words in studio quality audio, a majority involved content words in the speech-shaped noise condition for both Rev and Google. Not only was Amazon the highest performing system overall, it was also the least affected by the addition of background noise.

4.4 Phonological variables

Many errors within the West Yorkshire English data could be explained by a phonetic realisation deviating from what might be expected based on the assumed underlying acoustic models. This was especially true in the case of vowels where the phonology deviates markedly from SSBE. Given that previous studies suggest an overrepresentation of more ‘standard’ (in this context, Southern British) varieties in training data, we may expect to see the ASR systems struggling with some of the non-standard pronunciations of words by Yorkshire speakers. To explore this, four vowels which are well-known to differ in quality, length, or number of articulatory targets across SSBE and WYE were chosen for more in-depth analysis.

4.4.1 BATH

Words belonging to the BATH lexical set contain different vowels within the two accents: the long back vowel [ɑ:] in SSBE and, like many other varieties from the North of England, the short front vowel [a] in WYE. There were few occurrences of words belonging to the BATH lexical set in the Yorkshire data, though there were two utterances of the word “staff”, one by each of the Yorkshire speakers, which were produced with a short front vowel, i.e. [staf], rather than a long back vowel, i.e. [stɑ:f]. All three systems correctly transcribed this word for one speaker but not for the other. The pronunciations themselves were very similar but the surrounding context of the word was likely the cause of this issue. In the successful case, “staff” was uttered at the beginning of an intonational phrase but in the other occurrence it was preceded by a non-standard pronunciation of “with” [wɪʔ]. Omission of word-final fricatives, most commonly in function words, is a common process in some varieties of Yorkshire English (Stoddart et al., 1999). In this case, the voiced dental fricative /ð/ has been replaced with a glottal stop, resulting in the utterance [wɪʔstaf] which Rev and Google both analysed as one word, transcribing “waste” and “Wigston”, respectively. Amazon mistranscribed the word “staff” as “stuff”, a substitution which could be the result of the Yorkshire vowel being replaced with the closest alternative that creates a word in Standard Southern British English. Since [staf] in this case is not recognized as the word “staff”, the

closest SSBE alternative is the word “stuff” which contains a low central vowel [ʌ] that is closer within the vowel space to the uttered vowel than [ɑ:].

4.4.2 STRUT

There is a systemic difference between SSBE and WYE with regards to the number of phonemes in each accent's phonological inventory, whereby the SSBE STRUT vowel /ʌ/ does not feature in WYE. Instead, [ʊ] is produced in words belonging to both the STRUT and FOOT lexical sets. Many words containing this vowel were correctly transcribed within the Yorkshire data, though some occurrences resulted in phonologically-motivated substitutions. The word “bus”, pronounced [bʊs] by the Yorkshire speaker, was correctly transcribed by Amazon and Google but proved challenging for Rev which replaced it with “books”, a word containing [ʊ] in SSBE and belonging to the FOOT lexical set. A similar pattern was observed for the word “cut”, pronounced [kʊʔ], which Amazon and Google transcribed (almost correctly) as the present participle “cutting”, while Rev substituted it with a word from the FOOT lexical set, “couldn't”.

The word “muddy”, pronounced [mʊdɪ] by the Yorkshire speaker, proved challenging for all three systems. In both audio qualities, Amazon mistranscribed this word as “moody” /mu:di/, retaining the consonants but replacing the vowel with the closest alternative that creates a plausible word. Interestingly, Rev and Google both transcribed “much” in place of “muddy”, correctly recognizing the word uttered as belonging to the STRUT lexical set despite the high rounded quality of the vowel [ʊ].

Another example of a Yorkshire word belonging to the STRUT lexical set that proved to be challenging for the ASR systems was “haircut”, pronounced [ɛ:kʊʔ], though this was likely due to the h-dropping that takes places in word-initial position. Google semi-successfully transcribed “cut”, ignoring the first vowel in the word, while Amazon and Rev transcribed “airport” and “accurate”, respectively. The lack of /h/ at the beginning of “haircut” had a clear impact on the words consequently transcribed, since both begin with a vowel. This seems to have then had an effect on the vowel transcribed in the second syllable, as these systems transcribed final syllables containing the vowels [ɔ:] or [ʊ] in SSBE.

4.4.3 FACE

Words belonging to the FACE lexical set are subject to realizational differences across the accents; the FACE vowel is realised as the diphthong [eɪ] in SSBE but as the long

monophthong [e:] in WYE. Most words containing this vowel were transcribed correctly, e.g. “rains” and “place”, despite the monophthongal quality of the vowel produced by the Yorkshire speaker. However, some occurrences of [e:] proved challenging. For example, the word “potatoes”, pronounced [p(ə)te:ʔəz] with a glottal stop in place of the second alveolar plosive, was incorrectly transcribed as “tears”, “debt is”, and “date is” by Amazon, Rev and Google, respectively. While Google transcribes a word containing the correct vowel [eɪ] (“date”), the other systems transcribe words containing the vowels [ɛə] and [ɛ], which share similar vocalic qualities with the front mid vowel uttered by the speaker in terms of vowel height, frontness and steady state (or very little articulatory movement). Given that Rev and Google both transcribe words containing /t/ after the FACE vowel, it seems unlikely that the mistranscriptions are a result of the glottal stop, and are rather a direct result of the monophthongal realisation of the FACE vowel.

4.4.4 GOAT

Words belonging to the GOAT lexical set vary in their phonetic realisation across the two accents, such that the diphthong [əʊ] features in SSBE but a long monophthong features in WYE, which can be realised in a number of ways. Traditionally this was produced as a back vowel [o:] but it has undergone a process of fronting (Watt & Tillotson, 2001; Finnegan & Hickey, 2015) to [e:] for many younger speakers, including the two Yorkshire speakers in this study. Some words containing this vowel were transcribed without issue, such as “own” and “go”, though it should be noted that the latter was relatively diphthongal in quality given the phonological environment: the following word “in” begins with a vowel therefore a [w]-like sound is inserted, leading to movement during the vowel and creating a sound much closer to the SSBE diphthong [əʊ].

Other words containing the fronted monophthong proved more challenging for the systems, such as “drove” which was mistranscribed as “drew if”, “do if”, and “if” by Amazon, Rev, and Google, respectively. Amazon and Rev replace [e:] with words containing the vowel [u:], an alternative long monophthong produced in a relatively similar part of the vowel space, followed by [ɪ] and the voiceless version of the labiodental fricative. Google omitted the GOAT vowel, transcribing only the word “if” in studio quality audio and deleting the word completely in the speech-shaped noise condition. The word “road”, pronounced [ʁə:d], was also mistranscribed by two of the systems as “word” (/wɜ:d/), whereby the central monophthongal quality of the vowel was retained but the height was slightly adjusted to give [ɜ:].

4.5 Post-editing

In order to assess the possibility of incorporating an ASR output into the transcription process, it is necessary to assess the effort required to transform the ASR output into a more accurate (verbatim) transcript. The best performing system, Amazon, was evaluated in terms of the frequency and types of errors produced, as well as the difficulty of error identification within the data. Deletion and insertion errors may be more easily detectable than substitution errors, as suggested by Papadopoulou et al. (2021), in many contexts; in principle, these errors should stand out as missing or extraneous when the transcriber listens to the audio, while substitution errors may be more challenging to identify, especially if closely resembling the speech sounds in the audio recording.

In studio quality, 20 deletions were produced for SSBE speech and 13 for WYE speech, and in both cases, more than half of the deletion errors involved the omission of filled pauses. The rest of the deletions involved function words, and in almost all cases the transcription remained relatively unchanged in terms of semantic meaning, e.g. “I can’t **really** remember” → “I can’t remember” or “half an hour **something like that** depending on traffic” → “half an hour depending on traffic”. In the speech-shaped noise condition, 26 deletions were produced for SSBE and 16 for WYE. Fifty percent of the errors for SSBE involved filled pauses while the majority of WYE deletions (11/16) involved function words, and most deletions did not affect the semantic meaning of the utterance, e.g. “except **for** when it rains” → “except when it rains” or “he’s a tour guide **and** I knew **him** from secondary school” → “he’s a tour guide I knew from [a] secondary school”. Furthermore, some of the deletions occurred in instances where a pronoun or determiner, e.g. “I” or “a”, had been repeated, such that the transcript contained only one instance of each word.

Insertions were extremely rare within the data, particularly for Amazon which did not produce any insertions for SSBE and only inserted 1–3 words in the WYE transcripts. In studio quality, the only insertion to be made was “I knew him from secondary school” → “I knew [him] from **a** secondary school”, which is easily detectable given that the insertion of the determiner sounds unnatural in this context. The same insertion was made in the SSN condition, along with the insertion of first-person pronoun “I” and determiner “a”.

Substitutions may require more cognitive effort to identify, particularly in cases where the word in the transcript closely resembles the word that is uttered. First, the substitution of content words was assessed given that this type of mistranscription could lead to serious errors in forensic contexts, e.g. if a non-incriminating word such as “gum” is substituted with

an incriminating alternative like “gun”. In studio quality, six content words in SSBE and 16 in WYE were subject to substitution errors. The majority of SSBE substitutions in this case involved morphological alterations, such as a change in tense (e.g. “finish” → “finished”) or omission of an affix (e.g. “surnames” → “names”). Due to the phonetic similarity of the target and transcribed word, these substitutions could be difficult to notice in a post-editing phase, and an uncorrected change in tense could, in some circumstances, have a significant impact on the meaning of the utterance. However, the morphological alterations in the data were all relatively clear; either the change in tense was held in stark contrast to the tense used in the rest of the utterance, or it was coupled with another error which would indicate that the section needs closer review.

The remaining two errors were relatively easy to identify from the context of the utterance; the utterance-final phrase “I don’t **know**” was mistranscribed as “I don’t **you**” and “a really big **yew tree** right next to it” was mistranscribed as “a really big **utility** right next to it”. A much bigger proportion (11/16) of the WYE content-based substitutions involved non-morphological alterations, but the majority of these were easy to identify from context alone, such as the phrase “it’s bit uh cut and chop with staff” which was transcribed by Amazon as “it’s bitter cutting chocolate stuff”. The words directly preceding this part of the utterance referenced the frequent hiring of new staff, therefore the reference to “cutting chocolate” seems misplaced in this context. Other WYE substitutions included “airport” in place of “giving him an **haircut**” and “moody” in place of “when it rains it gets very **muddy**”.

In the speech-shaped noise condition, there were a very similar number of content-based substitutions in SSBE (5) while the number increased substantially for WYE from 16 to 29, only six of which involved morphological alterations. The rest of the errors were relatively clear from context, e.g. “I had a bit of **dessert**” → “I had a bit of **Giza**” when talking about lunch or “did have a **sack of potatoes**” → “did have a **sacrum tears**”, making them easy to identify when comparing the audio recording and the ASR transcript, and potentially even from simply reading the transcript through without audio.

The substitution of function words could be more difficult to detect in some cases as short grammatical words are generally paid little conscious attention and glossed over in reading tasks (Van Petten & Kutas, 1991; Chung & Pennebaker, 2007), and the meaning of the utterance often remains unchanged. For example, there is little difference between “go **in** get my drinks” and “go **and** get my drinks” in the context of visiting a pub. Substitutions involving function words featured around 10 times in SSBE and 16 times in WYE in both audio qualities, and the majority of these were relatively inconsequential, e.g. “**the** steak house” →

“a steak house” and “**that's** quite hard to see” → “**it's** quite hard to see”. However, a number of the errors involved the substitution of pronouns (see Table 6), which could be extremely difficult to notice due to similar pronunciations, but could be problematic within a forensic context if left uncorrected.

Accent	Reference transcript	Automatic transcript
SSBE	I couldn't put a name to a face	I couldn't put a name to her face
SSBE	I might have known them	I might have known him
YE	Uh can get a bit inebriated	You can get a bit inebriated

Table 6: Examples from the data of substitution errors involving pronouns. Words involved in substitution are highlighted in bold red text.

5 Discussion

5.1 ASR performance

The present study set out to investigate the reliability of ASR transcripts with simulated police interview recordings by exploring the impact of regional accent and audio quality on the transcription performance of three popular commercially-available ASR systems. Results revealed that the ASR system used and the audio quality of the recording had a significant effect on word error rate, and though regional accent was not found to significantly predict WER, clear differences were observed across the two accents in terms of the frequency and types of errors made.

5.1.1 ASR system and audio quality

With regards to the commercial ASR systems chosen for this study, Amazon Transcribe was clearly the best-performing system, consistently achieving the lowest WER in each condition: 13.9 and 15.4% for SSBE in studio quality and the speech-shaped noise condition, respectively, and 19.2 and 26.4% for WYE in studio quality and the speech-shaped noise condition, respectively. Google Cloud Speech-to-Text achieved the highest WER in almost every condition, and error rates for this ASR system were significantly higher than those for both Amazon and Rev, as well as consistently above 30%. Rev AI had the most variable performance, ranging from 19.0 to 42.5%. The patterns observed across accents and audio

qualities were relatively consistent within each system, but the specific reason behind the difference in performance across systems is not clear, especially given the “black box” nature of proprietary automatic systems. The addition of speech-shaped noise to the audio recordings was found to have a significant effect on word error rate, leading to a higher frequency of errors in almost every condition. However, it must be noted that Amazon Transcribe, the best performing system, was the least affected by the addition of speech-shaped noise, with WERs increasing by only 1.5% in SSBE and 7.2% in WYE between the two audio qualities.

5.1.2 Regional accent

Word error rate was not found to be significantly impacted by regional accent in this study, although this was likely due to variation between speakers and the small sample size. A clear pattern emerged whereby one speaker of each accent was favoured by the ASR systems, and performance for the best WYE speaker was roughly level with performance for the worst SSBE speaker.

Variation in system performance within an accent group has recently been investigated by Harrington and Hughes (2023), a study in which test data from a sociolinguistically-homogenous group was transcribed using Amazon Transcribe. Despite demographic factors such as age, accent and educational background as well as the content of the recordings being relatively controlled, a high level of variability was observed across speakers, with word error rates ranging from 11 to 33%. The variation across speakers observed in this study is therefore unsurprising, although the systematic effects of variety may emerge on a larger data set, as reported by Markl (2022).

Despite the lack of a statistically significant difference in WER across the accents, a higher number of errors were produced for the West Yorkshire English speech compared with the Standard Southern British English speech, and the majority of errors for the non-standard regional accent involved the substitution of words or phrases. Substitution errors can be extremely damaging in forensic contexts, particularly when the quality of the audio is poor. It is possible that deletion and insertion errors will be easier to identify alongside the audio within a transcript, but if the listeners have been “primed” by an alternative interpretation of a word or phrase (i.e. a substitution) then the identification of that error will in all likelihood be more challenging.

There are a number of factors likely contributing to the disparity in performance between accents. Modern ASR systems tend to involve two components, an acoustic model and a language model. Research on performance gaps between accent groups suggests that many ASR performance issues concerning “accented” speech stem from an insufficiently-trained acoustic model, which is caused by a lack of representation of non-standard accents in training data (Vergyri et al., 2010; Dorn, 2019; Markl, 2022). There were many errors within the Yorkshire data that can be attributed to a phonetic realisation deviating from SSBE, a large number of which involved vowels for which phonemic and realizational differences are observed across the accents. Numerous errors were likely the result of a combination of vocalic and consonantal differences between SSBE and WYE; for example, the combination of h-dropping and a Northern realisation of the STRUT vowel in “haircut” led to substantial substitutions by two of the systems.

Although the main focus of the fine-grained phonetic analysis was on errors seemingly caused by issues with the acoustic model, there were some errors that could not be attributed to acoustics and instead were likely a reflection of the language model. The language model calculates the conditional probability of words in a sequence, i.e. how likely is it that word D will follow on from words A, B, and C. Utterances containing non-standard grammar are therefore likely to cause problems for ASR systems, a few examples of which were observed in the Yorkshire data. The lack of a subject pronoun in the utterance “did have a sack of potatoes in front” led to the insertion of the pronouns “I” and “you” by Rev and Google respectively, both positioned after the verb “did”. The omission of the determiner in the phrase “in the front” led to the insertion of the verb “is” before this phrase, i.e. “is in front”, by both Rev and Google. Another example of an error likely resulting from the language model is the insertion of the indefinite article into the phrase “from secondary school”, transcribed by Amazon as “from a secondary school”. Having reviewed the audio, there is no phonetic explanation for this insertion given that the nasal [m] is immediately followed by the fricative [s], therefore this insertion is likely due to the language model calculating that the sequence of words including “a” is more probable.

5.1.3 Error analysis

A WER of 5% is generally accepted as a good quality transcript (Microsoft Azure Cognitive Services, 2022) but if the errors within that transcript lead to significant changes to the content, then that transcript could be harmful in a court of law. WER alone cannot indicate whether a system is good enough to use in a legal setting, such as the transcription of police-suspect interviews. Fine-grained phonetic analysis of the errors produced is a much

more informative approach that can highlight any major issues with a system such as high rates of substitution errors. This type of analysis could also help to identify common issues in ASR transcripts that could subsequently be built into training for police transcribers, if a computer-assisted approach to police-suspect interview transcription was adopted. However, this method of analysis is extremely labour-intensive in nature and is therefore not feasible for large data sets. A combination of the two approaches, in which WER is calculated for a large data set and a subset of the data is subject to more detailed analysis of the frequency, type and magnitude of the errors, may be more suitable.

5.2 Post-editing

One of the aims of this paper is to investigate the possibility of incorporating automatic transcription into the production of police interview transcripts. The transcripts produced by the three commercial ASR systems in this experiment would not be suitable for use without manual correction, which is to be expected given that this is a commonly acknowledged issue in the field of automatic speech recognition (Errattahi et al., 2018). The question to be addressed is therefore whether the automatic transcripts could act as a first draft which is then reviewed and corrected by a human transcriber.

Gaur et al. (2016) found that editing an ASR output actually takes longer than producing a transcript from scratch once the WER surpasses 30%. Given that the average WER for Google exceeded 30% in every condition, and in all but one condition the WER for Rev was more than 29%, neither of these systems would be adequate for the purpose of producing a first draft of a transcript to be corrected by a human transcriber. In contrast, WERs produced by Amazon ranged from 13.9 to 26.4%, falling into the range of “acceptable but additional training should be considered” according to Microsoft Azure documentation (Microsoft Azure Cognitive Services, 2022). Gaur et al. (2016) found that participants benefited from the ASR transcript provided the word error rate was low, i.e. below 30%. It is therefore possible that utilising the Amazon transcripts as a first draft to be edited could reduce the time necessary to produce verbatim transcripts.

Closer inspection of the transcripts produced by Amazon revealed that many of the errors should, in principle, be easy to identify or would be relatively inconsequential if left uncorrected. For example, over 50% of the deletion errors in studio quality audio involved the omission of filled pauses like “uh” and “um”, which is unlikely to have a substantial effect on the reader's perception of the speech and the speaker. Most deletions in speech-shaped noise audio involved short function words, and in almost all cases the meaning of the

utterance was unaffected by their omission. Insertions were very rare within the data but were quite easily identifiable from context or were paired with a substitution error. The substitution of content words, particularly for the Yorkshire English speech, was generally evident from context since the resulting transcript was often ungrammatical or nonsensical, and substitution errors involving function words generally made no difference to the meaning of the utterance. The exception to this was the substitution of pronouns and content words with morphologically-related terms (though cases of the latter in this data were relatively easy to identify); these errors would likely be much harder to spot due to the phonetic similarity between the word uttered and the substituted term.

5.2.1 Potential challenges

A potential challenge with the task of correcting a transcript is that post-editors could be “primed” (i.e. heavily influenced) by the content of the ASR output to such an extent that errors go unnoticed. Research in the field of forensic transcription has found that seeing an inaccurate version of a transcript can cause people to “hear” the error in the audio (Fraser et al., 2011; Fraser and Kinoshita, 2021). However, the quality of audio recordings in forensic cases is often extremely poor and the speech is “indistinct”, resulting in a reliance on top-down information such as expectations about the speech content (Fraser, 2003). In the case of police-suspect interviews, where the audio quality is often relatively good in comparison to forensic recordings, transcribers may be less susceptible to the effects of priming. It is also worth noting that many of the errors produced by the ASR systems were easy to identify from contextual knowledge or due to the nonsensical nature of the substitution. For example, one ASR transcript contained “giving him an airport” in place of “given him an (h)aircut” which, despite the similar phonetic content, is unlikely to influence a post-editor due to the implausibility of the utterance. Minor deletion errors, such as the omission of filled pauses, could be more challenging to identify in a transcript, though in many cases this would likely be inconsequential with regards to the readability of the transcript and the reader's perception of the speech and speaker.

Another potential issue is that errors in transcripts with a low WER may be more difficult to identify. As suggested by Sperber et al. (2016), post-editors may miss errors due to a lack of attention, and this effect would likely be increased in cases where the transcript is almost completely accurate and an excessive amount of confidence is placed in the performance of the automatic system. It is possible that the user interface employed could help to address this problem. Sperber et al. (2016) suggest two methods for focusing transcriber attention and therefore decreasing the chance of missing transcription errors: highlighting

low-confidence words in red, and typing from-scratch with the ASR hypothesis visible. Both methods were shown to improve the quality of the transcript (i.e. decrease WER) and reduce the time taken, and it was also found that different strategies work best for different levels of WER. Retyping (with the ASR output visible) gave the best results for segments with a high WER, while editing the ASR transcript text gave the best results for low WER segments.

5.3 Future work

This study used a small sample of commercially-available automatic speech recognition systems and has shown that not all ASR systems are suitable for the task of producing a “first draft” transcript, as evidenced by the frequency of errors produced by Rev AI and Google Speech-to-Text. However, promising performance was demonstrated by one of the systems tested and further analysis of the errors suggests that post-editing an ASR transcript, provided it is of adequate quality, is a worthwhile topic to explore in the context of police-suspect interviews. This approach could facilitate the production of verbatim transcripts of interviews without a substantially higher time requirement than the current practice of summarising the majority of the recording.

Future work on this topic should focus on two areas: ASR performance in a range of audio, speaker and speech conditions, and post-editing. In the present study, the addition of speech-shaped noise to the recordings may not have created an audio quality representative of real police-suspect interview data. It would therefore be interesting to use real recordings to investigate the capabilities of this technology. Other factors that may impact the system's performance and would be present in police-suspect interviews include different levels and types of background noise, multiple speakers, other regional accents, and longer stretches of speech.

More research is also required on the topic of post-editing. Papadopoulou et al. (2021) claims to be one of the first studies to employ qualitative analysis on automatic transcription errors and to evaluate the post-editing effort required in correcting ASR transcripts. Incorporating ASR outputs into the transcription process has been investigated by others, though these studies tend to focus on optimising efficiency (Sperber et al., 2016; Sperber et al., 2017) or simply report on the use of a computer-assisted transcription approach, e.g. for meetings of the National Congress of Japan (Akita et al., 2009) or for speeches in the Icelandic parliament (Fong et al., 2018). Transcripts have been found to be highly influential on the perception of speech content when the audio quality of the recording is extremely poor, but more research is required on the priming effects of ASR transcripts in the context

of post-editing police-suspect interviews, i.e. on comparatively better quality audio. Furthermore, it is crucial to investigate the practicalities of correcting an ASR transcript of a police-suspect interview. For example, how many errors are missed by post-editors, and what are the consequences of leaving those errors in the transcript? How long does it take to correct an ASR transcript of a full police-suspect interview, and how does this compare to the current time taken to create ROTIs? Future research should explore these questions as the incorporation of automatic speech recognition into the transcription process could be extremely beneficial.

6 References

- Akita, Y., Mimura, M., & Kawahara, T. (2009). Automatic transcription system for meetings of the Japanese national congress. *Tenth Annual Conference of the International Speech Communication Association*.
https://www.academia.edu/download/48519120/Automatic_transcription_system_for_meeti20160902-8378-1nvcdb7.pdf
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using “lme4.” *Journal of Statistical Software*, 67(1), 1–48.
- Boersma, P., & Weenink, D. (2022). *Praat: doing phonetics by computer [Computer program]* (Version 6.3.03). <http://www.praat.org/>
- Bokhove, C., & Downey, C. (2018). Automated generation of “good enough” transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, 11(2), 205979911879074.
- Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. *Social Communication*, 1, 343–359.
- Coulthard, M. (2013). The official version: Audience manipulation in police records of interviews with suspects. In *Texts and practices* (pp. 174–186). Routledge.
- DiChristofano, A., Shuster, H., Chandra, S., & Patwari, N. (2022). Performance Disparities Between Accents in Automatic Speech Recognition. In *arXiv [cs.CL]*. arXiv.
<http://arxiv.org/abs/2208.01157>
- Dorn, R. (2019). Dialect-specific models for automatic speech recognition of African American vernacular English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 16–20.
- Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128, 32–37.
- Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying Bias in Automatic Speech Recognition. In *arXiv [eess.AS]*. arXiv.

<http://arxiv.org/abs/2103.15122>

- Finnegan, K., & Hickey, R. (2015). Sheffield. *Researching Northern English*, 227–250.
- Fong, J. Y., Borsky, M., Helgadóttir, I. R., & Gudnason, J. (2018). Manual Post-editing of Automatically Transcribed Speeches from the Icelandic Parliament - Althingi. In *arXiv [eess.AS]*. arXiv. <http://arxiv.org/abs/1807.11893>
- Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech Language and the Law*, 10(2), 203–226.
- Fraser, H. (2020). Forensic transcription: The case for transcription as a dedicated branch of linguistic science. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 416–431). Routledge.
- 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.
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2). <https://doi.org/10.1558/ijsl.v18i2.261>
- Gaur, Y., Lasecki, W. S., Metze, F., & Bigham, J. P. (2016). The effects of automatic speech recognition quality on human transcription latency. *Proceedings of the 13th International Web for All Conference*, Article Article 23.
- Gillick, L., Baker, J., Baker, J., Bridle, J., Hunt, M., Ito, Y., Lowe, S., Orloff, J., Peskin, B., Roth, R., & Scattoni, F. (1993). Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 471–474 vol.2.
- Godfrey, J., & Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. *Linguistic Data Consortium*.
- Gold, E., Ross, S., & Earnshaw, K. (2018). The “west Yorkshire regional English database”: Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Interspeech 2018*, 2748–2752.
- Hain, T., Woodland, P. C., Evermann, G., Gales, M. J. F., Liu, X., Moore, G. L., Povey, D., & Wang, L. (2005). Automatic transcription of conversational telephone speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 13(6), 1173–1185.
- Harrington, L. & Hughes, V. (2023). Automatic speech recognition: system variability within a sociolinguistically homogenous group of speakers. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic 2023.
- Harrington, L., Love, R., & Wright, D. (2022, July). *Analysing the performance of automated*

- transcription tools for covert audio recordings*. Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.
- Harrison, P., & Wormald, J. (in press). Forensic transcription and questioned utterance analysis. In F. Nolan, T. Hudson, & K. McDougall (Eds.), *Oxford Handbook of Forensic Phonetics*. Oxford: OUP.
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, 22(4), 428–450.
- Haworth, K. (2020). Police interviews in the judicial process: Police interviews as evidence. In *The Routledge handbook of forensic linguistics* (pp. 144–158). Routledge.
- Hickey, R. (2015). *Researching northern English* (R. Hickey (ed.); pp. 1–493). John Benjamins Publishing. <https://doi.org/10.1075/veaw.g55>
- Hughes, A., Trudgill, P., & Watt, D. (2005). *English accents and dialects: An introduction to social and regional varieties in the British Isles*. London: Trans. Atlantic Publications, Inc.
- Jenks, C. J. (2013). Working with transcripts: An abridged review of issues in transcription. *Language and Linguistics Compass*, 7(4), 251–261.
- Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., & Dawson, L. (2014). A systematic review of speech recognition technology in health care. *BMC Medical Informatics and Decision Making*, 14, 94.
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7684–7689.
- Kowal, S., & O'Connell, D. C. (2014). Transcription as a crucial step of data analysis. *The SAGE Handbook of Qualitative Data Analysis*, 64–79.
- Lima, L., Furtado, V., Furtado, E., & Almeida, V. (2019). Empirical Analysis of Bias in Voice-based Personal Assistants. *Companion Proceedings of The 2019 World Wide Web Conference*, 533–538.
- Lindsey, G. (2019). *English After RP: Standard British Pronunciation Today*. Springer.
- Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1), 1–15.
- 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 <https://apps.dtic.mil/sti/citations/ADA414420>
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7. <https://doi.org/10.3389/fcomm.2022.803452>

- Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 521–534.
- Meyer, J., Rauchenstein, L., Eisenberg, J. D., & Howell, N. (2020). Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6462–6468.
- Microsoft Azure Cognitive Services. (2022). *Test accuracy of a Custom Speech model*.
<https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio>
- Mishra, A. (2021). *What is Rev AI's Accuracy?* Rev.
<https://help.rev.ai/en/articles/3813288-what-is-rev-ai-s-accuracy>
- Mošner, L., Wu, M., Raju, A., Krishnan Parthasarathi, S. H., Kumatani, K., Sundaram, S., Maas, R., & Hoffmeister, B. (2019). Improving Noise Robustness of Automatic Speech Recognition via Parallel Data and Teacher-student Learning. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6475–6479.
- Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*.
<https://journal.equinoxpub.com/IJSLL/article/view/10005>
- O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10), 2965–2979.
- Papadopoulou, M. M., Zaretskaya, A., & Mitkov, R. (2021). Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis. *Proceedings of the Translation and Interpreting Technology Online Conference*, 199–207.
- Paulus, T., Lester, J., & Dempster, P. (2013). *Digital Tools for Qualitative Research*. SAGE.
- Punch, K. F., & Oancea, A. (2014). *Introduction to Research Methods in Education*. SAGE.
- Siniscalchi, S. M., & Lee, C.-H. (2021). Automatic Speech Recognition by Machines. In R.-A. Knight & J. Setter (Eds.), *The Cambridge Handbook of Phonetics* (pp. 480–500). Cambridge: Cambridge University Press.
- Sperber, M., Neubig, G., Nakamura, S., & Waibel, A. (2016). Optimizing Computer-Assisted Transcription Quality with Iterative User Interfaces. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1986–1992.
- Sperber, M., Neubig, G., Niehues, J., Nakamura, S., & Waibel, A. (2017). Transcribing against time. *Speech Communication*, 93, 20–30.
- Stoddart, J., Upton, C., & Widdowson, J. D. A. (1999). Sheffield dialect in the 1990s:

- revisiting the concept of NORMs. *Urban Voices: Accent Studies in the British Isles*, 72–89.
- Stolcke, A., & Droppo, J. (2017). Comparing Human and Machine Errors in Conversational Speech Transcription. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/1708.08615>
- Strycharczuk, P., López-Ibáñez, M., Brown, G., & Leemann, A. (2020). General Northern English. Exploring Regional Variation in the North of England With Machine Learning. *Frontiers in Artificial Intelligence*, 3, 48.
- Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczyk, J., Augustyniak, L., Mizgajski, J., & Carmiel, Y. (2020). WER we are and WER we think we are. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2010.03432>
- Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59.
- 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*. Conference of the International Investigative Interviewing Research Group, Winchester.
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test*. Conference of International Association for Forensic Phonetics and Acoustics, Santander, Spain.
- Tüske, Z., Saon, G., & Kingsbury, B. (2021). On the limit of English conversational speech recognition. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2105.00982>
- Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, 19(1), 95–112.
- Vergyri, D., Lamel, L., & Gauvain, J.-L. (2010). *Automatic speech recognition of multiple accented English data*. www-tlp.limsi.fr.
http://www-tlp.limsi.fr/public/automatic_speech_recognition_of_multiple_accented_english_data_vergyri.pdf
- Walford, G. (2001). *Doing Qualitative Educational Research*. Bloomsbury Publishing.
- Walker, A. G. (1990). Language at work in the law. In *Language in the Judicial Process* (pp. 203–244). Springer US.
- Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. *Symmetry*, 11(8), 1018.
- Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140, 50–70.
- Watt, D., & Tillotson, J. (2001). A spectrographic analysis of vowel fronting in Bradford English. *English World-Wide*, 22(2), 269–303.
- Wells, J. C. (1982). *Accents of English* [3 v. : ill. ; 23 cm.]. Cambridge University Press.
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G.

(2016). Achieving Human Parity in Conversational Speech Recognition. In *arXiv [cs.CL]*.
arXiv. <http://arxiv.org/abs/1610.05256>

Zayats, V., Tran, T., Wright, R., Mansfield, C., & Ostendorf, M. (2019). Disfluencies and
Human Speech Transcription Errors. In *arXiv [cs.CL]*. arXiv.
<http://arxiv.org/abs/1904.04398>

Article 3 - Appendix A

Linguistic content of stimuli

Reference transcripts for stimuli used in experiment. 'D' represents speakers from the DyViS project and 'W' represents speakers from the WYRED project. The number reflects the speaker's identification number in the corpus.

Speaker	Accent	Reference transcription
D002	SSBE	Er no I don't think so actually my phone line has er been cut off
D002	SSBE	We sometimes go and eat together at the steak house he's also a barber there
D002	SSBE	Right well it could have been somebody from work that I was just giving a lift back maybe
D002	SSBE	Erm there's some other shops the er city tour bus leaves from there as well
D002	SSBE	Yeah I may have seen him around but I couldn't put a name to a face
D002	SSBE	And erm there's also a boat house but that's obviously that's quite hard to see from there
D002	SSBE	I really don't remember him he may have been at my school but I I can't really remember
D002	SSBE	Well she hasn't actually passed her test yet actually to be honest but erm she's quite environmentally conscious as well
D008	SSBE	Erm well I try and catch the news normally and er the weather for the next few days
D008	SSBE	She's oh she's got this adorable poodle uh it's very cute and er she drives her scooter to work
D008	SSBE	Not exactly I can't really remember their surnames but I might have known them I don't know

D008	SSBE	And erm on the right there's a big reservoir er and a a really big yew tree right next to it
D008	SSBE	Erm she lives on the same street as me and so sometimes we go for a drink after work
D008	SSBE	And then there's a deer park and a there's a boat house on the river
D008	SSBE	I start at 8 o'clock every day and I finish about 5 o'clock or just past
D008	SSBE	Oh I relaxed erm watched some tv there was something I wanted to watch on
W107	WYE	Er did have a sack of potatoes in front could have been that but erm
W107	WYE	Er just to get a bit of fuel you know some Doritos and that
W107	WYE	I don't know they hire a lot of lot of newcomers it's bit er cut and chop with staff
W107	WYE	Given him an haircut once or twice when he's come round but not er not seen him too much
W107	WYE	Not too much of a social butterfly me you know just go in get me drinks
W107	WYE	Er can get a bit inebriated sometimes so not all the time no can't say
W107	WYE	Yeah it's alright it's alright except for when it rains it gets very muddy
W107	WYE	Yeah I had me lunch I had er I had a bit of dessert let me food settle
W110	WYE	Erm he's a tour guide and er I knew him from secondary school er we regularly chat on Skype

W110	WYE	Ok no not that I'm aware of not that I drove I drive quite a lot on my own
W110	WYE	Just a main road erm and the city bus tour leaves departs from outside the shop
W110	WYE	I might do but I'm not in enough to recognise faces staff change don't they and
W110	WYE	She might do I don't think it's that one I've never seen her there
W110	WYE	Er I was but I was on my own just went for a quiet walk
W110	WYE	Ah he does but it's a it's a fairly big place you know you don't run into people that often
W110	WYE	Yeah quarter of an hour half an hour something like that depending on traffic

8. Discussion

The research in this thesis covers a broad range of topics concerning the production of transcripts within the criminal justice system, with a focus on different types of audio materials (poor quality evidential recordings, police-suspect interview recordings) and different transcribers (experts, non-experts, automatic systems). The common focus within all of the research is the methods employed to produce transcripts which will later be used in court alongside, or even in place of, audio recordings of forensic interest.

This discussion is centred on the aims presented in section 3. Each of the three aims will be explored in turn, considering the way in which the research in this thesis has addressed the aim and assessing the implications of the findings. After this, a summary of the overall findings is presented and recommendations are provided for a number of different target audiences.

8.1 Aim 1: to gain a better understanding of current practices

The first aim of this thesis is to gain a better understanding of the practices currently used to produce transcripts for use within the criminal justice system. More specifically, this aim concerns the transcription of evidential recordings, given that police-suspect interview transcription and the production of ROTIs (Record of Taped Interview; i.e. police interview transcripts) has been previously explored by researchers at Aston University (see Haworth et al., 2023). The research presented in Article 1 and the Additional Resource has explored the methods used to produce transcripts of poor quality evidential recordings by experts and non-experts, through both direct and indirect means.

8.1.1 Expert procedures

There are aspects of expert transcription procedures that have been discussed within the forensic speech science community, such as the fact that multiple transcribers working on a transcript may be better than an individual transcriber (Tschäpe & Wagner, 2012) and that contextual information can have an impact on the way in which speech is perceived (Lange et al., 2011). However, prior to this research, approach-level differences in the procedures employed by practitioners and different labs had not yet been explored, certainly not in the same way that the methods employed for forensic speaker comparison have been (e.g. Gold & French, 2011; Morrison et al., 2016; Gold & French, 2019).

This thesis set out to explore the procedures employed by expert practitioners in the production of transcripts of poor quality evidential audio recordings by conducting a survey of international forensic transcription practices. The findings of the survey demonstrate that different approaches are taken by different laboratories or individual practitioners, though some aspects are similar across the majority of respondents. For example, most practitioners produce multiple drafts of a transcript, and multiple practitioners are involved in the production of transcripts in almost all cases where this is possible. However, there were also areas of the methods which showed a clear divide between practitioners, namely the way in which drafts are produced and whether existing transcripts should be consulted; these areas in particular should be subjected to further research in order to establish which aspects contribute to the production of the most accurate, reliable and impartial transcripts.

What is most clear from the results of the survey is that forensic transcription is an area that is severely lacking the research needed in order to establish the most robust methods for producing transcripts of poor quality evidential audio materials. Practitioners are using the approaches that they believe to be best, synthesising findings from other areas of linguistic science and the forensic sciences in general, or that they have simply used for a long time. This is not the fault of the practitioners; without a considered effort within the research community towards the study of forensic transcription methods, this is the most that can currently be achieved.

The European Network of Forensic Science Institutes (ENFSI) has produced Best Practice Manuals for the Methodology of Speaker Comparison (2022) and Digital Audio Authenticity Analysis (2022), as well as Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition (2015) and Best Practice Guidelines for Electric Network Frequency Analysis in Forensic Authentication of Digital Evidence (2009)²⁷. However, forensic transcription is not an area that has received the same consideration, despite this task being frequently carried out by forensic practitioners according to the results of the survey presented in Article 1²⁸. Practitioners would likely benefit from an increased push towards developing guidelines or standards regarding how to produce the most reliable and impartial transcripts, given that this would entail much needed further research on the

²⁷ These documents are available under Forensic Speech and Audio Analysis on the following page from the ENFSI website: [Best Practice Manuals and Forensic Guidelines | ENFSI](#)

²⁸ Transcription is often considered as the second most frequently carried out task by forensic practitioners (following forensic speaker comparison). Transcription is involved in around one third of the casework load carried out by a forensic speech and audio lab in York, UK (Richard Rhodes, personal communication).

under-explored topic of forensic transcription. Relatively informal guidelines (rather than, for example, an official ENFSI best practice manual), which could be produced by a working group of the *International Association for Forensic Phonetics and Acoustics*, would be a very positive starting point.

Further research could establish which approaches to transcribing poor quality recordings have advantages over others in terms of efficiency, impartiality or overall accuracy, and would likely lead to the development of proficiency testing. One of the issues with forensic transcription is that the ground truth is very rarely, if ever, known, and practitioners therefore cannot know if the interpretation offered within their transcript was correct, even if it was accepted by the court. Proficiency testing could help to give confidence to transcribers, as well as the end users of the transcripts, in knowing that the practitioner is highly competent.

A number of resources are required for proficiency tests, including forensic-like audio recordings where the ground truth of what was said is known, and a framework for evaluating the transcripts produced; a novel method of analysing transcripts for use in forensic contexts has been developed within this thesis and could be used within proficiency testing as (a) a way of comparing the content and quality of transcripts across different demographics, e.g. experts versus non-experts, and (b) a training tool for experts, which would enable practitioners to assess where they made errors and what kinds of errors were made. Proficiency testing would also provide a foundation for developing and testing the validity of methods, which is of particular interest to practitioners working within England and Wales (see section 1.1.1 for a re-cap on issues related to the UK Forensic Science Regulator).

Finally, the findings of future research on expert transcription practices could also be used to demonstrate to non-experts, such as transcribers within police forces, the importance of the type of methods used to produce transcripts. Such findings could help to bring awareness to the issues surrounding transcription and methods of transcript production, and therefore incite positive change in the working practices of non-expert transcribers.

8.1.2 Non-expert procedures

Little is known about non-expert procedures for the production of transcripts of poor quality evidential recording. A focus interview with forensic practitioners carried out as part of this

thesis revealed that there is a lot of variability in the quality of these non-expert transcripts²⁹. Given the wide range of approaches with regard to the content, layout and transcription conventions (if there are any), it is reasonable to assume that there is no standardisation and little guidance, if any, provided to transcribers in non-expert contexts. Common issues revealed in the interview include a lack of time points within the transcript, inconsistent ways of representing speakers and speech transcribed in a big block with no separation between utterances and no indication of the time point or who is speaking. The provision of resources such as a transcript template and transcriber training could contribute to an improvement in the quality of non-expert transcripts; Appendix A contains a document presenting a template and transcription conventions for non-expert transcribers to use, as well as guidance on how to fill out the template and other transcription-related advice. The template(s) and conventions provided in Appendix A are based on those used by practitioners at the Forensic Voice Centre³⁰ in York, UK. For reference, a blank template is shown in Table 1.

Line	Time point	Speaker	Content
1			
2			
3			

Table 1: Blank transcript template that could be used by non-expert transcribers. See Appendix A for guidance on how to use the template.

The resource presented in Appendix A could be incredibly helpful for non-expert (or even expert) transcribers who are lacking guidance. The presence of a ‘Time point’ column would likely encourage more frequent records of the time at which utterances take place within the recording, and brief guidance on how to represent speakers in the ‘Speaker’ column could promote the use of systematic ways to represent speakers: e.g. “M1” and “M2” for unidentified male speakers, “F1” and “F2” for unidentified female speakers, “Call operator” or “Police officer” in cases where the identity of the speaker is clear and not questioned. This type of layout also encourages utterances to be represented in separate lines rather than in one large block. Transcription conventions and a short guide on how to fill out the template

²⁹ It is worth noting that practitioners are exposed to a limited sample of non-expert transcripts of poor quality evidential audio recordings. Sometimes the instructing party will provide an existing transcript, or an audio recording and accompanying transcript may be provided for use within a different type of analysis, e.g. speaker comparison or audio authentication. However, it is likely that the sample they have seen is representative of the wider state of non-expert transcripts.

³⁰ <https://forensicvoicecentre.com/>

could also lead to consistency in the symbols used, such as ellipses which can currently represent either pauses or speech omitted from the transcript, as well as the potential inclusion of levels of confidence and indications of summarising. One of the practitioners interviewed recently had an interaction with a police officer which suggested that there is currently no standard layout for transcripts produced by police but that there is a desire for it. The next step in this area of research is to work with police forces in England and Wales, surveying their attitudes towards transcription and current methods of transcript production, with an ultimate goal of developing national standards as well as providing guidance and training to non-expert transcribers.

The above suggestions are based on knowledge of the UK context, but similar templates and training could be provided to non-expert transcribers in other countries, where transcription of poor quality evidential recordings by police officers also seems to be common (e.g. French & Fraser, 2018; Cenceschi & Meluzzi, 2023). The templates provided in this thesis may also be of interest to forensic practitioners on an international scale, e.g. for comparison with their own and to initiate discussion regarding best practice regarding layout and content.

Much more is known about the production of police-suspect interview transcripts, also called ROTIs (Record of Taped Interview), thanks to the efforts of Kate Haworth and her team at Aston University (for a brief project summary, see Haworth et al. (2023)). Such transcripts are often viewed as a ‘copy’ of the audio recording of the police-suspect interview and therefore end up being used interchangeably with the audio. In many cases, the audio recording is not referred to at all, with the transcript instead essentially being presented as the evidence to the court. This is problematic for a number of reasons; a transcript cannot capture paralinguistic and extralinguistic features, such as the tone of voice, and ROTI clerks (the transcribers of police interviews) are careful to avoid including references to emotion (Haworth et al., 2023). Without the audio recording to convey such information, the jury has to generate their own interpretation of the speech content.

Furthermore, in many cases, the ROTI contains mostly summaries of the interview content, with some sections (e.g. parts that may be evidentially relevant) transcribed in full. Summarising is “a highly selective and subjective process” (Haworth et al., 2023, p. 2) and is problematic for a number of reasons:

- The speech content may not be accurately summarised, i.e. content may be omitted as it is not considered relevant.

- The use of non-neutral reporting verbs, such as “admitted” instead of “stated”, can imply guilt on the part of the interviewee.
- The choice of which parts to transcribe in full (that could help in an investigation) is made by ROTI clerks who do not have knowledge about, are not required to have knowledge about, and are not trained in legal matters.
- A summary is a transcriber’s interpretation of the speech content, and so readers of a ROTI may be exposed to someone else’s interpretation of what was said, rather than what was actually said.
- Transcribers employ different methods of summarising content, especially given the lack of guidelines and training for ROTI clerks, and so there is a lot of variability in the transcripts produced.

A solution to this is the production of verbatim transcripts for the interviews that are used in the evidentiary stage of the legal process. This approach would not be appropriate for all police-suspect interviews given that, in many cases, the interview record is a formality and is not used for further investigation or beyond. Verbatim transcription of police-suspect interviews is not necessarily desirable or appropriate for the initial production of police interview records. However, when the speech content (i.e. the specific words that are spoken) is of interest and plays an evidential role in a case, verbatim transcripts (as proposed in Article 3) would be preferable in the interests of justice, given that a verbatim transcript achieves a higher level of accuracy and impartiality than a summarised report.

Transcription is a time-consuming process, and so future methods could incorporate automatic methods as a way of saving time and effort on the part of the transcriber. Article 3 presented a preliminary exploration into whether the incorporation of such methods into the transcription of police-suspect interviews could be viable, and suggested an approach whereby a transcript produced by an automatic speech recognition (ASR) system could be used as a ‘first draft’ of the transcript and subsequently edited by a human transcriber. Findings showed that there may be potential for this type of hybrid method, given that many of the errors made by an automatic system were either relatively inconsequential if left in the transcript (e.g. “the” in place of “a”) or relatively easy to identify due to ungrammatical or nonsensical content (e.g. “giving him an airport” in place of “giving him a haircut”).

However, the study also demonstrated disparities in the types of errors produced across regional accents, with automatic systems producing more substitution errors for speech in non-standard regional accents; such variability in performance must be taken into consideration should automatic methods be employed for the purposes of transcribing police

interviews. It may be the case that current technology can only be used successfully for standard varieties such as Southern Standard British English. An automatic system could be developed for use specifically in the transcription of police-interview recordings, which is trained on relevant data from real speakers across the UK and on materials that are representative of the quality of recordings³¹. This is outside the scope of this research, but could be a solution to current commercial systems finding the accents and quality of recordings challenging.

8.1.3 Implications

Transcripts that are presented to the court must be as accurate as possible, particularly in cases where the specific words uttered by a speaker are of great evidential interest; the consequences of inaccurate transcripts being presented alongside evidential recordings can be very serious, given the powerful influence that exposure to a transcript can have on the perception of speech in poor quality audio recordings. Presentation of inaccurate interpretations of speech in poor quality audio can influence listeners to confidently hear the words in the incorrect transcript, even in the face of evidence presented by a phonetics expert (Fraser et al., 2011), and can distract from more plausible interpretations (Fraser & Kinoshita, 2021). Such misinterpretations could lead to serious miscarriages of justice, whereby innocent people are convicted and charged with years in prison for a crime they did not commit, or guilty people are wrongly acquitted.

Reliable and tested methods based upon linguistic and psychological principles are required in order to ensure that the transcripts are as accurate and as impartial as possible; however, the best methods for producing transcripts of poor quality evidential recordings and police-suspect interview recordings is a severely under-researched area. The research in this thesis has contributed to a better understanding of the methods that are currently employed by experts. Some positive findings are that the vast majority of practitioners have protocols in place to mitigate the effects of priming and cognitive bias, and there is a clear focus on these psychological phenomena throughout the procedures. There are a number of areas where different approaches are taken, namely the methods used for producing multiple drafts and the consultation of existing transcripts, which both concern the mitigation

³¹ One issue with commercial automatic speech recognition systems is that very little is known about the training data, yet this has been identified as the most likely cause for bias and disparities in performance (see section 8.2.2). It is, however, unlikely that the training data is representative of real police interviews in terms of the content and audio quality of the recordings, and it is probable that many speaker demographics are underrepresented, e.g. some interviewees have drug or alcohol dependencies which may affect their speech (Rhodes, 2012).

of priming effects. Much more empirical research on transcription procedures is needed, with a particular focus on the effects of priming and successful mitigation strategies.

A substantial issue that spans all three domains addressed in this thesis is the lack of standardisation with regard to procedures employed when producing transcripts; this results in huge amounts of variability in the content, layout, and overall quality of the transcripts produced.

In non-expert contexts, the complexity of transcription is often underestimated by those unfamiliar with the linguistic and psychological issues related to transcribing speech, and especially speech in poor quality audio; this misunderstanding leads to ‘ad hoc’ approaches to transcription, whereby each transcriber must decide on the method they believe to be best, despite their lack of knowledge about transcription or training concerning issues and different practices. As discussed in section 8.1.2, the introduction of standards and/or guidelines can help to bring attention to the task of transcription and encourage better methods that will ultimately lead to transcripts that are more accurate and impartial. A document has been provided in Appendix A that offers a template for non-expert transcribers to use, and it is accompanied by guidance on how to use the template and more generally about transcribing poor quality materials. This document can be used or further developed by police forces and other transcribers; it is hoped that the document will, at the very least, start conversations among non-expert transcribers regarding their practices and how they could be improved. In section 8.2.3, a more detailed description of potential topics for non-expert transcriber training is presented.

In expert contexts, an increased push towards investigating methods of forensic transcription and the production of guidelines would reassure experts that the methods they employ are reliable and valid. This is particularly important for practitioners working in the UK, given that they should follow the Forensic Science Regulator’s (FSR) Code of Practice, which states that all methods must be shown to be fit for purpose through validity testing (Code of Practice v1, Part D - Methods and method validation). Even for practitioners working in countries where similar regulation is not in place, the testing of methods and development of robust procedures founded on empirical research is a worthwhile endeavour. The development of proficiency testing could also reassure both the expert carrying out transcription and the users of the transcript (i.e. the customer or members of the court) that the expert is highly competent; furthermore, the UK FSR’s Code of Practice states that experts should carry out regular evaluation of their expertise (section 28.2.5). The European

Network of Forensic Science Institutes also recommends regular proficiency testing as a means of quality assurance and to develop best practice³².

More empirical research is required to investigate the best methods for producing transcripts of poor quality evidential recordings, but, in the meantime, expert transcribers (in the UK and internationally) can consider a number of factors relating to the content of their transcripts. The end user should always be considered when producing a transcript. If the transcript is produced by an expert with the purpose of only being used by the transcriber, as may be the case in forensic speaker comparison casework, then complicated conventions and transcriptions in the International Phonetic Alphabet can be included. However, if the transcript is to be used by lay people such as lawyers, judges and members of the jury, then such complicated conventions will likely cause confusion among readers, even if a key is provided explaining the conventions³³. Conventions should therefore be kept simple and, ideally, should be relatively intuitive to a reader.

8.2 Aim 2: to explore how factors relating to audio quality and regional accent can affect the content of transcripts

The second aim of this thesis is to explore how audio quality and regionally-accented speech, two factors commonly encountered within recordings of forensic interest, can affect the content of transcripts. Both factors have been shown to affect transcription performance in many studies (e.g. Lange et al., 2011; Clopper & Bradlow, 2008; Smith et al., 2014), with the findings demonstrating that poorer quality recordings and unfamiliar accents are more challenging to transcribe and therefore result in lower accuracy. It is also worth noting that these factors were considered the most influential on the perception and transcription of speech in poor quality evidential recordings by forensic practitioners in Article 1.

Many of the studies that have previously been carried out have focused their analysis on evaluating the degree of success achieved by the transcribers, i.e. word recognition accuracy scores. However, in forensic contexts, it is much more important to consider what is contained within the 'unsuccessful' parts of the transcript, i.e. the errors, given that the content of transcripts can be highly influential; this is explored in greater detail in section 8.3.

³² See ENFSI proficiency testing guidelines published in 2023: Framework for the Conduct of Proficiency Tests and Collaborative Exercises within ENFSI

³³ Tompkinson et al. (2023) found huge variability in the ability of non-linguists to identify the meaning of transcription conventions, even when a key was provided. A substantial number of respondents said that the use of additional transcription conventions actually made the transcript more difficult to read and understand.

The novelty of the experimental research in this thesis is the analytical focus on the content of the transcripts and, more specifically, the errors that are produced by transcribers (see section 8.3 for discussion of the novel approach to transcript evaluation).

Errors can be categorised into three groups:

- Deletions, whereby the transcriber has omitted a word that features in the reference material.
- Insertions, whereby the transcriber has added a word that is not featured in the reference material.
- Substitutions, whereby the transcriber has transcribed a different word to the one featured in the reference material.

In forensic contexts, substitutions can be considered as a much more problematic type of error than deletions on the assumption that “reduced information causes less damage than wrong information” (Tschäpe & Wagner, 2012). Insertions can also be considered as extremely problematic but, in practice, this type of error is produced very infrequently.

8.2.1 Audio quality

The audio quality of recordings can be affected by many things, such as transmission issues (e.g. telephone call data), compression (e.g. videos uploaded to social media), issues related to characteristics of the room (e.g. reverberation), the speakers’ distance from or orientation to the recording device and its microphone(s), and background or foreground noise. Many of these issues are encountered in recordings which are transcribed for use within the criminal justice system. For example, despite the requirement for police-suspect interviews to be audio recorded, there are often factors affecting the intelligibility of the speech such as rustling papers, the whirring of laptop fans, and reverberation as a result of the materials in the interview room (Richard Rhodes, personal communication). Evidential recordings are often poor quality due to the way in which they have been collected, e.g. via a covert recording device, in noisy environments, etc.

Given the prevalence of noise in audio recordings of forensic interest, the experiments in this thesis focus on audio quality as a function of the level of background noise. The background noise in both experiments is simulated via the use of speech-shaped noise, which shares the spectral characteristics of speech and is a commonly-used masker in psycholinguistic experiments (e.g. Adank et al., 2009; Clopper & Bradlow, 2008). Other types of masking

noise are available, such as randomly-varying cafeteria noise, babble in various languages or varieties, and amplitude-modulated speech-shaped noise; however, steady-state speech-shaped noise was chosen for the studies in this thesis to ensure that any differences in performance were a result of the regional accent of the listener (Article 2) or speaker (Article 3) rather than due to, for example, a particularly high amplitude section of masking noise.

Although steady-state speech-shaped noise was used to manipulate speech intelligibility in the experiments presented in Article 2 and Article 3, the signal-to-noise ratios (SNR) varied substantially across the studies. In Article 2, which involved human transcribers, the SNRs ranged from +6 dB to -3 dB, where a negative SNR demonstrates that the noise is louder than the target speech, and a positive SNR demonstrates that the target speech is louder than the noise. The highest SNR in this experiment, i.e. the least challenging listening condition, was +6 dB which presents only minor difficulty to the listeners, as demonstrated by average word recognition accuracy rates of around 90%. In Article 3, which involved automatic transcription systems, there were only two listening conditions, one of which comprised studio quality audio recordings with no added noise. The second (more challenging) listening condition employed a SNR of around +10 dB, such that the background noise was noticeable but, impressionistically, didn't have much of an effect on the intelligibility of the speech to human listeners.

The results of the studies presented in both Article 2 and Article 3 demonstrate that increased levels of background noise result in significantly worse performance, i.e. lower word recognition accuracy rates for humans and higher word error rates for automatic systems. This mirrors the findings of previous research on transcription in noise by both humans (Clopper & Bradlow, 2008; Lange et al., 2011) and automatic systems (Littlefield & Hashemi-Sakhtsari, 2002). It is worth noting that a significant decline in performance was observed in the performance of automatic speech recognition systems when relatively minimal noise was added to the files; the SNR of the 'challenging' listening condition in this experiment was around +10 dB, which likely would not have posed a problem for human transcribers, given that the average word recognition accuracy rates surpassed 90% in the (noisier) +6 dB SNR condition in the experiment presented in Article 2.

The decrease in word recognition accuracy is a direct result of an increased number of transcription errors in noisier conditions; and the increase in the number of errors was largely driven by an increase in deletions, particularly for the human transcribers in the study presented in Article 2. The majority of errors in all listening conditions in the human

transcription experiment were deletions, which made up an increasing proportion of the overall errors as the noise level increased (57% in +6 dB SNR to 64% in 0 dB SNR to 73% in -3 dB SNR). While the number of deletions increased in every condition in the automatic transcription experiment presented in Article 3, the effect was not as strong, which is likely the result of the higher (less challenging) SNRs in this experiment. Deletions made up the majority of automatic transcription errors for stimuli uttered in a Standard Southern British English accent (between 56 and 66% across the different ASR systems and noise conditions), but substitutions were the most frequently produced error for West Yorkshire English speech by the automatic systems. This will be explored in section 8.2.2.

The most common strategy employed to deal with increased noise therefore seems to be not attempting to transcribe the challenging material, i.e. a higher rate of deletions, at least for Standard Southern British English. The huge number of deletions produced by human transcribers as the SNR approached and passed below 0 dB suggests that the preferable option is to not attempt to transcribe the speech rather than to guess. It is possible that, even for West Yorkshire English (WYE), deletions would make up the majority of automatic transcription errors in similarly low SNRs; Loakes (2022) found that a commercial ASR system (Descript) transcribed only three words out of a possible 116 contained within a particularly challenging forensic-like audio recording, demonstrating that the strategy of not attempting to transcribe the speech in difficult recordings can be employed by ASR systems.

However, it is worth noting that a relatively large proportion of the errors in both experiments involved the substitution of words; in the human transcription experiment, substitutions made up between 23 and 34% of the overall number of errors across the listening conditions, and in the automatic transcription experiment, substitutions accounted for 33 to 50% of errors for SSBE speech and 40 to 70% of errors for WYE speech. This demonstrates a substantial number of substitution errors made in all listening conditions. Analysis of the types of substitutions that took place in the human transcription experiment showed that the majority of these errors were potentially ‘major’ in their impact, which was the case for all listening conditions.

In both studies, the types of words (i.e. function/content) involved in the substitution errors were analysed. In the human transcription experiment, there was a roughly even split between substitutions involving content words and function words, though function words were slightly more frequently involved, accounting for 52 to 55% of all substitution errors across the three listening conditions. However, in the automatic transcription experiment, the addition of background noise led to an increase in the proportion of substitution errors that

involved content words. The substitution of content words is more likely to change the meaning and perception of an utterance given that content words carry meaning while, for the most part, function words do not. Therefore, at least for automatic transcription, increased background noise can lead to more potentially damaging errors.

8.2.2 Regional accent

People may move from one part of the UK to another for reasons related to, for example, job opportunities, higher education and being closer to friends and family (Thomas, 2019). This means that people often live in locations where they did not grow up and where they do not share an accent with local speakers. It can often be the case, therefore, that speakers within evidential recordings or those taking part in a police interview talk with a regional accent³⁴ that is not shared by the transcriber. This could prove problematic, particularly when transcribing poor quality recordings, given that unfamiliar accents tend to be more challenging to transcribe in such conditions (Clopper & Bradlow, 2008; Smith et al., 2014).

With regard to the factor of ‘regional accent’, there are multiple aspects to be considered: (a) the accent of the speaker in the recording, (b) the accent of the transcriber and (c) the transcriber’s familiarity with the accent of the speaker. The last point is of particular interest in this thesis, with a human transcription experiment exploring the transcription of an accent judged to be ‘familiar’ to all listeners, and an automatic transcription experiment exploring the transcription of a ‘familiar’ standard variety and an ‘unfamiliar’ non-standard variety. Familiarity is not a term often used in the context of automatic systems, but it is broadly analogous to the assumed distribution of training data within these systems; given that ASR performance tends to be better for standard varieties, it is believed that there is an underrepresentation of non-standard accents in the training data (Vergyri et al., 2010; Dorn, 2019; Markl, 2022), such that non-standard varieties could be considered ‘less familiar’ to these systems than standard varieties.

The experiment presented in Article 2 explored whether native speakers of Standard Southern British English (SSBE) have an advantage over speakers of other varieties of British English when transcribing SSBE. All listeners in the experiment were judged to be ‘familiar’ with SSBE given the prevalence of this variety in the media and education (Lindsey, 2019); however, the results of a study by Smith et al. (2014) suggest that native SSBE

³⁴ It is worth noting that a wide range of accents - not just regional varieties of British English - can feature in recordings for transcription. This may include speakers of other World Englishes (e.g. American English, Indian English, etc.) and L2 speakers of English.

speakers may be better at transcribing their own accent than non-SSBE speakers (speakers of Glasgow English in the 2014 study). To the contrary, the results of the experiment in Article 2 reveal no significant differences in the transcripts produced by participants in the SSBE group and those in the non-SSBE group, in terms of both the overall word recognition accuracy rates and the number and types of errors made in the transcripts. This is somewhat unsurprising given the familiarity of the accent; Clopper and Bradlow (2008) carried out a transcription-in-noise task with accents of American English and found that participants from all accent backgrounds performed at a similar level and that all participants performed best for General American English compared with other accents of American English. This suggests that listeners rely on ‘standard’ acoustic-phonetic representations of words in poor listening conditions.

The next phase of experimentation on this topic should focus on the transcription of unfamiliar accents and the contents of such transcripts. Given that unfamiliar accents often involve phonetic realisations of words which do not align with the listener’s mental representation of those words, it is likely that the transcripts produced by listeners unfamiliar with the accent will be much worse than those produced by listeners who are familiar with the accent. This hypothesis is supported by multiple studies in psycholinguistics which have found worse performance in a range of speech processing and transcription experiments (Adank & McQueen, 2007; Floccia et al., 2006, Sumner & Samuel, 2009; Adank et al., 2009; Smith et al., 2014; Clopper & Bradlow, 2008). In terms of the errors within those transcripts, it may be the case that transcribers who are unfamiliar with an accent will produce more substitutions as a result of the differing phonetic realisations, given that they are more likely to mishear a word or phrase than someone who is familiar with that accent.

The automatic transcription experiment presented in Article 3 demonstrated an effect whereby more errors were produced for speech uttered in a non-standard regional accent (West Yorkshire English; WYE) than in a ‘standard’ variety (SSBE), and the distribution of errors varied substantially across accents. While the majority of errors made for SSBE speech were deletions, substitutions accounted for the majority of errors for WYE speech (65 to 70% in the two better ASR systems, Amazon and Rev³⁵). Disparity in automatic transcription performance across accents is often believed to be a result of an acoustic model that is not sufficiently trained on non-standard varieties (Vergyri et al., 2010; Dorn, 2019; Markl, 2022). As such, phonetic realisations deviating from the expected ‘standard’

³⁵ The transcription errors made by Google’s Speech-to-Text system did not follow the same pattern as the other ASR systems; the transcripts produced by Google’s ASR system contained an inflated number of deletion errors, a pattern also found by Harrington et al. (2022).

realisation of a word are unlikely to be correctly recognised, therefore leading to an increase in the number of substitution errors, as demonstrated in the findings of the automatic transcription experiment in this thesis. This is particularly worrying in forensic contexts, given that substitutions can be much more damaging than deletions.

In general, automatic systems are more likely to make transcription errors (particularly substitutions) than humans since they are unable to use contextual cues and information in the same way. This is demonstrated by a number of nonsensical transcriptions produced by the automatic systems in the experiment presented in Article 3, which were particularly prevalent in the transcriptions of WYE speech. For example, the WYE utterance “uh I was but I was on my own just went for a quiet walk” was transcribed by one of the ASR systems as “uh I lost but I was on my own swim for the quiet wall”. There were also some nonsensical transcriptions produced by human transcribers, such as “I’ve got generally goats there’s two pigs nothing special but I like it” in place of “I quite generally go it’s you know it’s cheap food it’s nothing special but I like it”, but these were much more infrequent, despite the larger number of transcribers in the human transcription experiment.

8.2.3 Implications

Given the significant impact that increased background noise and unfamiliar regional accent can have on the number and types of transcription errors contained within transcripts, these factors should be taken into consideration when deciding how certain recordings of forensic interest should be transcribed. Perhaps a ‘risk assessment’ stage could be implemented prior to recordings being transcribed, whereby the accent(s) of the speaker(s) are noted and a decision is made regarding who would be best suited to transcribe speech in that particular accent (whether that is someone within the police force, employees at another police force, or an expert). At this stage, the audio quality of the recording could also be noted, so that the transcriber is aware of the challenging nature of the recording.

Degradation of audio quality is a factor that deserves significant attention when transcribing recordings of forensic interest. Increased background noise tends to lead to higher numbers of deletions, given that the lack of bottom-up auditory information makes the task of transcription much more challenging; however, it should be noted that there was also a substantial amount of substitutions in these listening conditions, for both human and machine transcribers. In the human transcription experiment, there were almost double the number of substitution errors in the -3 dB SNR condition (the noisiest, most challenging listening condition) compared with the +6 dB SNR condition (the least challenging of the

listening conditions); furthermore, around 80% of substitutions were judged to be potentially 'major' in their impact in the -3 dB SNR condition compared with around 70% in the +6 dB SNR condition. Taken together, these findings show that the number of substitutions, and more specifically the number of substitutions that could potentially have a 'major' impact on the meaning of the utterance, increases substantially with a higher level of background noise. It is therefore crucial that transcribers are aware of the challenges of transcribing poor quality audio recordings as well as cognisant of the potential consequences of the errors contained within a transcript.

For non-experts, this could potentially be achieved through the delivery of training about transcription, including the issues surrounding it and what can be done to mitigate their effects. Such training could warn transcribers of the difficulties of transcribing poor quality audio, highlighting that caution should be exercised, as well as the consequences of transcription errors, particularly substitutions, highlighting that incorrect guesses could be extremely harmful in forensic contexts. A major problem that arises with poor quality audio is that top-down (i.e. contextual) information is more heavily relied upon given the reduction in bottom-up (i.e. auditory) information available, and this can lead to situations where the way in which transcribers perceive the speech in a recording is substantially affected by their contextual knowledge. This is one of the problems that French and Fraser (2018) highlight; often the person transcribing the audio is a detective on the case who knows details about the situation taking place in the recording and the suspect and their alleged role in the offence. This can affect their perception of the speech and therefore the transcripts they produce, such that they may essentially hear what they want to hear. However, this is probably an issue about which police transcribers in England and Wales (and elsewhere, i.e. in other countries or jurisdictions) are unaware, so a training course could also address priming and the potential issues around having knowledge of contextual information. The common strategy used by experts, i.e. linear sequential unmasking, could also be introduced to non-expert transcribers; at the very least, it is hoped that this type of initiative would encourage the reconsideration of current, potentially problematic practices.

In order to incite change in the way in which transcripts are produced, it is necessary to forge connections with those carrying out transcription within police forces. Such connections would allow researchers to survey current attitudes to transcription, particularly with regard to poor quality audio, as well as the specific methods currently employed and police attitudes towards formalised training in transcription issues. Many people who are not familiar with the issues related to transcription may not believe that additional training is required, but being

aware of the potential problems is very important for the reliability and impartiality of transcripts. Delivery of training to transcribers could address the following points:

- The potential consequences of incorrect transcriptions, i.e. errors not being identified or priming the jury
- The potential issues related to transcription itself, e.g. decision making and representing certain features
- The potential issues related to transcribing poor quality audio, e.g. being primed by contextual information
- The types of methods implemented by expert transcribers

The human transcription experiment in this thesis involved the transcription of a standard variety (Standard Southern British English) with which all listeners were judged to be familiar, even those who do not share an accent with the speakers in the stimuli. No significant differences were found across the performance of SSBE speakers and speakers of other varieties of British English when transcribing this variety, which suggests that the regional accent background of the transcriber does not necessarily need to be taken into account when transcribing speech in a standard variety. However, the results would have likely looked very different if the participants in the experiment had transcribed an accent they were unfamiliar with, given previous findings on the transcription of non-standard regional varieties of British English (e.g. Glasgow English in Smith et al. (2014)).

It would be worthwhile investigating the transcription of unfamiliar regional accents in forensic-like recordings. An issue with previous research, like the study in Smith et al. (2014), is the highly controlled nature of the stimuli and, more specifically, the use of highly ambiguous sentences which do not resemble the type of speech found in real speech. Furthermore, Smith et al. (2014) report that the transcribers who were unfamiliar with Glasgow English had lower word recognition accuracy than transcribers who were familiar with that variety; however, this evaluation metric reveals nothing about the types of errors that the transcribers from each accent group were producing. It may be the case that the types of errors most frequently made varied across the accent groups; perhaps the transcribers unfamiliar with Glasgow English made more substitution errors given that the phonetic realisations of many words will not have matched their own production or the mental representation of those words.

If it were the case that transcribers that are unfamiliar with an accent produce more errors, particularly if more of those errors involve word substitutions (i.e. those that could be much

more damaging in a forensic context), then the transcriber's familiarity with the speaker's accent should certainly be taken into account. Perhaps the best method would be to send the recording to be transcribed by a police force in the region where the speaker comes from, or an expert if necessary. Another advantage of forging connections with police forces and surveying transcriber attitudes is the ability to find out whether unfamiliar accents have been identified by non-expert transcribers as an issue and whether there are strategies, however informal, currently in place. Further empirical research investigating the transcription of unfamiliar accents is required prior to a formal recommendation, although it is likely the case that those familiar with an accent will be more successful in achieving an accurate transcription of the speech, particularly in poor quality audio recordings. A 'risk assessment' stage that takes place prior to transcription was suggested at the beginning of this section; this stage could involve a consideration of the speaker's accent as well as the transcriber's level of familiarity with that accent, to ascertain who would be best suited to that particular transcription task.

More research is needed into the potential incorporation of automatic methods into the transcription of better quality audio recordings of forensic interest, such as police-suspect interviews. The research in this thesis has shown that such an approach may be viable, although much more research is needed to explore the following aspects:

- The performance of automatic systems on a much larger data set, including speakers from different demographics and accent backgrounds
- The performance of automatic systems with real police-suspect interviews to investigate whether the quality of such recordings is sufficient for automatic systems to be successful in their transcription
- How successful human transcribers are in post-editing the automatic transcripts, i.e. whether the transcript has a significant priming effect on transcribers such that errors are not identified
- Whether a hybrid human-automatic approach is viable in terms of the amount of editing required and the associated financial costs

This research must be carried out in a timely manner, given that some police forces in England and Wales have already shown an interest in automatic transcription (Tompkinson et al., 2022) and it is not unlikely that more will turn towards automatic methods in the next few years. The use of automatic methods in the transcription of recordings that may go on to play an evidential role in court is a topic that must be approached carefully and transparently. For poor quality evidential recordings, automatic speech recognition (ASR) systems are not

yet capable of producing accurate transcripts (e.g. Harrington et al., 2022; Loakes, 2024); this is recognised by expert transcribers, none of whom reported the use of ASR systems in their transcription practices (Article 1). It is unknown whether police forces in other countries have started or are looking to employ ASR systems within their transcription practices, either for evidential audio recordings or for police-suspect interviews; if so, it is crucial that sufficient testing has been carried out, particularly on the performance of ASR systems on languages other than English given the comparative lack of available resources (e.g. freely available large training datasets; Milde & Köhn, 2018).

Police-suspect interview recordings tend to be of better audio quality than evidential recordings and so the performance of ASR systems may be at a good enough level for use, in some capacity, within the transcription of police interviews. However, it should be noted that, even in ideal conditions (i.e. studio quality audio and speakers of a standard variety), many transcription errors were produced by commercial automatic systems, and the number of errors substantially increased when the speech was uttered in a non-standard regional accent (Article 3). The transcripts produced by an automatic system will always need to be checked and/or edited by a human. It is particularly important that police forces do not employ automatic systems as a purely financial decision, without consideration of the factors that can affect automatic transcription performance as well as the issues surrounding the checking of automatic transcripts (e.g. editors being primed by the content of an automatic transcript and therefore not identifying errors).

8.3 Aim 3: to develop a method for analysing transcription performance for forensic purposes

Transcripts serve two main purposes within the criminal justice system: to serve as a referenceable record of the speech within a recording (which is particularly true of transcripts of police-suspect interviews) and/or to aid the listener in hearing and understanding the speech content of a poor quality recording (which is most often the case with evidential recordings). In both cases, it is extremely important that the content of the transcripts is accurate for the following reasons:

- In some cases, transcripts are presented without the audio and so the court is completely reliant on the transcript as a record of the content of the audio recording; this is often the case for police interview transcripts (see Haworth, 2018).
- The jury does not often get the opportunity to carefully listen to poor quality evidential recordings (e.g. due to bad audio playback procedures in courtrooms) and so errors

in the transcript may not be identified and will therefore be accepted as an accurate record of the speech content.

- Parts of a transcript may be contested and multiple interpretations of a word or phrase may be put forth; in this case, an inaccurate interpretation can be extremely distracting from a more plausible interpretation (see Fraser & Kinoshita, 2021).

In order to analyse the accuracy of transcripts, and for future research into transcription performance by different groups and across different conditions, it is necessary to have a method of evaluation that is specific to forensic contexts. The current methods used to analyse transcription performance vary by the application of the analysis; in psycholinguistic contexts, the main way of measuring performance is by calculating the word recognition accuracy, i.e. the percentage of words that have been correctly transcribed. Sometimes these measures are only interested in content words (e.g. Clopper & Bradlow, 2008) and chosen keywords (e.g. Walker, 2018). In speech technology contexts, the most frequently used method applied in the analysis of ASR performance is Word Error Rate (WER), which is the ratio of errors in a transcript to the total number of words spoken.

Both of the measures described above are inappropriate for use (alone) in forensic contexts, where the errors contained within transcripts are of particular interest given that they can affect the meaning of an utterance. The deletion of a word from a transcript could have a considerable impact on how an utterance is understood by the transcript reader. For example, “I was there” in place of “I was **not** there” could be incriminating in contexts where the presence of a suspect at the crime scene is under question. Substitution errors could have an even bigger impact on the meaning of the utterance, especially in cases of the substitution of content words. For example, “I got the **gun** out of my bag” in place of “I got the **gum** out of my bag” could be hugely incriminating in the context of a murder suspect. The added issue with substitutions is that, due to the powerful priming effects of transcripts which are particularly prevalent in poor quality audio, listeners may believe that they can hear the substituted term, even in the face of evidence to the contrary (Fraser et al., 2011).

For forensic purposes, it is therefore important to consider the types of errors made within transcripts and the implications of those errors. Word recognition accuracy only considers what is done successfully and gives no consideration to the errors made within the incorrect parts of the transcripts. While WER does consider transcription errors, it makes no differentiation between the different types of errors, and so the deletion of, for example, “erm” from a transcript is given the same weight as the substitution of “gun” in place of “gum”. A novel method for analysing transcription performance in forensic contexts is

therefore required; this method can then be applied to academic research on transcription performance using different methods and can also be used in proficiency testing for expert transcribers.

8.3.1 Methods applied in this thesis

The research in this thesis has employed a novel approach to analysing transcription performance in forensic contexts. Article 2 presents an experiment in which human transcription performance was compared across accent groups and listening conditions, and Article 3 presents an experiment in which automatic transcription performance was compared across speaker accents and audio qualities. The studies involved different amounts of data, such that the analysis in the automatic transcription experiment was mostly carried out manually, whereas an automated approach was utilised for the larger scale of the human transcription experiment.

The evaluative approach taken in both studies broadly involved three stages:

1. Measure of overall performance
2. Presenting the numbers and types of errors made
3. Detailed analysis of substitution errors

Firstly, an overall view of the performance was captured using a standard method of analysis for the type of experiment, i.e. word recognition accuracy for human transcribers and WER for automatic systems. This allowed for an overview (or ‘a snapshot’) of the results across the different conditions. Secondly, an analysis of the transcription errors was conducted, such that the total number of each error type was calculated and compared across conditions. The distribution of errors across different conditions was also explored. Thirdly, a more in-depth analysis was conducted on the substitution errors, whereby the types of words involved in the substitutions and the forensic implications of such errors were considered.

Additional phonological analysis was carried out on the small-scale data in the automatic transcription experiment, demonstrating that many of the substitution errors for the non-standard regional accent could be explained by phonetic realisations that deviated from a ‘standard’ production of the word. These findings are useful because they reveal which sounds prove to be most challenging for automatic systems and should therefore be at the forefront of post-editors’ minds when correcting automatic transcripts of accented speech. However, such an approach would be much more challenging for large data sets.

Additional analysis of the substitution errors was carried out on the human transcription data, whereby a classification scheme was devised to categorise errors into those which would have a relatively 'minor' effect if left within a transcript and those whose impact could be potentially 'major' on the meaning of the utterance. This scheme was developed during the data analysis stage of the study because many of the substitution errors made by transcribers were very small, e.g. "forwards" in place of "forward". While substitutions can be extremely impactful on the meaning of an utterance, others (like the "forwards" example) would have little to no effect on (a) the reader's understanding of the speech or (b) the reader's perception of the speaker. It would be inappropriate not to differentiate between 'minor' and 'major' substitution errors in forensic contexts, given that grouping all substitutions together would result in overinflated numbers of 'potentially damaging errors'.

8.3.2 Towards a systematic framework

The method of analysis employed within this research can be used for a number of purposes. Firstly, it can be used in proficiency testing for transcribers to prove competency³⁶ and to help them improve their own transcripts. Secondly, it can be used to assess both human and machine performance in future empirical research on forensic transcription, and to compare across different groups of transcribers (e.g. familiar with accent versus unfamiliar with accent). Thirdly, this method can be used for validation purposes, firstly by comparing expert or trained police transcriber performance with baselines from lay people, highlighting that transcription of poor quality evidential recordings is a task that should be carried out by those with appropriate training and expertise³⁷. Finally, the method presented in this research could be developed into a single validation metric for testing different transcription methods, after which an acceptance criterion can be set (e.g. a rate of substitution errors which is deemed acceptable, taking into account the magnitude of those errors). Though the method has been developed for use on English transcripts, it could easily be further developed for use on other languages.

³⁶ Within this discussion, there have been many references to proficiency testing for expert transcribers. Testing non-expert transcribers, who are also responsible for transcripts that are used within the criminal justice system, would also be an appropriate next step, particularly given that huge variability in the performance of lay people was observed in the experiment presented in Article 2.

³⁷ A recent pair of studies carried out by Basu et al. (2022; 2023) showed that their expert automatic speaker recognition system was much better at forensic speaker comparisons than lay people; this confirms that judges should not attempt to perform their own speaker identifications, and should instead rely on validated expert methods. A similar approach could be taken with regard to forensic transcription, comparing transcripts produced by lay people and trained transcribers, to demonstrate that this task should be carried out by those with the relevant training and expertise.

There are, however, areas in which the current method could be improved and turned into a systematic framework with, for example, a score given to each transcript. A number of issues for consideration, identified on the basis of experience using the evaluation method, are presented below.

The analysis in this research was conducted at a lexical level, and much of the initial data analysis was carried out using bespoke software designed and developed for this doctoral research. The software automatically aligns the words in a reference transcript and participant transcript by utilising a JavaScript implementation of the Hunt-McIlroy algorithm (Hunt & McIlroy, 1976). The algorithm matches up the words by finding the longest common subsequence between the two transcripts, but the software also allows manual correction; this is useful when multiple deletions have been made and substituted terms in the participant transcript are not aligned with the most likely reference word. The software also lets the analyst flag certain word pairs, such as those containing spelling errors and minor grammatical changes.

The benefit of analysing transcripts at a lexical level is the possibility of automation, as described above. When working with large datasets, a degree of automation is preferable as it is more time efficient and also reduces the risk of human error. However, this approach does require a certain amount of post-processing of the data; spelling errors, grammatical contractions and expansions, and the compounding of words (e.g. “steakhouse” in place of “steak house”) all require manual correction after the automatic alignment has taken place. Another issue with the word-level approach is that it does not take into account the reordering of words in the transcript. For example, one of the stimuli in the human transcription experiment contained “Thursday night, on Wednesday night, on Friday morning”, and multiple participants swapped the order of the days so that they were in chronological order (i.e. “Wednesday night, on Thursday night, on Friday morning”). These were marked as substitution errors of potentially ‘major’ impact in a transcript, despite the fact that the reordering in this particular context likely would make no difference to the understanding of the utterance.

A different approach to transcript analysis is investigating errors at a phonetic or syllabic level (Tschäpe & Wagner, 2012) or at a phrasal level (Fraser et al., 2023). A phrasal-level approach may be able to account for issues with the word order as well as dealing with the retention of meaning and categorisation of ‘minor’ and ‘major’ substitution errors in a more succinct way than a word-level approach. However, automation would be much more challenging for this type of approach, given that judging the accuracy of a phrase is a much

more subjective process than the matching of individual words between a reference transcript and participant transcript. A phrasal-level approach may therefore be more challenging for large datasets, but more informative for forensic purposes. Further research could compare the outputs of a word-level approach and a phrasal-level approach to explore their relative practicalities, as well as to deduce which type can capture more forensically-meaningful information.

Another issue to be considered is the way in which errors are categorised. The three main categories of errors (deletions, substitutions and insertions) are a good starting point but, as previously discussed, substitution errors can have very different impacts on a transcript and therefore should not be analysed as one group in forensic contexts. In Article 2, substitution errors were categorised as having a potentially 'minor' or 'major' impact on the meaning of a transcript, according to the forensic implications (i.e. potential consequences) of that substitution being contained within a transcript. The classification scheme utilised in Article 2 was produced impressionistically during the data analysis process as a way to generalise for forensic purposes; however, the scheme was not based on rigorous review of the literature surrounding different word types and how these substitutions may be perceived by readers of a transcript. A more formalised and linguistically-informed 'error typology' could be developed, building on the work carried out in this thesis.

8.4 Summary

The findings of this thesis can be summarised as follows:

1. There is a lot of variability in the methods employed in the production of transcripts for use in court, by both non-experts and experts. Such variability in methods leads to variability in the quality (in terms of accuracy, readability and impartiality) of transcripts, particularly for non-expert transcribers who receive no training and are provided with no guidelines. Expert transcribers (i.e. forensic practitioners) also have no guidelines to follow, and so each lab has developed its own procedures based on what they believe to be the best approach.
2. There is also variability in the content and therefore quality of transcripts (in terms of accuracy and the presence of potentially damaging errors) as a function of two factors commonly encountered in audio recordings of forensic interest: audio quality (with regard to the level of background noise) and regional accent (with regard to the transcriber's familiarity with the speaker's accent). These were rated as the two most influential factors on the perception and transcription of speech in poor quality audio recordings by forensic practitioners, but are unlikely to be considered in non-expert transcription procedures, given a widespread lack of appreciation for the complexity of transcription among lay people.
3. A possible solution to the issues outlined above is a push towards the standardisation of transcription methods within each domain. In the pursuit of developing standards, and guidelines and training to meet those standards, more (much needed) empirical research must be conducted concerning the best practices for producing transcripts for use in the criminal justice system. Guidelines for non-experts (e.g. in the form of transcript templates and transcriber training) would raise awareness of the issues related to transcription among those producing transcripts, and could also improve understanding across court users, e.g. lawyers, judges, the Crown Prosecution Service, concerning how transcripts should be used and evaluated. The findings from further empirical research can be used to incite positive change in the ways in which both experts and non-experts produce transcripts, with the ultimate goal of producing better, more reliable and more impartial transcripts for use within the criminal justice system.
4. Current methods for the evaluation of transcription performance are unsuitable for use in forensic contexts, given that no consideration is given to the types and magnitude of errors contained within the transcripts. A new method of evaluating transcripts for use in forensic contexts is presented in this research, whereby the

analysis is centred on the transcription errors and their implications in forensic contexts. The method can be used for a number of purposes, such as further research on transcription performance and proficiency testing for transcribers.

8.5 Practical recommendations

The practical recommendations made within sections 8.1 to 8.3 are summarised in the following sections.

8.5.1 For experts transcribers/forensic practitioners

- More research about the methods of forensic transcription should be conducted and is encouraged within the forensic speech science community, with a particular focus on validity testing and the effects of priming on particular aspects of transcription methods.
- Guidelines for producing transcripts of poor quality evidential recordings (even if relatively informal) could be produced by a new working group of the *International Association for Forensic Phonetics and Acoustics*.
- Experts should consider setting up proficiency testing within their labs, or a standard proficiency test for forensic transcription should be developed. The method of evaluation presented in this thesis can be used, or developed further, for evaluating the transcripts.
- Experts should consider the content of their transcripts in line with the needs of the end user. Even if clearly presented in a key, complex transcription conventions can confuse and ultimately be misunderstood by the lay people who eventually read the transcript.

8.5.2 For non-expert transcribers working with poor quality evidential recordings

- Police forces should collaborate with experts in forensic transcription so that information on police transcriber methods and attitudes can be collected, which can then inform the production of standards and/or guidelines for transcription.
- Police transcribers would benefit from training delivered by experts in forensic transcription with regard to both their methods of transcription and their understanding of the complexity of transcription as a task. Training could cover a range of issues related to transcribing poor quality audio, such as the priming power of contextual information and the potential consequences of transcription errors.
- A commonly encountered issue with non-expert transcripts is huge variability in the format and layout; a transcript template has been provided in Appendix A of this thesis that may be used by transcribers to give a consistent structure to transcripts. Guidance on how to use the transcript is also provided in Appendix A.

- A 'risk assessment' stage prior to transcription could be implemented such that the audio quality and factors such as the speaker's accent can be taken into account when deciding who is best suited for that particular transcription task and what precautions should be taken (e.g. the transcriber should have no case information).

8.5.3 For non-expert transcribers working with police-suspect interview recordings

- Transcripts that will be presented to juries alongside, or in place of, police interview recordings should be transcribed in a verbatim manner to ensure that the jury does not receive a subjective account of the defendant's speech and that parts of their speech are not omitted from the record.
- Automatic transcription systems may be able to produce a 'first draft' of a transcript, but the current performance of these systems is not good enough on its own. Such transcripts will always need checking and/or editing by a human, but more research is needed on the viability of this approach. It is recommended that automatic transcription systems are not implemented in the transcription of police-suspect interviews until the reliability and validity of this approach has been scientifically tested.

8.6 References

- Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, 35(2), 520–529.
- Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken-word comprehension. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, 1925–1928.
- Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. (2022). Speaker identification in courtroom contexts–Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, 341, 111499.
- Basu, N., Weber, P., Bali, A. S., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. (2023). Speaker identification in courtroom contexts–Part II: Investigation of bias in individual listeners’ responses. *Forensic Science International*, 349, 111768.
- Cenceschi, S. & Meluzzi, C. (2023). Transcription and voice comparison of noisy interceptions: remarks from an audio forensics report. *STUDI AISV*, 10, 99-111.
- Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: intelligibility and classification. *Language and Speech*, 51(Pt 3), 175–198.
- Dorn, R. (2019). Dialect-specific models for automatic speech recognition of African American vernacular English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 16–20.
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology. Human Perception and Performance*, 32(5), 1276–1293.
- 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.
- Fraser, H., Loakes, D., Knoch, U., & Harrington, L. (2023). *Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio*. Conference of the International Association for Forensic Phonetics and Acoustics, Zurich.
- Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2). <https://doi.org/10.1558/ijsl.v18i2.261>
- French, P., & Fraser, H. (2018). Why “Ad Hoc Experts” should not Provide Transcripts of Indistinct Audio, and a Better Approach. *Criminal Law Journal*, 298–302.

- Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, 18(2), 293–307.
- Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech Language and the Law*.
<https://doi.org/10.1558/IJSLL.38028>
- Harrington, L., Love, R., & Wright, D. (2022, July). *Analysing the performance of automated transcription tools for covert audio recordings*. Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, 22(4), 428–450.
- Haworth, K., Tompkinson, J., Richardson, E., Deamer, F., & Hamann, M. (2023). “For the Record”: applying linguistics to improve evidential consistency in police investigative interview records. *Frontiers in Communication*, 8.
<https://doi.org/10.3389/fcomm.2023.1178516>
- Hunt, J., & McIlroy, M. (1976). An Algorithm for Differential File Comparison. *Murray Hill: Bell Laboratories*, 9.
- 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.
- Lindsey, G. (2019). *English After RP: Standard British Pronunciation Today*. Springer.
- 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
<https://apps.dtic.mil/sti/citations/ADA414420>
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7.
<https://doi.org/10.3389/fcomm.2022.803452>
- Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 521–534.
- Milde, B., & Köhn, A. (2018, October). Open source automatic speech recognition for German. In *Speech communication; 13th ITG-symposium* (pp. 1-5). VDE.
- Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92–100.
- Rhodes, R. (2012). *Assessing the strength of non-contemporaneous forensic speech evidence*. [PhD thesis, University of York]. White Rose eTheses.
<https://etheses.whiterose.ac.uk/3935/>

- Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R.-A. (2014). Cross-accent intelligibility of speech in noise: long-term familiarity and short-term familiarisation. *Quarterly Journal of Experimental Psychology*, 67(3), 590–608.
- Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4), 487–501.
- Thomas, M. J. (2019). Employment, education, and family: Revealing the motives behind internal migration in Great Britain. *Population, Space and Place*, 25(4), e2233.
- Tompkinson, J. & Haworth, K. (2023). *The perception and interpretation of additional information in legally relevant transcripts*. Conference of International Association for Forensic Phonetics and Acoustics, Zurich, Switzerland.
- 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*. Conference of the International Investigative Interviewing Research Group, Winchester.
- Tschäpe, N., & Wagner, I. (2012). *Analysis of Disputed Utterances: A Proficiency Test*. Conference of International Association for Forensic Phonetics and Acoustics, Santander, Spain.
- Vergyri, D., Lamel, L., & Gauvain, J.-L. (2010). *Automatic speech recognition of multiple accented English data*. In Interspeech (pp. 1652-1655).
http://www-tlp.limsi.fr/public/automatic_speech_recognition_of_multiple_accented_english_data_vergyri.pdf
- Walker, A. (2018). The effect of long-term second dialect exposure on sentence transcription in noise. *Journal of Phonetics*, 71, 162–176.

9. Appendix

9.1 Appendix A

Transcript template with guidance

A transcript template, complete with guidance on how to use it, is provided within this document. This template may help those who produce transcripts as part of their work duties but are offered no guidance on how to do so, and is based on the structure and layout of transcripts produced by expert forensic practitioners. The template provided here can be used as it is, or developed further by police or other transcribers.

The structure of the document is as follows:

- An introduction to the template
- Guidance for use of the template
- An explanation of the transcript key
- An explanation of the transcript template

9.1.1 Introduction

An issue that arises from the provision of little to no training or guidance on how to produce transcripts of evidential audio is huge variability in the transcripts that are produced. Sometimes the transcripts lack structure, with the text presented in one big block and no reference to time points within the recording. Sometimes the transcripts contain symbols but it is not clear what those symbols mean. These things all make the transcript very difficult for future readers to follow.

In order to address this problem, this document contains a template that can be filled out when transcribing evidential recordings. The template is a four-columned table that can easily be replicated in Microsoft Word. Transcription conventions have been provided; this will ensure that transcribers are using the same symbols for certain phenomena. There are some instructions on how to fill out the template as well as some other guidance on transcription within this document.

9.1.2 Guidance on use of template

Some general guidance on the transcription of poor quality evidential recordings is provided below:

- **Be conservative** - you should not guess what a speaker is saying. It is likely that in poor quality recordings you will not be able to transcribe every word that is said. Transcription errors made as a result of guessing can be very damaging and can potentially lead to miscarriage of justice.
- **Provide verbatim transcriptions** - make a note of every word that a speaker says and do not try to 'clean up' someone's grammar. For example, do not write "I don't know" if the speaker clearly said "dunno".
- **Avoid case information** - people can be influenced to hear certain things according to the information that they know. It is best to take a 'blind' approach to transcription to begin with, where you do not know any of the details about the case or the speaker. This will help to make your transcripts more impartial. Information can then be introduced once a first draft has been produced.

More specifically to using the template:

- **Use brackets if uncertain** - if you are not completely confident about a word or phrase, you can put the word(s) in brackets to show a lower level of confidence. This does not mean that you think the word is wrong, but you acknowledge that it could be something else.
- **Include line numbers** - line numbers that can be used to refer to specific utterances will be very helpful for future users of the transcript, e.g. lawyers.
- **Include regular time points** - be sure to include regular time points, ideally one for each new line. This will be very helpful for people who will later use the transcript.
- **Indicate summaries** - it is quite common to have large portions of the recording (especially hours of covert audio) which do not require transcription in full. In such a case, make sure there is a clear indication that you are no longer transcribing the speech verbatim. You could use square brackets with an indication that you are now summarising, e.g. [Summary - the speakers spend the next ten minutes talking about football].

- **Avoid spelling errors** - make sure you use a Word processor to check for spelling mistakes.
- **Include the speaker** - make sure you include a label in the 'Speaker' column for each utterance. If you are not sure exactly who the speaker is, then use a label such as "M" or "F" to indicate their gender. If you are unable to identify the speaker's gender, you can use a question mark "?" in the 'Speaker' column.
- **Use ellipses for unintelligible speech** - if you cannot make out what the speaker is saying, do not guess. Instead, use an ellipsis "..." to show that the words are unintelligible and therefore cannot be understood. Make sure you do not use ellipses for other purposes.

9.1.3 Transcript key

A very important consideration when producing a transcript is the readability of the document. The transcript will often be used by 'end users' other than the transcribers themselves, e.g. the jury or other members of the court. It is important that a reader is able to understand the transcript without external help.

Transcripts may contain references to time points within the audio recording or may use symbols to represent certain features. Although these may seem obvious to the transcriber, the meaning of the symbol may not always be clear to readers of the transcript. A key of the transcription conventions should therefore always be included before the transcript. Table 1 shows a list of conventions that can be used within a transcript.

Convention	Explanation
00:00	Playback time - elapsed time from start of recording (MM:SS)
M1	Male speaker one
PO	Police officer
(M1)	Brackets indicate lower confidence in speaker attribution
M	Male speech - unattributed
F	Female speech - unattributed

Convention	Explanation
(Yes)	Brackets indicate lower confidence in words transcribed
Ye-	Hyphens denote incomplete or interrupted speech
...	Unintelligible speech
[Coughing]	Description of non-verbal sounds or other information

Table 1: List of conventions that can be used in a transcript.

The playback time is very useful to include in transcripts because it can help readers to locate the transcribed speech. This is particularly helpful for longer recordings, and should be regularly included within the transcript. It is recommended that the time is included for each utterance (i.e. each new line), as demonstrated in Table 4 and Table 5.

When attributing speech to different speakers, it is important that the transcriber does not make assumptions about who is speaking. In cases where it is very clear who the speaker is and this is not contested, then a descriptive label may be assigned to the speaker, e.g. “Operator”, “Caller”, “Police Officer”. It is not recommended to use names of the speakers, even if known.

For all other speakers, it is recommended that the speaker is identified by their gender and then assigned a number, e.g. “M1”, “M2”, “F1”, “F2”. If there is uncertainty about the identity of the speaker, their speaker label can be enclosed in brackets, e.g. “(M1)”, or the speech can be identified as male or female but left unattributed, e.g. “M” (with no number).

Transcribers may be fairly certain that a speaker said a particular word or phrase but there is some doubt. In this case, the transcriber may enclose the word(s) in brackets to demonstrate a lower level of confidence. Without this convention, all speech contained within a transcript seems to have the transcriber’s full confidence, but this is not always the case. The brackets do not mean that the transcriber is wrong, but simply that they are exercising some caution over their interpretation.

Often when people talk, they change what they are saying mid-word or mid-sentence. This can be very challenging to read without some kind of notation that the word or sentence has been interrupted. A hyphen can be used to demonstrate that the speech has been

interrupted or is incomplete. For example, “ambu-” in Table 3 shows that the speaker stopped halfway through saying the word “ambulance”.

Sometimes parts of the recording will be unintelligible, which means that it is very challenging or even impossible to make out what is being said. This is very common in poor quality evidential recordings due to the way in which the audio has been collected, e.g. using a covert recording device. In such cases, an ellipsis can be used to represent speech that is too difficult to transcribe. It is very important that transcribers acknowledge sections of unintelligibility and do not attempt to guess what is being said.

Finally, there are often sounds other than speech within audio recordings. In the transcript examples below, the ringing of the telephone can be heard at the beginning of the call and when the caller is transferred from the emergency services operator to the police operator. This type of information can be presented in square brackets, e.g. “[Ringing tones]”.

People may produce non-verbal sounds, such as coughing. It is not recommended that every non-verbal sound is included in a transcript because the transcriber may misinterpret the sound, e.g. the transcriber may believe they hear laughter when that noise is actually someone crying. However, very clear and uncontested sounds, which will help the transcript reader to follow the speech, should be included.

9.1.4 Transcript template

Below is a blank version of a template to be used for transcribing the content of audio recordings. The transcript is a table composed of four columns. The first column will contain the line number, which can be very useful for future readers of the transcript if they wish to discuss certain parts. The second column will contain the time point, i.e. the time elapsed from the beginning of the recording, in a HH:MM:SS format where “HH” represents hours (if necessary), “MM” represents minutes and “SS” represents seconds. The third column will contain reference to the speaker, which may be a descriptive label such as “Caller” or may be an indication of the speaker’s gender with no further attribution in cases of uncertainty. The fourth column will contain the transcription of the speech content.

Line	Time point	Speaker	Content
1			
2			
3			
4			

Table 2: Blank template of transcript. The transcript is a table made up of 4 columns.

Overlapping speech, i.e. multiple people talking at the same time, can be challenging to represent in the above template. An extra speaker column and content column can solve this issue, as presented in Table 3 below. In this way, the speech of two speakers who are talking simultaneously can be captured using a single time point.

Line	Time point	Speaker	Content	Speaker	Content
1					
2					
3					
4					

Table 3: Blank template of transcript that can be used in cases of overlapping speech. The transcript is made up of six columns.

9.1.5 Example transcripts

An example transcript is presented in Table 4, using the template design from Table 2 (with four columns). The example is a phone call to the emergency services in which there are four different speakers: the caller, the emergency services operator, a male speaker in the background of the telephone call, and a police operator.

Line	Time point	Speaker	Content
1	00:00		[Dialling tones]
2	00:02	Operator	Emergency, which service?
3	00:03	Caller	Please, ambu- no, police please.
4	00:04	Operator	Thanks, caller.
5	00:04	M1	Hey, put that phone down, now.
6	00:06		[Ringing tones]
7	00:06	Caller	Reg, stop it.
8	00:08	Police Operator	Police emergency, what's the address of the emergency?
9	00:08	M1	Get here (now, you twat).
10	00:09	Caller	It's one-one-five (Byron) Terrace in [coughing] Caerphilly.
11	00:09	M1	Get off- (give me that bloody phone) ...

Table 4: Transcript template that has been filled in.

One issue with the above layout is that overlapping speech is challenging to show. For example, the operator and male speaker are both speaking at four seconds into the recording (lines 4-5), but the linear formatting of the transcript does not clearly demonstrate this. A second example transcript is presented in Table 5, this time using the template design from Table 3 (with six columns) which more clearly shows the overlapping nature of the speech. The content of the transcript is the same as in the previous example.

Line	Time point	Speaker	Content	Speaker	Content
1	00:00		[Dialling tones]		
2	00:02	Operator	Emergency, which service?		
3	00:03	Caller	Please, ambu- no, police please.		
4	00:04	Operator	Thanks, caller	M1	Hey, put that phone down, now.
5	00:06		[Ringing tones]	Caller	Reg, stop it.
6	00:08	Police Operator	Police emergency, what's the address of the emergency?	M1	Get here (now, you twat).
7	00:09	Caller	It's one-one-five (Byron) Terrace in [coughing] Caerphilly.	M1	Get off- (give me that bloody phone) ...

Table 5: Transcript template for overlapping speech that has been filled in.