Applications of the Likelihood Ratio Framework in Forensic Speech Science Cases Involving Disputed Utterances, Tampering and Voice Lineups

Phil Rose

Emeritus Faculty, Australian National University

philjohn.rose@gmail.com philjohnrose.net

Abstract

Examples are given of the use of the likelihood ratio framework in real world case-work involving disputed utterances and tampering, and its theoretical application in voice lineups. Some of the problems in application are pointed out, especially with respect to estimation of priors.

Index Terms: likelihood ratio framework, Bayes' Theorem, disputed utterances, prior odds, tampering, voice lineups.

1. Introduction

About a quarter of a century has now passed since the first proposals to use likelihood ratios (LRs) in forensic speaker recognition. In that time it has become, after DNA, part of the new paradigm for the evaluation of forensic evidence [1] and now constitutes best practice for the *European Network of Forensic Science Institutes* [2: 7 *et pass*.]. However, forensic voice comparison, as it is now appropriately called, is not the only area of forensic speech science where evidence needs to be evaluated logically. The aim of this paper is to describe the use of LRs in three other areas of FSS: disputed utterance, tampering and voice-lineups – the first two from real case-work – and point out some of the problems involved.

The essence of the LR approach is to estimate the strength of evidence in favour of one hypothesis over another. For this one needs, of course, the two competing hypotheses and the evidence. For example, in forensic voice comparison the two hypotheses are often that the incriminating speech was said by the suspect; and that it was said by someone else. The evidence is usually some kind of quantification of the speech acoustics. The strength of the evidence is then straightforwardly the ratio of the probabilities of the evidence E under the competing hypotheses H_1, H_2: $p(E|H_1) / p(E|H_2)$. The LR is intended to indicate how much the evidence should make you rationally alter your belief in the hypotheses. Whatever your rational belief in a hypothesis before the evidence is adduced, the LR tells you how much you should strengthen or weaken that belief. The introduction of the notion of belief situates the LR in its proper place within Bayes' Theorem, where, in a mathematically simple and yet insightfully profound way, it provides the link between prior and posterior beliefs [3: 243, 4: 14-15].

2. Disputed Utterances

Did Abba really sing *see that girl, watch her scream, kicking the dancing queen*? Twenty-two percent of listeners apparently think so [5]. An inexhaustible list of hilarious mondegreens in many languages shows that people can mishear things when the sound is indistinct. Since recordings in real-world cases can get much less clear than pop-songs, it is sometimes disputed not *who* said something, but *what* was said. One of the best-known examples of disputed utterance comes from the New Zealand murder case *R v. Bain* where a prosecution witness claimed to hear Bain say the obviously incriminating *I shot the prick* in an emergency telephone call [6]. Alternative claims from defence experts were that Bain said *I can't breathe*, or *I can't help puking*, or *I can't touch it*, or that he was gasping for breath, or that it was not speech. This variety is what happens when you ask different unprimed people what they hear in a very indistinct recording. If you prime them, of course, you can, within reason, get them to hear what you want [7].

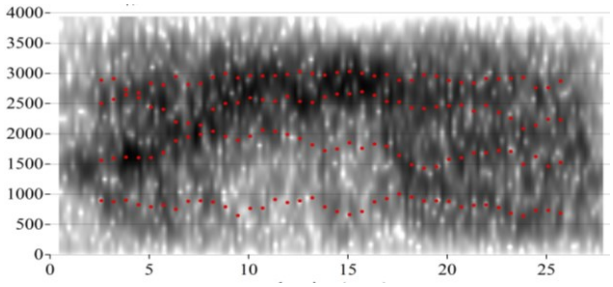

Figure 1: *Spectrogram with superimposed formants of* I shot/I can't *in R v. Bain. X-axis = duration (csec.), y-axis = spectral frequency (Hz)*

Perhaps the major benefit of a LR approach to disputed utterances is that it forces you to be precise about what is the evidence; and what the hypothesis. The temptation, of course, is to take what someone reports they heard as the evidence and then try to evaluate for example *what is the probability that x heard 'I shot the prick', given that that was indeed what was said*. But this is almost impossible to quantify. It is far better to take the observed acoustics as the evidence – that is, after all, the result of what was actually said – and the hypotheses as what someone reports they heard [8]. So the question becomes, for example: what is the probability of getting the observed acoustics, assuming that what was said was *I shot the prick*; and what is the probability of getting the observed acoustics, assuming that what was said was *I can't breathe*. Using the acoustics of the call as evidence, *not* what someone heard, has the major advantage that it is the same evidence for all hypotheses. The second major advantage is that the acoustics are quantifiable. In Bain it boiled down to a question of two hypotheses: the acoustics realize the /k/ in *can't* or the /ʃ/ in *shot*. Figure 1 shows the acoustics corresponding to *I shot ~ I can't*. The relevant portion can be seen to extend from about csec. 10 to csec. 15 and consists of narrowband noise between about 2 and 3 kHz for which *Praat* has extracted two resonances. It was pointed out in [8] that this portion neither sounds like, nor has the acoustic properties of the voiceless

postaveolar fricative [ʃ̟] that would be expected from /ʃ/. Rather it sounds like, and has the acoustic properties of a voiceless palatal fricative [ç]. Bain can be heard elsewhere in the call to have a fronted velar [k̟] for his /k/, so one plausible explanation of the palatal fricative is that he was saying /k/, but failed to make the appropriate closure under the stress of coming across his all family dead. In other words, you would be more likely to get the observed acoustics if Bain had not said *shot*, but a word beginning with /k/. If Bain had not said 'I shot the prick' that does not of course imply his innocence, but he was eventually acquitted on other grounds. A subsequent experiment [9] demonstrated how to estimate the strength of evidence from the acoustics of [ʃ] and [ç], and that Bain's acoustics were vastly less likely if he had said [ʃ̟].

In order to actually estimate the posterior probability that Bain said *I can't* rather than *I shot* we would still need the priors with which to combine the LR. Given the enormous amount of available data that allow our mobile phones to predict rather well our next texted word, it would probably be possible to estimate the probability of *can't* occurring after *I* relative to the probability of *shot* in emergency phone calls.

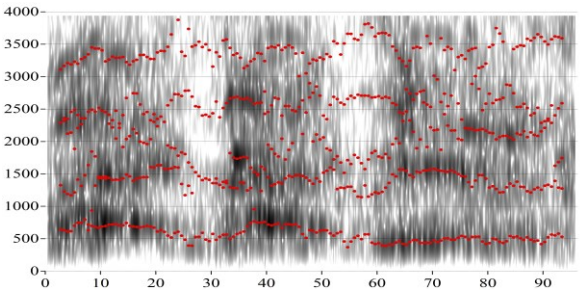

Figure 2: *Spectrogram with superimposed formants of chant 4.*

Normally a disputed utterance involves the speech of a single person. A case in October 2023, however, involved a whole crowd chanting during a pro-Palestinian protest in front of the Sydney Opera House. An accompanying video included captions stating *gas the Jews*! and went viral. The captioned video sparked understandable outrage and predictable schism and the media approached three forensic experts (I was one) [10]. It was mooted that the video chant might actually not be *gas the Jews!* but *where's the Jews?* In NSW, incitement to violence is an offence, and the police hired another forensic expert who concluded that the chant was 'with overwhelming certainty' *where's the Jews?* [11].

The fact that in this case we are talking about not one, but many people introduces several complications, e.g. that it would not be correct to persist with an all-encompassing hypothesis as *the crowd were chanting x*. The recording disseminated in the video would have picked up only a portion of the crowd, and – although that is not in the nature of a chant – a small number of them might have been chanting something different which by dint of their number was not picked-up (several individuals present reported hearing *gas the Jews*). There would have been other parts of the crowd that were not recorded simply because they were further away and out of range. It would therefore be more precise to formulate the competing hypotheses thus: *what was recorded on the disseminated video was* (1) *gas the Jews or* (2) *where's the Jews*.

In the captioned video played on her Sky News show by Sharri Markson as "proof" that *gas the Jews* was chanted [12], nine clear occurrences of the alleged chant can be made out. These are preceded by four full repeats of the chant *Allahu akbar* itself preceded by *takbiir!* (make great! – a conventional injunction to extoll Allah – by an individual leading the chant). As an example of Modern Standard Arabic the *Allahu akbar* chant is not phonetically accurate. For example the lateral seems to lack its appropriate gemination and pharyngealization; and the following long vowel should also have a pharyngealised (i.e low back) allophone. It is not possible to tell whether that is because some or all of the chanters were non-native speakers; or because of the act of chanting itself (c.f. the schwa vowel in *the* was lengthened to conform to the chant rhythm).

Figure 2 shows a reduced dynamic range spectrogram of chant number 4 with superimposed formants. Some of the formant structure has come out surprisingly clearly, which is remarkable, given that we are seeing the acoustics of many speakers. The segmental boundaries are not well-defined; neither is there any indication of putative /s/ or /z/ noise. It can be seen that both F1 and F2 in the first syllable have slightly lower values at their onset, but the F1 is subsequently at about the same height as the F1 in the *the*, whereas the F2 seems to have a slightly higher, rising trajectory than the F2 in *the*.

In order to get a handle on the probability of getting the questioned acoustics – in this case the F1 and F2 in the first word – under competing hypotheses of /we:/ in *where* and /æ/ in *gas*, we need to know what the F1-F2 monophthongal vowel space looks like. Thanks to the *Multicultural Australian English* project, we have F-pattern data for young speakers from several ethnically diverse sites in Sydney [13]. The site with the greatest connections to Islam is clearly Bankstown, with its ca. 17% Lebanese ancestry and ca. 21% of homes speaking Arabic as well as English [13]. The left panel of figure 3 shows an F1-F2 plot for the relevant vowels modified from [13]. It can be seen that the /e:/ has a very similar F1 location to the /ɜ:/, while having a higher F2; and that the /æ/ has a similar F1 extension to the /ɐ:/, while having an F2 a little higher than /ɜ:/.

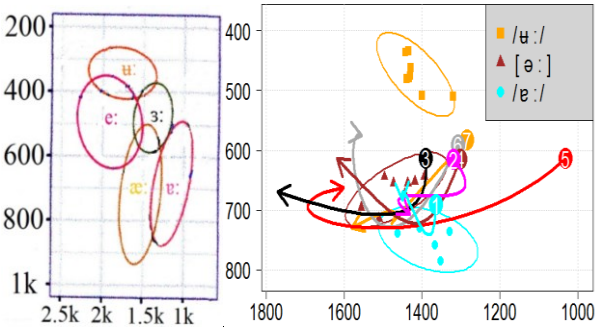

Figure 3: *Left = F1-F2 plot for young Bankstown male monophthongs (95% confidence ellipses). Right = F1-F2 trajectory of questioned syllables 1 to 7 plotted against known values. X-axis = F2, y-axis = F1.*

The right panel of figure 3 shows plots of the F1-F2 trajectory of questioned syllables in chants 1 to 7 against an F1-F2 plot of all the long chanted vowels (/u:/ in *Jews,* [ə:] in *the*, /ɐ:/ in *All**a**hu akb**a**r*). (The questioned vowel in the last two chants was obscured.) It can be seen that the trajectories involve glides from a retracted (low F2) position towards a space that would be expected for /e:/. Although their quantification is not clear, these characteristics would be very much more likely if the questioned word had been *where* rather than *gas*, which would be expected to have a short backing and falling perturbation from the velar stop. Assuming flat priors would

mean that the quantifiable chants were very likely not *gas the Jews*. However, unlike perhaps in Bain, it is not at all clear how one would estimate the priors in this case in order to arrive at the posterior probability quoted by the police expert.

3. Tampering

Sometimes it is claimed that a recording has been tampered with in some way. Perhaps something incriminating has been removed or inserted. In a 2000 case in the UK [14] it was alleged that the phrase *the… Law Society* (… indicates a pause) had been copied from the utterance *the … Law Society ringing me?* in an earlier part of a recording and pasted into an incriminating location later in the recording. An audio engineer retained by the prosecution noted the high degree of visual similarity in waveform between the two occurrences of *the… Law Society* and, adducing the phonetic maxim that one never says the same thing in an identical way, opined that the second occurrence had indeed been copied. Looking at the spectrograms and F0 of the two utterances in figure 4, one can have some sympathy with the engineer. Any phonetician, too, with experience in the variability of natural speech would think that kind of near-identity unlikely if the utterance had been repeated naturally. Phoneticians retained by defence argued, however, that the phonetic maxim upon which the engineer relied "had not been proved" and that the second occurrence of *the Law Society* had just been said in a very similar way to the first. One of the judges pertinently drew attention to the lack of quantification, asking: "… to what extent can the differences between the same word, spoken by the same person, be measured? Does it follow that where the word is electronically measured and compared to another word and there are marked similarities that must mean it is a digital clone?"

Indeed! So, how would one evaluate these claims logically, and quantitatively, within a LR framework? Call the first occurrence LS1 and the alleged copy LS2. The alternative hypotheses are (H_c) that LS2 has been copied and pasted from LS1; and (H_n) that LS2 does not so originate. The evidence E_a is the difference between the acoustics of LS1 and LS2. In a LR framework what has to be evaluated is the probability of getting E_a assuming that LS2 has been copied from LS1, and assuming that it has not. Note that this is not the question asked by either judge or engineer: they were typically asking about the probability of a hypothesis (digital cloning), given the high degree of similarity (evidence).

Since there is an infinite number of parameters one could use to estimate E_a, it is sensible to simplify. F0 was chosen as a suitable parameter to compare: it is easy to measure and its perceptual correlate of pitch is more straightforward to imitate than segmental and voice quality. When aligned with respect to the details of the waveform at the onset of the F0 in *Law*, the mean absolute difference between the F0 of the two utterances was 0.9 Hz. This is taken as E_a.

An experiment was then run to see how likely this value of 0.9 Hz is under competing hypotheses: when the phrase *the … Law Society* is digitally copied; and when it is repeated. The top panel of figure 5 shows the F0 of 30 attempts by me, not to just repeat, but actually imitate both intonational pitch and pause of *the…Law Society* as said in the recording. It can be seen that the F0 over the *the* shows a small amount of variation, but obviously getting the duration of the pause right introduces considerably more. Each of my 30 imitations was then digitally copied in *Praat* to different random downstream locations and the difference measured between original and copied F0.

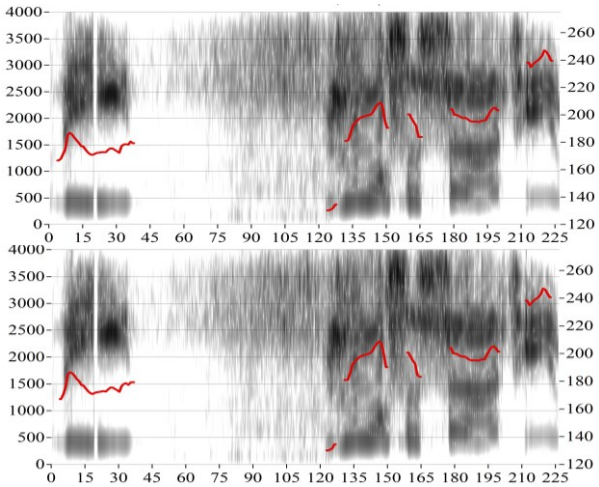

Figure 4: *Acoustics of* the…Law Society. *Top = original, bottom = alleged copy. X-axis = duration (csec.), y-axes: left = spectral frequency (Hz), right = F0 (Hz).*

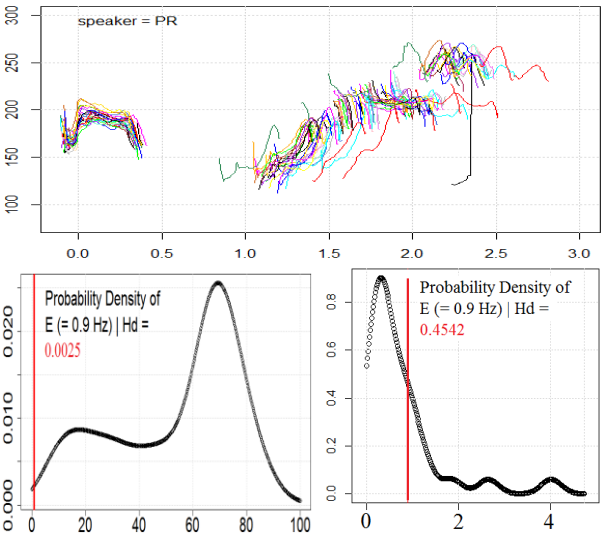

Figure 5: *Determining LR in tampering case. Top = F0 of 30 attempts to imitate* the…Law Society. *Bottom = distributions of absolute F0 differences between different (left) and copied imitations. Vertical red line = evidence.*

The bottom left panel of Figure 5 shows the probability density for all 435 pairwise differences in F0 between the imitated repeats. It shows that it is indeed possible to get an almost exact repeat – where the difference is close to zero – but most of the time the difference is much bigger (the reason for the distributional bimodality is not clear). As can be seen, the probability density of getting the value of 0.9 Hz for the evidence is a relatively low 0.0025. The bottom right panel of figure 5 shows the probability density for the digitally copied pairs. It is not, as one might at first expect, zero – there is for example one occurrence of a difference as big as 4 Hz. But of course the probability of getting the evidence assuming copying is relatively quite high: 0.4542. The difference between LS1 and LS2 is therefore (0.4542/0.0025 =) 180

times more likely if LS2 had been copied. This magnitude of LR already has some heft. But it will be much greater because the defence hypothesis was that it was a case of natural occurrence, not imitation. But there is still more, by way of contribution from the spectral properties. At the end of the spectrogram in the top panel of figure 4 can be seen the well-known anticipatory drop in F2 and F3 in the F-pattern for /i:/ in *society* caused by the following post-alveolar approximate [ɹ] in *ringing*. But the same perturbation can also be seen in the alleged copy, where *society* is not followed by an /r/ and there is therefore nothing to cause such a perturbation. The LR for this must be very big. The probability of getting the perturbation in /i:/ F-pattern given that it is followed by [ɹ] must approach 1. The probability of getting the perturbation with nothing following to condition it must be very low. (A devil's advocate might imagine a scenario where a speaker intends to say a word beginning with /r/, but then for some reason stops.) Perhaps one could estimate this probability by scouring recordings of natural speech for disfluencies of this type and estimate the amount of occurrences per unit time. But there is yet more: in figure 4 a dropout is evident in *the* in both LS1 and LS2 with the same duration and at precisely the same place. Again the LR for this must be vanishingly small. We now encounter a problem. Clearly, combining the LRs for the F0, the F-pattern and the dropout will result in an LR of enormous magnitude, but we cannot calculate it. Neither is it clear how to estimate prior odds for this scenario.

4. Voice Lineups

During an armed robbery someone hears the disguised robber speak. A suspect is arrested and the earwitness identifies the suspect in a voice lineup. Assuming the lineup was conducted fairly, what evidential strength should one give that identification? Is that good news for the prosecution?

A lot of empirical work has been done recently by the members of the UK project *Improving Voice Identification Procedures* (IVIP) to determine optimum parameters and procedures for voice lineups – for example, whether to tell the earwitness that the suspect might not actually be in the lineup; what the optimum number of foils should be; or how long the voice sample [15]. But surprisingly almost nothing has been done on arguably the more important question: how to evaluate the strength of evidence of a voice lineup identification, that is, where the earwitness picks out the suspect. Interestingly, attempts to address the question in visual line-ups e.g. [16] appear to have focused on signal detection parameters familiar in automatic speaker recognition like DET and ROC, but which relate to posterior probabilities, not the strength of evidence. In 2002, in an attempt to explain Bayes' Theorem to the legal profession, a judge of the Supreme New South Wales Court of Appeal who has also published widely on probability showed that the LR in a voice lineup is a function of the reliability of the earwitness and the number of foils; and how identification through a properly conducted lineup carries more evidential strength as the number of foils increases [17]. The argument below, presented in [18], is from that paper. A much later paper [19] has also approached the problem using Bayes Factors.

We assume the suspect is present in a lineup and the earwitness picks them out. In this scenario, the evidence E_s is the fact that the earwitness picked the suspect. What is the value of that evidence? The prosecution hypothesis (H_p) is that the suspect is the person the witness heard; the defence hypothesis (H_d) is that the suspect is not the person the witness heard. The strength of evidence will be the ratio of the probabilities of the evidence E_s under the competing hypotheses, viz: $LR = p(E_s \mid H_p) / p(E_s \mid H_d)$.

In order to estimate the probability that the witness would pick the suspect assuming the suspect was whom the witness heard we really need to know how good the earwitness actually is. Of course the best option is to test the witness; but *faute de mieux* one can estimate this from known naïve identification under circumstances similar to the crime. In [17] the hypothetical witness was assumed, completely unrealistically, to be 90% reliable. Experiments with simulated voice lineups in the *Improving Voice Identification Procedures* (IVIP) project report rates between ca. 45% and 30% when the suspect was present in the lineup. IVIP felt that, although this is poor, in reality the performance would be better, since real lineups would be more heterogeneous. Earlier experiments with naïve unfamiliar listeners have indeed shown better figures, but not that much better: between 40% and 50% [20: 202, 203]. Since the IVIP experiments were conducted under differing conditions to find optimum parameters, it makes sense to use their best performance – about 45% with 15 second duration speech and nine voices in the lineup – because the worst performances could have been caused by experimental choice of an adverse parameter. Assuming the witness is 45% reliable will give the value 0.45 for the probability of the witness picking the suspect assuming they had in fact heard them.

Estimating the other part of the LR – the probability that the witness would wrongly identify the suspect as the person they heard – can be broken down into two questions. We have to ask firstly, and obviously, what is the probability that the selection is wrong (i.e. the suspect is not whom the witness heard). With the earwitness' 45% reliability, this is 0.55. We also have to ask what is the probability that the suspect is selected. IVIP concluded that the nine-voice parade mandated in the UK should be maintained, so this is (1/9 =) 0.11… . The probability that the suspect is selected, and they are not whom the witness heard, is then (0.55 * 0.11… =) 0.061… . Thus the LR for this identification would be (0.45 / 0.061… =) ca. 7.4. Given the generally poor performance of naïve unfamiliar recognition this might surprise some, but as emphasised in [17] it is also a function of the number in the lineup, and shows the benefit of an explicit, quantifiable LR approach. But 7.4 is still not very strong evidence. In the absence of other evidence pointing to the suspect as the perpetrator, the biggest posterior probability achievable with a LR of 7.4 (with even priors, i.e. with only two possible people who could be guilty) is about 88%. But of course in conjunction with other available evidence it can be a powerful multiplier.

5. Summary

This paper has given an idea of what is involved in applying the likelihood ratio framework to cases of disputed utterances, tampering, and to voice lineups. In the disputed utterance cases it was shown that the evidence is the acoustics; not what someone reports hearing. In the tampering case there were potential problems with knowing the LR was inconceivably big but being unable to calculate it. Priors, already characterized in [21: 88] as an 'open-ended problem', were also seen to be a potential problem in both disputed utterance and tampering. Perhaps one can avoid this in tampering and disputed utterances with uninformative priors; perhaps not.

6. Acknowledgements

This paper builds on my keynote at the 2021 conference of the International Association of Forensic Phonetics & Acoustics. I thank Prof. Cox for providing from a preprint the indispensable data on Bankstown formants, and Prof. McDougall for providing references on evaluation of visual lineups. Many thanks also to my anonymous reviewers for their very useful comments, most of which I was able to address.

7. References

[1] Morrison, G. S., "Forensic voice comparison and the paradigm shift", Science & Justice 49: 298-308, 2009.

[2] Drygaylo, A., Jessen, M., Gfroerer, S., Wagner, I.,Vermeulen, J. and Niemi, T., "Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition", Verlag für Polizeiwissenschaft, 2015.

[3] Silver, N., The Signal and the Noise - The Art and Science of Prediction, Penguin, 2013.

[4] Ellenberg, J., How Not to be Wrong, Penguin, 2015.

[5] https://www.newsweek.com/50-famously-misheard-misunderstood-song-lyrics-explained-1561728

[6] Innes, B., "R v David Bain – A unique case in New Zealand legal and linguistic history", International Journal of Speech, Language and the Law, 18: 145–155, 2011.

[7] Fraser, H., "Enhancing' forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript?", Australian Journal of Forensic Sciences, 52(4): 465-476, 2019.

[8] Rose, P., "Evaluation of Disputed Utterance Evidence in the matter of David Bain's Retrial", http://philjohnrose.net/pubs/FVC_pubs/index.html, 2009.

[9] Morrison, G. S. and Hoy, M., "What did Bain really say? A Preliminary Forensic Analysis of the Disputed Utterance Based on Data, Acoustic Analysis, Statistical Models, Calculation of Likelihood Ratios, and Testing of Validity", 46th AES International Conference,

[10] Wilson, C., and Lattouff, A., "New footage and audio experts raise further doubts about Sydney Opera House protest video", https://www.crikey.com.au/2023/12/19/new-footage-audio-experts-sydney-opera-house-protest-video/ 2023.

[11] https://www.abc.net.au/news/2024-02-02/nsw-police-opera-house-protest-video-analysis/103418582

[12] https://www.youtube.com/watch?v=XGdM6FcSYXw

[13] Cox, F. and Penney, J., "Multicultural Australian English The new voice of Sydney", Australian Journal of Linguistics 1-20, https://doi.org/10.1080/07268602.2024.2380680, 2024.

[14] Rose, P**., "**Forensic speech science report in a case involving the alleged tampering of a recording**",** http://philjohnrose.net/pubs/FVC_pubs/index.html , 2009.

[15] McDougall, K., Nolan, F., Paver, A., Smith, H., Braber, N., Wright, D., Pautz, N., Goodwin., P., Robson,. J. and Mueller-Johnson, K., "Improving Voice Identification Procedures: Findings and Recommendations from the IVIP Project", IAPFA Webinar 18th April 2024.

[16] Wixted, J.T., and Mickes, L., "Evaluating eyewitness identification procedures: ROC analysis and its misconceptions", J. Applied Research in Memory and Cognition, 4(4): 318-323, 2015.

[17] Hodgson, D., "A Lawyer looks at Bayes' Theorem", The Australian Law Journal, 76: 109-118, 2002.

[18] Rose, P., "Evaluating Strength of Evidence in Voice Lineups", Symposium to honour Andy Butcher, Melbourne University, http://philjohnrose.net/pubs/FVC_pubs/index.html, 2017.

[19] Rosas, C., Sommerhoff, J. and Morrison, G.S., "A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker", Science & Justice, 59: 585-596, 2019.

[20] Hollien, H., The Acoustics of Crime, Plenum, 2002.

[21] Jaynes, E.T., Probability Theory – The Logic of Science, CUP, 2003.