The impact of reflection and retention intervals on earwitness accuracy: two experiments

Francis Nolan, Nikolas Pautz, Kirsty McDougall, Katrin Müller-Johnson, Harriet M.J. Smith and Alice Paver
IVIP: Improving Voice Identification Procedures

Project Team

University of Cambridge:

PI: Kirsty McDougall (phonetics)
Francis Nolan (phonetics)
Alice Paver (phonetics)

Nottingham Trent University:

Nikolas Pautz (psychology)
Natalie Braber (sociolinguistics)
David Wright (sociolinguistics)
Harriet Smith (psychology)

De Montfort University:

Jeremy Robson (law)

University of Oxford:

Katrin Müller-Johnson (criminology, psychology)

UK ESRC Grant Ref: ES/S015965/1
Simulating an earwitness situation

• Yet, in crime situations, if a witness realises they have heard a perpetrator, they’re likely to think back over the event

• RQ: Does explicit post-encoding (post target-exposure) reflection improve voice recognition accuracy?
### Experiment 1: distractor task, no delay

<table>
<thead>
<tr>
<th></th>
<th>target present</th>
<th>target absent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>with reflection</strong></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>no reflection</strong></td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

3 separate parades, each with its own target (if present), and 3 separate sets of foils, were used; but results will be combined here
Word-spotting distractor task

Try and find as many words as you can in 5 minutes.

When you find a word, type the coordinates \((X,Y)\) of the **first** letter in brackets after the word.

Please use a semi-colon (;) to separate your answers.

For example: APPLE (3,22); MANGO (1,1);
.....
Speakers

- Targets & foils from *DyViS* database (Nolan *et al.* 2009)
  - male, Standard Southern British English, aged 18-25 years
- 3 groups of 15 speakers, 1 per target
- Targets and foils chosen as in McDougall *et al.* 2021 IAFPA:
  - Listeners judged similarity of all pairings in a group of 15
  - Multidimensional scaling > pseudoperceptual space
  - 9 foils (for target-absent) or 8 (for target-present) speakers sounding most similar to the target selected from each group of 15
Speech material

• Exposure material for target voice: 60s sample from telephone call (‘perpetrator’ side; studio quality)
• Parade samples: 15s samples from simulated police interview task, using collage technique of Home Office guidelines
• Experiment conducted online using Gorilla
Participants

• N = 180 participants recruited via Prolific (randomly assigned to 1 of 3 targets)
  - born in and lived most of their pre-18 lives in England
  - 1st language English
  - No hearing loss or hearing difficulties
  - 88 male, 92 female, aged 18-40 years (M= 27.72, SD = 6.4)
  - Minimum approval rating of 90% on Prolific
Reflection condition:

“Imagine that the voice you have just heard is that of a criminal. You may be asked by the police to make an identification some time in the future. Take a few moments now to reflect on the voice.”

(20 seconds)
No reflection condition:

Simple attention task – participants push space bar when + appears on screen, several times (20 seconds)
Procedure

Voice (60 sec)  Reflection (or not)  5 min task  Voice parade  Decision

Retention interval:
- 5 min distractor task
  (word search, accompanied by lobby noise)
- exceeds short-term memory capacity; relies on long-term-memory
Results: effect of reflection (before distraction task)

- Poor accuracy percentages overall, but above chance (except in target-absent/no reflection)

- Target-present parades give best performance (consistent with previous findings, e.g. Smith et al. 2020)

- No meaningful differences in performance between with- and without-reflection (Bayesian mixed models)
Experiment 2: overnight delay

• Same design as Experiment 1, but 20-28 hour retention interval between exposure and parade instead of distractor task

• Participants (same recruitment and requirements):
  - N = 181
  - 87 male, 93 female
  - aged 18-40 years (M= 27.97, SD = 6.01)
Experiment 2: overnight delay, no distraction task

<table>
<thead>
<tr>
<th></th>
<th>target present</th>
<th>target absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>with reflection</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>no reflection</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

3 separate parades, each with its own target (if present), and 3 separate sets of foils, were used; but results will be combined here
Procedure

Voice (60 sec) → Reflection (or not) → Retention interval → Voice parade → Decision

Same reflection/no reflection manipulation
Procedure

Voice (60 sec)  Reflection (or not)  Retention interval  Voice parade  Decision

20-28 hour delay between exposure and parade instead of distractor task
Results: effect of reflection (before overnight delay)

- Again, poor accuracy percentages overall,
- Again, target-present parades give best performance,
- No meaningful differences in performance between with and without reflection (Bayesian mixed models),
- No interaction btw target presence and reflection.
The motivation for ‘reflection’ was that the immediate cognitive load of the word-finding task might hinder memory consolidation.

Disappointingly, neither with word-finding simulating a delay, nor with an actual overnight delay, did a period of reflection improve scores.
However, we don’t think this closes the case on ‘reflection’:

- our period of reflection was very short (20 seconds)
- it did not allow for repeated ‘rehearsal’ of the auditory memory, as might happen in a real event
- we had no check on whether online participants actually reflected on the target voice, rather than (e.g.) their shopping list

- Longer reflection, at least, will be worth exploring
Why is accuracy low? (1)

- Our design minimises ‘propitious heterogeneity’ (Wells 1993), in order that we can potentially see improvements when factors are varied. By design we use
  - tightly accent-controlled speaker population
  - rigorous selection of perceptually close foils and targets
- i.e., we make the participants’ task as earwitnesses as hard as we can
- Carlson et al. 2019 on visual parades:
  - ”empirical discriminability decreases as fillers [foils] become too similar to each other and the suspect”


Carlson et al.’s computer-generated faces

eyes only
eyes + nose
eyes + nose + mouth
Why is accuracy low? (2)

• We also suspect the (unavoidable) online presentation reduces participants’ engagement and motivation
  - McDougall, Nolan & Hudson (2015), in an in-person simulated parade, report 76% correct for target-present — and that after a week’s retention interval — compared to 30-40% here

• In the case of target-absent parades, results just emerging from another IVIP experiment suggest the strength of warning before the parade is crucial:

Three strengths of warning

• WEAK: “Remember that the voice you heard at the beginning of the experiment may or may not be present.”

• STRONG: “Remember that the perpetrator may or may not be present. Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction.”

• VERY STRONG: “Remember that the perpetrator may or may not be present. Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction. Voice recognition can be very difficult. Only make a positive identification if you are very sure.”
Effect of pre-parade warning

- Absent
- Present

Response Accuracy

Standard | Strong | Very strong

Reflection
Conclusions and further work

• Our two experiments failed to show an effect of a period of ‘reflection’ — either with simulated (word-task) or real (overnight) delay
  • the reflection allowed may have been too short
  • we had no check that participants really engaged

• Future work might test longer reflection, and check ‘engagement’

• What does improve ‘target-absent’ accuracy significantly, emerging results suggest, is the strength of warning
See IVIP website for updates

https://www.phonetics.mml.ox.ac.uk/ivip/