

Interpretation of a crisis call: persistence of a primed perception of a disputed utterance

Helen Fraser, Bruce Stevenson and Tony Marks

Abstract

This article describes an experiment designed to explore the effects of 'priming' (i.e. being exposed to a suggested interpretation of an audio signal) on how juries perceive disputed utterances in poor quality recordings used as evidence in legal cases. Using the actual disputed utterance from a real case, the experiment tracked how participants' perception of its content changed as 'evidence' about the case was gradually revealed to them. At a certain point, participants were randomly divided into two groups, each receiving systematically different 'evidence' – with significant effects on their interpretation of the section of interest. Results indicate that the dangers of juries being primed may be considerably greater than is generally recognised, and unlikely to be overcome by a mere caution from the judge. They also indicate that participants' propensity to consider the defendant guilty may be influenced by initial impressions of his trustworthiness based on his style of speech, rather than on objective assessment of the evidence presented to them. It is hoped these experimental results may form the basis for closer collaboration between the legal system and phonetics experts regarding issues of forensic transcription.

KEYWORDS FORENSIC TRANSCRIPTION, DISPUTED UTTERANCE, COGNITIVE PHONETICS, PRIMING

Affiliation

Helen Fraser is an independent researcher

Bruce Stevenson and Tony Marks: University of New England, Australia

email: helen@helenfraser.com.au

Introduction

Legal cases increasingly include evidence in the form of audio recordings obtained from telephone intercepts, covert listening devices and similar sources (Coulthard and Johnson 2007). Sometimes these recordings are of poor quality,¹ to the extent that prosecution and defence disagree as to the content of particular ‘disputed utterances’. In such situations, debate can surround not just the accurate interpretation of what words (if any) are spoken in the recording, but the admissibility of the recording as evidence for presentation to a jury. The case of David Bain, from New Zealand, offers a good example, where a very brief and barely audible section of a recorded crisis call was alleged by the prosecution to be the caller confessing to murder, and by the defence to be uninterpretable speech, perhaps not even speech at all.

How can the truth be decided in such a situation? One possibility is to play the recording to the jury, perhaps with advice from expert witnesses on each side, and let them decide. This approach raises a problem, very well known to the disciplines of phonetics and psycholinguistics (Fraser 2003, Byrd and Mintz 2010) and acknowledged to a certain extent by the legal system: the danger of ‘priming’ (i.e. the tendency, especially with a degraded signal, for the ear to hear words that have been suggested). This means that the very act of telling the jury one of the interpretations to be evaluated might cause them to hear that particular interpretation. Further, having heard one interpretation ‘with their own ears’, they might find it difficult to ‘unhear’ it sufficiently to consider an alternative interpretation with equal objectivity. Indeed, the primed perception might subconsciously affect their opinion of the defendant and the case as a whole, even if it is later shown to be based on an incorrect interpretation of the disputed utterance.

Based on arguments like these from the defence, the disputed utterance in the Bain case was not put to the jury, so we will never know for sure what would have happened. However, the present article offers findings from an experiment designed to track participants’ interpretation of the actual disputed utterance from the case, as it is progressively affected by evidence that might have been produced for the jury. The results show that the effects of priming may be even greater than is usually recognised.

Note: The recording of the crisis call is readily accessible on the internet, including on the first author’s website, where a mini-version of the experiment is available. Readers unfamiliar with the material are urged to try the mini-experiment before reading on so as to experience the audio ‘unprimed’.

Background

Innes (2011) describes the case in considerable detail, so here the background can be summarised briefly. In 1994 a crisis call was made by a young man reporting he had come home to find his entire family (parents, two sisters and brother) were all dead. Three days later he was charged with their murders, and, the next year, convicted and sentenced to 16 years in prison. There followed a lengthy and very high-profile series of appeals against the conviction. In 2007, during preparation for one of these appeals, a detective listening to the crisis call thought he heard Bain utter the words 'I shot the prick' under his breath. Interpreting this as a previously unnoticed confession to murder made the utterance relevant to the ongoing case.

Advice was sought, by both prosecution and defence, from several speech experts (see Innes 2011 for details). These provided somewhat differing opinions as to the content of the disputed section of the call, but experts on both sides were united in emphasising that it could not be transcribed with confidence, and in recommending caution in presenting any transcript to a jury, due to the danger of priming. This caused a great deal of legal argumentation about the admissibility of the disputed utterance as evidence. In the end, the appeal for retrial was allowed, but it was ruled that the crisis call was to be played with the disputed section excised.

The retrial found Bain not guilty, and he was freed after 13 years in prison. The entire case had been of considerable public interest, and as soon as the retrial was over, the media applied for the suppression order on the crisis call to be lifted. This was granted – followed, unsurprisingly, by headlines 'I shot the prick', and feverish public discussion of the case.

Regardless of opinion as to the justice of lifting the suppression order, it has had one good outcome: the audio recording is now in the public domain, allowing us to explore the question of just what a jury might have made of the material had it been presented in court.

The present article reports an experiment in which participants listened to the crisis call and the section of interest, then were invited to imagine they were on the jury as 'evidence' was presented to them in several stages. While this 'evidence' was loosely based on the Bain case, it did not seek to replicate the facts of the case in detail, as the intention was to design a statistically valid experiment with general implications, rather than to examine the Bain case in particular. At each 'Evidence Point', participants were asked to state what they thought was said in the section of interest, and provide their level of confidence. At a certain point, participants were randomly divided into two groups, each receiving parallel but systematically different 'evidence'. This design enabled

measurement of the effect of the ‘evidence’ (as opposed to the recording itself, which was the same for both groups) on participants’ interpretation of the section of interest.

Method

The experiment took the form of an online survey, presented via Qualtrics software (qualtrics.com). Though this method decreases the degree of control over listening conditions, it has the advantage of allowing large numbers of participants and a wide demographic reach.

Participants

Participants were recruited through invitations circulated via university and professional mailing lists, as well as to personal contacts from a range of demographic groups. Overall, 190 people took the survey, randomly assigned to two groups of 96 and 94 respectively. Demographic data are provided in Table 1. There is no difference between the two groups on any of the variables

Table 1: Demographic data for Groups A and B

		Group A (n = 96)	Group B (n = 94)
Age (years; mean, SD)		38.4 (15.0)	43.0 (15.4)
Gender (frequency)	Male	33	29
	Female	63	65
Education – max level achieved (frequency)	High School	16	12
	Diploma	29	20
	Undergraduate	15	20
	Postgraduate	36	42
Studied Phonetics (frequency)	Advanced	20	10
	Introductory	21	24
	None	55	60
English Dialect (frequency)	Australian	42	51
	NZ	5	6
	North American	27	18
	UK	17	13
	Other	1	0
	Non-native speaker	4	6
Experience with a Court Case (frequency)	Yes	27	24
	No	69	70
Know of this case? (frequency)	Yes	12	10
	No	84	84

measured (as indicated by *t*-test and chi-squared tests of independence), and indeed there were very few statistically significant differences (discussed later) in the behaviour of any of the demographic categories within each group.

Stimulus

The stimulus was the actual crisis call from the Bain case. It is about one minute in duration, and of average quality, with the gist of the conversation clearly audible, but some sections of lower intelligibility, mainly due to heavy breathing and evident stress on the part of the caller. A rough transcript is given below, with the location of the section of interest shown. Participants were able to play the whole call, the section of interest, or the section of interest in its immediate context (indicated in the transcript by italics), using a small 'Flash' player embedded in the survey.

Officer Yes, can I help you?
 Caller Help.
 Officer Yeah.
 Caller They're all dead.
 Officer What's the matter?
 Caller They're all dead. I came home and they're all
 dead.
 Officer Whereabouts are you?
 Caller Um, Every Street.
 Officer Every Street?
 Caller 65 Every Street. They're all dead.
 Officer Who's all dead?
 Caller My family. They're all dead. Hurry up.
 Officer *Okay. Every Street. And it runs off Somerville
 Street?*
 Caller Yes.

[LOCATION OF SECTION OF INTEREST]

Officer Telephone number you're calling from?
 Caller Four five four.
 Officer Mm.
 Caller Two five two seven.
 Officer Four five four?
 Caller Two five two seven.
 Officer Two five two seven. And your last name?
 Caller Bain.
 Officer Bain. Okay. We're on our way. Okay Mr. Bain?
 Caller Hurry up.
 Officer Yeah, we'll be there very shortly.

Procedure

Following presentation of an information and consent screen (including a warning participants would be asked to listen to a real crisis call which could potentially be distressing, but giving no further details about the call's content) and a test of the audio playback set-up, participants listened to the crisis call as a whole and answered initial questions to elicit their immediate impression of the call, how clear they found the speech, and how much they trusted the caller.

They then played the section of interest in its context, with a transcript to help them locate it accurately, and were asked what they heard, with a three-way choice: not speech at all, uninterpretable speech or specific words. If they heard specific words, they were asked to state what these were, using asterisks for unclear segments. They were also asked to rate their confidence on a five-point scale, and were given an opportunity to make comments if they wished.

Next they were told that the call was part of a murder case in which the section of interest was relevant, and were invited to imagine they were on the jury as evidence about the case unfolded. They were then given information about the case through several stages, or 'Evidence Points' (see Table 2), during each of which they could listen as often as they wished to the call, the section of interest or the section of interest in its immediate context. After each new piece of evidence, a following screen again asked participants what they had heard, via questions of identical form each time.

At the third Evidence Point, participants were randomly divided into two groups, each receiving systematically different evidence. Specifically, Group A was given a story roughly similar to the actual Bain story, with suspicion cast on the caller, and the section of interest alleged to contain the words 'I shot the prick', while Group B was given a story that cast suspicion on the caller's father, and alleged the section of interest contained the words 'He shot them all'.

At the end, participants from both groups were given the Full Story (a summary loosely based on the Bain case – see Appendix for details) and asked to provide a Final Verdict, by rating on a five-point scale each of a series of suggested interpretations for the section of interest (see Table 2). All suggested interpretations of the section of interest except 'He shot them all' were genuine suggestions made (tentatively) by experts. 'He shot them all' was invented as a foil, intended to be semantically similar to the alleged interpretation, but providing a poorer match to the actual acoustics of the section of interest. At this stage, participants indicated, also on five-point scales, their 'verdict' as to the guilt of each of: the caller, the father, someone else. As well, they were asked how interesting they found the survey. Their mean response to this question was 1.18 (SD = 0.448) on a scale of 1 (very interesting) to 5 (not interesting)

suggesting they had fully engaged with the material. Finally, demographic information was collected.

Table 2 summarises the information provided to each group at each Evidence Point. The exact wording is provided in detail in the Appendix.

Table 2: Outline of survey structure (see Appendix for more detail)

	Group A	Group B
Initial questions	Presentation of the crisis call as a whole, with questions about intelligibility of call and trustworthiness of caller	
Evidence 1/Baseline	Introduction to the section of interest, and indication it is part of a murder case	
Evidence 2	The crisis call is part of a murder case and the section of interest is relevant	
Evidence 3	The section of interest contains an admission of guilt by the caller	The section of interest contains an admission the caller knows his father is guilty
Evidence 4	The section of interest is alleged to contain specific words 'I shot the prick'	The section of interest is alleged to contain specific words 'He shot them all'
Evidence 5	An audio engineer backs up the allegation the words are 'I shot the prick'	An audio engineer backs up the allegation the words are 'He shot them all'
Evidence 6	A phonetician refutes the audio engineer's evidence	A phonetician refutes the audio engineer's evidence
Evidence 7/Final Verdict	<p>Both groups of participants receive the Full Story, including general expert opinion that the section of interest does not contain the words 'I shot the prick' and information that the caller was exonerated of the crime.</p> <p>Both groups rate each of the following section of interest interpretations</p> <ul style="list-style-type: none"> 'I can't touch it' 'I can't breathe' 'I can't help puking' 'I shot the prick' 'He shot them all' Uninterpretable speech Not speech at all Other (please specify) <p>Guilt ratings for all suspects are elicited</p>	

Analysis

Responses regarding participants' interpretation of the section of interest at each of the seven Evidence Points were coded to fit the eight Response Categories shown in Table 3 (with the forced-choice Final Verdict responses adapted to match the same categories as those used for the open-ended responses of earlier Evidence Points). This provided a spreadsheet of seven Evidence Points and eight Response Categories, which were subjected to a range of statistical analyses, testing null hypotheses of *no variation between Groups or Evidence Points*.

Table 3: Response categories and criteria for inclusion

Number	Name	Criteria
1	'I can't breathe'	this specific phrase, regardless of spelling and punctuation
2	'I shot the prick'	the specific phrase '(I) shot/killed the/that prick', not including ellipses like 'I shot the pr**' which would be category 4, or 'I sh** the pr**' which would be category 6
3	'He shot them all'	this specific phrase
4	'shot/killed'	any response (other than those covered by 2 or 3) which included the full words 'shot' or 'kill'
5	'prick'	any response (other than those covered by 2 above) which included the full word 'prick'
6	Other	any response other than those covered above, including any words or part words
7	Uninterpretable speech	used only where this option was selected AND no text at all was provided; any suggestion for any part of the utterance was coded as 6, even if participant had checked 'uninterpretable speech' at the forced choice part of the question
8	Not speech at all	used only if this option was selected AND no text was entered in the text box; any suggestion for any part of the utterance was coded as 6, even if participant had checked 'not speech at all' at the forced choice part of the question

Results

Initial questions

Answers to the initial questions indicated participants understood the speech in the call overall, interpreting it as a distressed call for help from a man who had come home to find his family dead. Fewer than 20% reported a change in their

initial impression when given the information the call was reporting a murder. Interestingly, in the context of recent discussion of ear-witness reliability (Watt 2010), eight participants mentioned in (optional) comments that they had assumed from his voice that the man was older than he really was, and that the 'family' referred to was his wife and children rather than father and siblings.

More than half of all participants (59%) indicated that some parts of the call were unclear. However, though some of these mentioned the caller's heavy breathing as a reason for lack of clarity, most comments nominated the phone number or the beginning of the call as the locations of unclarity. Importantly, no one indicated they had specifically noticed the section of interest. Indeed, responses at Evidence 2 and 3 indicated some participants did not notice the section of interest even after it had been pointed out in its context.

In answer to the question 'How much do you trust the caller?', the modal response was 4 (on a scale where 1 = don't trust him and 5 = trust him completely), with a mean of 3.92 (SD = 1.046). Around 20 participants (10%) indicated distrust, with a rating of 1 or 2. Comments referred to unnatural intonation and breathing patterns suggesting the call may be a hoax or fake of some kind.

Key findings

Change in perception

A snapshot of how participants were affected by the information given at each Evidence Point can be provided by considering responses of both groups at each Evidence Point to the Yes/No question 'Does this evidence change your interpretation of the section of interest?' As shown in Figure 1, Group A was greatly affected by the information provided at Evidence 4 (suggestion of specific words 'I shot the prick'), and at every subsequent Evidence Point significant numbers said they had changed their interpretation ($F(4,380) = 13.8, p < .001$). By contrast, there were no significant differences across Evidence Points for Group B's reported change in interpretation of the section of interest ($F(4,372) = 1.68, p = .15$).

This snapshot, while useful for its purpose, gives a limited view of the data. This is partly because the snapshot gives no indication of the direction of the change, and partly because, as discussed below, the relationship between participants saying they had changed their interpretation and an actual change in response category was far from perfect.

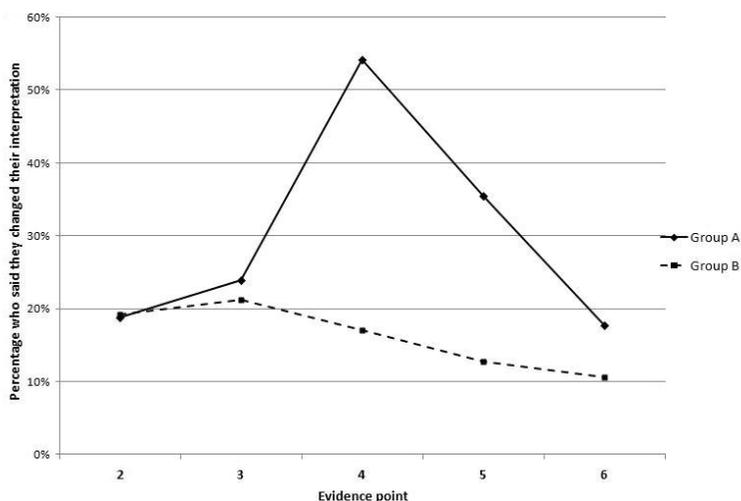


Figure 1. Percentage of participants by Group who said they had changed their interpretation of the section of interest at Evidence points 2 to 6

Direct effect of primes

A better overall view is given by the eight line graphs in Figure 2, showing the number of participants offering interpretations in each of the eight Response Categories at each Evidence Point. These graphs can be analysed statistically by testing the null hypothesis that *The information provided at various Evidence Points has no effect on Response Category* (reflecting a belief that people hear what they hear, unaffected by the context in which it is heard). If the null hypothesis were valid, we should expect to see relatively flat lines across the Evidence Points for all response categories, and both groups.

The greatest deviation from the null hypothesis of *No difference between Evidence Points* can be seen in Group A's response category 2, 'I shot the prick' ($\chi^2(6) = 86.1, p < .05$). This interpretation was heard by only one individual in Group A before Evidence 4. However, when this group was primed with the explicit suggestion 'I shot the prick' at Evidence 4, about one-third of participants reported hearing it. This number remained relatively constant through the next two Evidence Points (audio engineer and phonetics expert evidence), and dropped after the Full Story. Importantly, however, despite the strong indication given by the Full Story that that interpretation was incorrect, 17 participants still preferred 'I shot the prick' at Final Verdict.

Interestingly, Group B also deviated significantly from the null hypothesis in regard to 'I shot the prick' ($\chi^2(6) = 21.5, p < .05$), even though they were not explicitly primed with this phrase. Here, only three individuals heard the

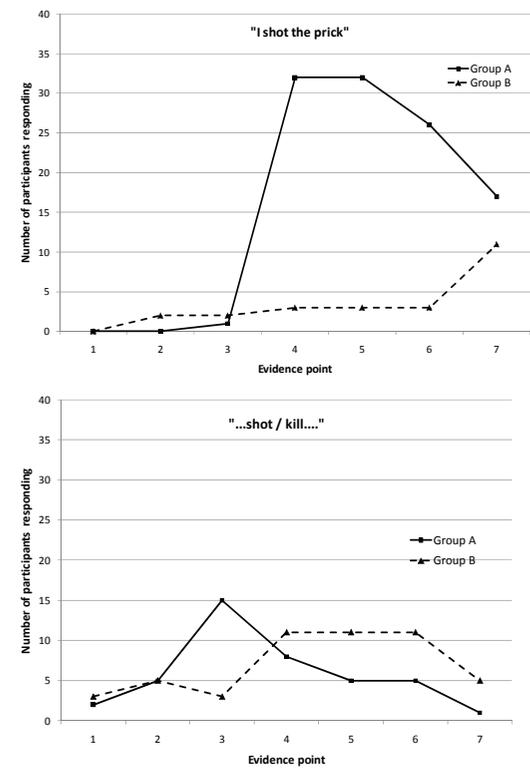
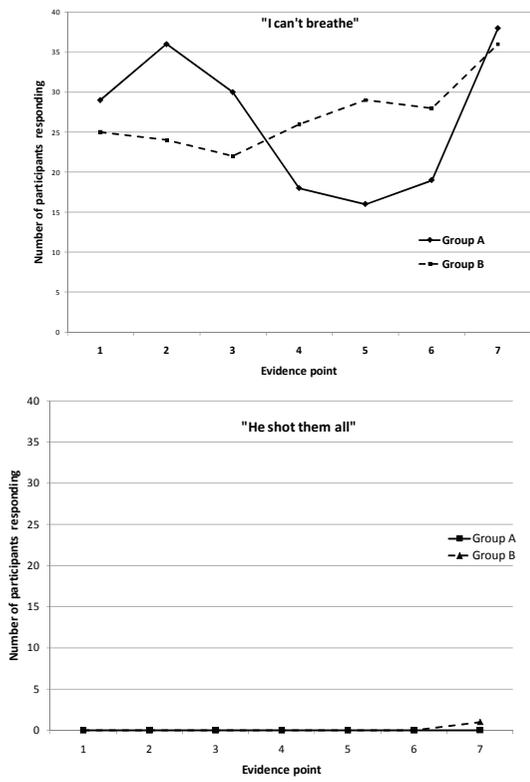


Figure 2 (part). Results by Response Category, showing number of participants in Group A and Group B offering each interpretation of the disputed utterance at each Evidence Point (see text for further detail).

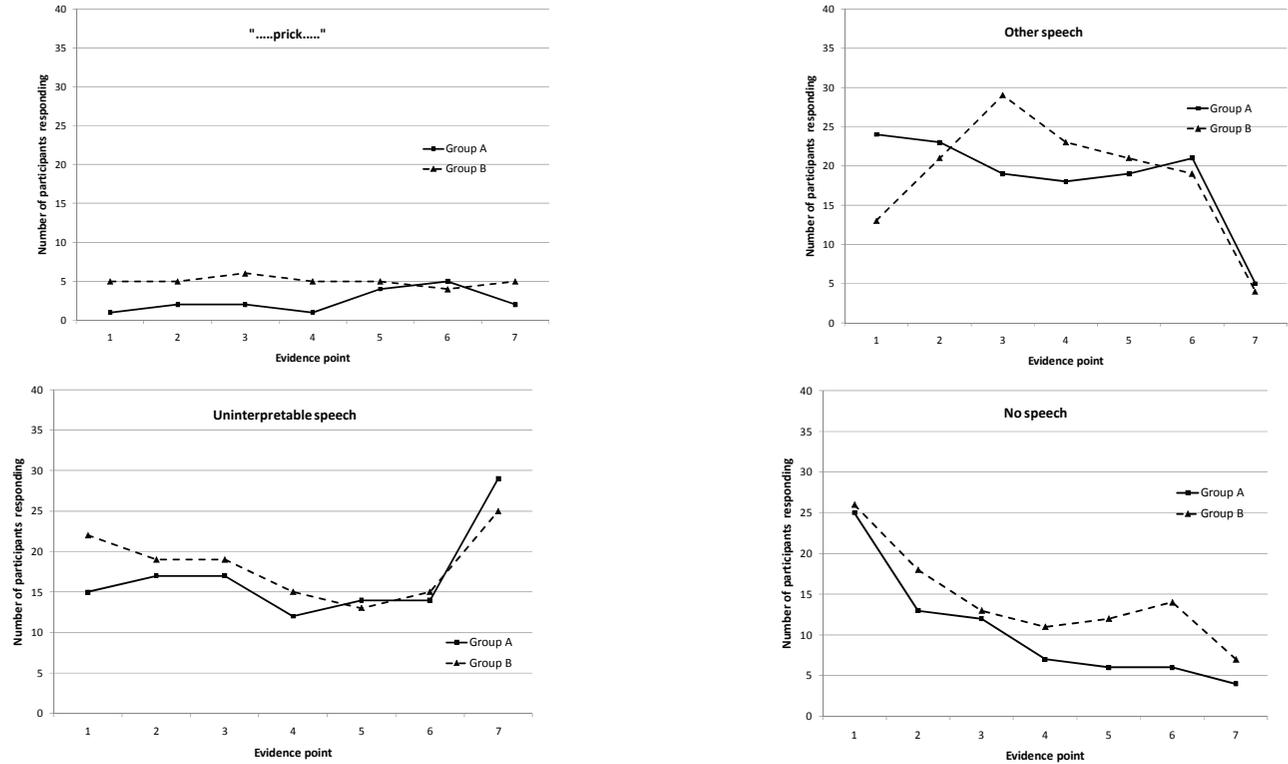


Figure 2 (cont.)

phrase up to Evidence 6. However, at Evidence 7, 11 participants (12%) offered 'I shot the prick' as their Final Verdict on the content of the disputed utterance. This was after the phrase had been mentioned (for the first time for this Group) in the Full Story, and, again, was despite the fact that this mention was in the context of an explanation that, though that interpretation had been alleged, it had been shown to be unreliable, and that the caller had been found not guilty.

These results indicate, first, how easy it is to prime perception of a brief, poor quality recording – even when this is done incidentally, as in Group B's reading of the Full Story; and, second, how persistent a primed perception can be, even in the face of evidence that it is incorrect.

It is interesting, then, to compare findings for 'He shot them all,' with which Group B were primed at Evidence 4, and to which Group A were exposed for the first time in the Full Story. Despite the apparently parallel situation, only one participant in either group ever gave the response 'He shot them all,' and that was only at Final Verdict. Since this person also checked the option 'uninterpretable speech,' which had been their consistent response through all previous Evidence Points, it seems plausible the other may have been a mistake.

These results demonstrate that, while priming is clearly an issue, priming alone is not sufficient to drive perception. The 'all' of 'He shot them all' flatly contradicts the acoustic signal, and was not heard despite the priming. Having said this, it is interesting to note there were several more subtle ways (discussed below) in which the 'He shot them all' prime did actually influence Group B. It is also worth noting, though there is not space here for proper discussion, that even 'I shot the prick' did not fit the (poor quality) acoustic evidence particularly well, as shown by the fact that virtually no one heard this phrase unprimed.

Effect on other response categories

For Group B, most responses other than 'I shot the prick' showed little deviation from the null hypothesis, with 'I can't breathe' starting and remaining as the strongest contender, heard by around 30–40% of participants throughout, outnumbered only briefly, at Evidence 3, by 'Other'.

Group A, on the other hand, showed deviation from the null hypothesis for several response categories in addition to 'I shot the prick'. Most important is 'I can't breathe' ($\chi^2(6) = 18.05, p < .05$). Though starting, like Group B, at high levels, this response dropped markedly during Evidence Points 4–6, presumably due to competition from 'I shot the prick,' then rose sharply at Final Verdict, when some (but not all) participants were released by the Full Story from the grip of the 'I shot the prick' prime.

These key findings indicate clearly that the null hypothesis is not valid, and that priming can indeed affect perception dramatically and lastingly. For more insight into exactly how this effect works, a detailed, step-by-step statistical analysis is provided below. First, however, we note another key finding.

Implications for Final Verdict

As part of the Final Verdict question, participants were asked who they thought was guilty (caller, father or someone else), with each suspect to be rated on a scale where 1 = guilty, 3 = not sure/can't tell, 5 = not guilty. Overall, all three suspects received a similar mean 'guilt rating' around 3.65 (mode = 4), with no statistical difference between groups or suspects.

Importantly, however, those participants (from both groups) who responded 'I shot the prick' at Evidence 7 were significantly more likely than others to find the caller guilty ($t(188) = 5.56 p < .001$).

Also very much worth noting, though not strictly a priming effect, is the strong correlation, across both groups, between participants' initial lack of trust in the caller (when the crisis call had been heard but no information about it had been given), and their final attribution of guilt to him ($r(190) = 0.26 p < .001$).

Detailed results by Evidence Point

This section looks more closely at the distribution of interpretations for each group at each Evidence Point, including the confidence with which they were given, elicited on a scale of 1 (not confident) to 5 (very confident), testing a null hypothesis of *No deviation from confidence level 3*.

Evidence 1 (Baseline)

At Baseline, responses did not differ significantly between the groups ($\chi^2(5) = 7.76, p = .17$). With groups combined, the most frequent responses were 'I can't breathe' and 'Not speech at all' at around 30% each, and 'Other' and 'Uninterpretable speech' at around 20% each. A potential misunderstanding about 'Other' responses should be addressed briefly at this stage. It is by no means the case that this category included a random assortment of unrelated interpretations. Rather, as discussed below, the offerings under this category were highly constrained.

Though categories 'I shot the prick' and 'He shot them all' were not heard at all at Evidence 1, the more general 'shot/killed' and 'prick' were heard by a few. The former was presumably influenced by the content of the call. Regarding the latter, some participants indicated in comments they believed the 'prick'

term was directed at the operator, in response to what participants considered to be his off-hand attitude to the caller's plight.

With respect to confidence, again there was no group effect ($F < 1$) and the combined group showed statistically significant variation from the null hypothesis, in an elevated level of confidence for both 'I can't breathe' (mean = 4.06, $t(53) = 7.77$ $p < .001$) and 'Not speech at all' (mean = 3.39, $t(50) = 2.74$ $p = .008$), and a slightly reduced level of confidence for 'Other' (mean = 2.41, $t(36) = 2.83$ $p = .008$).

Evidence 2 (murder)

At Evidence 2 (information provided to both groups that the crisis call was part of a murder investigation and the section of interest was relevant), again both groups were statistically similar ($\chi^2(6) = 6.67$, $p = .35$), despite the apparently higher number of 'I can't breathe' responses in Group A.

The only significant difference from Evidence 1 was a decrease in numbers of 'Not speech at all' responses ($\chi^2(1) = 4.88$, $p < .05$). However, it is notable that two people (both in Group B) heard 'I shot the prick' at Evidence 2. One of these later indicated prior knowledge of the case. The other was part of a subgroup to be discussed below.

'I can't breathe' and 'Not speech at all' remained the only two responses with confidence levels significantly above 3.00 (respectively, mean = 4.02, $t(59) = 9.23$, $p < .001$, and mean = 3.61, $t(30) = 2.91$, $p = .007$).

Evidence 3 (suspicion)

Evidence 3 was the first time the information provided to the groups was different (for Group A, suspicion fell on the caller, while for Group B, suspicion fell on the father), and the first time there was a significant difference between the groups ($\chi^2(6) = 13.8$, $p = .03$).

In Group A, there was a sharp increase in the number indicating a 'shot/killed' interpretation ($\chi^2(1) = 5.00$, $p < .05$), though 'I can't breathe' remained the most popular response category, and was now the only one with confidence levels significantly above 3.00 ($t(29) = 4.00$, $p < .001$). Other interpretations retained roughly constant numbers for Group A.

By contrast, Group B showed an increase in 'Other' responses ($\chi^2(1) = 4.10$, $p < .05$), with other numbers remaining similar. 'I can't breathe' and 'Not speech at all' continued to be the only two responses with confidence levels significantly above 3.00 (respectively, mean = 4.05, $t(29) = 4.00$, $p < .001$, and mean = 3.92, $t(12) = 2.98$, $p < .01$).

Evidence 4 (specific words)

This was the crux of the experiment, where each group was presented with a specific interpretation of the disputed utterance (Group A: 'I shot the prick'; Group B: 'He shot them all'). The two groups were again very different in their response profiles ($\chi^2(6) = 30.4, p < .001$).

For Group A, 'I shot the prick' became by far the most common response, heard by well over 30% of participants, and confidence in this response was immediately significantly above 3.00 (mean = 3.44, $t(31) = 2.08, p = .046$). 'I can't breathe' remained the only other response significantly above 3.00 (mean = 3.56, $t(17) = 2.40, p = .03$).

In contrast, Group B were unmoved by the 'He shot them all' interpretation suggested to them, with no one hearing the alleged phrase and several using the comment box to explicitly express their disbelief in this option.

There was, however, a significant increase from 3 to 11 'shot/killed' responses ($\chi^2(1) = 4.57, p < .05$). Interestingly, 4 of these 11 used the pronoun 'he' ('He shot' rather than 'I shot'), though this had never been heard before, and was never heard at all by Group A. It seems that even though they rejected the prime, it was affecting their interpretation. Also interestingly, all the new 'shot/killed' responses were 'shot', whereas prior to Evidence 4, the greater proportion of 'shot/killed' interpretations used the word 'killed'. This too suggests a subtle effect of the prime, to be discussed further below.

Evidence 5 (audio engineer)

The audio engineer's confirmation made no change to the response profile of either group. However, it did reduce Group A's confidence in 'I can't breathe', which for the first time was not statistically different from 3.00, and in 'Other', which was now significantly below 3.00 (mean 2.26, $t(18) = 3.07, p = .007$). This left 'I shot the prick', as the only response with confidence significantly above 3.00 (mean 3.71, $t(30) = 3.93, p < .001$).

For Group B, 'I can't breathe' still had confidence significantly above 3.00 (mean 3.97, $t(28) = 5.74, p < .001$), showing they remained unpersuaded by the interpretation suggested to them ('He shot them all'), despite the audio engineer's evidence. In one sense, this is as it should be, since the audio engineer's evidence supported a spurious interpretation of the section of interest. On the other hand, it indicates the extent to which juries might disregard expert opinion in favour of what they hear with their own ears'.

All other responses remained roughly similar, with a strong Group effect still apparent ($\chi^2(6) = 30.5, p < .001$).

Evidence 6 (phonetics expert)

The phonetics expert's evidence (that the disputed utterance was, after all, very unlikely to be 'I shot the prick') also had little effect on the response profile, and the difference between groups remained significant ($\chi^2(6) = 25.6, p < .001$).

For Group A, the only difference from Evidence 5 was a return of 'I can't breathe' confidence levels to significantly greater than 3.00 (mean = 3.68, $t(18) = 3.37, p = .003$). Confidence in 'I shot the prick' remained high, at 3.69 ($t(25) = 3.80, p < .001$).

For Group B, confidence remained significantly above 3.00 for 'I can't breathe' (mean = 4.18, $t(27) = 9.31, p < .001$), and there was a significant increase in confidence for 'Not speech at all' (mean = 4.14; $t(13) = 4.51, p < .001$), as well as, though only a few gave this response, for 'prick' (mean = 4.25; $t(3) = 5.00, p = .015$) (NB not 'I shot the prick' but Response Category 5).

Evidence 7 (Final Verdict)

After hearing the Full Story, the groups returned to being statistically indistinguishable in their response profiles ($\chi^2(7) = 7.50, p = .38$). The combined group had negligible numbers of 'Not speech at all', 'prick', 'shot/killed' and 'Other' responses, and strongly favoured 'I can't breathe', with around 40% overall giving this response. Another 30% responded with 'Uninterpretable speech', the response promoted in the Full Story. However, the third most frequent response at Final Verdict was 'I shot the prick' (14%). Importantly, as discussed above, this response was given in similar numbers across both Groups. No confidence data was collected at this point.

Reliability and Consistency

'Reliability' here refers to the relationship between a participant's answer to the Yes/No question 'Does this evidence change your interpretation of the Section of Interest?' and that participant's actual change in response category. If these corresponded, the participant received a Reliability score of 1 for that Evidence Point; if not, the score was 0. Figure 3 gives the percentage of reliable responses for both groups at each Evidence Point 2–6 (no change was possible at Evidence 1, and no 'change' question was asked at Evidence 7).

There was a significant Group by Evidence Point interaction ($F(4,752) = 3.86, p = .004$). With the suggestion of 'I shot the prick' at Evidence 4, Group A showed a significant drop in reliability (from 78% to 66%), which continued through Evidence 5, and recovered to 81% at Evidence 6 ($F(1,95) = 12.1, p < .001$). Group B, by contrast, showed a linear trend towards improved reliability with each successive Evidence Point ($F(1,93) = 3.57, p = .06$). This may be

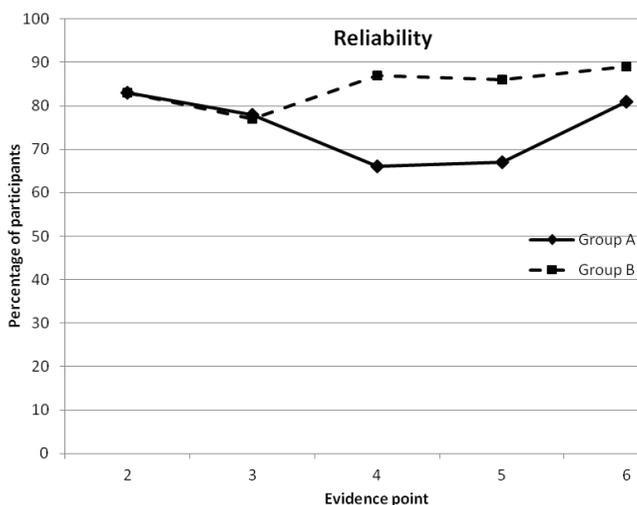


Figure 3: Reliability (see text for definition) of participants in Groups A and B over Evidence Points 2 to 6

taken to indicate a degree of confusion in the minds of Group A participants, caused by the suggestion of ‘I shot the prick’.

‘Consistency’ here refers to a participant’s propensity to give the same interpretation of the section of interest from one Evidence Point to the next, while ‘inconsistency’ is the converse – participants’ propensity to change their response at each Evidence Point. Figure 4 shows the percentage of participants in each group who gave inconsistent responses at each Evidence Point.

Figure 4 shows an underlying linear trend of increasing consistency (decreasing inconsistency) at each successive Evidence Point, interrupted for Group A at Evidence 4, where they became significantly inconsistent ($F(1,188) = 5.54, p = .02$).

The reason ‘inconsistency’ was plotted rather than ‘consistency’ is that it makes for a direct comparison with Figure 1, which plots the number of participants reporting a change in response at each Evidence Point (in answer to the Yes/No question ‘Does this evidence change your interpretation of what is said in the section of interest?’). It is interesting to note how the plots diverge for Group A, where significantly more participants than in Group B claimed to have changed when in fact they had not ($\chi^2(1) = 7.14, p < .01$), as opposed to roughly similar numbers in both groups claiming not to have changed when in fact they had. While the reason for this divergence is not immediately apparent, beyond it being a further indication of confusion in the minds of Group A, it does suggest a need for caution in interpreting psycholinguistic experiments based on participants’ reported changes as opposed to their actual changes.

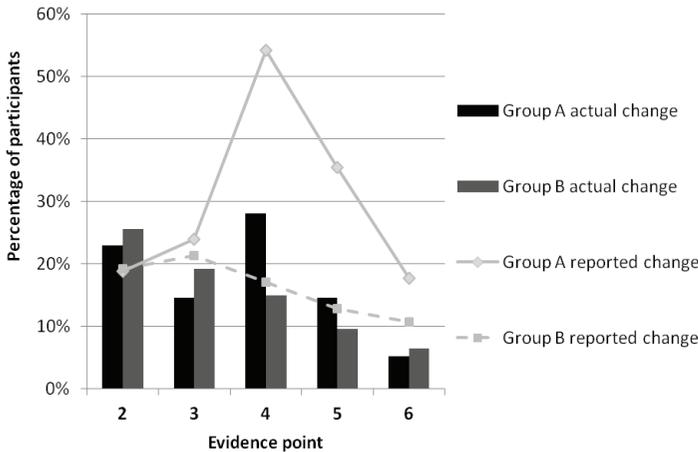


Figure 4: Bar graph showing percentage of participants at each Evidence Point giving a response in a different category from the prior Evidence Point, for Groups A and B, over Evidence Points 2 to 6, superimposed upon line graph showing participants' reported change in perception at each Evidence Point (cf. Figure 1, and see text for discussion)

Demographic analyses

Correlations of demographic categories with all Response Categories at each Evidence Point, and with Reliability and Consistency (see above), were calculated. In general, all demographic groups behaved similarly on all measures, with no overall effects of gender, age, educational background or dialect (allowing Groups A and B to be treated statistically as both similar and homogeneous in the analyses discussed above). However, there are a few demographic effects worth mentioning. First, however, it is maybe useful to emphasise one category that, perhaps surprisingly, had no effect.

Prior knowledge of the Bain case

The 22 participants stating they had prior knowledge of the case included all 11 speakers of New Zealand English, plus two speakers of North American dialects who indicated they lived in New Zealand. As far as can be judged from their comments, around 15 of the 22 knew of the case through the media, the others mainly through academic study. Overall, this group showed no statistically significant differences in their Response Categories at any Evidence Point, in their confidence at any Evidence Point, or in their Reliability or Consistency throughout the experiment, from those without prior knowledge.

Age and gender

Gender had some effect at Final Verdict ($\chi^2(4) = 9.15, p = .057$), with females more likely to respond 'I can't breathe' and males more likely to respond 'Not speech at all'. 'Not speech at all' is consistently associated with older participants ($F(6,183) = 2.48, p = .025$). (Note that the latter two effects are independent, as the mean difference in age between males and females was only 1.08 years ($t < 1$)). A possible reason for the effect of age is reduced auditory acuity, but this is by no means the only explanation, and it is relevant to note that in general, in both groups and across all demographics, 'Not speech at all' responses declined gradually throughout the experiment ($\chi^2(6) = 41.5, p < .001$). It seems that the more participants listened to the auditory material, the more they were inclined, rightly or wrongly, to hear it as speech.

Phonetics background

Phonetics background was elicited via the Yes/No question 'Have you ever studied phonetics?', and categorised into three levels (as best as possible based on comments in 'Please give brief details'). 'High' level included those who had studied phonetics to postgraduate level and indicated some use of their phonetics skills in their professional lives (not necessarily in the forensic context). 'Mid' level included those who had studied some phonetics, but not beyond an undergraduate degree. 'Low' level was those with no background at all in phonetics. Overall, there was not a clear effect of phonetics background on response category at Final Verdict ($\chi^2(14) = 22.38, p = .07$). However, it is notable that High-level phoneticians gave zero 'I shot the prick' responses throughout all Evidence Points, and also gave a higher percentage of 'Uninterpretable speech' responses at Final Verdict (36% for High level, 23% each for Mid and Low level).

Experience in a court case (police)

Answers to the Yes/No question 'Have you ever been involved in a court case?' showed no general correlation with responses at any Evidence Point. However, under 'please give brief details', six respondents (4 in Group A; 2 in Group B) indicated they were police officers (all speaking Australian English, and none stating any prior knowledge of the Bain case). Of these, three (50%) gave a Final Verdict interpretation of the section of interest as 'I shot the prick'. This is significantly higher than the 14% of participants overall responding 'I shot the prick' at Final Verdict ($\chi^2(1) = 6.13, p = .013$). Two of the three police who heard 'I shot the prick' at Final Verdict (discussed further in the next section) also found the caller 'definitely guilty' (i.e. after being told he had been exonerated).

This contributed to an overall higher propensity among police than participants overall to find the caller guilty (average 'caller guilt' score was 2.83 for police, compared to 3.65 for all participants). The numbers of police are too small to enable robust conclusions to be drawn, but these results suggest there may be value in a follow-up study of police understanding of issues in forensic transcription.

Hearing 'I shot the prick' before priming

Only four participants in the entire experiment heard the phrase 'I shot the prick' without it having been explicitly suggested to them (i.e. before Evidence 4 for Group A or Evidence 7 for Group B). Of these, one had prior knowledge of this interpretation via an undergraduate class in forensic science, while two (50%) were police officers. Though the numbers are small, the proportion of police in this group is significantly higher than the 3% of participants overall who revealed themselves to be police ($\chi^2(1) = 29.3, p < .001$). The only other demographic characteristic on which this group differed notably from the population overall was that 100% of the four indicated their highest level of education was diploma/certificate (compared to around 40% of participants overall – see Table 1).

It is also notable that this group of four were very much more likely to rate the caller as 'definitely guilty' at Final Verdict, with a mean 'caller guilt' score of 1.5 (mode = 1) as opposed to the overall mean of 3.65 (mode = 4).

For completeness, it is worth mentioning that the other member of this group of four was in Group B, and gave responses as follows. Evidence Point 1: 'I hit/hate **'; Evidence Points 2–3: 'I hate that prick'; Evidence Points 4–7: 'I shot that prick' – with a confidence level of 2 throughout (where 1 is 'not confident').

A more subtle influence of priming

It has already been noted (under 'Evidence 4' above) that the proportion of 'shot/killed' responses (Response Category 4) indicating 'shot' as opposed to 'killed' increased after the 'shot' prime was presented, and that the pronoun 'he' (as opposed to 'I') was only ever heard by Group B participants after Evidence 4 (when they were primed with 'He shot them all'). These observations suggest that even those who do not directly accept a prime may nevertheless be influenced by it. In this section, we consider this suggestion further by looking in a little more detail at the 'Other' response category. It will be recalled that this category was used for interpretations of the section of interest which considered it to be interpretable speech, but *not* 'I can't breathe' (Category 1), *not* 'I shot

the prick' (Category 2), and *not* any other phrase with 'shot' or 'prick' (Categories 3, 4 or 5). Here, we sub-categorise 'Other' responses according to whether they contained the phoneme /ʃ/, as in 'shot', 'should', etc., or /k/, as in 'I can't', 'I came', etc.

Of the total of 257 'Other' responses given by all participants at all Evidence Points, over 40% gave the /k/ phoneme, with over 70% of these being 'I can't' and a further 17% 'I can't believe'. By contrast, only around 27% of these 257 'Other' responses involved the phoneme /ʃ/, with around 27% of these attributable to the specific phrase, 'I should/shouldn't'.

More telling is a comparison of the pattern of these subcategories of 'Other' responses at early (less primed) versus later (more primed) Evidence Points. Figure 5 compares, within the 'Other' category, the proportion, at each Evidence Point and for each Group, of all /ʃ/ responses (including 'I should/shouldn't') to the proportion of 'I can't' responses (including 'I can't believe', but not including other /k/ responses). Group A shows a strong effect, with a clear crossover pattern centring around Evidence 4, when the prime was given. The proportion of 'I can't' responses decreases from Evidence 3 to 4 ($\chi^2(1) = 14.1, p < .001$) while the proportion of /ʃ/ responses increases from Evidence 2 to 5 ($\chi^2(3) = 16.3, p < .001$). A similar, though weaker, effect is seen in Group B, with a more gradual rise in /ʃ/ responses from Evidence 1 to 4 ($\chi^2(3) = 11.47, p < .01$), and the fall in 'I can't' from Evidence 3 to 5 not quite reaching significance ($\chi^2(2) = 5.03, p = .08$).

Discussion

Around 30% of participants in the group that received the incriminating prime 'heard' it after it was suggested to them, though virtually no-one in either group had heard it before it was suggested. Around half of these 30% continued to 'hear' the suggested interpretation of the section of interest at the end of the experiment, after being advised that experts were agreed that interpretation was not valid. Perhaps more surprisingly, in the group that did not receive the incriminating prime, 12% heard the suggested phrase at the end of the experiment even though they had only been exposed to it for the first time in the 'Full Story' – i.e. in the context of being told that experts had rejected the suggestion.

Further, there was a strong correlation, across both groups, between participants hearing the incriminating phrase in the section of interest and finding the speaker 'guilty' – again, despite having been told in the 'Full Story' that the caller had been found not guilty and released from prison. The evidence also shows a range of ways in which the prime offered to each group subtly affected even those participants who did not accept the suggested phrase as an accurate

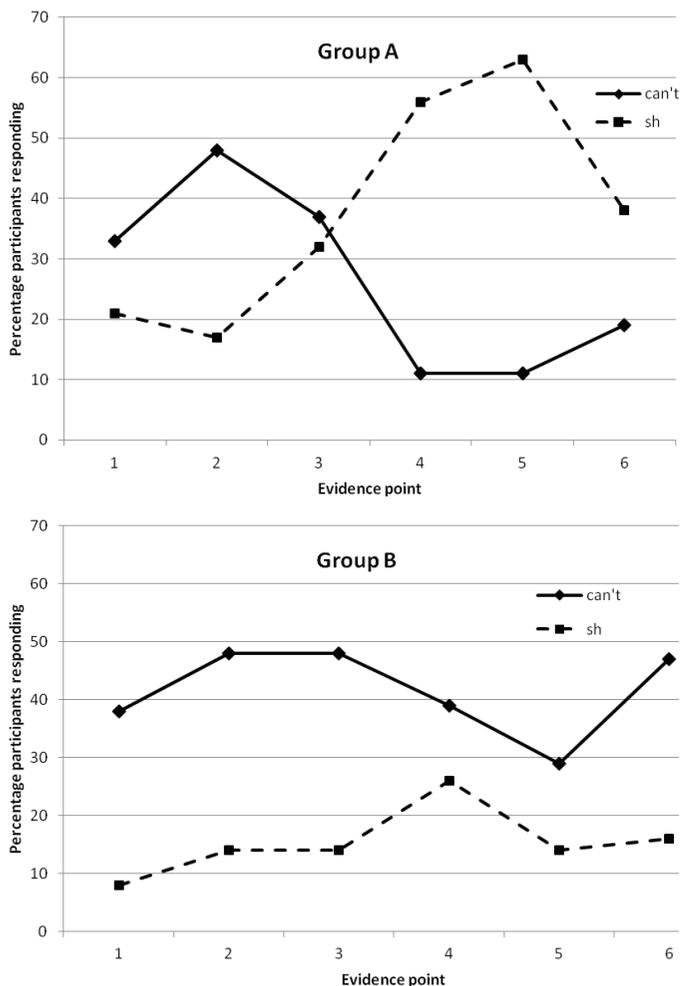


Figure 5: The proportion of responses including 'I can't' versus responses containing the phoneme /ʃ/, as a percentage of all 'Other' responses, across Evidence Points 1–6, for Groups A and B

interpretation. Finally, there was a strong correlation between participants stating they did not trust the caller at the very beginning of the experiment and finding him guilty at the end.

It is hoped that these results give an impression of the very real dangers involved in leaving evaluation of alternative transcripts of a poor quality recording to a jury, and demonstrate why phoneticians believe transcripts of poor quality recordings should be evaluated by genuine experts, who can produce objective evidence to support real expert opinions.

Conclusion

The experiment reported here provides strong support for the conclusion finally reached by the New Zealand Supreme Court in 2009, that the dispute over the section of interest should not be decided by a jury. This is in addition to the anecdotal support already given (Innes 2011) by the fact that release of the recording to the media created a frenzy of uninformed public vitriol after the trial had ended.

What is urgently needed now is widespread adoption of a standard practice whereby the legal system can reach similarly responsible conclusions in similar cases – with far lower costs in terms of time, money and personal suffering than in *Bain* and other cases (Fraser 2010). Hopefully this can be achieved in the near future through close cooperation between the legal system and experts in cognitive aspects of phonetic science (Gray 2010).

Acknowledgments

Thanks to the many people who assisted by piloting, publicising and participating in this experiment. Thanks also to Don Fraser for valuable statistical insight.

About the authors

Helen Fraser studied linguistics and phonetics at Macquarie University, Sydney, and the University of Edinburgh, Scotland, and taught at the University of New England, Australia, from 1990 to 2008. She is now an independent researcher, focussing on cognitive aspects of phonetics and phonology, and covering both theoretical and applied topics, especially second language pronunciation and forensic transcription. Please see helenfraser.com.au for further details.

Bruce Stevenson is a Lecturer in Cognitive Psychology at the University of New England, Australia, with research interests in language processing and working memory.

Tony Marks is a Senior Lecturer in Psychology and First-year Coordinator at the University of New England, Australia. His research background includes the misinformation effect in human memory and the impact of dual-process theories of cognition in applied social psychological contexts.

Note

1. 'Poor quality' is used here in a non-technical sense, to designate recordings in which the speech may be found 'unclear', whether this is due to recording factors, speech factors, or some other cause (cf. Fraser 2010).

References

- Byrd, D. and Mintz, T. H. (2010) *Discovering Speech, Words and Mind*. Chichester and Malden, MA: Wiley-Blackwell Publishing.
- Coulthard, R. M. and Johnson, A. (2007) *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.
- Fraser, H. 2003. Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *The International Journal of Speech, Language and the Law* 10: 203–226.
- Fraser, H. 2010. Transcripts in the legal system. In I. Freckelton and H. Selby (eds.) *Expert Evidence* chapter 100. Sydney: Thomson Reuters.
- Gray, P. R. A. (2010) The expert witness problem. *The International Journal of Speech, Language and the Law* 17: 201–209.
- Innes, B. (2011) R v David Bain: a unique case in New Zealand legal and linguistic history. *The International Journal of Speech, Language and the Law* 18: 145–155
- Watt, D. (2010) The identification of the individual through speech. In C. Llamas and D. Watt (eds.) *Language and Identities* 76–85. Edinburgh: Edinburgh University Press.

Appendix

Full text of experiment as presented to participants.

*At each Evidence Point 1–6 (below), the following question page was presented.

(Evidence 1) What did you hear in the section of interest?

(from Evidence 2) Does the evidence you have just been given change your previous interpretation of the section of interest? YES/NO

What do you now hear in the section of interest? This may be the same as last time you listened, or different.

(All) Click the choice below that corresponds with your impression.

Not speech at all; Speech but not clear; I heard specific words

If you heard speech, please write the word(s) in the 'What did you hear?' box below. If the speech was not clear, please write your best guess in that box, and add your comments in the Comment box at the bottom of this page.

What did you hear? (you might like to use *** to indicate parts that are unclear)
(It helps us analyse results if this box contains only the words you think you heard, so please reserve comments – e.g., 'I think it was XXX, but maybe I heard YYY' - for the Comments box)

How confident do you feel about your answer? Click the choice that corresponds with your confidence. (5-point scale, not confident to very confident)

Comments

INTERPRETATION OF A CRISIS CALL

Thankyou for agreeing to participate in this survey.

On the next screen you will hear a 000 crisis call. You will then be asked several questions, with the opportunity to listen to the call as often as you need to. In fact you will be hearing the same audio recording a number of times with different questions each time.

Please note that this is a genuine crisis call, and could potentially be distressing to hear. You may stop participating at any time, and if you feel upset in any way, we encourage you to contact a local counselling service.

Before we start, please click the Audio Test button (below) to be sure your audio works. Note that you need to remain connected to the internet, and have your computer sound turned up to a comfortable level.

Was there any part of the call where the words were not clear enough for you to understand what was being said?

Important note: Here and throughout, it is important to answer the questions as best you can from your memory of the call. You will always get another chance to hear the call and revise your answers if you wish to. If you find no option that precisely fits your answer, please choose the best available option and add a comment in the comment box provided.

If not all clear, in the box below please indicate (as best you can) which part(s) you found unclear.

What is your immediate impression of what is going on in this crisis call? Please provide a brief statement in the comment box below.

How do you feel about the caller? Indicate how much you trust him by clicking one of the choices below. (5 point scale)

One section of the call is of particular interest. The section is in the middle of the call. Click the button below to hear the relevant part of the crisis call. This short transcript should help you locate the section of interest:

Officer: "Somerville St?"

Caller: "yes"

SECTION OF INTEREST

Officer: "telephone number you're calling from".

You can listen as often as you wish. We will then ask you some questions as to what you heard.

EVIDENCE POINT 1*

This crisis call was part of a murder case.

We want you to imagine you are on the jury as the evidence in the murder trial unfolds. We will ask you how each piece of evidence affects your interpretation of the crisis call. There is no way to be 'right' or 'wrong' in your answers, so please just give your honest views based on the evidence you have at the time, without trying to second-guess what the questions are about. At the end we will provide a detailed account of the story. Here is the first piece of evidence.

The caller is a young man who went out early one morning on his paper round. He made the call shortly after he came home, to report that his mother, father, brother and two sisters were all shot dead in the house.

Does this evidence change your first impression of what is going on in the call as a whole? YES/NO

Comments

It turns out the section of interest is crucial to the case.

The evidence you have just been given may have changed your earlier interpretation of what you heard in the section of interest.

You may like to listen to the section of interest again before answering some questions on the next page.

As a reminder, the section is in the middle of the call:

Officer: "Somerville St?"
 Caller: "yes"
 SECTION OF INTEREST
 Officer: "telephone number you're calling from".

You can listen as often as you wish. We will then ask you some questions as to what you heard.

EVIDENCE POINT 2*	
Group A	Group B
<p>It turned out the young man's family had been a troubled one, and the father may have been abusing one or more of the siblings.</p> <p>Suspicion fell on the young man himself.</p> <p>It was alleged he came back from his paper round early, shot his family members, and then called 000 pretending he had just returned and found the bodies.</p> <p>In general, how plausible do you find this allegation?</p>	<p>It turned out the young man's family had been a troubled one, and the father may have been abusing one or more of the siblings.</p> <p>Suspicion fell on the father.</p> <p>It was alleged he waited for the young man to go out on his paper round, then shot the other family members, and finally took his own life.</p> <p>In general, how plausible do you find this allegation?</p>

<p>In fact, it was alleged the section of interest is actually an admission of guilt by the young man.</p>	<p>In fact, it was alleged the section of interest is an admission the young man actually saw the murders being committed.</p>
--	--

EVIDENCE POINT 3*	
<p>Specifically, the allegation is that the SECTION OF INTEREST contains the words 'I shot the prick'.</p>	<p>Specifically, the allegation is that the SECTION OF INTEREST contains the words 'He shot them all'.</p>

EVIDENCE POINT 4*	
<p>The next evidence is given by an expert witness. In summary, he says: 'I am an audio engineer. I have enhanced the audio and undertaken extensive technical analysis. I support the view that the caller in the SECTION OF INTEREST says "I shot the prick".'</p>	<p>The next evidence is given by an expert witness. In summary, he says: 'I am an audio engineer. I have enhanced the audio and undertaken extensive technical analysis. I support the view that the caller in the SECTION OF INTEREST says "He shot them all".'</p>

EVIDENCE POINT 5*	
<p>It now turns out there is conflicting evidence. Another expert is called and testifies as follows: 'I am an expert in phonetics, the science of speech. Audio engineers are skilled at analysing sound, but they do not have sufficient expertise about speech to determine what was said in the section of interest. The section of interest is so short, so unclear, and so isolated from the rest of the conversation that it is impossible to be certain exactly what the caller said.'</p>	<p>It now turns out there is conflicting evidence. Another expert is called and testifies as follows: 'I am an expert in phonetics, the science of speech. Audio engineers are skilled at analysing sound, but they do not have sufficient expertise about speech to determine what was said in the section of interest. The section of interest is so short, so unclear, and so isolated from the rest of the conversation that it is impossible to be certain exactly what the caller said.'</p>

EVIDENCE POINT 6*

THE FULL STORY (NB this was written before reading Innes 2011, but in any case was not intended as an accurate account of the Bain case but as part of an experiment)

The murders took place in 1994, in New Zealand. The crisis call was made by a young man after he returned from his paper round, as described. It emerged the young man's family had been a troubled one, and the father may have been abusing one or more of the siblings.

Suspicion fell on the young man himself. The prosecution alleged he had come back from his paper round early, shot his family members, and then called 000 pretending he had just returned and found the bodies. He was tried and convicted of the crime, and imprisoned.

However, there were many who believed the father was guilty, suggesting he had waited till the young man went out, then shot the other family members and finally himself. There were a number of attempts to exonerate the young man, and he was finally acquitted and released after more than a decade in prison.

It seems now to be accepted that the father killed the rest of the family and then himself while the young man was out, though this has not been proven and there are still many people who believe the young man was guilty. If you are interested, you can sign up at the end of the survey to receive further information.

At one stage of the investigation, a detective claimed he heard the words 'I shot the prick' in the emergency call, and interpreted it as a confession. He took the recording to an audio engineer, who analysed the section of interest and supported the detective's interpretation. However, though audio engineers are experts in audio, they do not have expertise in the linguistic aspects of speech. For that an expert in phonetics (the science of speech) is needed. In this case several phonetics experts were consulted. A range of suggestions were made as to what was actually said, none with much confidence, and all with recommendations that the material not be played in court. Eventually clear acoustic evidence was presented showing that whatever had been said, it was very unlikely to be 'I shot the prick'.

Fortunately in this case the judges decided not to put the call to the court and it played no direct role in the case. However, it gives an interesting opportunity for investigation of auditory perception and the role of different kinds of expertise in the legal system. In this study we are investigating the effects of context on people's interpretation of the section of interest. Half the participants are given the allegation the young man is guilty and the section of interest contains the words 'I shot the prick', while the other half are given the allegation the father is guilty and the section of interest contains the words 'He shot them all'.

<p>FINAL VERDICT (EVIDENCE POINT 7)</p>
<p>You have now listened to the call a number of times and have all the relevant information. It is time for you to give your final verdict. Below are listed all the suggestions that have been made as to what was said. You may agree with one of them or you may have your own interpretation. Listen again as much as you wish.</p>
<p>What do you think was said in the section of interest? Please rate each of the following options (5-point scale ranging from DEFINITELY NOT through NOT SURE/CANT TELL to DEFINITELY)</p>
<p>Who do you believe is guilty? (5-point scale) Caller Father Someone else</p> <p>Any final comments on the case?</p>
<p>How interesting did you find this survey?</p> <p>Any other comments on this survey?</p>
<p>Demographic data collection, including:</p> <ul style="list-style-type: none"> age gender education background dialect knowledge of phonetics prior knowledge of the Bain case experience in a legal case