frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/fcomm.2022.803452

Does Automatic Speech Recognition (ASR) Have a Role in the Transcription of Indistinct Covert Recordings for Forensic Purposes?

Debbie Loakes [1,2]*

[1] Research Hub for Language in Forensic Evidence, School of Languages and Linguistics, The University of Melbourne, Parkville, VIC, Australia, [2] ARC Centre of Excellence for the Dynamics of Language, Parkville, VIC, Australia

The transcription of covert recordings used as evidence in court is a huge issue for forensic linguistics. Covert recordings are typically made under conditions in which the device needs to be hidden, and so the resulting speech is generally indistinct, with overlapping voices and background noise, and in many cases the acoustic record cannot be analyzed via conventional phonetic techniques (i.e. phonetic segments are unclear, or there are no cues at all present acoustically). In the case of indistinct audio, the resulting transcripts that are produced, often by police working on the case, are often questionable and despite their unreliable nature can be provided as evidence in court. Injustices can, and have, occurred. Given the growing performance of automatic speech recognition (ASR) technologies, and growing reliance on such technologies in everyday life, a common question asked, especially by lawyers and other legal professionals, is whether ASR can solve the problem of what was said in indistinct forensic audio, and this is the main focus of the current paper. The paper also looks at forced alignment, a way of automatically aligning an existing transcriptions to audio. This is an area that needs to be explored in the context of forensic linguistics because transcripts can technically be "aligned" with any audio, making it seem as if it is "correct" even if it is not. The aim of this research is to demonstrate how automatic transcription systems fare using forensic-like audio, and with more than one system. Forensic-like audio is most appropriate for research, because there is greater certainty with what the speech material consists of (unlike in forensic situations where it cannot be verified). Examples of how various ASR systems cope with indistinct audio are shown, highlighting that when a good-quality recording is used ASR systems cope well, with the resulting transcript being usable and, for the most part, accurate. When a poor-quality, forensic-like recording is used, on the other hand, the resulting transcript is effectively unusable, with numerous errors and very few words recognized (and in some cases, no words recognized). The paper also demonstrates some of the problems that arise when forced-alignment is used with indistinct forensic-like audio—the transcript is simply "forced" onto an audio signal giving completely wrong alignment. This research shows that the way things currently stand, computational methods are not suitable for solving the issue of

OPEN ACCESS

Edited by:
Dominic Watt,
University of York, United Kingdom

Reviewed by:
Vincent Hughes,
University of York, United Kingdom
Michael Jessen,
Bundeskriminalamt, Germany

***Correspondence:**
Debbie Loakes
dloakes@unimelb.edu.au

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *28 October 2021*
Accepted: *28 April 2022*
Published: *14 June 2022*

Citation:
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

transcription of indistinct forensic audio for a range of reasons. Such systems cannot transcribe what was said in indistinct covert recordings, nor can they determine who uttered the words and phrases in such recordings, nor prove that a transcript is "right" (or wrong). These systems can indeed be used advantageously in research, and for various other purposes, and the reasons they do not work for forensic transcription stems from the nature of the recording conditions, as well as the nature of the forensic context.

Keywords: forensic linguistics, transcription, automatic speech recognition (ASR), phonetics, forced-alignment

INTRODUCTION

Covert recordings are "conversations recorded electronically without the knowledge of the speakers" — these are crucial records because "legally obtained covert recordings can potentially yield powerful evidence in criminal trials, allowing the court to hear speakers making admissions or giving information they would not have been willing to provide in person, or in an overt recording" (Fraser, 2014, p. 6). However, indistinct forensic audio is generally captured by hidden recording devices, with uncontrolled variables such as overlapping speech, background noise and distance from the microphone to name a few. As such, resulting audio is especially unclear, to the extent that a transcript is often needed to assist in determining what was said. While there are some moves toward improving the process of creating a transcription of indistinct forensic audio, especially by the Research Hub for Language in Forensic Evidence at The University of Melbourne (see e.g., Fraser, 2020), misconceptions abound in terms of what is possible as far as this type of audio is concerned.

A common question asked of people working with indistinct forensic audio, especially by lawyers and other legal professionals, is how the problem of what is said in indistinct forensic audio can be solved automatically, with artificial intelligence (AI) and specifically automatic speech recognition (ASR). This is a fair question, because automatic methods are useful for many real-world issues, but it is a question that needs to be explored experimentally to understand what the problem involves, the mechanisms of ASR, and also what happens when one attempts to solve the problem computationally — this will all be addressed in the current paper. In the paper, forced alignment is also analyzed because it is a way in which an existing transcript can be "overlaid" onto an audio file, effectively segmenting and aligning words (and even individual phonemes) to audio, yet there are many aspects of this which need to be properly understood to use forced alignment effectively and appropriately.

A working definition of AI is that it is intelligence demonstrated by machines instead of humans, and importantly, as noted by McCarthy (2007) "computer programs have plenty of speed and memory but their abilities correspond to the intellectual mechanisms that program designers understand". ASR specifically involves the recognition of speech, generally segmented orthographically into words. The following definition of ASR (from O'Shaugnessy, 2008, p. 2965) gives a good general introduction to what systems are attempting to do when faced with speech signals:

> As in any PR [pattern recognition] task, ASR seeks to understand patterns or "information" in an input (speech) waveform. For such tasks, an algorithm designer must estimate the nature of what "patterns" are sought. The target patterns in image PR, for example, vary widely: people, objects, lighting, etc. When processing audio signals such as speech, target information is perhaps less varied than video, but there is nonetheless a wide range of interesting patterns to distill from speech signals. The most common objective of ASR is a textual translation of the speech signal…

In their review of ASR systems, Malik et al. (2021, p. 9419–9420) describe that ASR performance architecture of ASR systems falls into four "modules". These are:

1) A pre-processing module–this is a stage in the process in which the signal-to-noise ratio is reduced (various methods are used such as end-point detection and pre-emphasis). While it makes sense that this would work to possibly enhance or make speech clearer, any pre-processing of a file in forensic situations needs to be considered extremely carefully (see e.g., Fraser, 2019).
2) A feature extraction module. Malik et al. (2021, p. 9421) describe how the most used methods for this are Mel frequency cepstral coefficients, linear predictive coding, and discrete wavelet transform.
3) A classification module, which outputs the predicted text. Malik et al. (2021, p. 9421) note that different methods can be used to do this, either using joint probability distribution (a generative approach), or a method that calculates predictions based on input and output vectors (a discriminative approach). Importantly, both make use of training data.
4) A language module — this contains language dependent rules about syntax and phonology. Malik et al. (2021, p. 9421) explain that many ASR systems now work without a language module, but they also note the improved performance that comes with using the language module.

Writing this research paper as a phonetician who has worked with forensic speech evidence, it seems obvious that there will be problems with an automatic approach, and that it is unrealistic to assume it would work, but what are these problems specifically? Using the definitions of both AI and ASR above from McCarthy (2007) and O'Shaugnessy (2008), who mention programme/algorithm designers respectively, it is evident that humans are also decision-makers — there are a whole host of

decisions and assumptions built in to the systems *by* humans. So it needs to be noted from the outset that these approaches are certainly not devoid of human intervention, and are thus not objective, despite common belief. Some biases in training data, for example, are discussed in research (e.g., Koenecke et al., 2020; Malik et al., 2021; Wassink et al., 2022) and this is expanded upon further in the next section of the paper. Additionally, O'Shaugnessy (2008), describes the fact that systems are taught to recognize "patterns", so perhaps one of the most obvious barriers expected in this research will be what kind of patterns (if any) are actually available in a noisy signal where speech can be less of an obvious feature than the noise. This issue will be explored in the current paper, which seeks to show what actually happens when ASR and forced alignment systems are used to help solve the problem of transcription of indistinct audio.

BACKGROUND

AI is particularly useful in various domains of our everyday lives, with cars that can center the vehicle in a laneway or brake before a collision can occur, facial recognition software that enables access to mobile phones, even spam filters on email systems that save time by automatically filtering emails that are not directly relevant. When it comes to speech, voice activated software is relatively commonplace–in smart phones, smart watches and in cars and homes to improve efficiency–for example people can ask their devices to turn on light switches, tell them the weather report, to find a location and direct them to that location, and so on.

In research, ASR, and forced alignment, have already proven extremely useful in the field of phonetics, sociophonetics and speech science more generally (some examples are Gonzalez et al., 2017; Mackenzie and Turton, 2020; Villarreal et al., 2020; Gittelson et al., 2021). Kisler et al. (2017) describe the "paradigm shift" that has occurred over recent years due to internet speed and connections being vastly improved, now allowing web-based platforms to be accessed and used easily by researchers. Automatic methods have also become very useful for language documentation purposes (e.g., Jones et al., 2019) and community members can also become involved due to accessibility (Bird, 2020). Such tools are also used very effectively in creating automatic subtitles, which can be done at very low cost, and even freely, with specific types of software. As many researchers have noted, the benefit of such tools lies in their efficiency, combined with the ability to analyse large amounts of data in order to better understand patterns in language. For example, one paper showed that it is possible to do 30 times the amount of analysis using automatic compared to manual methods (Labov et al., 2013), while another showed that depending on the task, automatic methods can improve efficiency of speech analysis by up to five times when compared with manual methods (for segmenting speech into utterances), or up to 800 times (for phonetic segmentation) (Schiel et al., 2012). This efficiency in processing, however, can also come hand in hand with a loss of precision. As noted by Coto-Solano et al. (2021, p. 17), for example, "in any scientific endeavor, there is a tradeoff between accuracy and speed, and each research project can determine what type of approach is appropriate". In forensics, however, there is no point at which speed is valued over accuracy due to the high-stakes nature of what is being analyzed.

This issue of efficiency also comes to the fore with forced alignment, which is a way of automatically aligning audio to a transcript (i.e., Jones et al., 2019), and is said to be "…highly reliable and improving continuously [yet] human confirmation is needed to correct errors which can displace entire stretches of speech" (Mackenzie and Turton, 2020, p. 1), and this is when clear recordings are used. In this paper, the analysis also focuses on how forced alignment fares with poor-quality recordings. This is of interest in the forensic domain, because a transcript can be created and then "matched" with an audio file—but there are various problems with this approach that need to be considered. Still on the topic of precision, in research contexts it has been convincingly argued that errors can be a risk worth taking. For example Evanini et al. (2009, p. 1658) state that "when very large corpora are used, errors in individual tokens and even individual speakers will not harm the analysis". Again, the same cannot be said for forensic situations, where what the speakers are saying is generally unknown and there is no definitive transcript to check the automatic version against. It is also often unclear who the speakers are, and even how many speakers there are (unlike in research situations). This is especially true in light of the fact that the success of systems comes with underlying assumptions which are explained well in the following quote "[i]n the cases of forced phonetic alignment and automated transcription … the technique rests on the assumption that there is some learnable, predictable pattern in the input that can be used to predict new cases" (Villarreal et al., 2020, p. 1); in forensic audio this condition is unlikely to be satisfied.

Before moving on further, it should be noted that most ASR systems work with HTK (Hidden Markov Model Toolkit) or Kaldi. HTK was developed at The University of Cambridge in 1993, and is described as "a toolkit for research in automatic speech recognition [which] has been used in many commercial and academic research groups for many years" (see e.g., Cambridge, 2021), while Kaldi is a more recently designed toolkit used for similar purposes (see e.g., Povey et al., 2011). MAUS, one of the systems used in this paper, uses HTK. Malik et al. (2021, p. 9417), explain that most ASR systems in use now also tend to use "long-short term memory (LSTM) … a type of recurrent neural network in combination with different deep learning techniques". Researchers are in agreement that ASR systems have shown vast improvements in a relatively short amount of time. For example Coto-Solano et al. (2021) explain the fact that this is due to the availability of training data, and deep learning algorithms, resulting in "important reductions in transcription errors". It is also important to note that ASR systems work differently due to "different feature extraction techniques and language models", yet this information is not always readily available to users seeking to understand and compare how the systems operate (see e.g., Malik et al., 2021). Even in "ideal conditions", then, ASR systems are certainly not error-free, and they are generally evaluated based on accuracy and/or speed, with "word error rate" and "word recognition

rate" being metrics used to determine accuracy (Malik et al., 2021).

Even the developers of automatic systems report that "transcriptions and annotations should undergo a final correction step"–internal validity is needed to keep improving system performance and ensure consistency–in other words, it is not expected to be error-free. Schiel et al. (2012, p. 118), reporting on internal validity of systems with human analysts, note that around 99% accuracy between humans performing orthographic transcription (of clear speech) has been observed, 97% for clear spontaneous speech, 95% accuracy for phoneme boundaries on read speech with a window of 20 ms, 85% accuracy for phonemic boundaries on spontaneous speech with a window of 20 ms accuracy, and quite poor agreement at 66% accuracy with prosodic labeling. This itself shows actually making decisions about language is not categorical due to the continuous stream of acoustic information that makes up the speech stream (see further Fraser and Loakes, 2020).

Another issue with respect to ASR performance is inherent biases that filter in at various stages. This is covered well in a paper by Wassink et al. (2022), who note that male speech is recognized better than female speech and also that effects on signal quality are different depending on gender, and that when dialectal differences are included in training data, dramatic improvements in performance can ensue. Racial biases are also shown to exist; in their "cross-ethnicity study" comparing white and non-white voices, Wassink et al. (2022) show that sociophonetic differences in ASR are involved in 20% of system errors. They note that if dialect forms were included in the language module, better performance would ensue. Aside from just the issues with accuracy, Wassink et al. (2022) note that "…it is, of course, clear that unevenness in the accuracy of ASR systems primarily occurs to the disservice of everyday people in these social dialect communities, who use voice assistants to accomplish a wide range of tasks, from interacting with mobile devices to paying bills, and many others". Their results support findings of a related study, which showed a word-error rate of 0.19 for white speakers, and 0.35 for black speakers, when comparing performance of five popular and widely-used ASR systems (Koenecke et al., 2020). Another broader issue to consider is, as pointed out by Malik et al. (2021, p. 9412) that "training models are available only for a handful of languages out of a total of ∼6,500 world languages".

So, errors with ASR are not unexpected due to the variable nature of the systems, the speech that is fed into such systems, and bias in training data. Forced aligners, too, have differing levels of accuracy. A research paper by Jones et al. (2019) compared the performance of two automatic forced-alignment systems using one transcription and one audio recording, and showed some of the issues that arise when using automatic methods not completely set up for the problem at hand, as well as some of the inherent merits of the systems. It is interesting because it shows that "tweaking" by humans can achieve some improvements in performance, but only because humans are aware of the source of the data and thus what it is possible to achieve. It also shows that performance will not be ideal. The speech data analyzed in Jones et al. (2019) is produced by five young adults conversing in Kriol, an Australian English-based lexifier creole. Jones et al. (2019) used two options within MAUS (a programme also used in the current paper). They used a language-independent model (i.e., one in which the system learns "from scratch" on the available data) as well as a language-specific model (one in which the system was trained on a major world language), noting that there are advantages and disadvantages of both approaches. For the language-independent model, the steps were relatively straightforward given that no assumptions are made by the system about which language the data (input) is in. The authors note that "[t]he more different the "small" language is from the world language, the more errors in orthography, phonology, and phonetics" in the resulting output. For testing with a language-specific model, Jones et al. (2019), on the suggestion of MAUS developers, tried Italian because like Kriol it has a transparent orthography, a similar number of vowels in the inventory, and relatively comparable data (i.e. spontaneous speech data was used in the Italian training model).

Comparing to a "gold-standard" human segmentation of the data, Jones et al. (2019) show that, for forced alignment, the language-specific model (using Italian) had an overall better accuracy than the language-independent model. Looking at the alignment boundaries for vowel onset and vowel offset, they showed that the language-dependent model was 41.4% accurate within 10 ms of a boundary, and 85.9% accurate within 50 ms; it should be noted, however, that in the context of a speech segment 50 ms is quite wide and so "accuracy" does not mean an exact match, simply that the system was in the vicinity of marking the correct segment. For the language-independent model, results showed accuracy of 31.8% within 10 ms of the vowel, and 75.4% within 50 ms of the vowel. They also noted that the system was better at determining vowel boundaries at the onset rather than offset.

The results in the Jones et al. (2019) study show that with relatively good audio, but mismatched modeling (i.e., the wrong language input), forced alignment systems can assist in analysis but errors occur, and this is when the system is fed a transcript to assist in the task. The benefits of automatic systems are said to be their increased efficiency as discussed above, but as noted by Jones et al. (2019, p. 296) the errors are "concerning because they tend to take even longer to manually edit the alignment" — in other words, efficiency is reduced.

Of interest for the current paper, Jones et al. (2019, p. 294) reflecting on some specific parts of their attempts to use AI for coding Kriol, note that:

> …neither MAUS Italian system nor MAUS language independent mode is originally designed for the forced alignment of north Australian Kriol. Unavoidably, there are missing, extra, and wrong phonetic labels … and misaligned segments. In this study, the tokens with missing labels were excluded before further analysis. In some extreme cases, the onset and offset time can be off for a few seconds compared with the manually-edited data [which occurs for other automated aligners as well (Mackenzie and Turton, 2020)]. In our dataset we noticed that completely misaligned tokens tended to involve long stretches of sonorous segments (e.g., vowels, nasals, liquids, and glides) where presumably MAUS lacked strong acoustic landmarks like stop-vowel boundaries to assist in the alignment.

Other papers have also compared how systems perform under various conditions. Kisler et al. (2017, p. 333) look at system validation, reporting that when the MAUS system is tested on forced alignment, there is a 97% "MAUS-to-ground-truth agreement" with three human labellers when spontaneous German speech is used, and accuracy with segmental boundaries is around 90% when compared with humans. Kisler et al. (2017, p. 333) also report on accuracy rates when an existing language model (Standard Southern British English) is used for a variety that the system has not been trained on (Scots English) finding in this case that "MAUS had an error rate twice that of human experts", which highlights the importance of using systems with inputs they have been trained on.

In a paper comparing the performance of forced aligners with Australian English, as well as a second human coder, Gonzalez et al. (2020) showed that the human coders were most alike and accurate in their performance, at around 80% agreement in this paper compared to between 65and 53% for the ASR systems. They also showed the ASR systems made errors depending on particular phonetic environments, whereas crucially, human coders were not prone to such errors. Gonzalez et al. (2020, p. 9) note that their "study lends empirical support to the common wisdom that humans are far more consistent in creating alignments than are forced aligners, indicating that regardless of the aligner used, alignment accuracy will be enhanced by manual correction".

The research discussed here highlights some important issues relating to good-quality audio, which need to be considered before exploring the usefulness of ASR with indistinct forensic audio. Coming from a position of knowing what the material involves in the first place (who recorded it, who the speakers are and what language/dialect they are speaking) is one of the key factors in effectively using these tools to recognize speech and perform a transcription. In other words, the ground truth needs to be accessible from the outset, which is not the case in forensic situations. In forensic cases, the stakes are high and errors are not a trivial matter.

The question addressed in this paper is how automatic transcription might assist in indistinct forensic transcription, whether via ASR or using a transcript and forced alignment. A common query in both academic and non-academic circles is whether this can be done — in Australia, automatic transcription is indeed sometimes used to assist with summarizing lengthy recordings collected for investigative purposes, while police in Australia and elsewhere are also actively looking at extending this technology for indistinct audio used as evidence. In recent years researchers have also been investigating the application of automatic methods in the forensic context, such as alignment of telephone tapped speech with an already existing orthographic transcription (i.e., Lindh, 2007). It is feasible that aside from simply making analysis easier, a transcript (whether correct or not) could be fed into to a forced alignment system — again while it may be intuitive that this is inappropriate, it does not take away the possibility that this method could be used.

AIM

This study has a specific aim of demonstrating how automatic systems work with forensic-like audio, in comparison with good-quality audio. As pointed out by Lindh (2017, p. 36) "if only limited work has been done on the combination of auditory and automatic methods in comparing voices and speakers, even less work has been done on combining automatic speech recognition and forensic phonetic transcription". In other words, relatively little is known about the best ways forward, or even if there *should* be a way forward.

The aim of this research is thus to analyse, experimentally, how two ASR systems perform when tasked with the transcription of indistinct forensic-like audio. It also aims to assess what happens when a transcript is fed into a system with indistinct forensic audio (i.e., a forced alignment system). Potential issues in forensic transcription which result from these demonstrations will be discussed.

METHODS

Data

This project used two recordings to test two ASR systems, and compare their performance. The number of recordings is minimal so that broad issues can be demonstrated[1]. The recordings are purposely different to replicate the forensic context where "mismatched conditions" are par for the course (e.g., Jessen, 2008, p. 700).

The recordings used are:

Audio

1. "*poor-quality" audio*. This is a 44.2 second stretch of audio from a recorded rehearsal by a singer and some musicians. This stretch of 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, there is overlapping speech, and there are also people present who were not recorded (but this was not recorded in the context of crime). This audio was recorded by one of the speakers via an iPhone and streamed to Facebook live, where it was retrieved with permission. We are in a fortunate position with the audio, because the speakers are known, access to an associated video was granted, nouns used have been checked, and the transcript has been verified with one of the speakers who organised and streamed this event. The recording used has one female voice and three male voices, and all speakers are using Australian English. In this case the speakers knew they were being recorded, but were focused on the task at hand and not attempting to be clear to the audience; they were sharing the file so fans could see what a rehearsal looked like, and so the audience could experience the music (in those parts of the file, microphones were being used). The content of the speech produced in between the songs was focused on planning the live music event, as well as general

[1]Another research project is currently underway using more data - real forensic audio, "fake" transcripts and recordings made on different channels (including telephone recordings).

conversation, and it is one section of speech in between songs used in this research[2].

For the poor-quality recording in the current experiment, a reliable transcript is as follows. Here we make no attempt to attribute the utterances to particular speakers.

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 practicing?
Run through
From the top
yeah

2. Unlike the poor-quality recording, the second audio file is termed *"good-quality" audio*. This was also recorded on an iPhone. In this case, there was a single speaker, the microphone was close to her mouth, there was little background noise, and the speaker was mindful of being understood. She was producing an utterance for a summer school for students learning about the programme MAUS, which is used in the current research paper (and described in the next section). This audio file is 8.4 s, and is spoken by an Irish English speaker recorded in Australia. The speaker has given permission to use this recording. The transcript for this file, separated into intonation units, is:

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.

It should be noted that these recordings, aside from being recorded on iPhones, are extremely divergent in nature — choosing divergent recordings is purposeful because it attempts to replicate forensic situations with their mismatched conditions. In the forensic domain, so-called "questioned samples" are compared with non-forensic "suspect" samples, and they are generally from extremely divergent sources — because forensic samples contain important speech evidence, it is often necessary for some kind of analysis to go ahead (i.e., simply discarding the samples due to these differences is not appropriate). This is discussed by, for example Rose (2002), and also see Jessen (2008, p. 685–686), who review some common technical differences across such samples, citing that forensic samples may be shorter, contain echo, have a mismatched sampling frequency compared to the suspect sample, be recorded via telephone, or have overlapping speech and/or background noises. The forensic sample in this recording is actually longer than the good-quality recording, but does indeed contain overlapping speech and background noise, with speech also at a distance from the microphone.

Software

There are three programmes used for the task of recognizing speech in the good-quality and poor-quality recordings respectively.

BAS SERVICES (Bavarian Archive for Speech Services)—ASR and WebMINNI

There is "a set of web services" at the Bavarian Archive for Speech Signals (BAS) in Munich that were developed for the processing of speech signals" (Kisler et al., 2017, p. 327). These include ASR, forced alignment, voice activity detection, speech synthesis and an online "labeller" which can be used to mark boundaries between linguistic events (syllables, intonation units) called EMU – these can all work together[3]. In this paper the focus is on two of these services.

Firstly, MAUS is used, and specifically "WebMINNI" because, as stated on the website, it "computes a phonetic segmentation and labeling based solely on the speech signal and without any text/phonological input". In this case, the result needs to be read back by reconstructing phonemes as there is no resulting orthographic transcription as such. This is effectively a forced-alignment tool which, in the words of Kisler et al. (2017, p. 331), uses

> [a] two-step modeling approach: prediction of pronunciation and signal alignment …. In the first step, MAUS calculates a probabilistic model of all possible pronunciation variants for a given canonical pronunciation. This is achieved by applying statistically weighted re-write rules to a string of phonological symbols. The language-specific set of re-write rules is learned automatically from a large transcribed speech corpus. The pronunciation variants, together with their conditional probabilities are then transformed into a Markov process, in which the nodes represent phonetic segments and the arcs between them represent transition probabilities. … In the second step, this Markov model is passed together with the (pre-processed) speech signal to a Viterbi coder … which calculates the most likely path through the model, and – by means of backtracking this path – the most likely alignment of nodes to segments in the signal.

The WebMINNI service does not have an Irish English model, so a UK model was used. It is acknowledged that this model probably included a majority of non-rhotic speakers, unlike the Irish English used by the speaker, but as the results will show this is not an issue for what is being focused on in the current study.

[2]Other sections of the audio which contain speech are being used for a separate experiment on the transcription of indistinct audio with human transcribers.

[3]https://www.bas.uni-muenchen.de/Bas/BasMAUS.html

The BAS services ASR system was also used, which requires only audio and returns an orthographic output[4]. For the ASR service there are many language models that can be selected, including both an Australian English and Irish English model which are used for the poor-quality and good-quality recordings respectively. As noted on the website for the BAS services, third party services are used for this service, including Google Cloud and IBM.

Descript

Descript is another programme used in this research[5]. It is described as "all in one video and audio editing" and has functions to assist with podcasting, screen recording, video editing and transcription (used in this research). It is freely available (up to 300 h per month) and has an ASR component, which works using "Google Cloud's Speech-to-Text technology" (Opiah, 2021), and in this way has some similarity with BAS Services (which uses Google Cloud, but other technology as well[6]). The mechanisms of Descript are less well-described, presumably as it is not normally a research tool in the way BAS services are, and is available for use to anyone without the need for explicit training.

RESULTS

BAS SERVICES: ASR

Firstly focusing on how the MAUS fared with the poor-quality recording, the ASR option was used within the BAS Webservices. The number of speakers was selected (four) and an Australian English model was used. Once we uploaded the file, this was unable to be read at all, the system returned the following error

> *StdErr: ERROR: callGoogleASR: can't find a transcript in server response; this means either a bad signal quality or empty signal–exiting*

Because we know it was not an empty signal, we can be confident that there was a bad signal, which is unsurprising. So in this case, the ASR failed for this recording.

When we tried the ASR service with the good-quality recording, and chose one speaker as well as an Irish English model, we had a successful result (with some errors, underlined).

> *Hello, my name is Chloe I live in Melbourne are from Ireland I met from Galway to 1/2 years ago and I love maths.*

This is a successful output, although there are some minor errors in the form of introduced sounds or wrong words, which are underlined. These are:

1. *name is* should be *name's*,
2. the word *are* should be *I'm*
3. *to 1/2* is almost correct (even though two *1/2* is technically more correct) but the words *and a* is missing, i.e. the speaker said two *and a 1/2;*
4. *maths* should be *MAUS*

The free "WebMINNI" service was also tried, which has the component allowing recognition of phonemes without any transcription. For the poor-quality recording, we found that almost no speech (no phonemes) were recognized at all–although the system did very well at finding silence intervals. To give some examples, **Figure 1** shows a screenshot of the waveform as well as the resulting phoneme tier which was the output from the WebMINNI system[7]:

As seen in the image, there are some sections that are labeled "<p:>" which means *silence interval*, and some labeled "<nib>" which means *non-human noise*. This image does not show the whole file. It is certainly not the case that the <nib> sections were non-human noise, in fact this is where the human speech was located in the file in many cases. The silence intervals, however, were relatively well captured.

As another example, and to be more specific about the kinds of errors observed, **Figure 2** shows some of the output from WebMINNI, which occurred later in the file after **Figure 1**. There is a small amount of overlap between the end of **Figure 1** and start of **Figure 2**.

Figure 2 shows more activity on the phoneme tier compared to **Figure 1**. Here it can be seen that the system attempted to find some speech segments, and while this is the case the actual identification of sounds was not successful.

Some specific examples are:

> <nib> at the left of **Figure 2** is an entire section of speech produced by the female speaker in which she says *are we fine to go there*, but is analyzed by the system as non-human noise.

For the first section marked "h" the female speaker is in fact saying "Mel" (so there are three segments, not just one, and the marked segment is wrong). The remaining four are trumpet noises (trumpet noise is also occurring in other sections).

In the section marked V (which technically represents an open vowel) the female speaker is saying the phrase *is OK no*.

Additionally, the first <p:> in **Figure 2** is in fact marked correctly as a silence interval–and while some activity can be seen on the waveform, this is background noise which is almost inaudible. The second <p:> (at the end of the **Figure 2**) is the speaker saying *it's relatively OK I I mean from the top of the s-* (the remainder of the word *set* is not shown). In this case, the <p:> is wrong.

WebMINNI then, has not been able to segment speech sounds in the poor-quality recording. It has identified some sections of speech as "non-human noise" and has incorrectly identified whole words and phrases as one speech segment.

On the other hand, the good-quality recording fared relatively well (but better when the ASR option was chosen). WebMINNI

[4]This requires a login via a Clarin account which can be accessed through education institutions.

[5]https://www.descript.com/

[6]While there is thus some similarity with BAS services and Descript, their differences lie in the specific language modules they use as well as different ways of applying feature extraction and prediction.

[7]The spectrogram is not visible in this Figure, nor in **Figure 2**, as the aim is to show "non-speech" category labels on the phoneme tier.

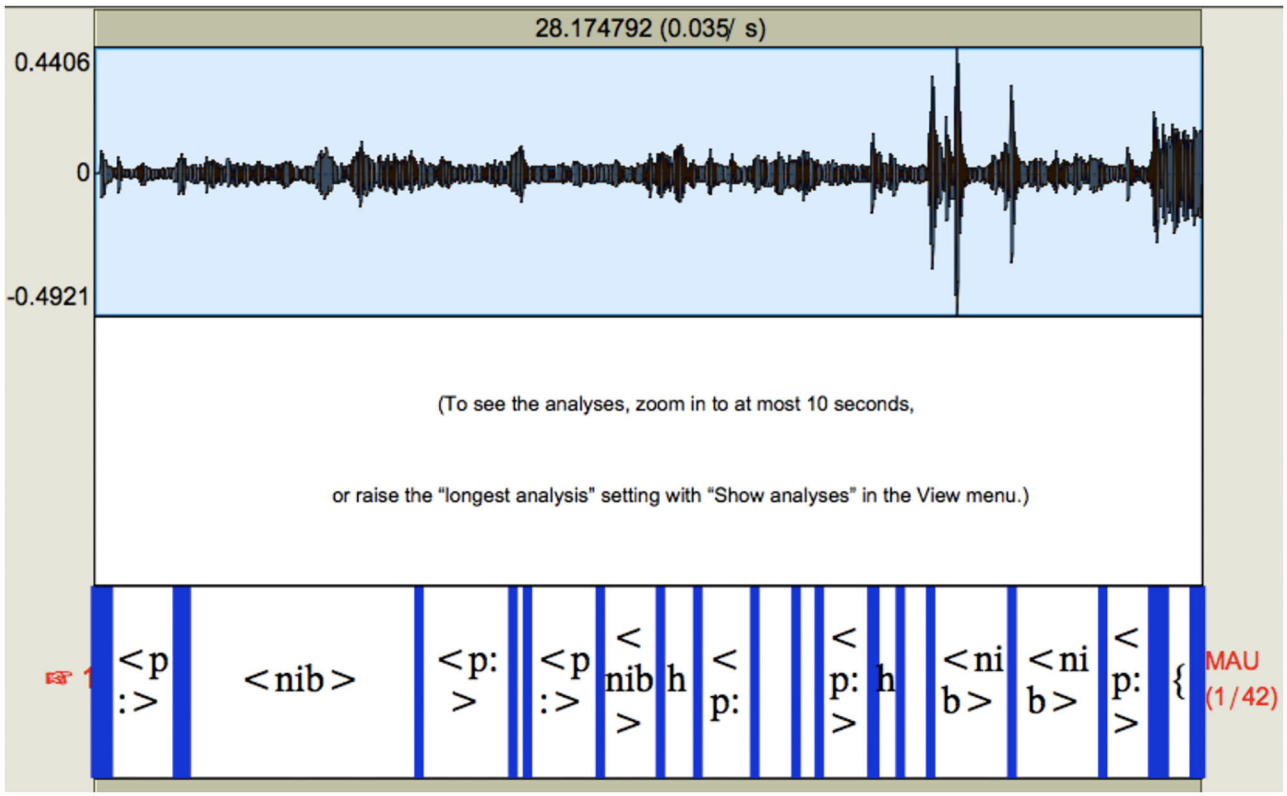

FIGURE 1 | Example 1 of system output with the poor quality recording using WEBMINNI, ASR.

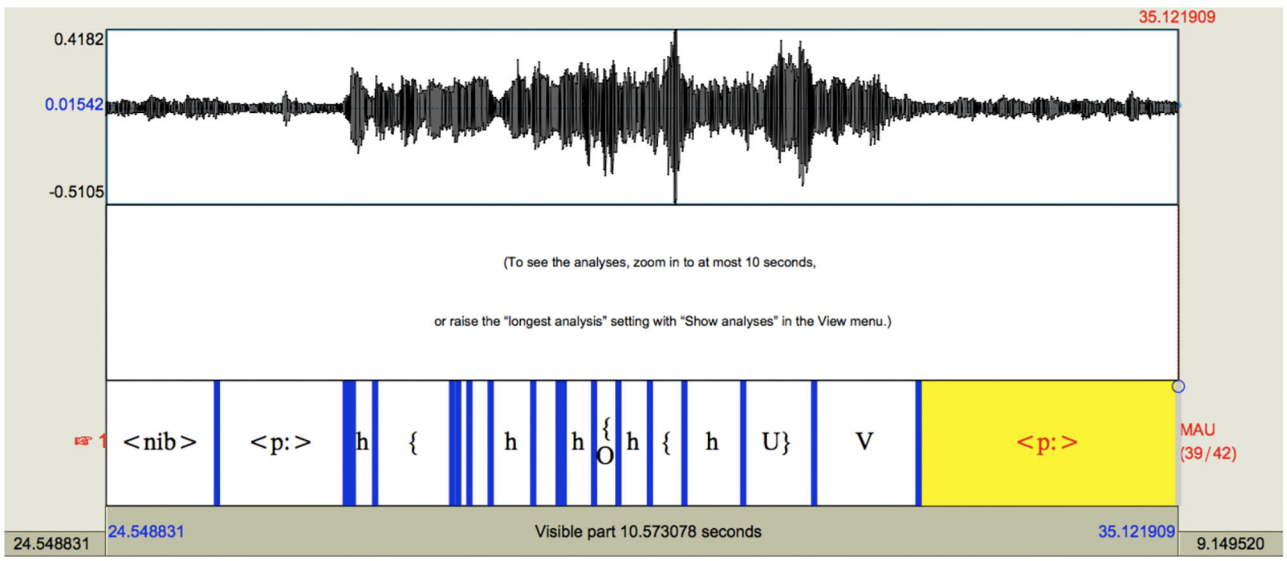

FIGURE 2 | Example 2 of system output with the poor quality recording using WEBMINNI, ASR.

was able to segment the speech segments but with some errors, and so it is possible from that to reconstruct what the speaker was saying. Using names as examples, some errors in the good-quality recording are:

Chloé is rendered /koʊɔaɪ/
Galway is /kaoɪeɪ/

This indicates there is some inability for the system to pick up the /l/ sound in the speaker's voice. Interestingly, the system appears

to have been making predictions about /l/ vocalization (replacing the speaker's relatively dark /l/ with back vowels), which may be because we are using the British English model, so anything /l/-like may be being converted to a back vowel for this reason. The best pattern recognition that the system could do in this case was a back vowel; in other words the system is interpolating from the available data and the assumptions being made about it. Across the file there are also some other minor errors, with some nasal sounds confused – i.e. /m/ sounds written as /n/. So, in this case, for the good-quality recording the ASR system worked better than WebMINNI, likely promoted by the Irish English model in the former – it is known that suitable training data, when it comes to sociophonetic and linguistic factors, boosts performance (i.e., Wassink et al., 2022). For the poor-quality recording, neither the ASR or WebMINNI was successful.

BAS SERVICES: Forced Alignment

Within the BAS services, the forced-alignment option was used, with an orthographic transcript. The important thing to note is that this was a reliable transcript — the subject matter is known, and the speakers are known, so the speech matter has been verified. This would not be possible to do in a forensic situation where there is no way of verifying anything that could be fed into the machine.

When the transcript was used with the poor-quality recording, WebMINNI was able to correctly segment (force-align) some of the words, although there were more errors than correct segmentations. The background noise and overlapping speech made the task difficult for the system because the noisy signal does not allow acoustic landmarks to be recognized. As an example, **Figure 3** shows a section of speech in which the speaker is saying *Just slowly building ener-* (not all of the word *energy* is visible in the figure shown). However, the system has force-aligned only the word *just* correctly, and none of the other words are correctly aligned. In fact the whole word *energy* is shown, as well as the word *and*, despite the fact that they are not present in this exact stretch of audio. Additionally, the poor-quality of the spectrogram is evident in this example.

As another example of WebMINNI's performance, in the following example shown in **Figure 4** the phrase (*From) the—from the top* is force-aligned onto a section of the recording that is actually drumming noise and laughter, but this was recognized as speech. This can be likened to what happens when software which is designed to recognize faces "believes" that clouds and trees are people. The system has attempted to match boundaries, or qualities observed in the signal, with phonemes / words—which it is designed to do but of course the trouble here is that there are no phonemes or words in this section.

In contrast, using a transcript with the good-quality recording is very successful as seen in **Figure 5**, although there are some errors which should be addressed. Because a non-rhotic model was used, the transcription of *Melbourne* and *Ireland* (of which the output does not contain /r/) are incorrect in this respect– in other words the system failed to recognize the rhotic in the speaker's pronunciation of these names because it is effectively trained to ignore them in the UK English model–presumably if we had tried an American English model the transcription would have been more reflective of the actual pronunciation of these items. Also, the second syllable of *Melbourne* is not transcribed with a schwa vowel (in the transcription system, schwa is the @ symbol) so the "O:" symbol, a long back vowel, is also technically wrong. Here, the system has inferred the statistically most likely pronunciation based on the "-ourne" spelling in this word. The remainder of the file, not shown here, was also relatively successfully transcribed.

Regarding alignment, the only errors visible in **Figure 5** are the boundaries between *Chloé I live in,* which are misaligned. The word *Chloe,* for example, is force-aligned onto just the onset segments of the /kl/ portion of the word. There are also alignment errors in the following words, but from *Melbourne* the alignment becomes accurate again.

DESCRIPT: ASR

Descript is a system which is designed for the general public, and so is very straightforward in terms of having an audio input and an orthographic output. When Descript was tried with the poor-quality recording, only three words were recognized by the system, the words *yes*, *yeah* and *okay*. While three words were identified, the word *yes* was not exactly correct (the speaker was actually producing another repetition of *yeah*). These words were recognized (or partially recognized) likely because they were somewhat louder, and so potentially "stood out" from the background noise. The Descript system did not recognize any other words. The total number of words uttered by the four speakers was 116, so this means the recognition rate was only 1.7%.

When Descript was tried with the good-quality recording, the output was almost entirely correct aside from the spelling of Galway (which was spelt with *Gallway*, but this is effectively inconsequential) and the very last word in the phrase *I love MAUS* which was recognized instead as *I love my house*. This recording was of course much shorter, but even if we say *Galway* is incorrect due to its spelling, and say that the error in *MAUS* is two errors, the recognition rate is 22/25 and effectively 88%. If we are more generous and say that *Galway* is correct, and *MAUS* is only one error (being an incorrect noun phrase) the recognition rate is 96%. Whichever way we decide to judge these errors, the performance of Descript is clearly superior when we use the good-quality recording. Mistakes are explainable due to predictability, which is especially low for the software *MAUS*.

DISCUSSION

This research shows that if we have clear, non-overlapping speech in a language variety that the system is familiar with, then ASR systems work very well. This is not surprising, as this is what the systems are designed to handle. However, if we have indistinct forensic-like audio, where speakers are not positioned near a microphone, or have overlapping speech with multiple sources of background noise, the systems perform badly. As shown with WebMINNI, even with a transcript, performance is far from ideal–forced-alignment does not accurately recognize word boundaries in most cases. However, this is not surprising, and not a criticism of developers of these systems, who have

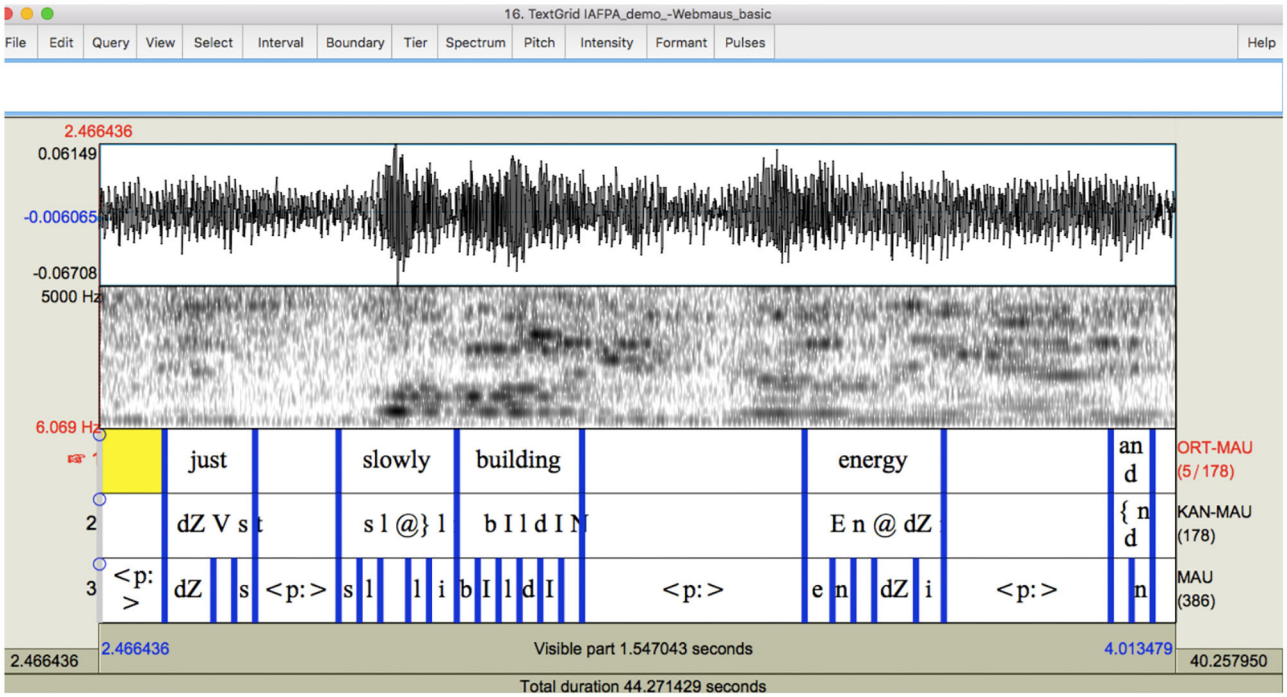

FIGURE 3 | Example 1 of system output with the poor quality recording using WEBMINNI, forced-alignment.

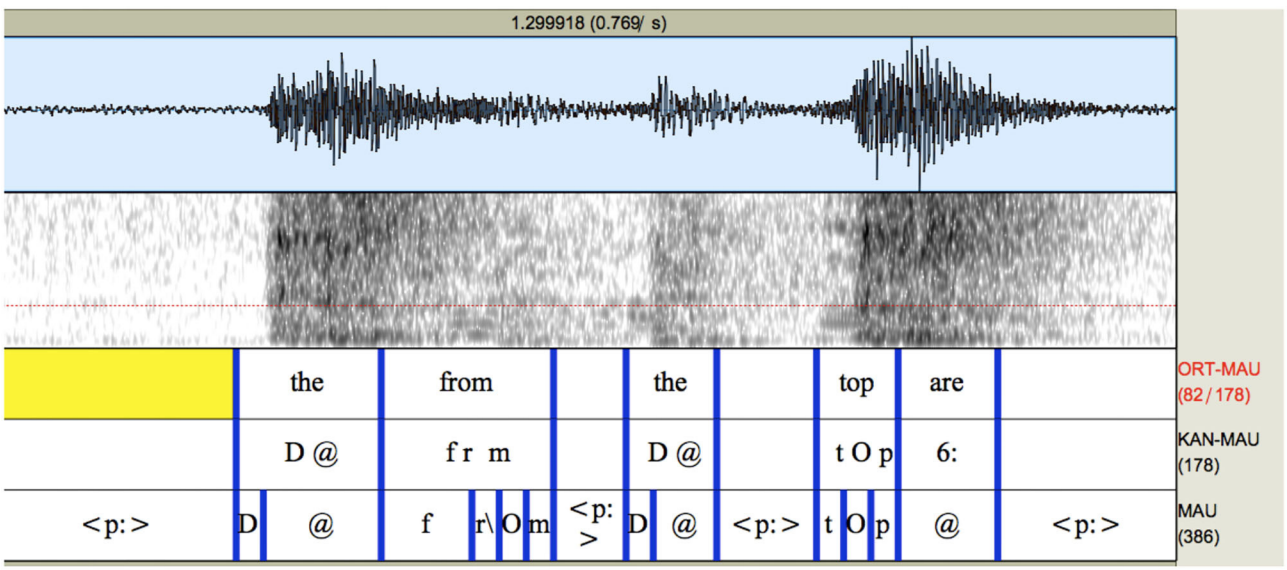

FIGURE 4 | Example 2 of system output with the poor quality recording using WEBMINNI, forced-alignment.

not advertised their systems as being made for the transcription of indistinct audio. It does, however, make clear why people working in the area of transcription of indistinct audio do not turn to computational methods to solve the problem.

It must also be acknowledged that automatic methods can be used to solve some issues in forensics–for example they can cut down significantly on manual work by an analyst, making tasks more efficient. One example is the segmentation of speech from non-speech, even if the recordings are very poor quality, as shown here with the poor-quality recording when it was run through WebMINNI.

Given the results of the research shown here, the cautions and concerns raised about automatic transcription in sociophonetic and sociolinguistic literature, where fine detail and "a

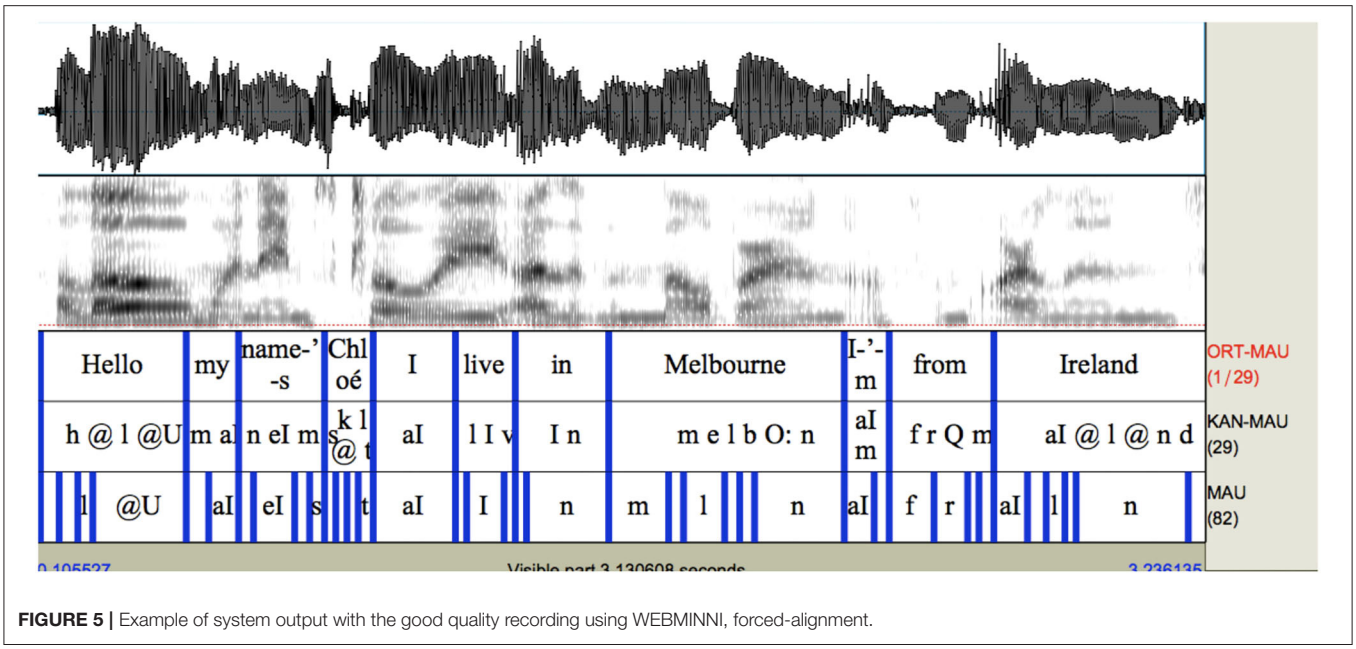

FIGURE 5 | Example of system output with the good quality recording using WEBMINNI, forced-alignment.

constellation of acoustic cues" are important and should not be factored out (Villarreal et al., 2020, p. 2), are even more pertinent for forensic purposes where the stakes are far higher. Returning to the quote from Mackenzie and Turton (2020, p. 1) though "…although forced alignment software is highly reliable and improving continuously, human confirmation is needed to correct errors which can displace entire stretches of speech." This human intervention raises the question of bias and priming which is unproblematic in research and language documentation situations, where acoustic cues are also clear and the ground truth can reasonably be established, and mistakes would regardless be occurring in a relatively low stakes environment. It is of course a concern for transcription of indistinct audio for use in court situations where stakes are far higher, and just like Lindh (2017, p. 58) reports for automatic speaker recognition contexts, it "would be unwise to presume that one can be a completely ultra-objective bystander feeding a system with the necessary inputs to decide the strength of the evidence".

As noted by Jones et al. (2019, p. 284), however, when evaluating whether to use a language-independent or language-specific model for Kriol within MAUS "the choice is always dataset-specific". This holds for indistinct forensic audio, but the very fact that the contents of the file are generally unknown (unlike in research) this means that any choices made about how to deal with the data effectively are simply guesswork, which is unsatisfactory.

Even though some people may expect better performance when computational methods are used, the requirement for human intervention can be *greater* when we use systems not designed for the task at hand (e.g., Jones et al., 2019). This is also clear in the current analysis, where using automatic methods offered arguably no benefit in assisting with the transcription of the poor-quality recording, with a refusal to read the signal when the BAS ASR service was used, nothing correct when using MAUS without a transcript, two words correct with Descript, and quite poor performance when forcing segmentation onto a transcription which we know to be a "gold standard" transcription. The good-quality recording, however, produced a useable transcript in the BAS ASR service and in Descript, although as shown there were some errors, especially where predictability was low, i.e., the word *MAUS* and some other cases in which small words were added or not recognized. However, when these automatically-produced transcripts are fed into MAUS, very little manipulation would be required at all. In other words, even though some manual intervention would be required for checking and correcting (especially for low-predictability items, as we saw), using ASR systems with data such as our good-quality recording is clearly more efficient than a fully manual method of analysis, as has been reported by other researchers.

CONCLUSION

As things currently stand, when recordings are poor quality and there is no definitive transcript (typical for forensic contexts), this research has demonstrated that automatic methods cannot solve the problem of what was said in indistinct forensic audio. The issue of what material ASR systems are trained on is unresolvable for many forensic contexts–the noisy conditions are problematic, as is the fact that speakers are often contested–therefore guesswork is needed to apply automatic methods and this is entirely unsatisfactory. It is also problematic that a transcript can be fed on to any audio and possibly *look* correct. Systems can appear to work on transcription data that is simply wrong, and just because a system error does not occur, it does not mean that an output is correct. These main points of the paper may perhaps be obvious to linguists and phoneticians, but the issues need to be demonstrated,

explored and acknowledged for a broader audience as has been achieved here. The demonstration in this paper has used data which is extremely mismatched to replicate common forensic situations, and has shown marked breakdowns in performance. Other experimental work that is planned on automatic methods will investigate the deterioration of ASR performance in a more stepwise manner, to better understand where these breakdowns in performance occur and why (focusing first on signal quality reductions and keeping speaker numbers equal, for example)[8].

In the new Research Hub for Language and Forensic Evidence at The University of Melbourne, we hope to work with others to find "solutions that allow maximal value of the intelligence contained in covert recordings, while reducing the risk of injustice through biased perception of indistinct audio" (Fraser, 2014, p. 5). This means taking a cautious and measured approach when it comes to the use of ASR (and forced alignment) in forensic phonetics, without discounting their effectiveness in every domain. We are engaged in experimental work which aims to better understand how well human transcribers (with an aptitude for transcription of indistinct forensic audio) handle forensic-like audio when producing transcripts. As mentioned in the background, and as can be deduced from comparing the research discussed here, we should expect that humans will perform better than machines, but also that it will take them longer (i.e., Schiel et al., 2012). This matter of efficiency should be subject to a risk-benefit analysis, and we argue that in forensics the risk of losing accuracy is too great, and that human intervention is entirely appropriate for this task – however, the specifics of how to do this in the best way is still an open question.

As noted by Watt and Brown (2020, p. 411) in their discussion of the role of automatic methods in speaker recognition, there is a clear need to "[develop] initiatives to stimulate broader and deeper dialogue among practitioners in … closely related fields" so that all parties understand the nature of indistinct covert recordings, as well as the capabilities of automatic systems–what they have been developed for, and their extension outside that realm.

[8]Thank you to reviewer 2 for explicitly pointing out this research focus.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Melbourne, Project ID 21285. Participants provided written consent to participate in this study, and written informed consent was obtained from the individual whose potentially identifiable data is included in this article.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The author receives funding support from the School of Languages and Linguistics and the Faculty of Arts, at the University of Melbourne. She also receives support from the ARC Centre of Excellence for the Dynamics of Language, Grant ID CE140100041.

ACKNOWLEDGMENTS

Thanks to Helen Fraser for assistance with ideas in this manuscript and discussion of issues surrounding the main themes within. Thanks are also due to Hywel Stoakes for discussion around ASR techniques.

REFERENCES

Bird, S. (2020). Sparse transcription. *Comput. Linguist.* 46, 713–744. doi: 10.1162/coli_a_00387

Cambridge (2021). *HTK–Hidden Markov Model Toolkit - Speech Recognition Toolkit.* Available online at: http://htk.eng.cam.ac.uk/HTK (accessed Sept. 21, 2021).

Coto-Solano,. R., Stanford, J., and, Reddy, S. (2021). Advances in completely automated vowel analysis for sociophonetics: using end-to-end speech recognition systems with DARLA. *Front. Artif. Intell.* 4, 1–19. doi: 10.3389/frai.2021.662097

Evanini, K., Isard, S., and Liberman, M. (2009). *Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora*. Brighton, UK: Interspeech. p. 1655–1658. Available online at: http://www.evanini.com/papers/evanini_INTERSPEECH09b.pdf (accessed April 28, 2022).

Fraser, H. (2014). Transcription of indistinct forensic recordings: problems and solutions from the perspective of phonetic science. *Linguagem e Direito.* 1, 5–21. Retrieved from: https://ojs.letras.up.pt/index.php/LLLD/article/view/2429

Fraser, H. (2019). Enhancing' forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript? *Aust. J. Forensic Sci.* 52, 465–476. doi: 10.1080/00450618.2018.1561948

Fraser, H. (2020). Introducing the research hub for language in forensic evidence. *Judicial Offic. Bull.* 32, 117–118.

Fraser, H., and Loakes, D. (2020). "Acoustic injustice: the experience of listening to indistinct covert recordings presented as evidence in court", in *Law, Text, Culture (special issue "The Acoustics of Justice: Law, Listening, Sound")*, eds M. San Roque, S. Ramshaw, and J. Parker (Wollongong: The University of Wollongong). p. 405–429.

Gittelson, B., Leeman, A., and Tomaschek, F. (2021). Using crowd-sourced speech data to study socially constrained variation in nonmodal phonation. *Front. Artif. Intell.* 3, 1–7. Article. *No.* 565682. doi: 10.3389/frai.2020.565682

Gonzalez, S., Grama, J., and Travis, C. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard.* 6, 1–13. doi: 10.1515/lingvan-2019-0058

Gonzalez, S., Travis, C., Grama, J., Barth, D., and Ananthanarayan, S. (2017). "Recursive forced alignment: a test on a minority language," in *Proceedings*

of the 17th Australasian International Conference on Speech Science and Technology, Epps, J., Wolfe, J., Smith, J., and Jones, C. (eds). *ASSTA Inc: Sydney*. p. 145–148.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistic Compass.* 2, 671–711. doi: 10.1111/j.1749-818X.2008.00066.x

Jones, C., Li, W., Almeida, A., and German, A. (2019). Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language. *Lang. Doc. Conserv*. 13, 281–299. Available online at: http://hdl.handle.net/10125/24869

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347. doi: 10.1016/j.csl.2017.01.005

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

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language.* 89, 30–65. doi: 10.1353/lan.2013.0015

Lindh, J. (2007). Semi-automatic aligning of swedish forensic phonetic phone speech in praat using viterbi recognition and HMM. *Proceed. IAFPA. 2007*. Plymouth, UK: The College of St Mark and St John.

Lindh, J. (2017). *Forensic Comparison of Voices, Speech and Speakers: Tools and Methods in Forensic Phonetics*. PhD dissertation. University of Gothenburg.

Mackenzie, L., and Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard.* 6. doi: 10.1515/lingvan-2018-0061

Malik, M., Malik, M. K., Mehmood, K., and Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimed. Tools. Appl.* 80, 9411–9457. doi: 10.1007/s11042-020-10073-7

McCarthy, J. (2007). *What is Artificial Intelligence?* Available online at: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (accessed Sept 14, 2021).

Opiah, A. (2021). *Descript Audio and Podcast Platform Review TechRadar Pro*. Available online at: https://www.techradar.com/au/reviews/descript (accessed October 11, 2021).

O'Shaugnessy, D. (2008). Automatic speech recognition: history, methods and challenges. *Pattern Recognit.* 41, 2965–2979. doi: 10.1016/j.patcog.2008.05.008

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). *The Kaldi Speech Recognition Toolkit.* Hawaii: Paper presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Rose, P. (2002). *Forensic Speaker Identification*. London, UK: Taylor and Francis. doi: 10.1201/9780203166369

Schiel, F., Draxler, C., Baumann, A., Elbogen T., and Steen, A. (2012). *The Production of Speech Corpora*. Available online at: www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook (accessed 21 Sept, 2021).

Villarreal, D., Clark, L., Hay, J., and, Watson K. (2020). From categories to gradience: Auto-coding sociophonetic variation with random forests *Laboratory Phonology* 11, 1–31. doi: 10.5334/labphon.216

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

Watt, D., and Brown, G. (2020). *Forensic Phonetics and Automatic Speaker Recognition. The Routledge Handbook of Forensic Linguistics.* London: Routledge. p. 400–415. doi: 10.4324/9780429030581-32

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.