Sound familiar?

Ady Cousins



John Wilding, Susan Cook *and* **Josh Davis**

present some curious findings from the field of 'earwitness' testimony.

IN 1935 Bruno Hauptmann was convicted of the kidnapping and murder of the infant son of the aviator Charles Lindbergh in the United States. Lindbergh claimed that Hauptmann's voice matched that of the man he had heard saying 'Hey, doctor, over here, over here!' when the ransom was paid three years earlier. This is the best-known example of voice identification in a legal case. There was virtually no relevant experimental evidence to aid the court, and the case prompted McGehee (1937) to undertake pioneering studies.

However, 60 years later a report in *The Guardian* (5 September 1997) suggested that British courts and police forces were still very uninformed on the issues:

> *…three Court of Appeal judges yesterday ordered a full hearing with leading counsel to explore the new police method of identifying suspects by 'voice parades'. They adjourned yesterday's hearing over a robbery conviction after the Crown's counsel said police forces were anxious for some guidance over voice identification parades.*

Research on 'earwitnessing' is meagre and unsystematic compared with the data available on eyewitnessing in general and face identification in particular. Sophisticated models have been developed that incorporate both visual identification of the individual and the extraction of visual information relating to emotional expression and linguistic utterance (lip reading). The role of the voice in identifying the person and emotions is acknowledged by adding some boxes and arrows to the models and assuming that the component processes operate in parallel with the visual channels. We argue that more complex models of person identification are needed, which may in turn have a bearing on the interpretation of witness evidence.

Factors affecting voice identification

We first survey the available findings and practical implications before considering theoretical issues.

Three preliminary points should be made. Firstly, recognition of familiar voices (friends or well-known personalities) and of previously unknown voices heard briefly have both been studied. Sometimes it is assumed that findings from studies of familiar voices are directly relevant to memory for unknown voices, but this is unjustified (Cook & Wilding, 1997a): matching an input to a poor trace from a single utterance differs from matching an input to a well-established template. The findings to be considered are concerned mainly with the former situation.

Secondly, research has employed either a design similar to a real-life identification situation (one or two voices, tested by line-ups after some delay) or a design convenient for a laboratory investigation, where many voices are employed and recognition is required after a brief interval. Results can differ markedly between these two situations (see below).

A third weakness arises from the difficulty in matching in the laboratory the stress level likely to be experienced by a witness to a crime, due to either personal threat or general emotional arousal. The most frequently cited evidence for a stress effect is the impairment in witness memory when a weapon is present in a scene; but Pickel (1999) has suggested that the effect occurs primarily when the weapon is an unexpected item, so it may be due to disruption of attention rather than stress. Most experimental findings may, therefore, only apply to situations

where a voice is heard in an apparently neutral context.

We divide the findings into factors relating to the original input, features of the line-up and characteristics of the witness.

Nature of the input Earlier research tended to conclude that increased length of the target utterance improved subsequent recognition only because it provided an increased variety of speech sounds. Using many voices and short delays before testing, Roebuck and Wilding (1993) controlled speech variety and length independently and found that only the former affected witness performance.

However, we have shown that only length was important when the method more closely resembled a police line-up (two voices, one male and one female, and a week's delay before testing with a line-up of six voices in each case) (Cook & Wilding, 1997a). We suggested that, when many voices are tested, variety helps in discriminating between them; when only two voices are tested, time to establish a template of the voice quality is more important. We also argued that the length effect was not due to increased opportunity to attend to voice characteristics with increased time, since instructions to attend to the voice did not modify the length effect.

Rating scales have been developed to assess distinctiveness of a voice (Yarmey, 1991) but Bricker and Pruzansky (1966) have argued that the features extracted may depend on what is said. Methods have also been devised for ensuring that line-ups are not biased by the inclusion of foils differing markedly from the target (see special issue of *Applied Cognitive Psychology*, vol. 13, November 1999).

Disguise by whispering or muffling the speech reduces identification drastically, even when spectrographic voice prints are available (e.g. Bull & Clifford, 1984; Reich & Duke, 1979). Changes in voice quality due to emotion of the speaker also reduce identification (e.g. Saslove & Yarmey, 1980), though the content of the message has no effect (Read & Craik, 1995).

We have explored the effect of providing several forms of context at the original utterance and restoring these at the line-up (Cook & Wilding, 1997b). Interestingly, when the speaker's face was visible at the first encounter the probability of a correct choice dropped from about half to one third. We called this the face overshadowing effect, which we discuss further below. Other forms of context (another voice, name, personal information) had no effect, nor did presenting the face again at the line-up.

As already suggested, no clear data are available on the effects of stress in the witness at the original exposure or at the line-up, because of problems in recreating ecologically valid and ethically acceptable parallels in the laboratory. Lawyers have strongly divided views on this issue. Deffenbacher (1983) cites an American survey showing that 82 per cent of defence attorneys believed that high arousal would reduce identification, while only 32 per cent of prosecution attorneys shared this view. Thus some resolution is urgently needed. One relevant finding from research in other contexts is that a narrowing of attention occurs under high arousal.

Line-up organisation and content As line-up length increases from four to eight voices, some decline in correct selections occurs (Clifford, 1980). If the target voice is later in the line-up, performance is worse (Doehring & Ross, 1972). Allowing a response of 'Don't know' or 'Target not present' reduces the number of false alarms (Warnick & Sanders, 1980). Performance seems to be relatively stable for line-ups presented during the first few days after encountering the target voice, with a decline thereafter (McGehee, 1937; Saslove & Yarmey, 1980).

As stated above, restoring context from the original encounter seems to have no effect. Also, Memon and Yarmey (1999) encouraged witnesses to think themselves back into the original situation (the 'cognitive interview' method, which is now widely employed in the USA) but found that this did not improve voice identification. Where beneficial effects of context on memory occur, this is because recreating context aids retrieval of stored information. If little information has been stored, as is likely with brief exposure to voices, and a recognition test is used that does not require the original input to be reproduced (i.e. retrieved), context effects are unlikely (Davies, 1988).

In most experiments the speech sample used at the line-up is the same as the original utterance and amounts to a single sentence. If the same words cannot be used (as may be the case in real situations), it is likely that performance will be worse, but no direct evidence is available. Nor have any principles been advanced for selecting speech samples to be used in the line-up. How important are variety and length, for example? Di Gregorio (1999) manipulated length by repeating the target sentence three times in the line-up (as done by Cook (1998) to increase the length of the original utterance). Substantial improvement occurred in identification performance, even though the original utterance had been heard only once.

The beneficial effects of repetition could

Ady Cousins

occur for several reasons. Repetition gives the witness an opportunity to remedy lapses of concentration or missing some of the words. The listener can also recheck a preliminary identification, focusing on critical features. In addition, some form of priming might occur, making the trace that does survive stronger or more accessible.

Witness characteristics Fragmentary evidence for gender differences has been reported. McGehee (1937) found a marked superiority in male listeners, but no other researchers have reported this. On the contrary, Roebuck and Wilding (1993) found that women were superior, but only when judging female voices. Exactly the same result appeared in the combined data of five experiments (728 participants) reported by Cook (1998). No obvious reason for this effect is apparent. Performance improves up to the age of 14 then levels off, though there may be a drop from age 10 before the rise continues up to 14 (Mann *et al*., 1979). Elderly listeners show reduced performance (Yarmey *et al*., 1984).

One study reported superior performance in listeners who had been born blind (Bull *et al*., 1983), and Shirt (1984) found that phoneticians tended to be better than untrained listeners, though some of the latter were as good as the best phoneticians. This suggests that the need to rely on voices as a primary source of identification and practice in analysing voice characteristics are important variables, but also that there may be other individual differences in ability.

However, there was no evidence in Cook's extensive data that participants who identified one voice correctly were more likely to be successful on the other voice than those who failed on the first voice. This finding suggests that there are no consistent individual differences in voice recognition ability other than those due to special experience.

Practical implications In summary, many factors affect accuracy of voice identification but few of them can be manipulated in a line-up. In general, voice identification is poor, particularly if the target utterance is brief or factors differ between the original situation and the line-up (muffling, emotion, etc.). Identification can be improved by repetition in the line-up, and erroneous selection of an innocent member of the line-up when the target voice is missing can be reduced by allowing 'Don't know' responses.

Theoretical issues

Few of the above findings have obvious implications for developing theories of person perception, apart from the face overshadowing effect (FOE). Voice processing is sometimes added as an independent channel in theories of face perception — but the FOE shows that these processes are not independent. The result is somewhat reminiscent of the finding of Schooler and Engstler-Schooler (1990) that verbally describing a face after seeing it reduced subsequent ability to recognise it, but the FOE differs from this effect in being an interaction between two simultaneous streams of information.

The FOE contradicts the expected effects of context, and the expectation that person recognition should combine face, voice and other information, rather than these sources interfering with each other. A possible explanation is that faces are dominant for person identification and speech is processed mainly for comprehension, unless it is the only source of person identification.

Cook (1998) tested this explanation. In one experiment, participants were instructed that voice was the feature of interest, but this did not affect the FOE. This implies either that attention could not be voluntarily redirected to the voice or that the explanation of the FOE in these terms is incorrect. However, longer exposure to face and voice and pre-exposure to the face before hearing any speech did reduce the FOE, suggesting that habituation to the face might permit more attention to be paid to the voice.

Abbott (1999) introduced a condition in which the face was covered by a stocking mask, predicting that this would remove the FOE because the face would not be identifiable. However, the FOE survived, probably because this type of mask still permits many useful facial features to be extracted.

A specific face-processing system has been postulated. One might expect a similar self-contained voice-processing system functioning independently and relatively automatically in the skilled adult. Since faces and voices of known acquaintances are associated with each other, information from both sources is combined at some stage of processing, together with additional facts. For adults, face information is clearly more important for identification, but the developmental literature suggests that this may not be true early in life. An examination of this literature may help to provide some clues to the operation of the adult system that produces the FOE.

ADY COUSINS

Neonates prefer to listen to recordings of their mother's voice rather than another female (DeCasper & Fifer, 1980), at an age when the visual system is unable to differentiate faces. Fifer and Moon (1995) found that at birth babies preferred a filtered version of the mother's voice that had the same acoustic properties as those available in the womb. DeCasper and Prescott (1984) also found that neonates could discriminate between their father's voice and that of a male stranger; and Fernald (1993) found that infants generally preferred listening to female rather than male voices.

All these results show an early ability to identify individual characteristics of voices. Though infants do prefer looking at their mother's face rather than a stranger's (Walton *et al*., 1992), their abilities are limited since obscuring the mother's hairline eliminates ability to discriminate

(Pascalis *et al*., 1995). Caron *et al*. (1988) suggest that in the early months more reliance is placed on the auditory system for identification.

Babies also rapidly develop abilities to match visual and auditory information on the basis of gender (Walker-Andrews *et al.*, 1991) or emotion (Walker-Andrews, 1986). Murray & Trevarthen (1985) showed that 6- to 12-week-old infants became distressed when viewing a video of their mothers in which the speech and visual content were discrepant. Hence voice characteristics are analysed early in life, and auditory and visual information combined.

Interactions between the different streams of information become more complex as comprehension of speech develops. Speech content and speaker identification are processed in different brain systems. In phonagnosia, speaker identification is impaired and speech comprehension intact (van Lancker *et al*., 1988) while in pure word deafness the reverse holds. Whereas speech comprehension is based in the left hemisphere of the brain, the right hemisphere is implicated in speaker identification (Hornack *et al.*, 1996; van Lancker *et al.*, 1988).

The speech comprehension and speaker identification systems co-operate when both are intact. Recognition of words as having been heard previously is facilitated when they are tested with the same voice as on the first encounter (Palmeri *et al*., 1993), or when they are spoken by a familiar voice (Nygaard *et al*., 1994). Whereas speaker identification requires idiosyncratic features to be extracted, speech processing requires these to be disregarded; knowing a speaker's individual characteristics will aid speech processing (Nygaard & Pisoni, 1998).

Obviously, fairly lengthy exposure to a speaker is needed before a 'template' of that speaker's individual characteristics can be constructed. Then, once a speech input has been identified as originating from a known speaker, the 'cleaning-up' process can be applied at an early stage in processing in order to aid word identification. Early processing of this kind tends to reveal its effects on memory indirectly, for example through speed and ease of subsequent perceptual identification (so-called *implicit* memory), rather than through *explicit* tests of recall or recognition (as in a line-up).

We are suggesting, therefore, that in the adult listener individual speech characteristics are extracted largely to identify a known speaker and to use this identification as an aid to speech perception. Though such information does become established in memory and added to the set of known voices after longer exposure to a previously unknown voice, no very detailed representation is likely to be extracted or retained after only a brief exposure. In Cook's experiments all speakers were unknown and the speech samples were short, so there was minimal opportunity for extracting features of the voice. Also the listeners knew beforehand that this would be the case.

Visual processing can interact with both these speech-processing systems. When a face is present, a more useful source of information about the speaker's identity than the voice becomes available, and the face also provides an aid to interpreting speech. Lipreading aids speech perception and can even override it (McGurk & MacDonald, 1976), so presence of the face makes it less important to attend to individual speech characteristics.

Whether both these effects are operating to produce the FOE can only be ascertained by further experimentation. The effect of pre-exposure observed by Cook (1998) does, however, suggest that the attentional effect is the main factor. Interference from lipreading should occur whenever the face is present during the speech, but Cook found that, if the face was exposed before the speech began, then continued to be present during the speech, no FOE occurred. This suggests that lipreading does not contribute to the FOE.

Conclusions

We suggest, therefore, that individual characteristics of a speaker's voice may be extracted primarily to aid speech perception rather than individual identification. Use of the voice for identification is further reduced when the speaker's face is available, but the dominance of the face

is reduced after a few seconds exposure.

This argument implies that the elaboration of models of face processing into models of person perception that are applicable to witnessing situations cannot be achieved simply by adding parallel voice and speech-perception systems. These different streams of information interact with each other and the nature of these interactions may vary depending on other aspects of the situation.

We have indicated some ways in which interactions occur. To improve our understanding of the conditions that affect voice identification (and face identification and speech perception) in witnessing situations, more detailed investigations of these interactions will be needed. From the legal point of view, our results suggest that extreme caution is needed when using voice identification by witnesses in legal contexts, unless the voice is already familiar or a fairly long utterance occurs.

■ *Dr John Wilding, Dr Susan Cook and Josh Davis are at the Department of Psychology, Royal Holloway, University of London, Egham, Surrey TW20 0EX. Tel: 01784 443519; e-mail: j.wilding@rhbnc.ac.uk.*

References

Abbott, C. (1999). Investigating the face overshadowing effect in earwitness testimony. Unpublished student project, Department of Psychology, Royal Holloway, University of London.

Bricker, P .D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40*,1441–1449.

Bull, R., & Clifford, B.R. (1984). Earwitness voice recognition accuracy. In G. L.Wells & E. F. Loftus (Eds.), *Eyewitness testimony, psychological perspectives* (pp. 92–123). Cambridge: Cambridge University Press.

Bull, R., Rathborn, H., & Clifford, B. R. (1983). The voice-recognition of blind listeners. *Perception, 12*, 223–226.

Caron, A. J., Caron, R. F., & MacLean, D. J. (1988). Infant discrimination of naturalistic emotional expressions: The role of face and voice. *Child Development, 59*, 604–616.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behaviour, 4*, 373–394.

Cook, S. A. (1998). *Earwitness testimony: Length effects, familiarity effects and the role of context with special reference to faces*. Unpublished PhD. thesis, University of London.

Cook, S., & Wilding, J. (1997a). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*, 95–111.

Cook, S., & Wilding, J. (1997b). Earwitness testimony 2: Voices, faces and context. *Applied Cognitive Psychology, 11*, 527–541.

Davies, G. (1988). Faces and places: Laboratory research on context and face recognition. In G. M. Davies & D.M. Thomson (Eds.), *Memory in context: Context in memory* (pp. 35–53). New York: Wiley.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mother's voice. *Science, 208*, 1174–1176.

DeCasper, A. J., & Prescott, P. A. (1984). Human newborns' perception of male voices: Preference, discrimination and reinforcing value. *Developmental Psychobiology, 5*, 481–491.

Deffenbacher, K. A. (1983). The influence of arousal on reliability of testimony. In S. M. A. Lloyd-Bostock & B. R.Clifford (Eds.), *Evaluating witness evidence* (pp. 235–243). Chichester: Wiley.

Di Gregorio, L. (1999). The effect of repetition of unfamiliar voices upon the identification of speakers. Unpublished student project, Department of Psychology, Royal Holloway, University of London.

Doehring, D. G., & Ross, R. W. (1972). Voice recognition by matching to sample. *Journal of Psycholinguistic Research, 1*, 233–242.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development, 64*, 657–674.

Fifer, W. P., & Moon, C. M. (1995). The effects of fetal experience with sound. In J. P. Lecanuet, W. P. Fifer, N. A. Krasnegor & W. P. Smotherman (Eds.), *Fetal development: A psychobiological perspective* (pp. 351–366). Hillsdale, NJ: Lawrence Erlbaum.

Hornack, J., Rolls, E. T,. & Wade, D. (1996). Voice and face expression identification in patients with emotional and behavioural changes following ventral lobe damage. *Neuropsychologia, 34*, 247–261.

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology, 17*, 249–271.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology, 27*, 153–165.

Memon, A., & Yarmey, A. D. (1999). Earwitness recall and identification: Comparison of the cognitive interview and the structured interview. *Perceptual and Motor Skills, 88*, 797–807.

Murray, L., & Trevarthen, C. (1985). Emotional regulation of interactions between two-month-olds and their mothers. In T. M.Field & N. A. Fox (Eds.), *Social perception of infants*. Norwood, NJ: Ablex.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*, 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning Memory and Cognition, 19*, 309–328.

Pascalis, O., de Schonen, S., Morton, J., Deruelle, C., & Fabre-Grenet, M. (1995). Mother's face recognition by neonates: A replication and extension. *Infant Behaviour and Development, 18*, 79–85.

Pickel, K. L. (1999). The influence of context on the 'weapon focus' effect. *Law and Human Behaviour, 23*, 299–311.

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*, 6–18.

Reich, A., & Duke, J. (1979). Effects of selected vocal disguises upon speaker recognition by listening. *Journal of the Acoustical Society of America, 66*, 1023–1028.

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology, 7*, 475–481.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65*, 111–116.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36–71.

Shirt, M. (1984). An auditory speaker recognition experiment. *Proceedings of the First Police Conference on Applied Speech and Tape Recording Analysis*, 71–74.

van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195–209.

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviours: Relation of eye and voice. *Developmental Psychology, 22*, 373–377.

Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants bimodal perception of gender. *Ecological Psychology, 3*, 55–75.

Walton, G. E., Bower, N. J. A., & Bower, T. G. R. (1992). Recognition of familiar faces by newborns. *Infant Behaviour and Development, 15*, 265–269.

Warnick, D. H., & Sanders, G. S. (1980). Why do eyewitnesses make so many mistakes? *Personality and Social Psychology Bulletin, 8*, 60–67.

Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society, 31*, 421–428.

Yarmey, A. D., Tressillian Jones, H. P., & Rashid, S. (1984). Eyewitness memory of elderly and young adults. In D. J. Muller, D. E. Blackman & A. J.Chapman (Eds.), *Psychology and law* (pp. 215–228). Chichester: Wiley.