Producing ‘good enough’ automated transcripts securely: Extending Bokhove and Downey (2018) to address security concerns

Joseph Da Silva

Abstract
Interviews are an established research method across multiple disciplines. Such interviews are typically transcribed orthographically in order to facilitate analysis. Many novice qualitative researchers’ experiences of manual transcription are that it is tedious and time-consuming, although it is generally accepted within much of the literature that quality of analysis is improved through researchers performing this task themselves. This is despite the potential for the exhausting nature of bulk transcription to conversely have a negative impact upon quality. Other researchers have explored the use of automated methods to ease the task of transcription, more recently using cloud-computing services, but such services present challenges to ensuring confidentiality and privacy of data. In the field of cyber-security, these are particularly concerning; however, any researcher dealing with confidential participant speech should also be uneasy with third-party access to such data. As a result, researchers, particularly early-career researchers and students, may find themselves with no option other than manual transcription. This article presents a secure and effective alternative, building on prior work published in this journal, to present a method that significantly reduced, by more than half, interview transcription time for the researcher yet maintained security of audio data. It presents a comparison between this method and a fully manual method, drawing on data from 10 interviews conducted as part of my doctoral research. The method presented requires an investment in specific equipment which currently only supports the English language.

Keywords
Interviews, transcription, automatic speech recognition, qualitative data, data security, information security

Introduction
The use of automated methods to produce transcriptions of recorded audio from interviews has been explored by a number of researchers, as discussed below. Recent methods have involved the use of cloud-based services which, while effective, introduce challenges to the security of the data and present ethical concerns. As a practicing cyber-security professional, specifically a Chief Information Security Officer (CISO), I am acutely aware of security and ethical considerations with regard to data access as my professional field has a significant focus on managing the risk associated with the processing and storage of data in many different forms. However, such a concern should be shared by all researchers who collect and process data; risks to the security of data, whether caused by malicious or accidental action, are ever-present, particularly given the increased reliance on software and online services in many aspects of everyday life (Levy, 2019: 9). The 2019 UK National Cyber Security Centre report authored by Dr Ian Levy specifically references University accounts as one of the top 10 targets of criminal phishing attacks that year, with multiple attack groups responsible (Levy, 2019: 15–16). As a successful compromise of a University email account is likely to also provide access to that user’s electronic documents,
researchers in particular should be concerned about management of their electronic files.

My doctoral research explores the purpose of dedicated cyber-security organisational functions within businesses, in order to understand why businesses employ specialist teams that focus on this risk. I am interviewing others who perform my role, as well as senior leaders within those organisations, using semi-structured interviews to inquire into any differences in perception and expectations as well as motivations and contexts. As semi-structured interviewing has become established as a core, and arguably the default, qualitative data-gathering method across the social sciences, for example, Coronel et al. (2011) and Rainford (2020), and related disciplines, including business, for example, Maznevski and Chudoba (2000) and Peterson (2004), and cyber-security, for example, Ashenden and Sasse (2013), Singh et al. (2013) and Van der Kleij et al. (2017), researchers, like myself, that gather data using this method are faced with the task of preparing that data for analysis. For novice researchers, in particular, the predominant guidance from the literature, for example, Hammersley and Atkinson (1995) and Bryman (2012: 482) is for manual, verbatim (Seale, 2000: 148) transcriptions of audio-recorded interviews to be produced in ‘orthographic’ (Braun and Clarke, 2006: 88) form for subsequent analysis, despite the acknowledged limitations of transcripts themselves (DeVault, 1990; Green et al., 1997; Lapadat and Lindsay, 1999; ten Have, 1990). However, manual transcription is, as others have identified, for example, Bird (2005) and Matheson (2007), time-consuming and, while undoubtedly valuable, somewhat tedious as a process. Like other researchers, for example, Bokhove and Downey (2018), following my first experiences of manual transcription and reflecting upon the number of interviews I would be conducting in future, I began to explore automated transcription software and wish to share, as encouraged by other qualitative researchers (Matheson, 2007: 558), what I believe is an innovative method that improves upon prior work in this field with regard to the security of interview data. I build upon Bokhove and Downey (2018) in particular regarding the production of ‘good enough’ transcripts, extending this by responding to the security challenges posed by the use of automated cloud-based transcription services, presenting a method that provides greater security and addresses ethical limitations arising from the potential for third parties to access interview data.

Privacy and ethical considerations

Maintaining confidentiality of participant data and involvement is an established principle of sociological research, for example, Walford (2005) and Wiles et al. (2008) and, if provided, confidentiality assurances given to research participants must be honoured (British Sociological Association, 2017: §37). If participants have previously been advised of confidentiality and anonymity at the point of their involvement, then the subsequent use of a service that threatens this would be unethical. It was apparent from even a cursory reading of the terms and conditions of the predominant cloud-computing transcription offerings, for example, Google (n.d.a ‘Your Content in our Services’), Otter.ai (n.d.b, n.d.c) that these zero-cost offerings could not provide adequate assurances around the confidentiality of the data I was seeking to transcribe, and would not satisfy the ethical requirements of the research that I was performing, the anonymity of which I had agreed upfront with the participants. My interviews would certainly contain both confidential and identifiable information and, therefore, any risk of a third party being able to access these recordings, however slight, was unacceptable. Bokhove and Downey (2018: 11–12) identified such data privacy challenges, particularly with regard to data being processed and stored on third-party computing resources, but did not provide a solution beyond ethics committee approval. Ethics committees may not approve the use of services that provide no assurance of confidentiality for processing data that has been agreed to be treated anonymously, or for any data shared in confidence. Therefore, it is my opinion that in a considerable number of cases, the specific method used by Bokhove and Downey (2018) could not be utilised. However, it is acknowledged that this is a question of risk; this risk may be considered acceptable by some, particularly within wider research constraints.

Related work

Automated transcription of interview data has previously been explored by a number of researchers. Prior to Bokhove and Downey (2018), Moore (2015) and Matheson (2007) utilised transcription software that did not rely on cloud computing. Moore identified that such software was inconsistent in terms of accuracy but, similar to Bokhove and Downey (2018), was effective in generating a ‘first-pass transcript’ (Moore, 2015: 253). Moore (2015: 269–270) predicted that capabilities in this field would continue to advance, and also suggested that ‘partial automation’, in particular, would be more desirable than full automation, which aligns with the method I explore below. Matheson (2007) identified an improvement in transcription time although significant improvements were dependent upon mastery of the technique and its specific software over time, and even then, the ratio of transcription time to recorded time was still four or five to one; for inexperienced users, efficiency improvements were only ‘modest’ (Matheson, 2007: 558).

A number of researchers have identified limitations of automated transcription software, including the impact of noise levels on accuracy, for example, McGurk et al. (2008); the inability to distinguish between different voices, for example, Matheson (2007); and multiple biases affecting software performance, for example, Cramer et al. (2018), Howard and Borenstein (2018) and Tatman (2016).
Both Moore (2015) and Matheson (2007) identified the need for ‘training’ the software involved, both initially and on an ongoing basis, which is not a necessity with the software utilised by Bokhove and Downey (2018) or with that used in the method I relate below. The software used by Matheson (2007) also improves in accuracy ‘the longer it is used by the same user’ (p. 553) which, while valuable for the specific cases they explore, would not be of benefit in situations where only one recording of a specific interviewee exists. Matheson (2007: 553) also highlights the importance of maintaining the confidentiality of interview data and the improved security of digital recordings over the use of analogue tape. The potential for third-party access to interview data was not discussed; as the software used in their research was not specifically cloud-based, this may not have been an obvious concern. However, although not explicitly cloud-based, the exact mechanism by which the software involved processed data was not discussed; there would have been the potential for the software to ‘talk back’ to a remote server in order to process data and no mention of network connectivity is made.

Although discussion of confidentiality in interview-based research is common, for example, Kaiser (2009) and Tolich (2004), limited literature was identified relating to the security of interview recordings. Myers and Newman (2007: 10) also note a lack of discussion of these considerations and describe the need for researchers to ‘[keep] transcripts/records/and the technology confidential and secure’ (p. 23), although they do not discuss how that may be achieved. Others have discussed limiting risks regarding data retention, for example, McLellan et al. (2003: 71) but little discussion around ‘in-life’ protection of interview data was found.

Clearly seeing the value in Bokhove and Downey’s (2018) approach to producing ‘good enough’ transcripts, I sought automated transcription software that did not rely on cloud computing but was more effective than that used by Moore (2015) and Matheson (2007). I identified a modern automated transcription offering that solely relied on local device capabilities – Google Recorder (Google, n.d.c). This software is based on recurrent neural networks (Google AI Blog, n.d.a; He et al., 2018). This technology appears similar to that used by YouTube for transcription which are also recurrent neural network based (Google AI Blog, n.d.b, n.d.c), however, it is unclear whether or not they are exactly the same technologies. In order to explore this software, there was one further hurdle; although this software is designed for an Android-powered device (and I had an Android device), it is only officially available for certain specific, Google-manufactured, hardware, namely the Pixel-branded devices. Unless I owned one of these devices, I was stymied. The compatible devices ranged in cost from £399 to £829; I discounted the purchase of a secondhand device for a number of reasons. First, the devices themselves were quite new to the market and therefore secondhand availability would be both limited and unlikely to present much of a saving. Second, a previously used device presents fewer assurances than a new one in terms of both security and reliability.

I calculated the cost-per-transcription of the cheapest device based on the minimum number of interviews I was planning to complete from that point (December 2019) forward. As I was planning at least 15 further interviews, this worked out at a worst-case cost of £26.60 per transcription (and a best-case of less than £15). Given the time, I anticipated that this process would save, I considered this a worthwhile investment; there were certainly times during my manual transcriptions where I would happily have paid more than this amount to speed up the process. I decided to purchase the cheapest device (Pixel 3a), not just on the basis of cost but also because, unlike the more expensive device (Pixel 4), it featured an inbuilt 3.5 mm audio socket that offered the potential for line-level recording – this was seen to be attractive from both a security and potentially an increased accuracy perspective, as well as being easier to use than the inbuilt microphone on the device. However, as I discovered, the software only records using the latter; it does not process incoming audio from the 3.5 mm audio socket.

**Method**

**Background to my research**

This method was trialled on data arising from my doctoral research into the purpose of cyber-security functions within commercial businesses. The management of risk is fundamental to businesses and cyber-security risk, in particular, has become a board-level concern in recent years (Hooper and McKissack, 2016), even being considered existential by many businesses (Bevan et al., 2018). Cyber-security incidents affecting businesses achieve significant coverage in mass media, for example, Voreacos et al. (2019), Kuchler (2017) and Thomas (2016) and as well as managing threats to their own operations, businesses are also seen as performing a role in securing nations from cyber-security threats (UK Government, 2016: 34). As well as the use of risk committees at board-level as explored by, for example, Akbar et al. (2017) and Beasley et al. (2009), many businesses operate dedicated teams that focus on cyber-security, as they do for other forms of risk, for example, Legal, Compliance, Tax. However, it is not clear what the purpose of such a function is and, in my professional experience, there is a wide range from scapegoat to educator to business enabler, and from operator of controls to assurer of controls; diversely differently understandings of each of these can exist within a single business. Without a clear purpose, and certainly without a consistent understanding of purpose, then it is difficult for a business to be confident that it is managing that risk effectively.

As I am exploring differences in human perceptions regarding purpose, a concept which is socially constructed and highly interpretive, I have adopted an interpretivist, social constructionist epistemology. Cyber-security as a domain is inherently social; it relies on, and affects, ‘human communities’. Businesses, as organisations, are equally
social (Greenwood et al., 2008: 31), as are the technologies that they rely on (Czarniawska, 2008: 774; Eason, 2003: 169; Ihde, 1990: 20). Therefore, an exploration of cyber-security within business cannot ignore the social dimensions, and in particular, how those dimensions affect the meanings attributed by individuals within the participating communities. As well as being social, cyber-security is interpretive: there is no one objective method of ‘doing’ cyber-security. As mentioned earlier, I have not experienced there being, an objective, or consistent, understanding of the purpose of cyber-security functions across the industry and its stakeholders; rather, that understanding is jointly, and continually, constructed by its participants.

Qualitative methods are recommended as appropriate for situations in which theory is, as Edmondson and McManus (2007: 1158) describe, ‘nascent’; I considered my research to fall into this category. As part of the first phase of my fieldwork, I performed a number of semi-structured interviews with different participants representing a range of different businesses, following an initial pilot study with three organisations. My interviews were conducted with participants who either performed the role of CISO or were senior executive or non-executive directors. The participants, therefore, represented both ‘elites’ and practitioners (who could arguably also be considered as elites, or at least sub-elites)

At this stage, I had performed 10 interviews, with the first five being transcribed manually. My manual transcription had taken a lot longer than I thought (5–8 hours per interview over five interviews at this stage). Although tedious, it was useful in helping me understand both my data and the qualitative interviewing process as a whole – and also prompted a lot of exploration around associated aspects, for example, transcribing ‘um’s and ‘er’s, which I had done for the pilot interviews but not for subsequent transcriptions, following the arguments in Bucholtz (2000, 2007) and Collins et al. (2019). I felt better about the transcriptions as they progressed (as captured in my fieldnotes) but began to consider the use of automated transcription software for subsequent interviews, at least to get the core text down. As someone who works in a heavily technology-dependent profession and industry, I was aware of advances in the field of voice-to-text as summarised by Bokhove and Downey (2018: 3). However, as a security professional, I was also very aware that these newer software offerings relied predominantly on cloud-based capabilities, for example, as used by Bokhove and Downey (2018) and were predominantly zero-cost; these factors led me to consider such services cautiously and to review the terms and conditions and privacy policies of these in detail. Cloud-based solutions are not necessarily any less secure than non-cloud alternatives; however, any system that is under the administrative control of others presents de facto data privacy challenges, given those administrators have access to the system (Kamarinou et al., 2016: 188–189). In addition, zero-cost cloud services typically rely on business models in which data are monetised (even data in ‘raw’ form; Spiekermann, 2019: 209) and, in order to be monetised, data must be accessible. There is also a further risk of data being intercepted during upload to cloud-based services.

**Explanation of method**

Upon purchasing the Pixel 3a device, I signed into it with my existing Google account and downloaded the Google Recorder software, as well as the Microsoft Word and Microsoft Outlook applications from the Google Play app store. Automated transcription using the device was tested initially using existing interview recordings stored on my computer; playback was attempted using both the in-built speakers on my computer and with the computer running through a Hi-Fi amplifier and Hi-Fi speakers. A rough comparison between transcriptions of the same interview performed using these two different playback set-ups yielded no significant differences; some words had been captured more accurately using the Hi-Fi speakers, others more accurately using the computer’s speakers (see Table 1).

Table 1 shows little difference between the two playback methods in terms of word recognition; neither seems more or
Table 2. Examples of raw automated transcription versus manual transcript.

| Raw automated transcription                                                                 | Manual transcript                                                                 |
|---------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| yeah Thursday line, they will escalate a third line issue at the I’ve said tonight.        | Yeah, first and second line, or they will escalate a third line issue. At, I’ve   |
| I’d rather you escalate more until you understand better what so somebody wants low      | said to them I’d rather you escalate more until you understand better what’s, so,  |
| climbing rights that eat compositions now. So once a pest a few times then it will go    | somebody wants local admin rights, the default position’s no, (,) erm so once they’ve  |
| to one of my team to understand the [judgification sic] and then approval of the eye     | pestered a few times then it will go to one of my team to, understand the          |
| right I’ve empowered the team say look.                                                    | justification and then approve or deny.                                           |
| You know, you’re technical experts. If you say no have a very good reason why you’re     | I’ve empowered the team to say look, you know, you are                           |
| saying no but give them an alternative and so I’ve been paired the team to say, all right!| the technical experts, if you say no, have a very good reason why you’re saying   |
| Yes we can do this and I’ve opened the door a lot more to be able to say yes.             | no but give them an alternative, and so I’ve empowered the team to say, alright,  |
|                                                                                           | we can do this and I’ve opened the door a lot more to be able to say yes.         |

less successful than the other. I decided to use the Hi-Fi speaker set-up from this point forward because this offered a more controllable environment; due to the set-up of my home Hi-Fi, which is in an attic room with a well-insulated door, I was able to ensure no one else in the house14 was able to overhear the audio playback. Taking the computer out of the equation for future transcriptions also removed the risk of high processor overheads or, more likely, unplanned shutdowns of the computer (e.g. to install updates) negatively impacting upon playback.

My most recent interview recording (from 2 days previous to this first test) was used as the sample data; this was played straight from the encrypted recording device15 using a Hi-Fi amplifier and Hi-Fi speakers.16 The transcription device (Pixel 3a) was placed on a platform 80 cm from the floor and 57 cm from one speaker, the base of which was 85 cm from the floor. The device was set to ‘airplane’ mode in order to reduce any risk of automated synchronisation of the recording data; although Google advise that data from this application is only ever stored on the device and is not synchronised (Google, n.d.b), I wanted to make sure. Once the recording had completed, the text of the transcript was copied into a new Word document on the device and saved onto the device itself, following which the document was attached to a new email using the Outlook app on the device (configured to use my University account; no other email accounts were configured on the device) and sent to myself at my University email address. Extra care was taken to double-check the recipient address to ensure data were not inadvertently sent to the wrong address. The recording was then deleted from the device. Next, the device was taken out of airplane mode and connected to my secure home Wi-Fi network17 to allow the emails to be sent. I then downloaded the transcribed document onto my computer and deleted the email, including the sent email with the attachment, from my email account.18 I then opened the document and commenced playback of the original interview recording using headphones, cutting and pasting the raw transcribed text into a blank version of my standard interview transcription template (a tabular format that separates different participants’ speech into separate cells), correcting the transcription as I went. This process required close listening to every word of the interview, as per my previous fully manual transcriptions, but took significantly less time (see ‘Findings’ section). As per my previous manual transcriptions, all identifying or confidential information was redacted.

Findings

The method achieved what I considered to be a ‘good enough’ first-pass transcript. Table 2 provides examples of the raw automated transcript (excluding timecodes) when compared to a previously produced fully manual transcript of the same interview.

This table provides two examples; the first shows a low degree of accuracy (WER 31%) when the raw automated transcript is compared to the manual transcript while the second shows a high degree (WER 15%). Note that these WER figures only relate to the excerpts earlier, rather than the whole transcript. It was clear to me that, although this speech recognition capability was promising, relying solely on a fully automated transcription would not be feasible.

Table 3 compares a raw automated transcription (including the timecodes automatically inserted by the software) to the resulting semi-automated transcript, which was achieved by close listening to the audio with simultaneous manual correction.

Table 3 provides an example of the amount of effort required to transform the raw output into a finished transcript; in this particular example, it is apparent that the effort required was minimal. Note the insertion of bracketed dots in the finished transcript; in my transcription notation, this indicates a pause of more than 1 second, which was not captured in the raw transcription.

Tables 4 and 5 summarise the time taken to transcribe 10 interviews, 5 using manual transcription and 5 using the method summarised earlier; a significant reduction in transcription time was observed. Admittedly, the interviews are different; I did not see the value in attempting to perform a semi-automated transcription of interviews that I had already transcribed and even if there were seen to be value in

Examples of raw automated transcription versus manual transcript.

Table 2.

| Raw automated transcription | Manual transcript |
|-----------------------------|-------------------|
| yeah Thursday line, they will escalate a third line issue at the I’ve said tonight. | Yeah, first and second line, or they will escalate a third line issue. At, I’ve said to them I’d rather you escalate more until you understand better what’s, so, somebody wants local admin rights, the default position’s no, (,) erm so once they’ve pestered a few times then it will go to one of my team to, understand the justification and then approve or deny. |
| I’d rather you escalate more until you understand better what so somebody wants low climbing rights that eat compositions now. So once a pest a few times then it will go to one of my team to understand the [judgification sic] and then approval of the eye right | I’ve empowered the team to say look, you know, you are the technical experts, if you say no, have a very good reason why you’re saying no but give them an alternative, and so I’ve empowered the team to say, alright, we can do this and I’ve opened the door a lot more to be able to say yes. |
Comparing the two methods using the same interview data, given the manual transcription had already been done (before exploring the other method), the semi-automated method would have been biased by prior knowledge of what was said. Therefore, it was decided to progress with the semi-automated method on interview data arising from subsequent interviews (which were already scheduled).

These tables show that there is a clear difference in the time taken to transcribe using the semi-automated method. Note that the times stated in Table 5 do not include the real-time playback of the original interview recording so that the transcription device could re-record and transcribe it. As this activity was performed in an unattended mode, in other words, where I could perform other productive tasks while it was running, this was not considered to be part of the transcription time. It should be noted that, had I used the transcription device in a real-time manner during the original interview, no subsequent playback and re-recording would have been necessary to produce the raw transcript. This was not performed at the time as, at that point, I had not developed this method.

**Discussion**

The ‘rule of thumb’ ratio between time taken to transcribe versus length of recording appears to be 5:1 as expressed by Bokhove and Downey (2018: 2) and Cohen et al. (2013: 539), both citing Walford (2001), as well as by Bryman (2012: 484); my experience was around 7:1 as shown in Table 4. The use of the semi-automated method described earlier appears to have reduced this ratio considerably, in these cases at least, to roughly 3:1, as shown in Table 5. However, it should be acknowledged that there are likely to be order effects at play, in the sense that my interview transcription performance should have improved over time and the interviews in Table 5 did take place after those in Table 4. There is a clear improvement between the first two interviews I performed (Interviews 1 and 2 in Table 4) which suggests an order effect. I consider it unlikely, however, that greater experience at transcription would have resulted in the scale of the improvement shown between Tables 4 and 5.

I observed that the software automatically excluded ‘um’s and ‘er’s which was viewed as beneficial, not least from the perspective of how such conversational ‘fillers’ (Collins et al., 2019: 664) may be interpreted, as identified by other researchers (Bucholtz, 2000, 2007; Collins et al., 2019). Collins et al. (2019), in particular, suggest that the inclusion of these may be incorrectly, or unfairly, interpreted as uncertainty and encourage a thoughtful consideration as to whether or not they should be included in a transcript. Given the software tested earlier automatically makes that decision, however, that greater experience at transcription would have resulted in the scale of the improvement shown between Tables 4 and 5.

The software inserts timecodes at various points and there was no apparent pattern to these other than them appearing at breaks in speech; they appeared roughly every 30–60 seconds. The software also automatically inserted commas,
percentage ampersands and Arabic numerals although both commas and numerals were not inserted consistently; in some cases, numbers were captured as text and there was no clear pattern as to why the software would sometimes choose one option rather than the other. Slang such as ‘gonna’ was recognised and transcribed verbatim by the software, which also constructed non-existent words, presumably based on phonetics, such as ‘judification’ (the word enunciated by the participant was ‘justification’). All words transcribed used Americanised spellings, and this was not configurable.

My experience of using this method has been overwhelmingly positive. Once I had established its efficacy, I noticed that I was no longer approaching the task of transcription, or the preceding interviews, with a sense of dread (Bird, 2005). Somewhat ironically, the use of automation made performing the transcription feel less mechanical; I had noticed that towards the end of my 4- or 5-hour manual transcription marathons that I was becoming somewhat of an automaton myself, processing the words I was hearing through my ears into words on a page through my fingers but not really listening to what was being said or the meaning of it. With the semi-automated method, I felt that I was paying much more attention, partly because of the reduced time period and the associated reduction in fatigue but also because not only was I listening and typing, I was also reading; reading the automated transcript, correcting it and organising it as I went along meant that I was listening much more closely than when I was transcribing manually. The time reduction has also allowed me more opportunity to add in comments to the transcript as I went along; when transcribing manually, I did not feel I could afford the time to do this (particularly when I knew there was at least another 3 hours to go on a Sunday evening when I had work in the morning). This method has helped me get a lot closer to my data and I have not lost control of the transcripts to a piece of software; quite the opposite.

This article supports the argument that there is significant benefit to qualitative researchers from using automated transcription; this argument itself is not novel, for example, Bokhove and Downey (2018) and Moore (2015). However, I have presented a method that is not just effective but also addresses the security concerns arising from the use of cloud-based services and can be achieved at relatively low cost (assuming multiple interviews are planned). As transcripts are, by nature, ‘incomplete’ (ten Have, 1990: 33) and Lapadat and Lindsay (1999: 71) and imperfect (DeVault, 1990: 108), they do not objectively represent the interview scenario; they are ‘not just talk written down’ (Green et al., 1997: 172). The manner of their production, whether fully manual or semi-automated, does not alter this fact; the production of the transcript remains ‘interpretative and constructive’ (Lapadat and Lindsay, 1999: 72) and ‘theory laden’ (Lapadat and Lindsay, 1999: 64). Importantly, and in line with more experienced researchers, I believe that the resulting transcripts will ‘adequately serve my analytic purposes’ (Silverman, 2005: 27). Following what I consider to be a successful definition of this method, I am proposing to apply it to my future interviews. I believe this will still achieve a close relationship with the data, and the time saved versus fully manual transcription will allow me to focus more time on analysis.

Limitations of the method and future research directions

The restriction of the software to specific hardware is a significant limitation. As discussed earlier, although installation on other Android-powered hardware is technically possible, it introduces significant security risks and may have legal implications. Therefore, achieving this method securely requires specific equipment that comes at a cost which many researchers will be loath to incur. If researchers already own a suitable Android-powered device, they may be tempted to forgo the security risks of ‘sideloading’ in order to achieve the transcription benefits, undermining the security advantages of this method.

Beyond the limitation of the upfront investment required, the software used by this method is limited to the English language which is a significant drawback for non-English speakers; it is unclear whether language support will be expanded in future iterations of the software. It is also possible (although there is no indication of this) that future versions of this software will not operate in a solely offline manner; if so, this would undermine the security benefits of this method. Importantly, this research has not explored whether there are any differences in the effectiveness of this method with regard to native versus non-native English speakers; other researchers, for example, McGurk et al. (2008) have identified higher error rates in other (albeit more dated) voice recognition software with regard to non-native speakers and, therefore, this could be a significant limitation of this method. It was noted that the software described earlier successfully transcribed a number of different English regional accents, although not with a fully representative sample by any means, as well as Scottish and American accents. Accuracy appeared to be reduced when transcribing a Scottish accent and improved when transcribing an American one, although clearly there was a very limited sample, with only one example of each. Further cases would need to be explored in order to identify any differences in performance in this regard; notably, when discussing voice recognition more generally, Cramer et al. (2018) identified ‘underperform[ance] for users with certain accents’ (p. 59). It should also be noted that, although there was variation in accents, all of the participants (and myself as the interviewer) had well developed oral communication skills and could be, relatively speaking, considered either elites or sub-elites. They also had a similar educational background, all having at least a Bachelor’s degree.

There was no discernible difference in transcription accuracy between male and female voices, however, there was only one female interviewee in this sample and, therefore, this
is far from conclusive. Gender bias in voice recognition systems has been previously established, as described by Howard and Borenstein (2018: 1525), and such bias has been identified by Tatman (2016) in the specific software used by Bokhove and Downey (2018). While Bokhove and Downey (2018) and I utilise different specific software applications, both are provided by Google and may be based on the same technology and therefore both may suffer from the same bias Tatman (2016) identifies, which could further limit this research. This limitation is likely to extend beyond Google-authored applications to other voice recognition software, particularly as 'the single most popular speech corpus in the Linguistic Data Consortium, is just over 69% male' (Tatman, 2016).

Such limitations could indicate, whether intentional or not, elitist aspects to voice recognition which could be linked to other forms of bias in artificial intelligence software as discussed in detail by Howard and Borenstein (2018). If data corpuses used in the development of such software are overwhelmingly male, native speakers with a limited range of dialects then this software will continue to be most effective for interviewees who reflect that bias. However, eradication of all bias may not be achievable; Olteanu et al. (2019) even question whether removal of that bias is advisable, suggesting that biases, however ethically questionable, may improve performance of artificial intelligence (AI) applications (Olteanu et al., 2019: 23–24), pointing to other research that supports this argument. Nevertheless, the question of ‘progress at what cost?’ remains unanswered.

McGurk et al. (2008) also identified that noise levels in the immediate environment affected the error rate in their voice recognition study; my research has been conducted in fairly controlled circumstances, always indoors and in a closed room. In a different environment with more background (and likely foreground) noise such as the hospital environment studied by McGurk et al. (2008), or outdoors, then this method may not be as effective. Notably, in a subsequent interview transcribed using this method with a participant who had a particularly deep and booming voice and where the room was quite reverberant, the level of accuracy appeared to be reduced. Therefore, it should be recognised that this method, while appearing to be effective under a particular set of conditions, may not be effective for all and even in controlled conditions could have variable results. At best, this study indicates that with predominantly male interviewees and a male interviewer, all of whom are well educated with English as their first language and in a controlled indoor office environment, the particular software and method utilised was effective in reducing transcription time.

It was also observed that the software could not distinguish different voices; all speech was transcribed as if it were one voice. This is a limitation of other automated transcription software, as experienced by Matheson (2007), whose method to counteract this was not applicable in this instance. This was merely a mild irritation to me; as I was assembling the transcript through close listening to every word, and my transcription format already required separate tabular rows for each speaker, it was not much more effort to move the relevant text adjacent to the appropriate name. However, my interviews only involved two people; in an interview environment with more than two participants, or in another setting such as a focus group, the inability to distinguish between different speakers could be a more serious limitation.

A further limitation of this method is that it requires playback of an interview recording rather than the uploading of a file. This could be a significant drawback in instances where a high volume of audio data requires transcription, in which instance, a solution that permits uploading of audio files may be more suitable, albeit with the security implications discussed earlier.

As well as exploring the biases and environmental factors mentioned earlier, future research could consider measuring the accuracy of the software, particularly if seen as an alternative to (less secure) bureau-based human transcription services. Quantifiable accuracy was not an objective of this particular study, as this method focuses on generating ‘good enough’ results.

Conclusion

The method described in this article provides, in certain controlled circumstances, an effective means of generating ‘good enough’ transcripts with greater security than that described by Bokhove and Downey (2018), addressing the ethical issues associated with the use of cloud-based transcription services. By removing the dependence on external computing resources that could be accessed by third parties, as well as removing the risk of data interception during the uploading of interview recordings to such resources, this method addresses a number of security concerns associated with cloud-based transcription. The method has a number of limitations as discussed earlier but it offers significant value to researchers and students both from the perspective of saving time and avoiding the ‘dread of transcription’ (Bird, 2005: 233) but also, more importantly, by potentially enabling greater engagement with interview data.

Acknowledgement

The author would like to thank the anonymous reviewers for their valuable feedback.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship and/or publication of this
article: At time of writing (January 2020), the author has a small shareholding (held for approximately 2 years, through a private pension) in Alphabet Inc, which is the parent company of Google.

Funding
The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD
Joseph Da Silva https://orcid.org/0000-0002-4153-6237

Notes
1. Although interviewing as a technique is not without criticism, for example, Whitaker and Atkinson (2019).
2. For example, ‘[w]e may select segments of raw or partially processed audio and associated transcribed text from multiple recorded conversations from different users to be used as training data . . . ’ (Otter.ai, n.d.a)
3. It is also worth noting that it can be unclear to users what the exact terms and conditions applicable to individual cloud services are, and furthermore that these can change without warning (Kamarinou et al., 2016: 186–188).
4. These are discussed in more detail in the ‘Limitations’ section.
5. The Google Recorder software can be ‘sideloaded’ on to certain other Android-powered devices, however, this introduces significant security risks (Samsung Business Insights, 2019) not least as it requires changes to the devices security settings to allow ‘Unknown Sources’ and there is considerable opportunity for malicious actors to modify applications before users download them. Sideloaded also has legal implications (Android Central, 2017; Google Play, n.d.). The use of unsupported devices may also provide an inconsistent user experience when compared to use on an ‘official’ device.
6. I have emailed the Google app development team to ask them to clarify the reasoning behind this but have had no response.
7. A broad seam of literature exists in relation to the specific experiences of elite interviewing, for example, Gilding (2010), Conti and O’Neill (2007), Platt (1981).
8. As a practicing Chief Information Security Officer (CISO) at a senior level, I arguably fall into this category myself; Boucher (2017) refers to research that classifies ‘functional experts who hold expertise even if they are of not high organizational stature’ (p. 100) as elites.
9. My first fieldnote relating to transcription says ‘Hard work!’; a subsequent note says ‘[later in the day] Has taken ~5 hours to do 35 mins of recording. Hard work! Exhausted and still not done – will have to do more tomorrow . . . took another 3 hours to complete’.
10. After the second transcription I noted ‘Feeling more positive about write up. Think this will improve as well’.
11. Securely, in an encrypted format.
12. Note that all interviewees had signed consent forms and had given their permission for the use of anonymous verbatim quotations.
13. It was clear from reading the transcripts that they were not sufficiently accurate to render manual adjustment and correction unnecessary.
14. Which has only two inhabitants including myself.
15. Olympus DS7000.
16. Arcam Solo Mini amplifier and Monitor Audio Silver RX2 speakers.
17. I have not inserted a SIM card into this device and do not intend to; it cannot access cellular networks as a result.
18. It is worth noting that a more secure method of transferring the transcribed document would be to connect the device directly to the computer, for example, through USB; this would remove the need for any emailing of the file and, therefore, remove the risk of inadvertently sending it to the wrong recipient. This option was not available to me due to computer device restrictions that prevented direct connection of devices.
19. Word Error Rate = 100 (Insertions + Substitutions + Deletions)/total words (Moore, 2015: 258)
20. Note: in this calculation, I excluded the ‘er’ captured in the manual transcription as the software deliberately excluded this; given I am no longer transcribing ‘ums’ and ‘ers’ as mentioned earlier, I do not consider this an inaccuracy.
21. Note that I disabled automatic software updates for this very reason.
22. As briefly discussed in the ‘Related Work’ section earlier.

References
Akbar S, Kharabsheh B, Poletti-Hughes J, et al. (2017) Board structure and corporate risk taking in the UK financial sector. International Review of Financial Analysis 50: 101–110.
Android Central (2017) Is sideloading APK and files considered piracy? [WWW document]. Android Central. Available at: https://www.androidcentral.com/ask-ace-sideloading-apk-files-piracy (accessed 3 May 2020).
Ashenden D and Sasse A (2013) CISOs and organisational culture: Their own worst enemy? Computers & Security 39: 396–405.
Beasley MS, Carcello JV, Hermanson DR, et al. (2009) The Audit Committee oversight process. Contemporary Accounting Research 26: 65–122.
Bevan O, Boehm J, Manocaran M, et al. (2018) Cyber risk strategy for cybersecurity/McKinsey[WWW document]. Availableat: https://www.mckinsey.com/business-functions/risk/our-insights/cybersecurity-and-the-risk-function (accessed 2 December 2018).
Bird CM (2005) How I stopped dreading and learned to love transcription. Qualitative Inquiry 11: 226–248.
Bokhove C and Downey C (2018) Automated generation of ‘good enough’ transcripts as a first step to transcription of audio-recorded data. Methodological Innovations 11: 790743.
Boucher A (2017) Power in elite interviewing: Lessons from feminist studies for political science. Women’s Studies International Forum 62: 99–106.
Braun V and Clarke V (2006) Using thematic analysis in psychology. Qualitative Research in Psychology 3: 77–101.
British Sociological Association (2017) BSA statement of ethical practice (WWW document). Available at: https://www.brit-soc.co.uk/media/24310/bsa_statement_of_ethical_practice.pdf (accessed 5 January 2020).
Bryman A (2012) Social Research Methods (4th edn). Oxford: Oxford University Press.
Bucholtz M (2000) The politics of transcription. Journal of Pragmatics 32: 1439–1465.
Bucholtz M (2007) Variation in transcription. Discourse Studies 9: 784–808.
Cohen L, Manion L and Morrison K (2013) Research Methods in Education. New York: Taylor & Francis.
Platt J (1981) On interviewing one’s peers. The British Journal of Sociology 32: 75–91.

Rainford J (2020) Confidence and the effectiveness of creative methods in qualitative interviews with adults. International Journal of Social Research Methodology 23: 109–122.

Samsung Business Insights (2019) What are the risks of side loaded android applications? (2019) Samsung Business Insights, 26 November. Available at: https://insights.samsung.com/2019/11/26/what-are-the-risks-of-sideloaded-android-applications/ (accessed 3 May 2020).

Seale C (2000) The Quality of Qualitative Research, Introducing Qualitative Research. Thousand Oaks, CA: SAGE.

Silverman D (2005) Doing Qualitative Research: A Practical Handbook. Thousand Oaks, CA: SAGE.

Singh AN, Picot A, Kranz J, et al. (2013) Information security management (ISM) practices: Lessons from select cases from India and Germany. Global Journal of Flexible Systems Management 14: 225–239.

Spiekermann M (2019) Data marketplaces: Trends and monetisation of data goods. InterEconomics 54: 208–216.

Tatman R (2016) Google’s speech recognition has a gender bias. Available at: https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/ (accessed 24 January 2020).

ten Have P (1990) Methodological issues in conversation analysis. BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique 27: 23–51.

Thomas D (2016) TalkTalk chief signals change after cyber attack (WWW document). Financial Times. Available at: https://www.ft.com/content/1e03001c-e22b-11e5-96b7-9f778349aba2 (accessed 15 November 2018).

Tolich M (2004) Internal confidentiality: When confidentiality assurances fail relational informants. Qualitative Sociology 27: 101–106.

UK Government (2016) National cyber security strategy 2016 to 2021 (WWW document). GOV.UK. Available at: https://www.gov.uk/government/publications/national-cyber-security-strategy-2016-to-2021 (accessed 17 January 2019).

Van der Kleij R, Kleinhaus G and Young H (2017) Computer security incident response team effectiveness: A needs assessment. Frontiers in Psychology 8: 2179.

Voreacos D, Chiglinsky K and Griffin R (2019) Merck cyberattack’s $1.3 billion question: Was it an act of war? Bloomberg.com, 3 December. Available at: https://www.bloomberg.com/news/features/2019-12-03/merck-cyberattack-s-1-3-billion-question-was-it-an-act-of-war

Walford G (2001) Doing Qualitative Educational Research: A Personal Guide to the Research Process. New York: Continuum.

Walford G (2005) Research ethical guidelines and anonymity. International Journal of Research & Method in Education 28: 83–93.

Whitaker EM and Atkinson P (2019) Authenticity and the interview: A positive response to a radical critique. Qualitative Research 19: 619–634.

Wiles R, Crow G, Heath S, et al. (2008) The management of confidentiality and anonymity in social research. International Journal of Social Research Methodology 11: 417–428.

Author biography

Joseph Da Silva is a part-time PhD researcher within the Information Security Group at Royal Holloway, University of London and is performing multidisciplinary research into the purpose of cybersecurity functions within commercial organisations. He works full-time as a Chief Information Security Officer (CISO).