Law Text Culture

Volume 24 Article 16

2020

Acoustic injustice: The experience of listening to indistinct covert recordings presented as evidence in court

Helen Fraser
University of Melbourne, helen.fraser@unimelb.edu.au

Debbie Loakes
University of Melbourne, dloakes@unimelb.edu.au

Acoustic injustice: The experience of listening to indistinct covert recordings presented as evidence in court

Abstract

Audio recorded by hidden listening devices can provide powerful evidence in criminal trials. Unfortunately these covert recordings are often indistinct, to the extent the court needs a transcript to understand the content. Australian law allows police to provide transcripts as 'ad hoc experts'. Legal procedures incorporate safeguards intended to ensure the transcripts are not misleading. The problem is that these safeguards have been shown to be ineffective, with multiple examples of inaccurate transcripts being provided to 'assist' the jury in determining what is said and who is saying it. The present paper explains the problem, provides an accessible overview of the nature of speech and how speech perception works, and outlines the solution proposed by the Research Hub for Language in Forensic Evidence to the 'acoustic injustice' embodied in current legal procedures.

Acoustic injustice: The experience of listening to indistinct covert recordings presented as evidence in court

Helen Fraser and Debbie Loakes[1]

1. Introduction

One of the most important acoustic experiences in court is that of listening to covert recordings. Covert recordings are conversations captured by telephone intercept or a hidden listening device, without the knowledge of the participants. When used as evidence in a criminal trial, covert recordings allow the court to have the experience of 'hearing with their own ears' as speakers make admissions they would not make openly. This can provide powerfully persuasive evidence, exerting a strong influence on anybody exposed to it: lawyers, judges, even journalists and other commentators – but most importantly, on the jury.

The very power of covert recordings also gives them the potential to be powerfully misleading, if the evidence they present is erroneous (Fishman 2006). For this reason it is essential that the jury reach a reliable understanding regarding who is speaking, and what they are saying. Hearing spoken language accurately can be difficult under any circumstances (Burridge 2017). Yet a common characteristic of covert recordings creates a particular problem: since it is hard to record good quality audio in secret, the content is often indistinct, to the extent it

cannot be understood without assistance.

In such cases, Australian law allows the jury to be provided with several kinds of assistance, the most important of which is a transcript setting out relevant utterances, and attributing each to a speaker. Of course, the law recognises that it is essential to ensure the court is not inadvertently misled by the 'assistance' of an unreliable transcript. To avoid this, a number of safeguards have been developed (discussed in more detail below). However, these safeguards rely heavily on lawyers and other listeners experiencing a sense of personal confidence that they hear the content represented by the transcript in the audio. Unfortunately, as the present paper explains, this experience of personal confidence is known to be surprisingly unreliable, creating actual and potential injustice (Fraser 2018a).

These and other issues prompted Australian linguistic scientists to raise a Call to Action – a 2017 letter, endorsed by all four national linguistics organisations, asking the judiciary to review and reform the handling of indistinct covert recordings in four main areas: transcription of English language utterances, translation of non-English utterances, attribution of utterances to speakers, and admission of 'enhanced' versions of the audio (Fraser 2018b). The present paper focuses on just two of these areas: identifying *who is speaking* and *what is being said* in indistinct English audio. It starts with a brief overview of issues related to covert recordings, first from a legal perspective, and then from a linguistic science perspective. Next it calls attention to the mismatch between these perspectives, and the problems that result from that mismatch, first in relation to determining what is said, and then in relation to attributing relevant utterances to particular speakers. Finally, it describes the 'acoustic injustice' arising from the mismatch, and indicates the way forward recommended by Australian linguists.

2. Covert recordings in legal perspective

A basic legal principle is that understanding spoken or written English requires only common knowledge. This means that, whereas DNA, fingerprints and other kinds of forensic evidence require specialist

interpretation, covert recordings are expected to be understood directly by the jury – much as they understand the statements of witnesses testifying in person. This is part of what makes covert recordings such powerful evidence: admissions the jury may hear can be considered to be direct, as opposed to circumstantial, evidence (see for example Martin 2014: 340).

An obstacle arises when (as often happens) recordings are so indistinct that the jury cannot understand the content simply by listening as the audio is played in court. This is where current Australian law, following the landmark High Court decision in *Butera v DPP* (1987), allows assistance to be provided in the form of a transcript, setting out what is said in the recording and attributing each utterance to a speaker. For English language recordings, the transcript is typically created by detectives from the case, who are deemed to be 'ad hoc experts' on the grounds that they have gained specialised knowledge in relation to the audio by listening to it many times (Edmond et al 2009). Of course the risk is recognised that police transcripts might not always be fully accurate, and important precedents have established a number of safeguards to mitigate the risk (for example *Eastman v the Queen* [1997]; *R v Cassar* [1999]).

The first, crucial, safeguard is the expectation that the defence will listen to the audio carefully, check the police transcript critically, and bring any differences of opinion to the attention of the prosecution. Another safeguard is the expectation that, in the event of differences not resolvable between the parties, the judge will listen personally at a voir dire. If the judge detects specifically misleading elements, the transcript can be excluded. However, the normal expectation is that evaluation of competing transcripts should be left as a matter for the jury, on the grounds that, since understanding English language requires only common knowledge, the jury is in the best position to determine the content of an indistinct covert recording, taking into account all the evidence and advice provided throughout the trial.

To ensure they do this properly, the final and most important safeguard is the judge's instruction to the jury that they should not

simply accept the transcript, but must listen carefully to the audio and reach their own conclusion regarding what is said and who says it, using the transcript(s) only as an aid (this is the 'aide memoire instruction').

These procedures have been routine for more than thirty years, and are familiar and uncontroversial from the perspective of Australian law. Nevertheless, from the perspective of linguistic science, they are deeply concerning.

3. Linguistic science perspective

3a) Common knowledge vs linguistic science

Over recent decades, linguistic science has found that many widely held beliefs about language and speech are false (Bauer et al1998). Unfortunately, relevant findings have been slow to percolate through to the broader community. This means that confident false beliefs about language and speech remain widespread, even among educated professionals, including lawyers, and scientists from other disciplines. This section gives a brief informal overview of findings relevant to the present discussion (for more detail see Fraser 2014 and references therein).

3b) The nature of speech

An important characteristic of spoken language is that it is ephemeral: utterances disappear before listeners have a chance to study them in detail. Nowadays we have recording technology that lets us capture and analyse speech, but for many centuries, the only way to preserve a record of what was said was via writing. Written language developed in many forms around the world (Daniels et al 1996), with profound consequences, not just for the societies that use it (Olson 1994) but for their conception of speech itself (Harris 1986). In our society, alphabetic literacy has promoted the belief that speech is much like printed text: a sequence of discrete words, each made up of a sequence of discrete, invariant 'sounds' or, more technically, 'segments' (Linell 1988). However this belief, though widespread, is quite false (for engaging and accessible discussion, see Port 2007).

Through careful auditory analysis, linguists have long known that individual segments actually have a far greater range of variation than listeners realise (Bloomfield 1933). This variation is usually explained as being due to 'coarticulation' between neighbouring sounds (for example the /s/ in 'sue' is subtly different from the /s/ in 'sea', due to coarticulation with the subsequent lip-rounded vowel).

During the 1940s-60s, development of technology for recording speech and analysing its acoustic structure showed that segmental variability was far greater than previously recognised (Shankweiler et al 2015). Further research spurred by this discovery demonstrated the 'alphabetic conception' of speech to be thoroughly misleading (Appelbaum 1999). In reality, speech is not a sequence of discrete units, but a continuous stream of sound, reflecting the dynamic and highly variable processes by which it is articulated (Ladefoged et al 2012). A small impression of its nature can be given by considering it to be more like 'running writing' than printed text. However, even the clearest of speech is like extremely messy running writing – with no spaces even between words, let alone between letters. This means that, amazing as it might seem, there is no set of acoustic features that is always and only associated with any particular segment (in more technical terms, there is no 1:1 relationship between any unit of acoustics and any unit of perception).

This raises the question of how listeners perceive the words and segments they are so keenly aware of in experience. While there are many competing theories, there is broad agreement regarding basic principles (Magnuson et al 2013).

3c) How speech perception works

For proficient speakers, understanding spoken language seems so effortless it is often assumed to be a simple process of recognising 'sounds' ('c-a-t') and putting them together into words ('cat'). However, the discovery, discussed in the previous section, that segments do not exist independently even in clear speech, shows that this assumption must be false. Actually, speech perception involves not recognising sounds but constructing them, via a suite of complex (though almost

entirely unconscious) mental processes. Its effortlessness is testament not to its simplicity, but to the immense proficiency speakers develop over years of practice (Cutler 2012).

The ability to recognise and discriminate segments is strongly dependent on this proficiency (this is why it is so difficult to identify sounds in a foreign language). Even local dialect proficiency can be relevant. For one example, even in Australia, where dialect differences are minimal, listeners from southern Victoria may be unable to discriminate words like 'celery' and 'salary', which those from northern Victoria readily hear as different (Loakes et al 2014).

Even for proficient speakers, however, the acoustic information in the continuous stream of speech is not enough on its own to enable words to be recognised. This is often found hard to believe, but is easy to demonstrate by excising words or phrases from a recorded conversation. Words that are perfectly clear when heard in the context of the conversation are typically unintelligible when played on their own (Shockey 2003). In order to understand speech, listeners must combine the acoustic information in each word with information from other sources – and everyday conversation offers many other sources of information. Within speech itself, there is a great deal of 'suprasegmental' and 'paralinguistic' information – rhythm, intonation, voice quality and other characteristics that extend beyond individual words, making speech a far richer and more complex signal than can be captured with a segmental (alphabetic) representation (Clark et al 2007)

In addition, listeners can use visual information to see who is speaking, and, by following speakers' gaze or gestures, identify objects or events they refer to. Information from the facial expressions that accompany speech are particularly salient (Diehl et al 2005). One interesting example is the 'McGurk effect', whereby perception of the same acoustic information can be radically changed by superimposing video of speakers articulating /ba/, /ga/ or /fa/ (it is worth experiencing this in multimedia).[2]

The need to juggle so much information from disparate sources means that speech perception is prone to a surprising number of errors

(Shockey et al 2014). Most are barely noticed, since they are corrected on the fly, as part of the perceptual process. As the unfolding speech creates its own internal context, listeners simply update their understanding, only mentioning transient errors if they happen to be humorous (for example 'I thought you asked me to send an all-star female – but then I realised you must have meant all-staff email'). On the rare occasion that hearing-errors create genuine misunderstanding, they are readily corrected via intervention from the interlocutor, allowing the conversation to continue seamlessly.

Considering the nature of speech, however, what is surprising is not that so many hearing-errors occur, but rather that errors are not more frequent, and more disruptive. Why, for a simple example, do we readily seize on the phrase 'recognise speech', without even noticing that 'wreck a nice beach' is an equally plausible way to interpret the acoustic information? The reason is found in what is perhaps the most intriguing, though least observable, source of information needed for speech perception: the listener's tacit and constantly updating expectations about what the upcoming speech is likely to be about. These expectations guide or 'prime' perception in ways that are hard to recognise except via experimental studies that control listeners' access to contextual information (Warren 2012).

In short, speech perception is far from the 'direct' observation of sounds and words assumed by the legal perspective. It is an active, dynamic, predictive, collaborative process, in which segmental acoustics plays an important but relatively minor role. Its key characteristic is listeners' use of inference, interpretation and feedback in constructing a meaningful and contextually relevant message. The only difference from other inferential reasoning is that it mostly occurs without conscious awareness.

3d) Listening to recorded speech

A major benefit of recording technology is the ability to make speech available to listeners who are not physically present (for example via radio broadcast). At first this was limited to prepared monologues. Producing recordings of speakers in spontaneous conversation had to

await technology to record and mix high quality audio from multiple microphones. Even now that the technology is available, recording intelligible audio requires management. For example, a radio talk-show host has to ensure participants speak one at a time, minimise overlaps and interruptions, mention speakers' names frequently, and so on. Without such intervention, conversation is typically difficult to follow from a recording, even if it was perfectly clear to the participants. The reason is that the recording takes speech out of its context, denying listeners the visual information they use in everyday conversation to identify speakers and resolve overlapping utterances. This is why the well-known 'cocktail party effect', whereby listeners can focus their attention on speech in a noisy environment does not work for recorded speech (Arons 1992).

Another benefit of improved recording techniques has been its enabling of advanced research on discourse and conversation analysis, which has given insight into the structure of conversation (Sidnell et al 2012), and the nature of spontaneous speech (Shockey 2003). One key observation from this research is that conversational speech is highly elliptical. Since speakers know that listeners can retrieve information from the context, they do not bother to specify every detail. In a recording, however, the omitted information is no longer retrievable, making the speech hard to understand. To extend the analogy with running writing, it could be said that listening to a (high quality) recorded conversation is like reading a Twitter thread rendered in extremely messy handwriting with no gaps between words or letters – incomprehensible without contextual information.

The interesting thing is that such a recording is less difficult to understand than the analogy might suggest. The reason is that, where listeners do not have access to the real context, they use more abstract contextual information obtained from external sources. This has been demonstrated by numerous experiments. In one example, from Germany, researchers played the same recording to several groups of listeners, priming each group with different information regarding the speaker's regional dialect (Jannedy & Weirich 2014). Though

the priming was very subtle (the region was not referred to explicitly, merely noted on the response sheet), participants heard the speech in line with expectations about how speakers from that region pronounce particular words. Similarly strong perceptual effects have been shown for Australian and New Zealand listeners, using even subtler priming (Hay et al 2006).

Experiments like these show the strong assistance provided by contextual priming, without the listener's conscious awareness. Other research has added an important new concept. While contextual priming is generally helpful in everyday life, listeners' heavy reliance on its assistance creates a perceptual vulnerability in understanding recorded speech. Priming with inaccurate contextual information does not, as might be expected, reduce understanding. Rather it encourages confident but inaccurate understanding. The effect is that listeners can be easily and unwittingly manipulated into confident but erroneous perception, especially if the audio has any degree of indistinctness (Fraser 2014 discusses all this in detail, while videos at forensictranscription.net.au provide compelling multimedia examples).

These and many other demonstrations show that the risk of perceptual error is far higher for recorded conversation than for in-person conversation – and of course the most important feature of a recording is that the speaker is not available to provide immediate feedback regarding errors that cause misunderstanding.

3e) Transcribing recorded speech

One further benefit of recording technology is the possibility for official events such as court or parliamentary proceedings to be captured in full detail. Useful as such recordings are, however, they are not nearly so convenient as the written documents traditionally provided by stenographers. For this reason, audio recordings are usually transcribed, to provide a 'verbatim' record.

Transcribing any recorded spoken interaction is difficult. Depending on the various factors discussed above, simply understanding the content of the discourse can be hard enough. Writing down each and every word can be extremely challenging (further demonstration, if

needed, that understanding does not result from simple 'bottom up' recognition of readily observable segments). However, the range of difficulty spans a long continuum. While it definitely requires more than common knowledge, producing transcripts of recorded court proceedings is at the 'relatively easy' end of that continuum.

For one thing, court recordings are generally of at least fair quality, and the speech is monitored to ensure that only one voice is heard at a time (the judge in a courtroom performing a similar function to that of the radio host mentioned earlier, by reminding witnesses to answer verbally rather than with gestures, asking for difficult words to be spelled out, and so on). For another thing, the transcriber is trained in the use of appropriate equipment, and provided with background information to assist with difficult material (for example names of speakers and technical terms). Interestingly, even with all these advantages, initial versions of court transcripts often contain errors. Fortunately, they can usually be checked by the participants, and corrections made.

Importantly, however, even the best court transcript is far from being truly 'verbatim'. This was discovered by sociolinguistics researchers who sought to use existing transcripts as the basis for analysis of courtroom discourse (Eades 1996). The very 'tidiness' that is valued in court transcripts is a disadvantage for research, as it omits or alters many of the very features of natural spoken language that researchers most want to study (Voutilainen et al 2019). Various researchers' need for transcripts that include all the detail relevant for their studies has led to development of many different forms of transcript, suited to the objectives of different disciplines (Heselwood 2013). All are valuable for their purposes. However, no transcript, no matter how detailed or how accurate, is ever equivalent to the recording (much less to the original discourse) – as can easily be demonstrated by reading the transcript aloud and comparing the result with the audio (Komter 2019 takes up this topic in a highly relevant study). This is not a criticism of transcribers. It merely demonstrates that any transcript requires abstracting from the rich and complex recording to create an

artefact suitable for a particular purpose – in much the same way as any map requires abstracting from the terrain it represents (Dibiase 2018).

Lack of recognition of the abstract nature of transcripts creates major problems in court, even for overt recordings (Haworth 2018). Covert recordings raise far more serious issues.

4. Forensic transcription

4a) A highly specialised task requiring independence, skill and process

With the background above, we can turn to consider transcription of indistinct covert recordings. Clearly, the task is far harder than transcribing court recordings. On first acquaintance, the main reason appears to be the poor quality of the audio. It is true that indistinct covert recordings are typically of far worse quality than anything that is normally transcribed for court (or other) purposes – and 'enhancing' techniques are rarely, if ever, fully effective (Fraser 2020a).

However, audio quality is only one of the features that make forensic transcription different from court transcription, and not the most important. Another difference is that covert recordings usually feature unmonitored conversation, often with multiple participants who share enough contextual knowledge to let them use highly elliptical expressions. As we have seen, even in a good quality recording, such material is hard to understand (our analogy likened it to a Twitter thread, rendered in messy handwriting with no spaces). A poor quality recording greatly exacerbates the difficulty (as if the messy handwriting was rendered in pale ink on thin paper that has become soiled and damaged). In order to understand material like this, listeners must have contextual information. Without it, as discussed earlier, the audio is simply unintelligible.

This highlights one of the most important differences between court transcription and forensic transcription: in the latter, the context may not be known, or if it is, relevant aspects may be contested, or simply wrong. This creates the risk of the transcriber being exposed to unreliable contextual information. As we have seen, unreliable

contextual information is liable to induce inaccurate perception – resulting in an inaccurate transcript. Unfortunately, even ensuring the transcriber receives only reliable contextual information does not guarantee accurate perception. The only way to be absolutely sure a transcript is reliable is by checking it against 'ground truth' (accurate, uncontestable knowledge of what was really said). This raises the most crucial difference between forensic and other transcription: ground truth is not available – that is why it is necessary to task the jury with determining what was said, using the transcript as assistance.

For all these reasons and more, forensic transcription is not just harder than court transcription. It is actually harder than the transcription undertaken by advanced researchers in conversation analysis or phonetic science. We end with one final consideration: the consequences of error. In forensic transcription the stakes are far higher than in academic research. To avoid serious injustice, it is essential to avoid any possibility of misleading the jury with unreliable 'assistance'.

From the perspective of linguistic science, then, it is clear that achieving appropriate reliability for forensic transcription requires independent transcribers with demonstrable skill, following an evidence-based process that provides them with necessary, relevant and reliable contextual information, while shielding them from misleading or biasing assumptions, at least until they have formed their own preliminary interpretation of the audio (cf Dror et al 2015). Clearly there is a major mismatch between this practice, required from the perspective of linguistic science, and the practice currently used in Australian courts as outlined in Section 2 above.

4b) Mismatch between legal perspective and linguistic science perspective

While the discussion above is far from a full explanation of issues in forensic transcription, it may give some insight into why linguistic scientists are so concerned about current legal practice: it embodies a paradoxical situation, whereby unusually difficult audio, used in contexts where the consequences of error are unusually severe, is transcribed and evaluated by personnel with unusually low qualifications ('listening

many times' is necessary but not sufficient: it is quite possible to be wrong every time).

The reason for the mismatch can be traced to a major error in the fundamental principle on which legal practice is based: the concept that understanding spoken language requires only common knowledge. This may be reasonable in relation to some forms of speech in some contexts (for example, listening to court proceedings). However, extending it to the assertion that understanding indistinct covert recordings requires only common knowledge involves a serious fallacy.

Ensuring reliable understanding of indistinct covert recordings requires advanced expertise of a level similar to that required for DNA or other expert evidence (Fraser 2018b). The difference is that, where DNA evidence is opaque to the court in the absence of expert interpretation, listening to an indistinct covert recording may give listeners an experience of understanding the content for themselves. However, this ability to 'hear with one's own ears' does not make the evidence 'direct' – it merely hides crucial inferential reasoning below the level of conscious awareness (Section 3c). This makes it even more essential with audio than with DNA to protect juries from misleading 'assistance', by ensuring they are only ever exposed to reliable transcripts.

Current law, however, allows police transcripts to 'assist' juries. This is problematic, not just because police lack genuine expertise in transcription. Police transcribers are further hindered by having contextual information that is potentially unreliable (having not yet been tested by the trial process). As we have seen (Section 3d), while contextual information can confer useful insight regarding the interpretation of particular indistinct phrases, it is equally possible that it will mislead perception. And indeed it is known that police transcripts are frequently inaccurate (French and Fraser 2018).

So while police suggestions should certainly form part of the input, at an appropriate stage, to an expert transcription process, ensuring their suggestions are not misleading requires specialist evaluation. The law, however, leaves evaluation to lawyers. Most of the safeguards (Section

2) involve lawyers experiencing a sense of personal confidence that the police transcript assists their own perception (Gray 2018). The problem, as explained above, is that this experience is highly unreliable. The effect is that listeners are unlikely to detect and correct all relevant errors.

That in turn means it is common for juries to be given erroneous police transcripts, with only the aide memoire instruction to protect them from being misled (Fraser 2018b). Unfortunately, this final safeguard too is unrealistic. As explained in Section 3d, a transcript inevitably has a powerful and lasting effect on perception of indistinct audio, even if it is inaccurate. It is very unlikely listeners will successfully 'reset' their perception to give equal consideration to alternative interpretations – making the aide memoire instruction one more way in which juries are asked to 'do the impossible' (cf Tiersma 2009). The overall effect is extraordinary privilege for the police interpretation of indistinct covert recordings.

Lacking insight into the factors that affect speech perception, the law has invested unwarranted confidence in police transcripts, both in specific cases and in general. Since the transcript is not considered to be evidence, but only 'assistance' in understanding the evidence, police 'ad hoc experts' are subject to none of the scrutiny that genuine expert evidence is given (Roberts 2020). There are known examples of demonstrably inaccurate police transcripts having been admitted as assistance to the jury despite careful checking by lawyers and even judges – and it is certain there are more, as yet unknown (Fraser 2018a). The threat to justice is evident.

5. Speaker attribution

This section turns from the question of *what was said* to consider difficulties in establishing, with reliability appropriate for a criminal trial, *who said it*.

5a) Linguistic science perspective

We all have the daily experience of recognising the voices of people we know. The fact that we are usually right gives confidence in listeners' ability to recognise speakers by their voices. It also gives confidence

in the 'common knowledge' explanation for the ability: voices have unique features which listeners come to recognise, through familiarity. However, both of these concepts have been shown to be incorrect.

First, speakers do not have unique voices (Rose 2002, Watt et al 2020). This is one reason that responsible experts deprecate the term 'voiceprint', and its suggestion of an analogy with fingerprints (the technical term is 'spectrogram'). There is of course a great deal of variation between the voices of different speakers. However, voices are highly complex signals that also vary greatly within the speech of any individual: consider the difference in the voice of the same person speaking angrily to a colleague, formally to a boss or lovingly to a child. It is well known, from twin studies and other research on similar-sounding speakers, that within-speaker variation can be as great as or greater than between-speaker variation (Loakes 2008, in press) – and that is without even considering the possibility of deliberate disguise.

This counter-intuitive fact is confirmed by the observation that there are currently no methods that allow reliable identification of a voice from an open population in anything remotely like what is possible with fingerprints (Foulkes et al 2012) – and it is useful to recall that even fingerprint identification is far from infallible (Campbell 2011, Walvisch 2017). The increasing success of voice verification services (for example for lodging tax claims by telephone) depends on reducing the population of possible speakers by requiring input of personal details such as date of birth and tax file number. It also simplifies the task by requiring the speaker to cooperate in producing standard phrases for comparison. While the results are impressive (though not foolproof), such constraints make the methods unsuitable for forensic purposes.

Second, while listeners certainly do recognise speakers' voices, they do not recognise speakers *by* their voices. As with speech perception, speaker recognition relies far more heavily than listeners realise on priming by contextual expectations. Experiments that deprive listeners of contextual information, forcing them to use only the voice itself, show surprisingly poor performance, and, importantly, poor correlation between listeners' confidence and accuracy (Kreiman et al 2011). This

is why earwitnesses, who identify speakers based on memory of having 'heard that voice before', are even less reliable than eyewitnesses (Fraser 2019) – and eyewitnesses, as is well known, are notoriously fallible (Gould et al 2012).

Third, while having a recording of an offender's voice reduces the need to rely on a listener's memory, it is by no means a panacea for the problems of unreliable speaker identification. The handwriting analogy we have been developing might help explain this. If clear speech is like messy running writing with no gaps, then identifying voices is at least as problematic as identifying authorship of a handwritten text – a task for which common knowledge notoriously overestimates reliability, and even experts are surprisingly fallible (Found et al 2013). In fact, even in clear recordings, recognising voices is far more problematic than recognising handwriting, due to listeners' poor ability to provide valid descriptions even of apparently basic features like pitch or accent (Tomkinson et al 2018). The fact that covert recordings are often indistinct only adds to these problems (making it like recognising the author of messy handwriting in pale ink on thin, soiled paper).

With this brief background, we can review the legal perspective on identifying speakers whose voices are heard in covert recordings.

5b) Legal perspective

Utterances in covert recordings are generally attributed to specific speakers as part of the police transcript. In addition, investigators may testify as 'ad hoc experts' that they recognise voices in covert recordings, on the grounds that familiarity with the voices, gained through their work on the investigation, gives them specialised knowledge. However, as discussed above, no one can reliably identify voices from an open population. Familiarity may assist in some cases, but is no guarantee of accuracy (Yarmey 2004), especially in cases where cognitive bias may be a factor (Smith et al 2014).

For many types of covert recordings, there is also external evidence regarding the identity of the speakers, for example from personal surveillance, or from information about time, location and phone numbers in intercepted calls. While these kinds of external evidence

are obviously useful in themselves, it is important not to misunderstand their relationship to the evidence obtained from the voices themselves. It is often assumed that such external evidence confirms investigators' voice recognition. However, this could only be (potentially) correct if investigators first recognised the voices independently, and then reviewed the external evidence. Of course, this is not what happens in practice: the whole basis of investigators' 'ad hoc expertise' is (putative) familiarity with the voices gained through their work on the case as a whole. It seems clear that in many or most cases, what really helps investigators identify speakers is the external evidence, with the voice characteristics adding confirmation. This is not necessarily a problem in itself. The problem is the fallacious assumption, which gives unwarranted confidence in the ability of police to identify speakers, both in specific cases and in general. The effect of this circular reasoning is to give yet more privilege to the police interpretation of covert recordings.

Recent years have seen more frequent use of expert witnesses in relation to speaker identification. Unfortunately, however, it can be difficult for the courts to distinguish real expertise from pseudoscience, and even genuine expert opinion is seen only as an alternative to the police identification, with the choice between them left to the jury (Edmond et al 2011). As a consequence, multiple cases of 'acoustic injustice' are known (Catanzaro 2015).

5c) Capabilities and limitations of expert analysis

While voices cannot be uniquely identified from an open population, there is still a great deal of useful evidence that expert analysts can provide (Foulkes et al 2012), usually by comparing voice samples from an 'unknown' or 'disputed' recording (typically featuring one or more offenders' voices) with a sample from a 'known' recording (typically from a police interview, or some other context where the speaker can be unequivocally identified).

An essential consideration in making this kind of comparison (Rose 2002), as in any forensic comparison evidence (Aitken et al 2010), is to avoid the common pitfall of focusing on *similarities* between the

known and unknown samples, with insufficient consideration of the *distinctiveness* of the similar features. For a simple example, consider an unknown sample featuring a male voice with an average pitch around 120 Hz. The fact that the known sample features a male voice with a similar average pitch is not, in itself, of much forensic value, since many male voices have this pitch. The principle is the same as in the intuitively more obvious case of an offender and a suspect both having brown hair: this similarity is of limited value – unless the population of possible offenders can be reduced in some reliable way to include very few people with brown hair, making this a distinctive characteristic.

However, forensic voice comparison is substantially more problematic than other types of superficially similar expert evidence. One reason is the very large overlap of within-speaker and between-speaker variability already mentioned, which affects almost every characteristic of voices. This makes it essential, but difficult, to ensure samples are fully commensurate (i.e. that known and unknown samples compare like with like). It also makes it very difficult to collect population statistics for specific voice characteristics, and to use them in meaningful ways (Morrison et al 2016).

For these and other reasons, despite extensive research, we still have nothing like a standard method for forensic voice comparison, certainly not one that can be applied in a context-independent manner (Gold et al 2019). In most cases, reliable information about other evidence in the case is needed to narrow the population of possible speakers. Unfortunately, it is still not entirely clear how best to enable experts to make use of reliable information about the case, while minimising the risk of cognitive bias (Kinoshita et al 2015). This and many other issues are still under active discussion by researchers.

Another issue is the difficulty of determining the 'defence hypothesis'. Statistical comparison of samples requires determining a 'likelihood ratio' representing the likelihood of observing the data under competing hypotheses. While the prosecution hypothesis is generally straightforward ('the known and unknown samples were produced by one and the same speaker'), specifying the defence hypothesis in a

valid way is surprisingly problematic, since 'not produced by the same speaker' is insufficient for statistical analysis (Hughes et al 2018).

A further problem is the difficulty of determining which specific utterances should rightly be included in the 'unknown' sample. In many cases, the unknown sample comprises more than one utterance – sometimes spread over multiple recordings, and often occurring within complex conversations featuring numerous voices in indistinct audio – all of which makes speaker attribution highly problematic. Yet it is surprisingly common for expert analysts simply to accept the police-attributed voices as the unknown sample (Fraser 2018c).

Finally, it is worth remembering that the point of expert evidence is not for the expert to reach a reliable conclusion, but for the jury to reach a reliable conclusion – or at least, perhaps more importantly, to prevent the jury from reaching a misleading conclusion. Communicating complex scientific results involving advanced statistics to a jury is a major hurdle for any field of forensic science (Martire 2018). Again, these difficulties are magnified for speaker comparison evidence, especially by the fact that it is so easy for listeners to 'reach their own conclusion' about speech evidence, with insufficient recognition of how often their 'own conclusions' are confident but wrong.

Expert evidence is a useful and necessary, but partial, corrective to problems of current legal practice regarding identification of speakers in covert recordings (McGorrery et al 2016). It might be preferable to aim for practices that reduce the mismatch between legal and linguistic perspectives on voice evidence – via the collaborative research recommended by proponents of the Call to Action (Fraser 2018b).

6. Conclusion

This paper has demonstrated several kinds of acoustic injustice arising from misconceptions within the law that allow a transcript to be treated merely as 'assistance' rather than as evidence in its own right. We end by reflecting on a quote from the *Washington Law Review* more than a decade ago:

> Rather than treating a transcript as a non-evidentiary "aid to understanding" the recording […], a transcript of a recording should be recognized for what it is, i.e., opinion evidence as to the contents of the recording, and its admissibility should be governed by the same rules and procedures that apply to opinion evidence generally. (Fishman 2006: 523)

While we agree wholeheartedly with the first point, we note that the powerful effect a transcript can exert on listeners' perception of *who is speaking* and *what is being said* means that 'procedures that apply to opinion evidence generally' may not be enough to solve the problems identified above. Finding an adequate solution requires linguistics, law and law enforcement working together to develop and implement transparent, evidence-based procedures that ensure all covert recordings are provided with a demonstrably reliable transcript before they enter the trial process. The Research Hub for Language in Forensic Evidence, established as a direct result of the successful Call to Action discussed above, seeks to develop a collaborative research program to achieve this (Fraser 2020b).

Endnotes

1. Though the first author wrote the text, both authors contributed equally to the content.
2. Experience the McGurk effect here: https://www.youtube.com/watch?v=2k8fHR9jKVM.

References

Aitken C, Roberts P and Jackson G 2010 *Fundamentals of probability and statistical evidence in criminal proceedings* Royal Statistical Society London

Appelbaum I 1999 'The dogma of isomorphism: A case study from speech perception' *Philosophy of Science* 66: S250-S259

Arons B 1992 'A review of the cocktail party effect' *Journal of the American Voice I/O Society* 12/7: 35-50.

Bauer L and Trudgill P 1998 *Language myths* Penguin Books London New York

Bloomfield L 1933 *Language* University of Chicago Press Chicago

Bohn OS and Munro MJ eds 2007 *Language experience in second language speech learning: In honor of James Emil Flege* J Benjamins Publishing Co Amsterdam

Burridge K 2017 'The dark side of mondegreens: How a simple mishearing can lead to wrongful conviction' *The Conversation* 6 June 2017 https://theconversation.com/the-dark-side-of-mondegreens-how-a-simple-mishearing-can-lead-to-wrongful-conviction-78466

Campbell Sir A 2011 *The fingerprint inquiry report* APS Group Edinburgh

Catanzaro M 2015 'Speech forensics: When Hollywood seldom mirrors real-life court cases' *Euroscientist* <https://www.euroscientist.com/speech-forensics-when-hollywood-seldom-mirrors-real-life-court-cases/>

Clark JE, Yallop C and Fletcher J 2007 *An introduction to phonetics and phonology* Blackwell, Oxford

Cutler A 2012 *Native listening: Language experience and the recognition of spoken words* MIT Press Cambridge

Daniels P and Bright W eds 1996 *The world's writing systems* Oxford University Press Oxford

Dibiase D 2018 *The nature of geographic information: An open geospatial textbook* Penn State University Press University Park

Diehl RL, Lotto AJ and Holt LL 2005 'Speech perception' *Annual Review of Psychology* 55: 149-179

Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane DE, Saks MJ and Risinger M 2015 'Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making' *Journal of Forensic Sciences* 60: 1111-1112

Eades D 1996 'Verbatim courtroom transcripts and discourse analysis' in Kniffka 1996: 241-254

Edmond G and San Roque M 2009 'Quasi-justice: Ad hoc expertise and identification evidence' *Criminal Law Journal* 33: 8-33

Edmond G, Martire KA and San Roque M 2011 'Unsound Law: Issues with "expert" voice comparison evidence' *Melbourne University Law Review* 35: 52-112

Fishman CS 2006 'Recordings, transcripts, and translations as evidence' *Washington Law Review* 81: 473-523

Foulkes P and French P 2012 'Forensic speaker comparison: A linguistic-acoustic perspective' in Solan et al 2012: 557-572

Found B and Ganas J 2013 'The management of domain irrelevant context information in forensic handwriting examination casework' *Science & Justice* 53: 154-158

Fraser H 2014 'Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science' *Language and Law/ Linguagem e Direito* 1: 5-21

Fraser H 2018a 'Forensic transcription: How confident false beliefs about language and speech threaten the right to a fair trial in Australia' *Australian Journal of Linguistics* 50: 129-139

Fraser H 2018b 'Thirty years is long enough: It's time to create a process that ensures covert recordings used as evidence in court are interpreted reliably and fairly' *Journal of Judicial Administration* 27: 95-104.

Fraser H 2018c Review of 'Forensic Communication in Theory and Practice: A study of discourse analysis and transcription' *Language and Law/Linguagem e Direito* 5: 103-108

Fraser H 2019 'The reliability of voice recognition by ear witnesses: An overview of research findings' *Language and Law/Linguagem e Direito* 6: 1-9

Fraser H 2020a 'Enhancing forensic audio: What works, what doesn't and why' *Griffith Journal of Law and Human Dignity* 8/1: 85-102

Fraser H 2020b 'Introducing the Research Hub for Language in Forensic Evidence' *Judicial Officers' Bulletin* 32/11: 117-118

French P and Fraser H 2018 'Why "ad hoc experts" should not provide transcripts of indistinct forensic audio, and a proposal for a better approach' *Criminal Law Journal* 42: 298-302.

Gold E and French P 2019 'International practices in forensic speaker comparisons: second survey' *International Journal of Speech Language and the Law* 26: 1-20.

Gould JB, Carrano J, Leo R and Young J 2012 *Predicting erroneous convictions: A social science approach to miscarriages of justice* National Institute of Justice Washington DC

Gray P 2018 'Lost in transcription: covert recordings' *Victorian Bar News* 163: 51-53

Harris R 1986 *The origin of writing* Duckworth London

Haworth K 2018 'Tapes, transcripts and trials' *International Journal of Evidence and Proof* 22/4: 428-450

Hay J, Nolan A and Drager K 2006 'From fush to feesh: Exemplar priming in speech perception' *The Linguistic Review* 23: 351-379

Hay J and Parnell E eds 2014 *Proceedings of the 15th Australasian International Conference on Speech Science and Technology* ASSTA Canterbury, New Zealand

Heselwood B 2013 *Phonetic transcription in theory and practice* Edinburgh University Press Edinburgh

Hughes V and Rhodes R 2018 'Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice' *Science & Justice* 58: 250-257.

Jannedy S and Weirich M 2014 'Sound change in an urban setting: Category instability of the palatal fricative in Berlin' *Laboratory Phonology* 5: 91-122

Kinoshita Y and Ishihara S 2015 'Background population: how does it affect LR based forensic voice comparison?' *International Journal of Speech Language and the Law* 21: 191-224

Kniffka H ed 1996 *Recent developments in forensic linguistics* Peter Lang Frankfurt

Komter M 2019 *The suspect's statement: Talk and text in the criminal process* Cambridge University Press Cambridge

Kreiman J and Sidtis D 2013a 'Identifying unfamiliar voices in forensic contexts' in Kreiman et al 2013: 237-259

Kreiman J and Sidtis D 2013b *Foundations of voice studies: An interdisciplinary approach to voice production and perception* Wiley-Blackwell Hoboken

Ladefoged P and Disner SF 2012 *Vowels and consonants: An introduction to the sounds of language* 3rd edition Wiley-Blackwell Hoboken

Linell P 1988 'The impact of literacy on the conception of language: The case of linguistics' in Säljö 1988: 41-58

Loakes D 2008 'A forensic phonetic investigation into the speech patterns of identical and non-identical twins' *International Journal of Speech Language and the Law* 15: 97-100

Loakes D in press 'Twin research' In McDougall K and Nolan F (eds.) *Oxford Handbook of Forensic Phonetics* Oxford University Press Oxford

Loakes D, Hajek J, Clothier J and Fletcher J 2014 'Identifying /el/-/æl/: A comparison between two regional Australian towns' in Hay et al: 41-44

Magnuson JS, Mirman D and Myers E 2013 'Spoken word recognition' in Reisberg 2013: 412-441

Martin BR 2014 *Inquiry into the conviction of David Harold Eastman for the murder of Colin Stanley Winchester* Report of the Board of Inquiry

Martire KA 2018 'Clear communication through clear purpose: Understanding statistical statements made by forensic scientists' *Australian Journal of Forensic Sciences* 50(2): 166-182

McGorrery PG and McMahon M 2016 'A fair "hearing": Earwitness identifications and voice identification parades' *International Journal of Evidence and Proof* 21: 262-286

Morrison GS, Enzinger E and Zhang C 2016 'Refining the relevant population in forensic voice comparison – A response to Hicks et al (2015) The importance of distinguishing information from evidence/observations when formulating propositions' *Science & Justice* 56: 492-497

Olson DR 1994 *The world on paper: The conceptual and cognitive implications of writing and reading* Cambridge University Press Cambridge

Port RF 2007 'The graphical basis of phones and phonemes' in Bohn et al 2007: 349-365

Reisberg D ed 2013 *Oxford handbook of cognitive psychology* Oxford University Press Oxford

Roberts A 2020 'Knowledge, reliability, and the admissibility of forensic science evidence' *Australian Journal of Forensic Sciences* 52/3: 269-274

Rose P 2002 *Forensic speaker identification* Taylor & Francis London

Säljö R ed 1988 *The written world: Studies in literate thought and action* Springer-Verlag Berlin

Shankweiler D and Fowler CA 2015 'Seeking a reading machine for the blind and discovering the speech code' *History of Psychology* 18: 78-99.

Shockey L 2003 *Sound patterns of spoken English* Blackwell Oxford

Shockey L and Bond ZS 2014 'What slips of the ear reveal about speech perception' *Linguistica Lettica* 22: 107-113

Sidnell J and Stivers T eds 2012 *The handbook of conversation analysis* John Wiley & Sons Chichester

Smith HMJ and Baguley T 2014 'Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings' *Journal of European Psychology Studies* 5: 59-68

Solan L and Tiersma P eds 2012 *The Oxford handbook of language and law* Oxford University Press Oxford

Tiersma PM 2009 'Asking jurors to do the impossible' *Tennessee Journal of Law and Policy* 5: 105-148.

Tomkinson, J and Watt D 2018 'Assessing the abilities of phonetically untrained listeners to determine pitch and speaker accent in unfamiliar voices' *Language and Law/Linguagem e Direito* 5: 19-37.

Voutilainen E and Inoue M 2019 'Sociolinguistic and sociotechnical approaches to official transcripts' presented at the *16th International Pragmatics Association meeting* Hong Kong

Walvisch J 2017 'Fingerprinting to solve crimes: Not as robust as you think' *The Conversation* 24 October 2017 https://theconversation.com/fingerprinting-to-solve-crimes-not-as-robust-as-you-think-85534

Warren P 2012 *Introducing psycholinguistics* Cambridge University Press Cambridge

Watt D, Harrison P and Cabot-King L 2020 'Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker' *International Journal of Speech Language and the Law* 26: 137-180

Yarmey AD 2004 'Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations' *International Journal of Speech Language and the Law* 11: 267-277

CASES

Butera v DPP (1987) 164 CLR 180

Eastman v the Queen [1997] FCA 548

R v Cassar [1999] NSWSC 436