

The effect of voice sample duration and lineup size on voice identification performance

Nikolas Pautz, Harriet Smith, Katrin Mueller-Johnson, Francis Nolan, Alice Paver, Kirsty McDougall



Economic and Social Research Council









Voice identification?



- There are **crimes** that occur where **the perpetrator is not seen**, but may be **heard** (masked attack, telephone fraud)
- In such cases, **the voice** of the perpetrator may be the **only evidence** available to confirm their identity
- Voice parades involve an "earwitness" trying to identify the perpetrator from a series of voices
- In official VPs, the voice samples of the suspect are excerpts from their police interviews, and the foils being similarly-sounding suspects from unrelated cases



What are the issues?



- Low hit-rate, high false alarm rate (Smith et al., 2020; Kerstholt et al., 2006; Ohman et al., 2013)
- Juries find a confident earwitness extremely convincing (Van Wallendael et al., 1994), but confidence is **unlikely to be diagnostic of accuracy** (e.g., Smith et al., 2020, Perfect et al., 2002)
- Where guidelines for conducting voice parades exist, they are heavily adapted from face ID procedures
 - In comparison to faces, voices provide relatively weak cues to identity (Stevenage & Neil, 2014)
 - Memory for voices is also **subject to higher levels of interference** than memory for faces (Stevenage et al., 2011)
- Unfamiliar voice ID is under-researched, especially when it comes to system variables (i.e., procedure focused manipulations)



Guidelines for VP construction

- Guidelines published by the Home Office (2003) England and Wales is the only common law jurisdiction to have established VP-specific guidelines
- Guideline on voice sample duration: should be at least 60s
- Guideline on the size of the parade: 8 foils + target
- No empirical evidence that these are the best parameters
- Alternative guidelines offer different parameters (Broeders & van Amelsvoort, 2001)
 - Also developed partly on 'tried and tested' procedures that Dutch police use for visual ID
- Our experiments focused on testing if these guidelines are most suitable parameters for voice parades



Experiment 1

- Can we reduce voice sample durations without performance cost?
- 9 * 60s samples is cognitively demanding (esp. with serial procedure)
- Some evidence that listeners can extract basic identity information from much shorter durations (Bestelmeyer et al., 2010)
- Theoretical reasons: based on temporal-ratio models of memory (Bjork and Whitten, 1974; Brown et al., 2007); shorter voice samples may be more distinctive
- Practical reasons: parades with shorter samples will reduce the overall time needed to develop the parade; may reduce the delay between exposure and identification, maintaining a stronger memory of the perpetrator's voice





Speaker / Stimulus selection

- In 'real' VPs recordings of the suspect and foil voices are taken from police interviews
- We constructed six parades in our experiments taken from forensically-orientated speech databases
 - DyViS SSBE speakers (x3 parades)
 - YORVIS York speakers (x1 parade)
 - WYRED West Yorkshire (x2 parades)
- Randomly selected 15 voices for each parade, extracted 3s utterances and got listeners to rate the similarity of each voice pairwise
- Used MDS analysis to determine which set of 10 speakers had the shortest perceived 'distance' to the designated target (McDougall, 2013)

Speaker	Perceived similarity "distance"
Target S10	0.0
S1	1.81
S13	1.94
S5	2.7
S7	2.87
S9	2.91
S14	3.19
S4	3.29
S11	3.3
S3	3.3
S8	3.4
S6	3.41
S2	3.9
S12	3.94
S15	4.08





Experiment 1(N=271; 135 F)

- 2 (target presence: absent, present) X 3 (duration: 15s, 30s, 60s) between-subject design
- Outcomes: accuracy and self-rated confidence



- Recruited via Prolific.co (18-40, native English speakers, normal hearing)
- Hosted using *Gorilla.sc* (headphone checks, attention manipulations)



Overall accuracy



- Overall accuracy in line with previous outcomes (e.g. Smith et al., 2014, 2020; Kerstholt et al., 2004, 2006; Ohman et al., 2013)
- No statistically meaningful interactions
- Main effect of target presence
- No main effect of sample duration
- At-chance levels for TA parades





Sensitivity and Criterion



- Moderate evidence that
 15 & 60s have sensitivity >
 0
- Moderate evidence that
 15s has response criterion
 < 0 (liberal)
- No pairwise differences



Conclusion



- Consistent with previous research, performance is low (hit rate < 50%, high FA rate)
- No evidence to suggest that shorter voice sample durations are more distinctive than longer durations (i.e., temporal model; Brown et al., 2007)
- No evidence to suggest that there is any benefit in using lineup samples of 60 s. These initial results suggest that the voice identification procedure currently recommended in England and Wales can be safely adapted by reducing the duration to 30 s or even 15 s
 - In line with alternative guidelines which recommend shorter sample durations (Broeders & van Amelsvoort, 2001)





Experiment 2

- Does ID performance change if there are fewer voices in the parade?
- Fewer voices may reduce the overall cognitive load required of a listener
- Greater auditory attentional demands increases the risk of erroneous processing when comparing an incoming signal to long-term information (Zimmerman et al., 2016)
- From an applied perspective, the benefits would be similar to shorter durations (i.e., reduced resources required to prepare -> reduced delay between exposure and testing)





Experiment 2 (N=270; 136 F)



- Largely the same as E1 but we reduced the parade size to 6 instead of 9
- We removed the three least similar voices based on the MDS results
- We kept all 3 sample duration conditions to see if the pattern of responses changed when using fewer voices
- 2 (target presence: absent, present) X 3 (duration: 15s, 30s, 60s) betweensubject design

	Speaker	Perceived similarity "distance"
	Target S10	0.0
	S1	1.81
	S13	1.94
	S5	2.7
	S7	2.87
	S9	2.91
	S14	3.19
5	S4	3.29
	S11	3.3
	S3	3.3
	S8	3.4
	S6	3.41
	S2	3.9
	S12	3.94
	S15	4.08



Nottingham Trent University Psychology

Overall accuracy





- Slight numerical differences to the 9voice parade, but overall the same outcome pattern when adjusting for the increased chance level (1/7 vs 1/10)
- No statistically meaningful interactions
- Main effect of target presence
- No main effect of sample duration
- At-chance levels for TA parades





Sensitivity and Criterion



Sample Duration

- Moderate evidence that 15s and 60s d' above chance
- Moderate evidence that 30s and 60s have liberal criterion
- No pairwise differences



Conclusion



- Performance patterns are largely the same as when using a 9-voice parade (E1)
- No evidence that fewer voices reduces attentional demands to such an extent that it aids ID performance
- Slight numerical increase in TA performance, but still at chance-levels when adjusting for fewer voices (14.2% vs. 10%)
- Increased chance of an innocent suspect being randomly selected with no statistical improvement in accuracy suggests that Home Office recommendation of 9-voice parades should be maintained (as opposed to the shorter parade recommended by alternative guidelines)





Summing up

- Relatively poor performance overall, but
 - A difficult task (real voice parades would have greater heterogeneity)
 - Despite difficulty, still possible at above-chance levels when target is present
- Voice samples used in parades can be reduced from 60s to 30s or even 15s without adversely affecting identification performance
- Evidence suggests that 9-voice parades should be maintained over 6-voice parades to protect innocent suspects



Future directions



- We know that false alarm rates are high (a chance of an innocent suspect getting selected), so the next stage of our research programme focuses on mitigating this danger:
- Manipulating instructions to promote a criterion shift initial results of this experiment suggests a strong warning about the difficulty of the task is effective at reducing false alarm rate
- Alternative options for presenting the stimuli Home Office recommends a serial presentation, but what about options such as elimination, last-yes counts sequential, and short term repeated identification?







Economic and Social Research Council







Nottingham Trent University Psychology





Monday, 5 December 2022

Thanks for listening 🕲

References



Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. Applied Cognitive Psychology, 25(1), 112–118.

Van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). 'Earwitness' voice recognition: Factors affecting accuracy and impact on jurors. Applied Cognitive Psychology, 8(7), 661–677.

Smith, H. M. J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: Optimising witness performance. *Memory*, 28(1), 2–17. https://doi.org/10.1080/09658211.2019.1673427

McDougall, K. (2013) 'Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades.' International Journal of Speech, Language and the Law 20.2: 163-172.

Home Office. (2003). *Home Office circular* 057/2003: Advice on the use of voice identification parades. <u>http://webarchive.nationalarchives.gov.uk/20130308000037/http://www.homeoffice.gov.uk/about-us/corporate-publications-strategy/home-office-circulars-2003/057-2003/</u>

Broeders, A., & van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: Constructing a voice line-up. Problems of Forensic Sciences, 47, 237–245.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 38. https://doi.org/10.1037/0033-295X.114.3.539

Zimmermann, J. F., Moscovitch, M., & Alain, C. (2016). Attending to auditory memory. Brain Research, 1640, 208–221. https://doi.org/10.1016/j.brainres.2015.11.032

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. Psychologica Belgica, 54(3), 266-281.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(2), 187–197.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013a). Angry voices from the past and present: Effects on adults' and children's earwitness memory. Journal of Investigative Psychology and Offender Profiling, 10(1), 57–70.

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. Applied Cognitive Psychology, 16(8), 973–980. https://doi.org/10.1002/acp.920



Monday, 5 December 2022