Evaluating Strength of Evidence in Voice Lineups

phil rose

The Topic

- The teller during an armed bank robbery hears the disguised robber speak.
- A suspect is arrested.
- The teller later identifies the suspect in a voice lineup.
- What weight do you give that identification?
- Cf. a blood stain is found at the scene of the crime which matches the blood group of the suspect. What weight do you give that match?

The Inspiration

- A 1996 paper on voice line-ups by one A. Butcher:
- GETTING THE VOICE LINE-UP RIGHT: ANALYSIS OF A MULTIPLE AUDITORY CONFRONTATION. *Proc. 6th Australian Intl. Conf. Speech Science & Technology*: 97-102.
- AB describes the evaluation of the recordings used by police in a voice line-up in an armed robbery case.
- AB presented *voir-dire* evidence from both auditory evaluation by a group of voice professionals and acoustics (mean, sd F0; mean F-pattern in /ae/ /i:/ /u:/ [schwa], tempo) that two of the 12 voices stood out, one of which was also the suspect.
- The judge ruled the line-up results inadmissible.
- Correct, and useful: A demonstration of forensic speech science at its best.

This presenatation …

- I will be concerned however not with the proper constitution of a voice line-up, but on the evaluation of its outcome.
- Cf. 2002 paper by D. Hodgson (Judge of the NSW Court of Appeal):
- A LAWYER LOOKS AT BAYES' THEOREM. *The Australian Law Journal* 76: 109-118.
- attempt to explain to the legal profession how to properly (= rationally) evaluate prosecution and defence hypotheses, given the (numerical) evidence adduced in their respective favour.
- H uses a *gedankenexperiment* of a visual line-up to suggest identification through a properly-conducted line-up can yield strong evidence, which becomes more powerful as the number in the line-up increases.

Real-world application

- Ongoing case of large-scale ice importation.
- Police use “ad-hoc expert” witness (professional Cantonese translator) to identify offender’s voice from the voice of the suspect she had heard 2 years previously (i.e. naïve non-familiar SR).
- Line-up of 1 (!).
- Obvious question for the trier-of-fact:
- How reliable is that?
- Obvious answer from forensic speech scientist:
- What properly quantified evidential weight do you give it?

Structure of presentation

- The variability of strength of evidence and how to estimate it
- Role of SoE in decision making
- Estimating SoE from a (properly conducted) voice-line up
- How useful is it?

Evidence ain't just Evidence

- A murder is committed in Peking …
- A black hair from the offender is found at the scene of the crime …
- A suspect is arrested. They have black hair.
- Prosecution claims that the suspect is the offender because of the match in hair colour.
- The evidence (E) is the match in hair colour.
- Is this evidence any good – value/strength?
- Why?

Suppose the hair was blond and the suspect had blond hair (you're in Peking, remember)

Probabilities of Evidence

- Formal reason underlying the intuition:
- If the offender's black hair had come from the black-haired suspect (Hypothesis$_p$), then the match in colour (Evidence) would be highly likely:
- $P(E \mid H_p)$ = high
- If the offender's black hair had come from someone else in Peking (Hypothesis$_d$), then the probability of the colour match (Evidence) would also be highly likely:
- $P(E \mid H_d)$ = high
- This is thinking probabilities of Evidence, not Hypothesis!

How Strong is the Evidence?

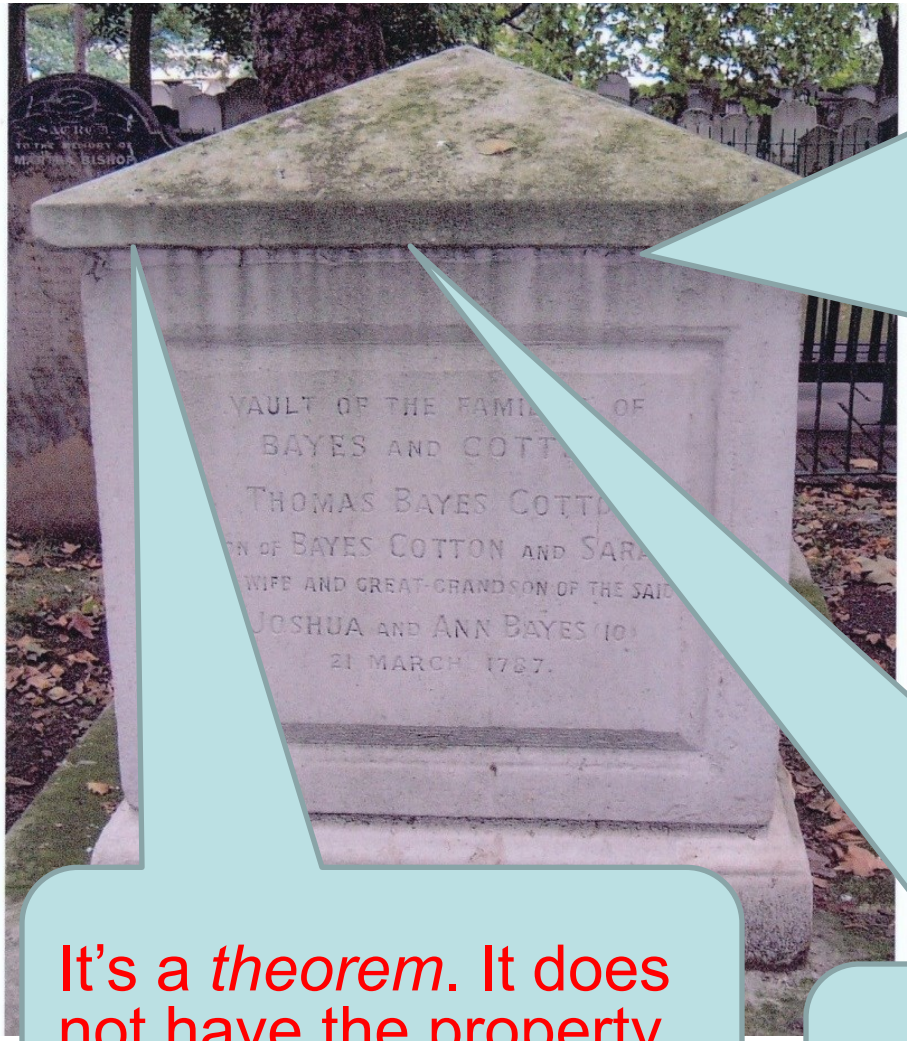
- Given the locality (Peking) you are just as likely to get the evidence of a colour match in black hair under both hypotheses:
- Ratio of $P(E|H_{\boldsymbol{p}})$ to $P(E|H_{\boldsymbol{d}})$ is unity.
- > Such evidence has no strength; it's useless.
- What about the blond hair case?
- Ratio of $P(E|H_{\boldsymbol{p}})$ to $P(E|H_{\boldsymbol{d}})$ is much bigger than unity.
- You are much more likely to get the evidence under the prosecution hypothesis H***p*** than under H***d***
- > This evidence has strength (in favour of H***p***); it's useful.

Likelihood Ratio

- Ratio of probabilities of *evidence* under competing hypotheses:
- $P(E|H_p)/P(E|H_d)$
- Likelihood Ratio, quantifies strength of evidence (proportional to deviation from unity)

How Strong *is* the Evidence?

- NB LR nothing to do with the probability of the hypotheses! *P*(quadruped|cow) = *P*(cow|quadruped)?
- How likely the *hypotheses* are to be true is given by strength of evidence AND prior probability/odds of hypotheses by Bayes' Theorem.



So you want to know the *Probability of a Hypothesis given the Evidence?* My theorem tells you it is proportional to the *Strength of your Evidence* and the *Prior Probability of your Hypothesis.*
[Bayes' Theorem 1763]

It's a *theorem*. It does not have the property of being wrong.

Nowadays they call it "*...the* fundamental formula of forensic science interpretation"
Evett, UK Forensic Science Service

A voice line-up scenario

- Witness hears voice of offender
- Witness is tested with properly constituted (= …) voice line-up
- 10 foils + suspect
- Witness picks out suspect
- What is the strength of the evidence (E = witness has picked the suspect)?

LR applied to voice lineup

- Strength of Evidence = LR = $P(E|H_p)/P(E|H_d)$
- E = witness has picked suspect in properly constituted lineup (10F 1S)
- H_p = prosecution hypothesis = suspect is the person witness heard
- H_d = defence hypothesis = suspect is not the person witness heard
- LR = ?

Voice line-up scenario $P(E|H_p)$

- $P(E|H_p)$ = if the witness did indeed hear suspect, what is the probability that they would pick them out?
- To estimate that, you need to know how good they are at recognising (unfamiliar, we assume) voices under conditions comparable to the case …

Naïve voice recognition performance P(E|Hp)

- Naïve *unfamiliar* voice recognition varies as function of many factors, e.g. identifier's ability, distinctiveness of the voice, expectation, elapsed time … (Hollien 2002*:* 200-203, Rose 2002: 97-104).
- This variability will need to be modelled as a continuous pdf, but we assume for demonstration purposes an extremely generous point value for the witness of 75%.
- If they are 75% reliable, and the suspect was whom they heard, *P*(witness picks out suspect | suspect was whom they heard) = 0.75
- One term of the LR.

Naïve voice recognition performance P(E|Hd)

- The other LR term is more tricky …
- What is the probability that the witness would pick out the suspect, assuming that the suspect was NOT whom they heard (and that the offender was not one of the foils!)?
- Hodgson: If witness was 75% reliable, then if the suspect was not the person they heard, they would (incorrectly) pick someone from the lineup 25% of the time…
- And if there are 10 foils and 1 suspect, the probability of picking the suspect is 1/11.
- Thus *P*(witness picks suspect | suspect not heard) =
- 1/4 * 1/11 = 0.0227… (NB this is quite low)

Another way of thinking about it:
what is the probability that the suspect is selected? 1/11.
What is the probability that the selection is wrong? 1/4

LR from voice lineup

- Strength of evidence = $P(E \mid H_p) / P(E \mid H_d)$
- = 0.75/ 0.023
- = ca. 33
- [with the given parameters] “The witness is 33 times more likely to identify the suspect if the suspect is the person they heard than if they were not the person they heard.”
- SoE *aka* LR is function of number in lineup and reliability of witness:
- LR = r /[(1-r) * 1/n], where r = reliability of witness, n = number in lineup

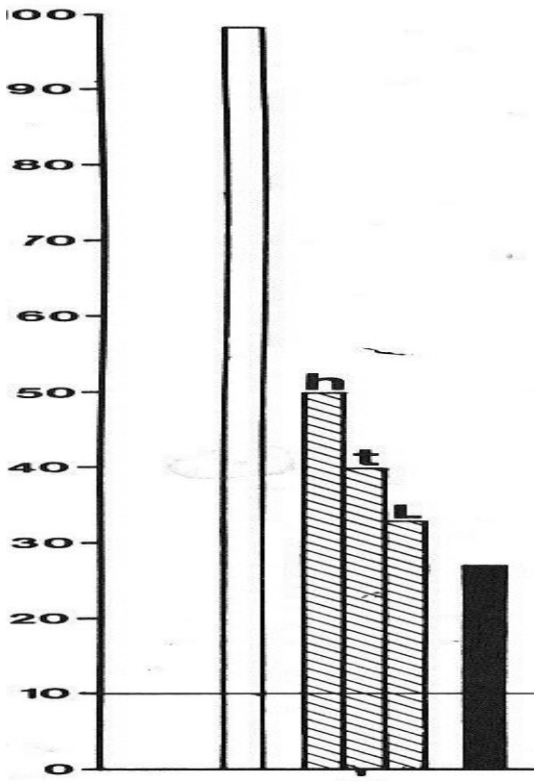
Meaning of LR

- What does a LR of 33 mean?
- Can only be given meaning in conjunction with prior …
- Suppose suspect was one of five people who could have been the offender …
- By Bayes' Theorem, LR of 33 means posterior odds that the suspect was the offender of : prior odds * likelihood ratio = 1:4 * 33:1 = 33:4
- Equivalent to post. probability of 33/(33+4) = 89% likely it was the suspect.
- If suspect was one of 100 other people the *same* strength of evidence will give you a very different value: … 25%

Experimental results (1)

- McGehee's 1937, 1944 research inspired by Lindberg son kidnapping (Lindberg claimed to identify the voice of the kidnapper 2 years after he heard it).
- Voice identifications quite good at first – 85% correct on day following exposure –
- Dropped off gradually with time –
- 13% correct after 5 months.

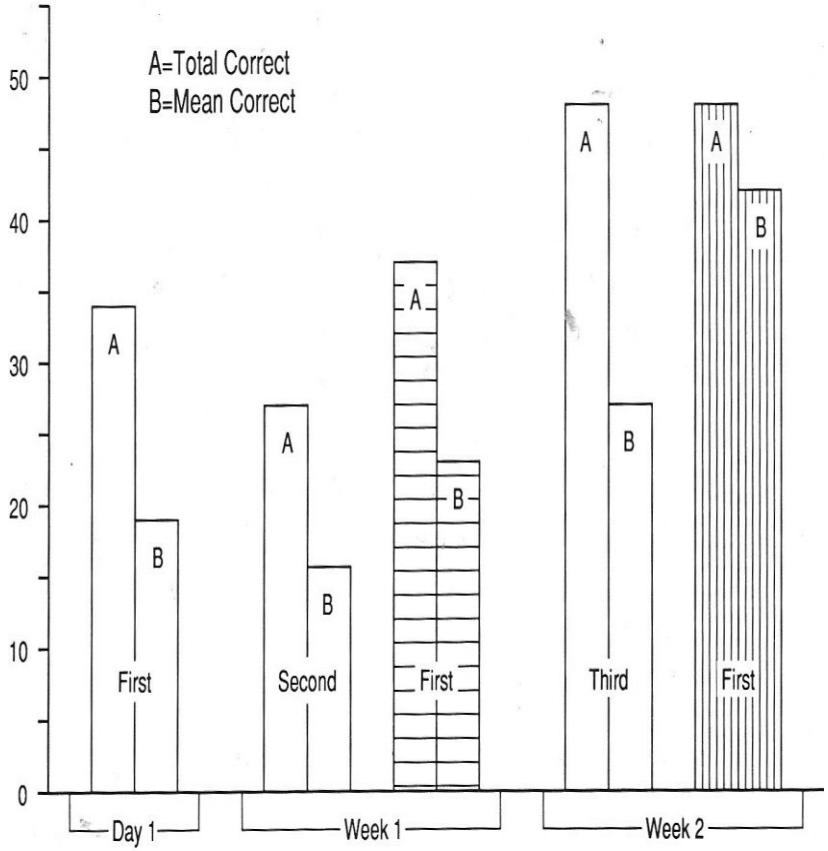
Experimental results (2)



Performance of 3 different groups of naïve subjects in an aural speaker identification task with 10 targets (from Hollien 2002 *Acoustics of Crime*: 202, left part of figure 9-2).
unshded = familiar listeners;
solid = non-english speakers;
cross-hatched = unfamiliar listeners trained on the voices to be recognised! (H = good performers, L = weak performers) .
Vertical axis = %correct identification, horizontal line = chance performance.

Familiar wildly better than unfamiliar. The very best performance of the best-performing (H) unfamiliar group is 50%. The mean of the unfamiliar group is 40%

Experimental results (3)

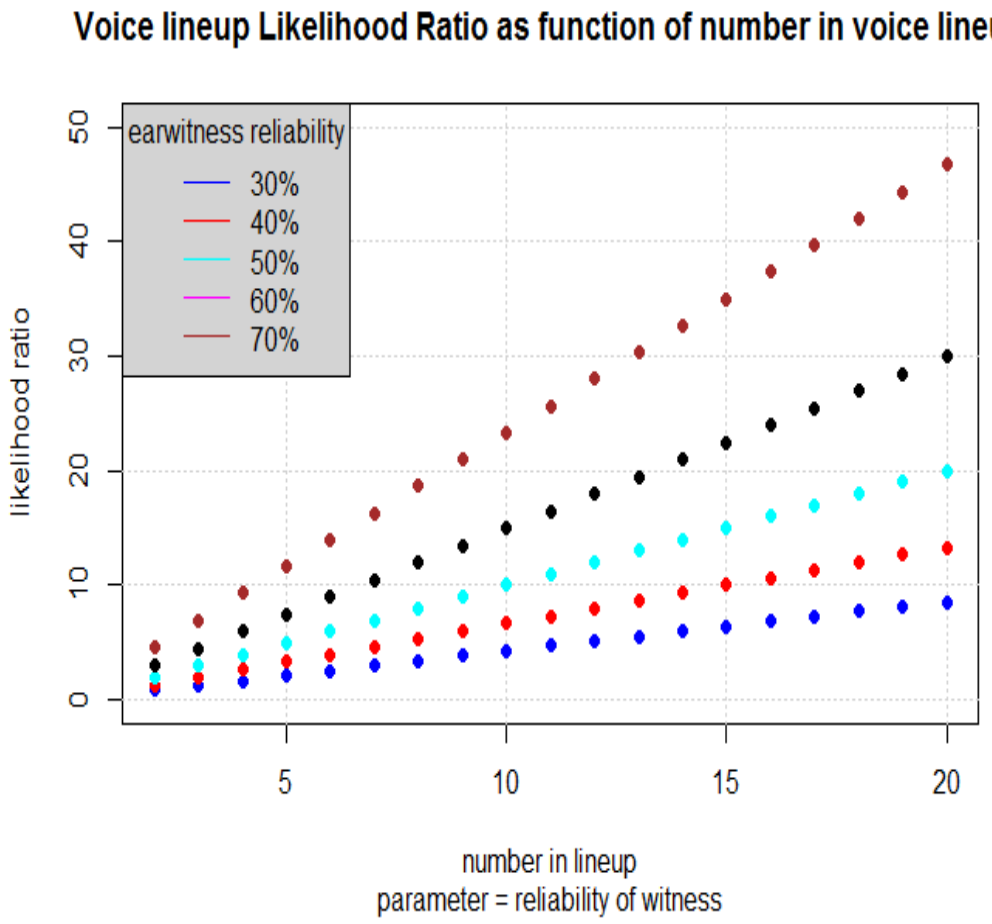


Unfamiliar naïve identification performance of three groups of listeners on a simulated crime. Unshaded = identifications after one day, one week and two weeks. Horizontal shading = another group after one week. Vertical shading = another group after two weeks. Vertical axis = % correct identification. From Hollien 2002 *Acoustics of Crime*: 203, figure 9-3)

Performance is best after two weeks.

The best mean correct id is still only ca. 43%

System performance



With realistic naïve earwitness reliability (40% - 50%) and realistic lineup size (ca. 10 < problems of amassing a fair selection of voices) LR is expected to be around 10, which would require a very advantageous prior to be much use on its own.
Earwitness reliability is obviously the important term.

Summary/Conclusions

- Presentation looked at the logically correct evaluation of the strength of evidence in voice lineups.
- Demonstrated SoE is a simple function of the number of subjects in lineup and the reliability of the witness, with reliability the dominant term.
- Therefore, the reliability of the witness needs to be tested before you can say anything about the SoE
- (or model the uncertainty in reliability with a pdf)
- Given performance in naïve unfamiliar voice recognition is poor, one cannot rationally expect voice lineups to provide very good strength of evidence (unlike, perhaps, visual lineups).

To (septugenarian) AB:

HOCH SOLL ER LEBEN

DREIMAL HOCH!!!