Author/s:
Loakes, D

Title:
Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?

Date:
2024

Citation:
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.

Persistent Link:
https://hdl.handle.net/11343/345908

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 14 February 2024
DOI 10.3389/fcomm.2024.1281407

OPEN ACCESS

EDITED BY
Dominic Watt,
University of York, United Kingdom

REVIEWED BY
Chiara Barattieri Di San Pietro,
University Institute of Higher Studies in Pavia, Italy
Vincent Hughes,
University of York, United Kingdom
Cristina Aggazzotti,
Johns Hopkins University, United States

*CORRESPONDENCE
Debbie Loakes
✉ dloakes@unimelb.edu.au

RECEIVED 22 August 2023
ACCEPTED 09 January 2024
PUBLISHED 14 February 2024

CITATION
Loakes D (2024) Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?
Front. Commun. 9:1281407.
doi: 10.3389/fcomm.2024.1281407

Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?

Debbie Loakes*

Research Hub for Language in Forensic Evidence, School of Languages and Linguistics, The University of Melbourne, Parkville, VIC, Australia

This study provides an update on an earlier study in the "Capturing Talk" research topic, which aimed to demonstrate how automatic speech recognition (ASR) systems work with indistinct forensic-like audio, in comparison with good-quality audio. Since that time, there has been rapid technological advancement, with newer systems having access to extremely large language models and having their performance proclaimed as being human-like in accuracy. This study compares various ASR systems, including OpenAI's Whisper, to continue to test how well automatic speaker recognition works with forensic-like audio. The results show that the transcription of a good-quality audio file is at ceiling for some systems, with no errors. For the poor-quality (forensic-like) audio, Whisper was the best performing system but had only 50% of the entire speech material correct. The results for the poor-quality audio were also generally variable across the systems, with differences depending on whether a .wav or .mp3 file was used and differences between earlier and later versions of the same system. Additionally, and against expectations, Whisper showed a drop in performance over a 2-month period. While more material was transcribed in the later attempt, more was also incorrect. This study concludes that forensic-like audio is not suitable for automatic analysis.

KEYWORDS

forensic linguistics, transcription, automatic speech recognition (ASR), phonetics, artificial intelligence

1 Introduction

This study provides an update on Loakes (2022), which aimed to demonstrate how automatic speech recognition (ASR) systems work with indistinct forensic-like (poor-quality) audio, in comparison with good-quality audio. The original study was motivated by misunderstanding, particularly within the law, around the problem of what is said in indistinct forensic audio being solved automatically. As discussed in that study, this is a question that needs to be explored experimentally, and the current study is intended as confirmation that the unsuitability of ASR for forensic transcription remains, despite recent improvements.

Forensic audio is audio that is generally captured in high stakes and often criminal contexts. This type of audio is defined by Fraser (2022): 8, as

> *…speech that has been captured, typically in a covert (secret) recording obtained as part of a criminal investigation, and is later used as evidence in a trial. Such recordings provide powerful evidence, allowing the court to hear speakers making admissions they would not make openly. One problem, however, is that the audio is often extremely indistinct, to the extent of being unintelligible without the assistance of a transcript.*

The original idea behind the research in Loakes (2022) was to address the fact that computational methods are sometimes seen as a solution to solve the issue of what was said in indistinct recordings. This is part of a wider belief system, dubbed as *technosolutionism* (Morozov, 2013) in which technology is seen as the solution to any problem. Loakes (2022) looked at a poor-quality recording, which was livestreamed via an iPhone, and contained multiple voices with overlapping noise and variable distances from the microphone. That study analysed a good-quality recording by comparison, also recorded via an iPhone, but containing only one speaker who was specifically focussed on being understood. Based on the results of Loakes (2022), it was concluded that AI systems work well when applied to tasks they are designed for—with non-overlapping speech in a language variety the system is familiar with—but poorly when there is background noise, speakers who are not stationary, and when the signal is indistinct, which is all characteristic of forensic audio.

In the short time since that study was published, the availability of more advanced AI systems, especially Open AI's ChatGPT, has changed the artificial intelligence[1] landscape. ChatGPT in particular has received swathes of attention in academic and popular literature. The availability of ASR systems has also risen rapidly, again in particular Open AI's *Whisper*. While there are some critical analyses of artificial intelligence and its role in society (Bender, 2022; Preston, 2022; Bridle, 2023; Perrigo, 2023), there is also still much, less critical, attention on how well these ASR systems work and how much time they save. For example, there are popularly available articles citing Whisper as being 'an ASR model that shows human levels of accuracy and robustness' (Rodriguez, 2022), yet this itself assumes human accuracy is infallible, and anyway the accuracy and robustness appear true only in some limited circumstances.

This study aimed to critically assess the use of Whisper, and some other ASR systems using large language models (e.g., Kallens et al., 2023), to determine how accurate they are in transcribing a section of poor-quality forensic-like audio. Specifically, the aim was to provide new data to compare how the current generation of ASR systems performs when tasked with the transcription of indistinct forensic-like audio (e.g., Loakes, 2022).

2 Background

Automatic speech recognition is not designed for forensic transcription, yet it is often seen by legal professionals as a possible option for the solution of what is being said in indistinct recordings (see, e.g., the discussion in Loakes, 2022). This belief assumes that automatic methods are somehow free from bias and should be more objective than human transcription. However, these automatic systems are of course designed by humans and have in-built biases in their training data (e.g., Koenecke et al., 2020; Wassink et al., 2022). In fact, the more advanced these systems are becoming, the more these inherent biases are also coming to the fore. Talking about ChatGPT's predecessor, GPT-3, Perrigo (2023) notes that its outputs originally involved inappropriate and offensive content, which was then later screened to improve usability by very low-paid workers so its 'huge training dataset was the reason for GPT-3's impressive linguistic capabilities, but was also perhaps its biggest curse'.

In automatic speech recognition, biases are in the direction of better recognition of 'standard' accents (Markl, 2022; Wassink et al., 2022; Harrington, 2023), one or other of male or female voices depending on the system (Markl and McNulty, 2022) as well as non-pathological voices (Benzeghiba et al., 2007; Markl and McNulty, 2022). Additionally, as noted by Benzeghiba et al. (2007), children's voices and elderly voices are also generally not modelled well and cause performance issues with ASR.

It is, nevertheless, important to continue to investigate the issue experimentally to determine limits in ASR performance, as the current study aims to do. In Loakes (2022), the good-quality recording was transcribed well, while the poor-quality recording was not. For example, one of the commercially available systems, Descript, had approximately 96% correct recognition of the good-quality recording and 1.7% correct recognition of speech in the poor-quality recording. That study also demonstrated that using a transcript and trying to align it with speech events using a forced-aligner is replete with problems—the system forces boundaries onto speech events that are not present and may look correct to non-linguists even when it is clearly not. For example, in that study, drumming noise and laughter were aligned with speech events (Loakes, 2022): 9.

Since Loakes (2022) addressed the issue of how ASR copes with indistinct forensic-like recordings, some new work in this space has been conducted with the newer generation of ASR systems which further demonstrates some of the issues discussed above; however, this new research has not made use of *Whisper*. Similar to Harrington et al. (2022), Loakes (2022) carried out a comparison of various ASR systems with recordings known to be difficult for human transcribers, as reported by Love and Wright (2021). They used 18 British English utterances of which they could be certain of the content and used 12 commercially available ASR systems to compare how well the systems transcribed forensic-like audio. They found extreme variability in system responses, ranging from a 70% match across the *Microsoft Transcribe* and the ground truth transcripts, compared to 13.9% for Sonix [which also had low performance in Loakes (2022) for the data analysed in this study]. Harrington et al. (2022) note that errors relate to a degree of phonetic similarity between the error and the actual word spoken, as well as predictability errors from training data. Examples are the word *worrying* mistaken to be *varying* and *chicken tikka masala* (likely low frequency) mistaken to be *she can take*. The authors conclude that managing and interpreting the output of such systems is more effortful than having a human transcribe the data in the first place.

Harrington (2023) considered the use of ASR for police-suspect interviews, with a view to making the process more efficient and

1 In Loakes (2022) *artificial intelligence* was defined as "intelligence demonstrated machines instead of humans" (c.f. McCarthy, 2007). Other researchers have also noted that artificial intelligence nevertheless has origins in "human contrivance and ingenuity" (Fetzer, 1990).

potentially using human post-editing. She compared three commercial ASR systems (*Rev AI, Amazon Transcribe*, and *Google Cloud Speech to Text*) to assess how they performed across accents and recording qualities. She looked at audio from the DyViS database (Nolan et al., 2009), with Standard Southern British English and West Yorkshire English speakers, using both studio quality files and files with speech-shaped noise added to degrade the signal. Harrington (2023) observed three main kinds of errors with the systems, which involved insertion of material (extra words in the output compared with the transcript), deletion (missing words), and substitution (a mismatch between the reference transcript and the output). She also describes varying levels of success across the systems, noting that errors were higher with West Yorkshire English speakers, whose accents were likely represented less in the training materials, and she also noted different kinds of errors across the accents. Unsurprisingly, Harrington (2023) found that the audio quality affected the performance of the systems. She found that *Amazon Transcribe* had the lowest error rates regardless of whether it was focussed on the studio condition or the speech-shaped noise condition, while *Rev AI* was the most variable. Similar to findings from Loakes (2022), she found that even the best performing system did not accurately transcribe all of the material. Harrington (2023) showed a 13.9% word error rate (WER) for *Amazon Transcribe* with Standard Southern British English in the studio condition and 15.4% WER in the speech-shaped noise condition. The worst performance was for *Rev AI*, which had an error rate of 42.5% with the degraded speech for the West Yorkshire accent.

Another recent study by Harrington and Hughes (2023) looked at the variability of the ASR system. Using *Amazon Transcribe*, which was the best performing system in the study by Harrington (2023), the aim was to look at variability in performance with a homogenous group of speakers, and whether the errors observed correlated with particular phonetic properties. Using the DyViS database (Nolan et al., 2009) and focussing on 'homogenous' speakers with the same accent, Harrington and Hughes (2023) observed that for 99 speakers, WERs ranged between 11.2 and 33%, with a mean of 20% errors across the entire sample. They analysed various phonetic properties which included F0, formants, articulation rate, and voice quality to determine which features predicted performance and found that only articulation rate predicted WER. Taking all results together, Harrington and Hughes (2023) discuss how phonetic reasons for performance issues (even in clear speech) are not clearly predictable, and identifying causes of variability is also problematic. Harrington and Hughes (2023, 3134) note that the number of errors they observe in their homogenous sample of clear speech recordings (with 11.2% WER being the best performance) is 'worrying given the favourable [speech and recording] conditions… and raises issues about the general utility of ASR for many applications'.

The findings of Loakes (2022) and Harrington (2023) in particular are entirely consistent with known issues in automatic speaker recognition when degraded audio is used. For example, in a review paper about trends and developments in ASR, O'Shaughnessy (2023, 2) describes how sponsored challenges address the matter of 'noisy, far-field multi-speaker conversations' being difficult for systems, having up to 50% word error rates for automatic systems; other studies have shown approximately 15% word error rate for automatic systems in which humans can understand speech well. However, as has been shown in this section, even clear speech recordings (Harrington and Hughes, 2023) can have relatively high error rates without a predictable cause.

Analysis of the performance of ASR also brings into question how well human transcribers perform in forensic-like transcription tasks, and while this is not the focus of the study, it is important to address how humans perform in comparison. As mentioned earlier, Harrington (2023) analysed the output of 12 ASR systems, and this same audio was transcribed by professionally trained human transcribers in the study by Love and Wright (2021). While neither the humans nor the systems were able to provide accurate transcriptions of the entire recording, Harrington (2023) concluded that 'at present, it is more effective for humans to transcribe indistinct audio 'from scratch' as opposed attempting to manage and interpret the output of such systems'.

There is a similar finding to this in a recent experiment (Fraser et al., 2023), in which our team focussed on transcription performance from human transcribers who were presented with a section of audio from the same recording as the one used in this study (as well as in Loakes, 2022). Fraser et al. (2023) focussed on how well transcribers performed and saw that overall accuracy was relatively low, but still the top 11 transcribers (of a total of 40) were able to accurately transcribe between 50 and 62% of the material.

The new generation of automatic speech recognition systems needs to be tested because they are iterative and predictive and have access to masses of data compared to systems available only a year ago (e.g., Kallens et al., 2023). Any discussion of how automatic speech recognition performs with poor-quality forensic-like audio, therefore, needs to include these updated systems, because they have the potential to perform better than the older systems which do not draw on large language models, but their performance nevertheless needs to be analysed critically.

3 Aim

The aim of this study was to continue to update knowledge of ASR, and how it performs when applied to indistinct forensic-like audio. This research report is a direct update on a previous article (Loakes, 2022) which looked at forced alignment and smaller language model ASR systems and how they transcribed good-quality audio compared with poor-quality audio. It is also an update of some work by other teams which has looked at the matter of how naturalistic forensic-like audio is handled in modern ASR systems (e.g., Harrington et al. 2022). Given the rate of rapid technological advancement, it is imperative to test the new generation of automatic speech recognition systems, which have to date not been included in work on forensic transcription. In total, eight different systems using deep-learning and large language models are tested in this study—and taking into account updated versions and different file types there are 14 different ASR attempts on both the good- and poor-quality audio files.

The scope of this study is purposefully limited in experimentally providing an initial focus on how the newer generation of ASR systems performs on a sample of poor-quality audio, compared to a sample of good-quality audio. This means that only broad conclusions about the efficacy of automatic speech recognition in forensic contexts can be drawn, but this study nevertheless aims to contribute to the ongoing conversation about this rapidly advancing technology and how it is used and understood in forensics. The ensuing conclusions

of this approach may indeed be obvious to linguists, but the goal of this study was to inform a broader audience about the issues.

4 Methods

To give more detail about the audio files used in this study (also used in Loakes, 2022), these are as follows:

4.1 Poor-quality audio

This is a 44.2-s stretch of audio from a recorded rehearsal by a singer and some musicians. This audio includes speech and instrument noise and is forensic-like in that there are varying background noises, there are multiple speakers who are at a distance from the microphone, and there is overlapping speech. This audio was recorded by one of the speakers via an iPhone and streamed to Facebook live, where it was retrieved with permission. The reference transcript has been verified with one of the speakers who organised and streamed this event, and the researcher's access to the accompanying video meant that the sample was clearer than the audio-only version (Loakes, 2022). The recording used has one female voice and three male voices, and all speakers are using Australian English. The speakers knew they were being recorded but were focussed on the task at hand and not attempting to be clear to the audience.

A transcript of this audio is provided below.

4.1.1 Poor-quality audio transcript

> *Yeah so just slowly building energy and nnnn and then I yeah*
> *What about what about another big drum fill will you let us know when you*
> *Yeah*
> *Alright*
> *Nah nah*
> *You gonna give us a hand signal or tell us what you do*
> *I I can't [laughter] ok*
> *From the from the top are we fine to go there*
> *Mel you don't need to do it so you know*
> *I mean this song I think is OK no it's relatively OK I I mean from the top of the set just marking it out what do you think yea nay care*
> *Sorry my brain just*
> *What song are we practising?*
> *Run through*
> *From the top*
> *yeah*

The good-quality audio file was also recorded on an iPhone, by the author. This file is shorter than the poor-quality audio, at 8.4-s duration. The speaker is an Irish English speaker, who knew she was being recorded and was specifically speaking into the microphone with the aim of being understood. The quality of the file is stable, with no background noise or overlap. The context of this recording is a short greeting, where the speaker introduces herself and also refers to a speech programme called *MAUS*, which was used in Loakes (2022) but is not used in the current study. The aim of this research was to deliberately stretch the systems, to determine whether ASR performance is better using systems with large language models.

4.1.2 Good-quality audio transcript

> *Hello*
> *my name's Chloé*
> *I live in Melbourne*
> *I'm from Ireland*
> *I moved from Galway*
> *two and a half years ago*
> *and I love MAUS*

There are eight commercially available systems used in this report. Unless otherwise stated, the files inputted are .wav files. The systems are as follows:

Descript[2]—This is the system used in Loakes (2022), and the results from previous research are also reported here. In November 2022, Descript upgraded and began using large language models (see, e.g., Plumb, 2022), so a new Descript attempt is also made here with both .wav and .mp3 files.

Sonix[3]—This is an automated transcription service that is described on its website as 'fast, accurate, and affordable'. This was a system used by Harrington et al. (2022).

Google Cloud[4]—This is a suite of services using Google infrastructure, also including speech-to-text based on generative AI. This was another system used by Harrington et al. (2022).

Assembly AI[5]—This is a service for speech-to-text, described on the company's site as a system that 'makes up to 43% fewer errors on noisy data'. It is also described as being trained on over 1.1 million hours of data.

Deepgram[6]—This service is considered on the company website to be a 'world-class speech and domain-specific language model'.

Amazon Transcribe[7]—This is a speech-to-text transcription platform within Amazon Web Services. It is described as a platform for developers who want to add speech-to-text to their applications. It is often used for call centre and medical transcription. Amazon Transcribe is the best performing system in Harrington (2023), when compared with Microsoft Azure and Rev.

Microsoft Azure[8]—This is speech-to-text software that now operates within Microsoft Word 365. It uses a 'Universal Language Model' and also allows customisation. This method was also used by Harrington (2023).

Whisper[9]—This is run by the company Open AI. It is described as a system that 'approaches human level robustness and accuracy on English speech recognition' and 'has been trained on 680,000 hours of multilingual and multitask supervised data collected from the web'.

2 https://www.descript.com/
3 https://sonix.ai/
4 https://cloud.google.com/speech-to-text
5 https://www.assemblyai.com/
6 https://deepgram.com/
7 https://aws.amazon.com/transcribe/
8 https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-speech-to-text
9 https://openai.com/research/whisper

There are multiple versions of Whisper available, and this research used those with large language models. Aside from the 'Whisper AI March 2023' attempt, Whisper was run through a third-party app. Whisper was used with different audio files, so there is a 'Whisper June 2023 .wav' and 'Whisper June 2023 .mp3' version as well.

5 Results

Turning attention first to the good-quality audio file, Table 1 shows the system used, the number and % of words transcribed, the number of these words *correctly* recognised, the proportion of the entire attempt which was correct, and the WER (the % of errors compared to the total words spoken). This breakdown shows performance and where there are trouble spots in the outputs of the systems.

With the good-quality audio, the worst performances were from the older version of Descript reported in Loakes (2022) and from Google Cloud. Assembly AI and Descript (the August attempts) performed best with the good-quality recording. These systems correctly identified all of the material in the audio—the low predictability *MAUS* was transcribed *mouse*, which is not technically incorrect because the phonemes are exactly the same for both. Amazon Transcribe and Microsoft Azure also recognised *mouse*, while Sonix produced *my house* and Whisper (in all three attempts) produced *mouths*.

The other systems had some other minor errors, such as *two and a half* transcribed as *to ½* (Descript and Google Cloud), and one larger error in Deepgram's output with the place name *Norway* used instead of *Galway*. Additionally, some systems (the later Whisper attempts, and Microsoft Azure) also used *name's* instead of *names* as uttered by the speaker, which may be an attempt at producing a more readable transcript, but technically introduces an error. Google Cloud had the greatest number of errors overall—also transcribing *mouse* as *maths* and missing the word *I'm* entirely.

To sum up how the systems responded to the good-quality audio, we can see that this audio is largely recognised by these systems, retaining the sense of what the speaker was saying in almost every case. While there is not full accuracy in recognition for most systems, despite the clear quality of the file, these transcriptions can still be classified as being useable overall, and in some cases error-free, or almost error-free.

Turning now to the poor-quality audio, the results in Table 2 show a clear reduction in the number of words transcribed by the systems, as well as a reduction in their accuracy.

Comparing Tables 1, 2, it is clear that less of the material is attempted, and less is correct, for the poor-quality audio. Better performance of a system is indicated by the results in the final two columns—both the proportion of the attempt correct and the WER.

Where we see '100% accuracy' for two of the systems in the second last column, it is important to remember that this is showing *% of attempts correct*. For example, Descript, before the large language model upgrade, only recognised three words in total, and these words happened to be correct, but this is by no means a good performance as can be seen by the error rate of 97.4%. Arguably though, this poor performance could be seen as useful forensically, because the lack of content eradicates the issue of whether the material is correct or not (also see Harrington et al., 2022).

Later attempts using both a .wav and .mp3 file had marked improvement as should be expected, with 57 and 59 of the total 116 words transcribed. While around half of these attempts were correct (52.5% with the .wav file, and 49.1% with the .mp3 file), the total word error rates were still very high, at 73.3 and 75.9%, respectively.

Whisper (March 2023), on the other hand, recognised 21 words, but this is only 18.1% of the total number of words used in the audio (meaning around 82% of the audio is not transcribed). This version of Whisper may be considered to perform relatively well in the sense that of the 21 words recognised, there are no errors, as shown below:

> *Yeah, so just slowly building energy. And then I… Yeah?*
> *It's relatively okay.*
> *I'm just marking it out.*
> *What do you think?*
> *Okay?*

However, this is far from ideal because a closer look at the performance of the system shows that the material comes from different parts of the audio and only from the female speaker, with

TABLE 1 Results for good-quality audio.

System	No. (and %) of words recognised (n = 25)	No. of words correct	% of attempts correct	% errors (WER)
Descript (Loakes, 2022)	24 (96%)	19	76%	14%
Descript (August 2023) *.wav* and *.mp3*	25 (100%)	25	100%	0%
Sonix	26 (104%)[a]	25/26	96%	4%
Amazon Transcribe	26 (104%)	25/26	96%	4%
Microsoft Azure	25 (100%)	24	96%	4%
Google Cloud	24 (96%)	18	75%	18%
Assembly AI	25 (100%)	25	100%	0%
Deepgram	25 (100%)	22	88%	12%
Whisper AI (March 2023)	25 (100%)	24	96%	4%
Whisper AI (June 2023) *.wav* and *.mp3*	25 (100%)	24	96%	4%
Whisper AI (Aug 2023) *.wav* and *.mp3*	26 (104%)	24	92%	8%

[a]In some cases such as this, an additional word *name is* instead of *name's* was recognised, so 26 words (of the original 25) have been counted.

TABLE 2 Results for poor-quality audio.

System	No. (and %) of words recognised ($n = 116$)	No. of words correct	% of attempts correct	% errors (WER)
Descript (Loakes, 2022)	3 (2.5%)	3	100%	97.4%
Descript (August 2023) *.wav*	59 (50.8%)	31	52.5%	73.3%
Descript (August 2023) *.mp3*	57 (49.1%)	28	49.1%	75.9%
Sonix	53 (45.7%)	20	37.7%	82.8%
Amazon Transcribe	29 (25%)	11	37.9%	90.6%
Microsoft Azure	33 (28%)	17	51.5%	85.4%
Google Cloud	0 (no attempt)	0	(no attempt)	(no attempt)
Assembly AI	32 (27.6%)	21	65.60%	81.9%
Deepgram	29 (25%)	14	48.30%	88%
Whisper AI (March 2023)	21 (18.1%)	21	100%	81.9%
Whisper AI (June 2023) *.wav*	80 (68.9%)	58	72.5%	50%
Whisper AI (June 2023) *.mp3*	82 (70.69%)	49	59.7%	57.6%
Whisper AI (Aug 2023) *.wav*	96 (82.7%)	55	57.3%	52.6%
Whisper AI (Aug 2023) *.mp3*	97 (83.6%)	57	58.7%	50.9%

large amounts of material from her speech (and all of the speech from male speakers) ignored. The WER for this Whisper attempt was 82%.

A number of the other systems do not perform well at all with this audio, for example, Google Cloud made no attempt, with an error message stating 'we could not process your audio with this model', which was presumably because of the audio quality given that the good-quality audio worked with this system. Of the attempts made by the systems, Sonix recognised the most words (53/116) but also made the most errors (37.7% accuracy). This system also performed poorly in the study by Harrington et al. (2022). Looking more closely at the output from Sonix in this study, some totally incorrect phrases are used in the output. For example, in the poor-quality audio, this section of speech:

> *You gonna give us a hand signal or tell us what you do*
> *I I can't [laughter] ok*
> *From the from the top are we fine to go there*

Is transcribed as:

> *How are you gonna go through this with the High Court, huh?*
> *OK*
> *from the from the top are we fine to go that.*

Here, there are some sections that are relatively accurate and some that have no resemblance to the original. It is worth noting that when processing this file using Sonix, the system came back with an error message warning about the 'low accuracy potential' due to the nature of the audio, so the poor performance is not unexpected.

Amazon Transcribe, which performed well in Harrington (2023) and was then used for a more in-depth analysis in Harrington and Hughes (2023), performed poorly for this data. The values reported above are actually from the United States-English model because the Australian English language model transcribed only *Wait. What* of the entire 116 words. For Amazon Transcribe, neither using the American English model, nor the Australian English model, have given a good outcome. Microsoft Azure also performed relatively poorly with this audio. To give some further examples of the performance of these systems, this is the entire output for Amazon Transcribe (using United States-English):

> *Yes, I just got all your building in. Yeah.*
> *Signal or, uh, OK.*
> *To talk we find because there's nothing.*
> *It's relatively unpayable said just marking it up, okay?*

Some words and phrases are recognisable from the ground truth transcript, but even phrases that are almost correct are still wrong in some way. For example *it's relatively ok no?* is transcribed as *it's relatively unpayable* and *just marking it out* is transcribed as *just marking it up.*

The best performance of all systems tested was Whisper (the June 2023 .wav file attempt) in which almost 69% of the 116 words were transcribed—of that attempt 72.5% was correct. However, 'good performance' is relative; the overall WER is still 50%. This attempt also correctly recognised some speech produced by the male speakers, unlike the March 2023 version. Interestingly, the .mp3 file of the exact same audio had a similar rate of words transcribed in the June 2023 attempt, but this version had more errors, with an error rate of 57.6%. The later August 2023 Whisper attempts, with both the .mp3 and .wav file, had the most words transcribed of all systems used but had slightly higher error rates than the June attempts.

It is also worth noting that the sections of transcripts correctly transcribed were different across all of the Whisper attempts, sometimes completely different, and sometimes just slightly different. For example, the (arguably) low predictability phrase *What about what about another big drum fill will* was correctly transcribed, minus the repetition, in the August 2023 .wav attempt as *What about another big drum fill?* The later August 2023 .mp3 attempt transcribed this as *What about not being comfortable with my weight?* The June 2023 attempts both produced *What about now being comfortable?* As shown above, the March 2023 attempt did not recognise any of the content

from this phrase. Another example is the phrase *This song I think is okay, no? it's relatively okay* was transcribed correctly in the August 2023 .wav attempt, and almost the same transcription was produced using the .mp3 file except the word *no* was transcribed as *now.*

When there are incorrect transcriptions, there are also some similarities across how systems dealt with this; for example, the phrase *yea nay care* (which is asked with questioning intonation for each word) is transcribed by Descript in both August attempts and the June Whisper .wav attempt as *Gay, no? Gay?*, by the August 2023 .mp3 Whisper attempt as *Gay enough? Gay?*, and by the August 2023 .wav Whisper attempt as *Okay? Okay.*

6 Discussion

The aim of this research was to determine how well automatic speech recognition works on indistinct forensic-like audio with the new generation of systems that have large language models. Here, we have seen that the good-quality audio file had 24 or 25 (of 25) words recognised (and in some cases one extra word) with error rates between 0 and 18%. As demonstrated, the new generation of ASR systems largely perform well with that audio, despite some errors. This is unsurprising, as the older generation of ASR systems used in Loakes (2022) also had very good performance for this particular audio file, and as mentioned in that study each system is responding to a task it is designed to do.

For the poor-quality audio, the results were much more variable. The best performance was with Whisper (the June 2023 .wav file attempt) in which almost 69% of the 116 words were transcribed, and of that attempt 72.5% was correct. While this is a better performance than seen in the other systems and in Loakes (2022), this still leaves one quarter of the attempt either wrongly transcribed or missed by the system which is problematic for forensic contexts—equally problematic is the total word error rate of 50%. However, the better performance of this system compared to what was observed in Loakes (2022), and compared to other studies such as Harrington et al. (2022) and Harrington and Hughes (2023), needs to be acknowledged—this speaks to the fact that the audio used in the training of large language models is so diverse and so the systems can indeed respond better to new types of data (e.g., Kallens et al., 2023). Comparing back to the literature discussed earlier, the WER of 50% obtained for Whisper is exactly the same error rate mentioned earlier for sponsored competitions for ASR on multi-party speech in noise (e.g., O'Shaughnessy, 2023), so at this point, Whisper appears to be performing as well as any other system currently reported for this kind of audio recording.

The least accurate performance in this research is technically Descript (reported in Loakes, 2022) in which only three words were recognised by the system. While those words were correctly transcribed, there was no usable transcript. Later Descript attempts using large language models had a superior performance in comparison but still had error rates of approximately 75% for both .wav and .mp3.

Another result that should be noted is the earlier Whisper attempt from March 2023, where only 21 of 116 words were recognised, and these were all correct. While that appears to be a cautious response in the sense that if words could not be recognised no attempt was made to transcribe them, 18% accuracy is not a usable output.

The Google Cloud system had poor performance overall for these data overall, not recognising any of the poor-quality audio and having only 75% accuracy for the good-quality file. As seen in Table 2, Amazon Transcribe, Sonix, and Deepgram also had relatively low levels of recognition for the poor-quality file. Assembly AI, touted as performing well on noisy data, performed as well as a number of other systems using this data.

While Whisper in particular, using a .wav file, worked well *compared to the other systems tested*, in terms of correct transcriptions for low predictability items, its performance was not accurate enough to use for forensic transcription. Additionally, problems such as a correct transcription of a phrase in the June attempts being wrong in later attempts are a cause for concern in situations where the need for accuracy is so important. Finally, comparing the .wav files, the error rate increased across the June and August attempts of Whisper but *decreased* slightly when a .mp3 file was used.

Before concluding, this difference in transcription output when an .mp3 file is used compared to a .wav file is worthy of note. In the poor-quality condition, this study showed differences in the transcriptions depending on which file type was used, but there were no differences in performance for the good-quality audio files. With both Descript and the June 2023 Whisper attempts, the .wav files resulted in more accuracy (lower WER)—the difference was small for Descript, but for the June 2023 Whisper attempts there was a 7.6% difference in WER. However, in the August 2023 Whisper attempts, the .mp3 file had a slightly better performance than the .wav file. This variability is likely due to the fact that mp3 audio is compressed; one study looking at the effect of this compression on automatic speech recognition has shown .mp3 files can reduce transcription errors in some types of noise and induce transcription errors in other types—and that the effects are not consistent (Andronic et al., 2020). This result is simply one to be mindful of when working with ASR systems, and this highlights a topic worthy of further study so that better predictions can be made about how automatic speech recognition systems respond to the various types of audio that users may feed in.

7 Conclusion

This study compared the performance of a number of ASR systems, looking at how well they transcribed spoken language from a good-quality recording and a poor-quality recording. Taking into account the study as a whole, Open AI's Whisper performed far better than the other systems, having the lowest error rates overall. The study also showed that different versions of the same system (used at differing time points) do not always have equivalent outputs, and the later Whisper attempts are not necessarily the best attempts.

While Whisper performed best amongst the systems tested, it also needs to be remembered that forensic transcription is a task that is necessarily done without any ground truth to compare against. The potential for such a large error rate (50% WER at best) is not appropriate for forensic contexts; a transcription in which only 50% is correct is not useable. While the results of this study, for Whisper in particular, are a marked improvement in performance compared to the systems trialled on the same audio in Loakes (2022), this study advocates for the use of human transcription done in a measured and systematic manner (e.g., Fraser, 2022, also Loakes, 2022; Harrington, 2023) and for keeping ASR methods limited to tasks they are designed for. This aligns with the

findings from Harrington (2023) discussed earlier, who observed that it is more efficient to do a transcription from scratch than to try and use the output of ASR systems which contain relatively high error rates.

Another important finding of this study was that .mp3 and .wav files can induce different outputs from ASR systems. With a good-quality recording, the ASR outputs were the same, while for the poor-quality recording, the results were variable. While the differences may not be large between them, it is nevertheless an important consideration when using ASR systems with noisy data. More generally, it is not apparent from the outset whether there are key similarities or differences across the ASR systems in terms of how they function and exactly which differences might predict variable performance. However, parameters can be adjusted in some systems (including Whisper), and the amount of material the systems are accessing is constantly changing, so at the very least we can predict variable performance, and be mindful of the inevitable variability in resulting outputs, even if it is not clear exactly what the variability will be linked to. Given this, it is likely that the variable performance demonstrated by the different versions of Whisper will happen almost every time one of these systems is used, even with the same audio. The lack of information and full transparency about the exact architecture of the systems, and the resulting lack of certainty about what causes differing levels of performance, is another reason that ASR systems are currently not useful or suitable for the forensic domain.

Finally, the fact that the data used in this study are forensic-like, and not from a real forensic case, does not *per se* limit its implications for forensics. The issues about recognition of particularly infrequent lexical items, background noise, speakers being at variable distances from the microphone, overlapping speech, and background noise still remain and (as noted by O'Shaughnessy, 2023) have hindering effects on speech recognition. In this study, however, these variables are conflated, and future work should focus on specifically controlling variables such as the degree of background noise. Arguably, it could be expected that Whisper, with its particularly large language model (not entirely trained on studio quality audio) and iterative processing, should be one of the best performing systems on the market, and we have seen that is indeed the case in this study.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the author upon request.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

DL: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft.

Funding

The author declares financial support was received for the research, authorship, and/or publication of this article. The author received funding support from the School of Languages and Linguistics and the Faculty of Arts, at the University of Melbourne.

Acknowledgments

The author thanks Helen Fraser for assistance with ideas in this manuscript and discussion of issues surrounding the main themes within and Yuko Kinoshita for running the March 2023 Whisper attempt.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Andronic, I., Kürzinger, L., Chavez Rosas, E.R., Rigoll, G., and Seeber, B.U. (2020). "MP3 compression to diminish adversarial noise in end-to-end speech recognition" in *Speech and Computer: 22nd International Conference, SPECOM 2020*. St. Petersburg, Russia, October 7–9, 2020, Proceedings 22, 22–34; Springer International Publishing.

Bender, E. (2022). Resisting dehumanisation in the age of AI. Cognitive Science Society YouTube. Available at: https://www.youtube.com/watch?v=wuU-5rGPbyg (Accessed July 11, 2023).

Benzeghiba, M., Mori, R. D., Deroo, O., Dupon, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: a review. *Speech Comm.* 49, 763–786. doi: 10.1016/j.specom.2007.02.006

Bridle, J. (2023). The stupidity of AI. The Guardian. 16 March. Available at: https://www.theguardian.com/technology/2023/mar/16/the-stupidity-of-ai-artificial-intelligence-dall-e-chatgpt (Accessed July 11, 2023).

Fetzer, J. H. (1990). "What is artificial intelligence?" in *Artificial Intelligence: Its Scope and Limits. Studies in Cognitive Systems, Vol. 4* (Dordrecht: Springer), 3–27.

Fraser, H. (2022). A framework for deciding how to create and evaluate transcripts for forensic and other purposes. *Front. Commun.* 7:898410. doi: 10.3389/fcomm.2022.898410

Fraser, H., Loakes, D., Knoch, U., and Harrington, L. (2023). "Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio" in *Presentation at the 31st IAFPA Conference*. Zurich, Switzerland, p. 21.

Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Front. Commun.* 8:1165233. doi: 10.3389/fcomm.2023.1165233

Harrington, L., and Hughes, V. (2023). "Automatic speech recognition: system variability within a sociolinguistically homogeneous group of speakers" in *Proceedings of the 20th International Congress of Phonetic Sciences Guarant International*. (eds.) R Skarnitzl & J Volín, Paper ID: 593; 3131–3135.

Harrington, L., Love, R., and Wright, D. (2022). "Analysing the performance of automated transcription tools for indistinct audio recordings" in *Poster presented at the 2022 Conference of the International Association for Forensic Phonetics and Acoustics*,

Prague, Czech Republic. Available at: https://robbielovelinguist.files.wordpress.com/2022/07/harrington-et-al._iafpa.pdf

Kallens, P., Kristensen-McLachlan, R., and Christiansen, M. (2023). Large language models demonstrate the potential of statistical learning in language. *Cogn. Sci.* 47:e13256. doi: 10.1111/cogs.13256

Koenecke, A., Nam, A., and Lake, E. (2020). Racial disparities in automated speech recognition. *PNAS.* 17, 7684–7689. doi: 10.1073/pnas.1915768117

Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Front. Commun.* 7:803452. doi: 10.3389/fcomm.2022.803452

Love, R., and Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. *Front. Commun.* 6:797448. doi: 10.3389/fcomm.2021.797448

Markl, N. (2022). "Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition" in *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. ACM Association for Computing Machinery Seoul. 521–534.

Markl, N., and McNulty, S.J., (2022). Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. arXiv [Preprint]. doi: 10.48550/arXiv.2202.12603

McCarthy, J. (2007). What is Artificial Intelligence? Available online at: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (Accessed January 23, 2024).

Morozov, E. (2013). *The Folly of Technological Solutionism*. New York: Public Affairs.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int. J. Speech Lang. Law.* 16, 31–57. doi: 10.1558/ijsll.v16i1.31

O'Shaughnessy, D. (2023). Trends and developments in automatic speech recognition research. *Comput. Speech Lang.* 83, 101538–101522. doi: 10.1016/j.csl.2023.101538

Perrigo, B. (2023). OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic time magazine. Available at: https://time.com/6247678/openai-chatgpt-kenya-workers/ (Accessed July 25, 2023).

Plumb, T. (2022). How descript's generative AI makes video editing as easy as updating text Venturebeat.com. Available at: https://venturebeat.com/ai/how-descripts-generative-ai-makes-video-editing-as-easy-as-updating-text/ (Accessed August 2, 2023).

Preston, L. (2022). Becoming a chatbot: my life as a real estate AI's human backup. The Guardian. Available at: https://www.theguardian.com/technology/2022/dec/13/becoming-a-chatbot-my-life-as-a-real-estate-ais-human-backup (Accessed July 11, 2023).

Rodriguez, J. (2022). OpenAI's new super model: Whisper achieves human level performance in speech recognition medium. September 27. Available at: https://medium.com/@jrodthoughts/openais-new-super-model-whisper-achieves-human-level-performance-in-speech-recognition-78289d48f9a1 (Accessed July 11, 2023).

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Comm.* 140, 50–70. doi: 10.1016/j.specom.2022.03.009