

Identification of voices in shouting

Helen Blatchford and Paul Foulkes

Abstract

Two experiments were carried out to assess the ability of lay listeners to identify familiar voices from shouted samples. The experiments were conducted after a murder case in which a witness claimed to recognise the voice of a masked gunman based on two shouted words: get him! Our experiments were designed to explore the extent to which a listener can identify a known voice from a shouted sample, and also whether a two word sample is sufficient for identification. Recordings were obtained from a group of nine females who formed a close social network, plus six foils. Two shouted utterances were extracted for listening tests: for Test 1, 'get him!', and for Test 2, 'face down on the ground and hands behind your back now!' Thirteen listeners from the same social network participated in the tests. Listeners correctly identified familiar speakers in 52% of cases in Test 1 and 81% in Test 2. There was considerable variation in results across listeners, and also with respect to individual voices. The results suggest that recognition of shouted voices is far from perfect, even in closed tests carried out among a close network. The variability in performance emphasises the need to subject a witness's ability to identify a voice in a forensic case to formal testing wherever this is feasible.

KEYWORDS: LAY WITNESS, SPEAKER IDENTIFICATION, SHOUTING, FAMILIAR VOICES

Affiliations

Helen Blatchford: City University, London

Paul Foulkes: J. P. French Associates and University of York

Correspondence: Paul Foulkes, Department of Language and Linguistic Science, University of York, York YO10 5DD, UK

email: pf11@york.ac.uk

Introduction

In 2003 two young women, Letisha Shakespeare and Charlene Ellis, were shot and killed in a drive-by shooting in Birmingham, UK. The incident was apparently gang-related, although the women were not the intended victims. The car from which the shots were fired continued along the road, and further shots were fired at Leon Harris, who at the time was sitting in a parked car. Harris got out of the car and ran away. As he did so the masked gunman shouted 'get him!' Harris gave a description of the characteristics of the voice and named Tafarwa Beckford as the perpetrator. The two had spent time together in a young offenders' institution some time previously, but that appears to have been their only form of contact.

J.P. French Associates were subsequently engaged by Beckford's defence team to produce a report on Harris's evidence (French and Harrison 2005). Analysis was undertaken to assess whether Harris's description was a good match for Beckford's voice. Tests were also made of the reliability of the identification given Harris's fairly limited experience of Beckford's voice, and the fact that it was made on the basis of only two shouted words. Eventually Harris's evidence was withdrawn by the prosecution. Following problems with other evidence against Beckford the jury were instructed to return a verdict of not guilty for him.

The experiments we report in this article were carried out to explore some of the key issues in the Beckford case in a more general context. Our aim was to provide further evidence on the reliability of witness identification of voices when hearing short and shouted samples. We begin with a brief summary of previous research on speaker identification by lay listeners, focusing on the effects of familiarity with the speaker, sample length, and shouted stimuli. We then describe the experiments we performed, before finally drawing conclusions on the forensic implications of our results.

Speaker identification by lay listeners

Speaker identification is a task performed as a matter of routine, for example when answering the telephone or hearing voices outside one's office. However, it is a well known fact, at least among linguists, that the task of identifying a familiar voice is not always performed successfully. When answering the phone, for example, we may fail to recognise the voice of someone we know, we may mistakenly believe we recognise the voice of a stranger, or we may attribute a familiar voice to someone else we know rather than to the actual speaker.

It is of great forensic importance that we should understand how reliable listeners are at speaker identification, and that we isolate the factors which have an impact on the identification process, in order that courts can assess

evidence such as that described in the Beckford case earlier. Evidence suggests that lay-people – thus potential jurors – overestimate the ability of listeners to identify familiar voices and to reject unfamiliar ones (Yarmey *et al.* 2001, Yarmey 2004). In reality everyone makes mistakes from time to time, for a number of reasons (reviewed by Bull and Clifford 1984). Contributory factors to variation in performance include the age and hearing ability of the listener, the length of delay between initial exposure to a voice and the identification task, the nature and extent of any disguise in the voice sample, and the degree to which the listener is familiar with the voice. The two factors of particular relevance here are the duration and nature of the sample upon which identification is performed.

With respect to duration, there is some debate as to whether identification rates correlate with length of the stimulus. Bull and Clifford (1984: 108) conclude that duration is not particularly important provided that ‘at least one sentence’ is presented. Studies which have used long samples report relatively high rates of successful identification of familiar voices. In a closed identification task using samples of 2.5 minutes, for example, Hollien *et al.* (1982) found that a group of ten listeners identified familiar voices in 98% of cases. Far lower rates have been reported when very short stimuli have been used. For example, Ladefoged and Ladefoged (1980) report a success rate of 31% in a test using the single word *hello*, in which Peter Ladefoged (in)famously failed to recognise his own mother. Varied sample lengths of similar sounding voices were used in a set of tests by Rose and Duncan (1995). Their results support the conclusion drawn by Bull and Clifford (1984). When the stimuli consisted of *hello*, correct identification averaged 57%. Identification rates were 85% when stimuli contained either a four word sentence or 45 seconds of speech.

Identification rates are further affected by the nature of the sample, including its specific content, the medium of transmission, and whether it consisted of modal or non-modal speech. Yarmey (2004) carried out a series of closed tests using four speakers. He found that short samples consisting of linguistic material elicited higher rates of identification than non-linguistic ones such as laughter, coughing or moaning. Several studies, including Rose and Duncan (1995) and Foulkes and Barron (2000), have noted that listeners are able to identify some voices more readily than others in the same test. This suggests that listeners attend to salient acoustic or linguistic cues in the samples such as reflexes of voice quality or regional accent. The medium of transmission has also been shown to affect identification. In particular, transmission through a telephone results in lower performance levels compared with tests using speech that has been recorded under optimal conditions (Rathborn *et al.* 1981). In a study with an open task and nine second samples of telephone speech, Foulkes and Barron (2000) found identification rates to average 68%.

Finally, rather little research has been carried out on the accuracy of speaker identification when the sample of speech is shouted. While this was not a central concern in the study by Yarmey (2004), it is noteworthy that his participants scored 47% with the shouted utterance *help me!* compared with 68% for *hello* spoken normally. However, it has been established that the acoustic characteristics of shouting differ systematically in a number of ways from those of normal speech (Rostolland 1982, 1985, Traunmüller and Erickson 2000). For example, segmental durations, formant values and suprasegmental features are all affected by a change from modal speech to shouting. In short, therefore, shouting is not simply 'loud speech'. It has been suggested that shouted voices are less discriminable due to increased uniformity in f_0 compared with modal speech (Rostolland 1982). We can therefore also hypothesise that listeners who are familiar with a particular voice might not be as accurate in identifying that voice when they hear it in a different mode of vocal effort such as shouting. Support for this hypothesis is provided by Brungart *et al.* (2001). They trained listeners to identify eight previously unfamiliar voices, four male and four female, in three modes of speech: normal conversation, whisper and shouting. When listeners heard test stimuli in the same mode as the training stimuli, correct identifications were high (90%-86% for conversational speech, 87%-95% for shouting, 69-67% for whisper; the first figure in each case for male talkers and the second for females). Identification rates were far lower when the test stimuli were in a different mode from that used in training. For example, when shouted stimuli were given to listeners who had been trained with conversational material, identification rates dropped to 70% for male talkers and 43% for females.

In summary, previous research has shown that short samples can pose a serious problem for speaker identification. Less clear is the effect of shouted samples, but the available evidence also indicates that shouting may impair listener performance. We turn now to our own experimental investigation. In light of the evidence in the Beckford case, the main issue to be addressed was the accuracy of speaker identification by lay listeners when presented with short, shouted samples of familiar voices.

Experimental design

Participants

Previous studies have shown that degree of familiarity affects listener performance in identification tasks (e.g. Hollien *et al.* 1982). All things being equal, the more familiar the voice the better the identification rate. Some studies have controlled for this factor by training listeners to identify a set of voices

which were, prior to the test, unfamiliar. However, while this strategy gains in comparability across listeners, it loses in terms of the naturalness of the task. We therefore elected to recruit a pre-existing group of participants to act as speakers and listeners. The group, henceforth referred to as the network, consisted of 14 (nine female, five male) final-year undergraduate students of French at the University of York. This network, which included the first author, comprised the entire cohort for their year. At the time of the experiment the network members had known each other for three and half years. They had studied as a group for at least four hours a week for the duration of their acquaintance. They had also formed close social bonds, with all members of the group having had extensive social interaction with all other members. All were native British English speakers, from a variety of regional backgrounds. None reported any history of speech or hearing problems. Most had studied linguistics and phonetics as part of their degrees but none (apart from the first author) had specialised in subjects directly relevant to the task at hand.

Naturally the social contacts between some members were stronger than others, but all appeared much more firmly established than that between Leon Harris and Tafarwa Beckford. We were therefore confident that our network members could justifiably be described as having a high degree of familiarity with each others' voices.

Speakers

The nine female members of the network were asked to provide materials to be used in the listening tests. Six female foils were also recruited who were unknown to any of the participants other than the first author. The foils were all native English speakers, and included two Americans.

Listeners

Thirteen participants undertook the listening tests: eight of the nine women who had acted as shouters, and the five male members of the network. The only network member who did not participate, for obvious reasons, was the first author.

Voice samples

All voice samples were recorded in a sound-proof studio with a Neuman V87 microphone and Adobe Audition software at a sample rate of 44.1 kHz. The speakers were first asked to read aloud both a short text and a list of sentences, three times and in a normal voice. The sentences included *get him!*, the phrase

used in the Beckford case, and a number of longer items. All sentences were constructed to be feasible in a criminal context. The act of reading at a normal level was designed to enable the speakers to get accustomed to the materials and to the setting of the recording laboratory. They were then asked to shout the sentences as loudly as possible, standing approximately one metre away from the microphone.

Some participants showed understandable reluctance to shout on cue (and a seventh potential foil in fact refused to do so, withdrawing from the experiment). They were therefore asked to repeat the exercise until the administrators were happy that the shouts were of a reasonably natural amplitude. Successful elicitation was aided by the fact that most participants were close friends of the first author, who in some cases demonstrated the appropriate level of shouting in order to help participants lose their inhibitions. The material eventually obtained undoubtedly varied across participants in terms of vocal effort, but we were satisfied that all materials could be classified as genuine shouts.

Sound levels were monitored during the practice shouts in order to set the recording levels as high as possible without circuit overload. Once the recording levels had been adjusted participants were recorded shouting each sentence three times.

Listening tests

Sections of the recordings were extracted for use as stimuli in two listening tests. The original .wav sound files were edited using Audacity 1.2.4 software. For Test 1, tokens of the phrase *get him!* were extracted. Test 2 was designed to assess the effect of a longer sample for comparison. The twelve syllable item *face down on the ground and hands behind your back now!* was chosen for this purpose. For each test 39 tokens were obtained: three each from the nine network members and two each from the six foils.

Stimuli were not normalised further for amplitude, on the reasoning that variability in the level of shouting might have served as a cue to speaker identity. However, exploratory analysis of responses to the stimuli with highest and lowest amplitude revealed no distinct patterns. Given the limited sample of stimuli and number of voices available in these experiments, we elected not to consider relative amplitude any further.

For presentation to listeners the stimuli were compiled into a PowerPoint file. The use of PowerPoint allowed them to be played automatically, supported by visual information on the structure of the test and the identification number of the stimulus being heard. The playback order was random, except that no adjacent pair was taken from the same speaker. A delay of eight seconds was

inserted between each stimulus. A set of six stimuli was also used to provide a practice run before listeners participated in the main tests. Listeners were given an answer sheet containing the nine names of the possible shouters. They were also offered two other choices: *unknown* to indicate an unfamiliar speaker, and *don't know*. The nature of the task was explained at the beginning of the PowerPoint presentation. Listeners participated individually in the tests via a Dell Inspiron 510m laptop and Sony MDR-V250 headphones in a quiet room. Test 1 was administered three weeks after the recordings were made, and Test 2 was carried out two weeks after Test 1.

The tests were therefore closed, which marks a departure from the forensic scenario in which Leon Harris found himself. However, the nature of the network meant that members of the group were aware that others were participating in the recordings and it was therefore not practicable to construct an open test.

Results

A summary of the results is provided in Table 1. Further details are given in the discussion that follows.

Table 1 Summary of results

	Syllables	Familiar shouters			Foins	
		Correct %	Range for listeners	Range for voices	% rejected	% false positives
Test 1	2	52	22–96%	19–83%	27	45
Test 2	12	81	48–100%	31–100%	64	19

Test 1: *get him!*

As indicated in Table 1, the overall proportion of correct responses to familiar voices was 52% in Test 1. There was, however, considerable variation across listeners. The results for the 13 individual listeners are given in Figure 1. The shaded bar to the right represents the overall mean. The results for eleven of the listeners were close to the overall mean, at c. 40–60%. One listener, A, registered an almost perfect performance, with just one error for the 27 stimuli. Listener F, on the other hand, scored only 22% (6/27).

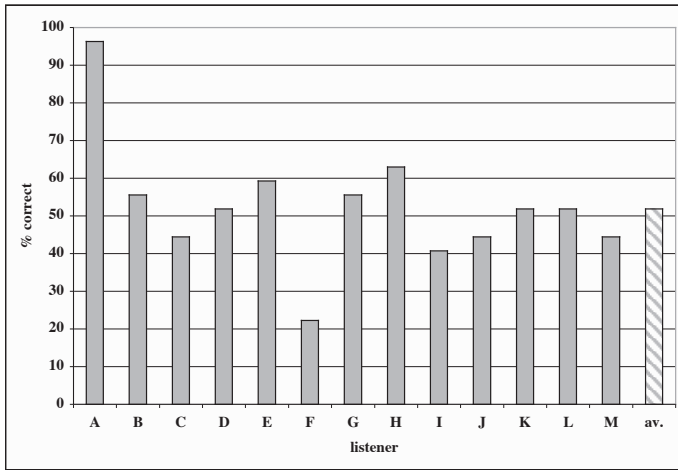


Figure 1 Results for individual listeners, Test 1 (n = 27 per listener, total N = 348).

Results also varied for individual voices. Figure 2 shows the relative proportion of correct responses for each voice, with responses from all listeners pooled (black bars). Also shown are the proportion of false positives (i.e. incorrect naming of another network member) and the proportion of ‘don’t know’ responses. The rightmost bar represents the mean scores for all listeners. It is apparent that the voices can be split into two groups. Voices 1–5 were identified more accurately than voices 6–9. Correct responses for voices 1–5 were at least 60%, and as high as 83%, while voices 6–9 were identified in at most 33% of cases. False positives and ‘don’t know’ answers were thus also higher for the latter group. Across all nine voices 33% of the responses were false positives.

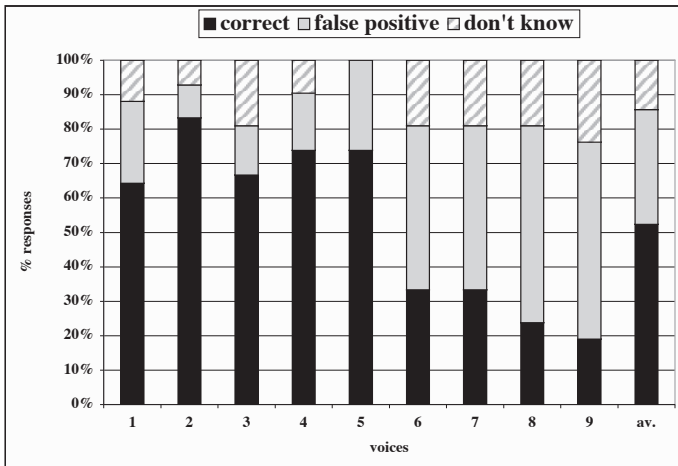


Figure 2 Results for individual voices, Test 1 (n = 42 per voice, total N = 348).

Foils were rejected in only 27% of cases. The best performance by a listener was to reject 75% of foils (9/12). One particular foil was rejected much more frequently (61%) than the other five (7–32%). Almost half the foils (45%) were incorrectly attributed to network members. The two American foils were misidentified just as frequently as the British foils, and of particular note is the fact that one foil was incorrectly identified as the same network member no less than 15 times.

Test 2: *Face down on the ground and hands behind your back now!*

The performance of listeners improved markedly in Test 2, in which they heard the longer samples. Correct identifications rose to 81% compared to the 52% found in Test 1. Figure 3 shows the results for all listeners individually (black bars). The scores from Test 1 are repeated for the sake of comparison (grey bars). It can be observed that 11 of the 13 listeners made a higher number of correct identifications in Test 2. The exceptions to this were A (the best performer in test 1) and L, but the performance of these two did not fall significantly. Six listeners showed a statistically significant improvement in performance, as indicated by asterisks (comparisons made with chi square tests). Variation across listeners was again observed, however. One listener, C, attained 100% correct identifications, eight managed scores over 80%, but the worst performance (L) was 48%.

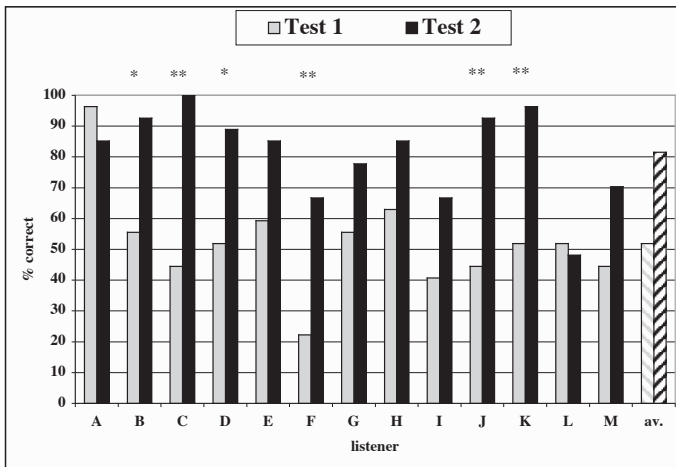


Figure 3 Results for individual listeners, Tests 1 and 2 ($n = 27$ per listener, total $N = 348$; significant differences across tests: * $p < .05$, ** $p < .01$).

Figure 4 illustrates responses to the nine familiar voices. For this test we can observe that seven of the voices were handled well by listeners. Two voices (2 and 5) were correctly identified in all instances. Scores for voices 8 and 9 were markedly lower than those for the rest of the network. False positives accounted for 15% of responses overall.

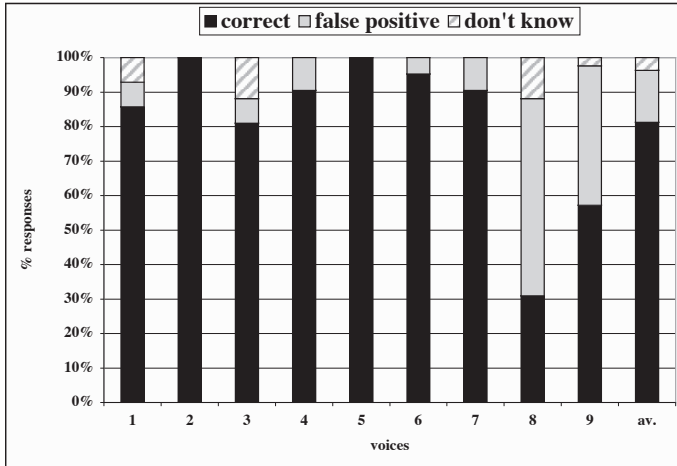


Figure 4 Results for individual voices, Test 2 ($n = 42$ per voice, total $N = 348$).

The rejection rate for foils showed a substantial increase in Test 2, to 64%, with 19% false positives. One listener managed to reject all foil stimuli. The American foils were this time the most consistently rejected (89% and 93% compared with 18–71% for the others).

Discussion

The proportion of correct identifications of familiar voices in Test 1, 52%, was well above chance (11%, given that there were nine candidates), and closely comparable to the results reported in similar studies where short and/or shouted stimuli were used (e.g. Rose and Duncan 1995, Yarmey 2004). Results for Test 2 were predictably better. Again, the proportion of correct responses was comparable to that reported in similar studies (e.g. Bull and Clifford 1984, Rose and Duncan 1995). The longer samples presumably provided listeners with more material to process in order to make their decisions. The difference in duration enabled several listeners to register significant improvements in their performance with familiar voices, and also led to a far higher rejection rate for foils.

To focus solely on mean scores in the tests, however, is to underplay the complexity apparent in the findings. The results suggest that robust identification *can* be made on the basis of short, shouted samples, but successful identification depends very much on both the listener and the voice. Even in Test 1, where the material presented was simply *get him!*, one listener managed to identify the shouters in all but one case. In Test 2 one listener scored perfectly and three others made only one or two errors with 27 stimuli. Two voices in Test 2 were correctly identified in all instances by all listeners.

Two particular factors can be isolated which may contribute to the variation in results. First, we should recall that although participants were undoubtedly familiar with each others' voices, the degree of familiarity was inevitably variable across the network. It could be observed informally that the best performer in Test 1 was a particularly prominent member of the network who, by her own account, based her social life around the group. The best recognised voices also belonged to particularly popular members of the network, while the two voices which proved most difficult to identify (8 and 9 in Test 2) were those of relatively peripheral group members. Degree of familiarity might therefore be of relevance in assessing listener performance, as other studies have concluded (e.g. Hollien *et al.* 1982).

A second important factor in explaining the results is the relative salience of individual features in the voices tested. The better recognised voices were those which contained marked sociolinguistic or regional features relative to the rest of the group (cf. Rose and Duncan 1995, Foulkes and Barron 2000). The participant who was most frequently identified in both tests was originally from Liverpool, a city renowned for its distinctive accent (Wells 1982). She had a very noticeable regional accent compared with the other participants. In Test 2 the other voice identified accurately in 100% of cases contained a strongly labiodental realisation of /r/, apparent in the word *ground*, which is an emergent variant in many dialects in England (Foulkes and Docherty 2000). Note also that the American foils were consistently rejected in Test 2.

Concluding comments

The experiments described here were designed to explore the accuracy and reliability of listeners in identifying familiar voices from shouted samples. As expected, the short duration and non-modal nature of the stimuli created problems for a group of listeners who knew each other well, although the samples did not present an insurmountable obstacle to listeners. However, an overall rate of 52% in Test 1 is of course well below the level of reliability that could reasonably be tolerated by a court in a case, like that against Beckford, where identification evidence of this type is central. Even the correct rate of

81% in Test 2 surely still falls below a judicially acceptable level (cf. Künzel 1994, Rose and Duncan 1995). Familiarity with a voice in everyday interaction therefore does not guarantee successful identification of that voice when it is heard shouting.

Although the experiments were conducted in light of the Beckford case, we refrain from making direct comment on the evidence which was to be given by Leon Harris since there are potentially significant differences between the tasks carried out in our listening tests and the identification made in the original incident.

First, we cannot accurately assess how much variation there was in degree of familiarity with the voices concerned. However, our participants certainly knew each other well, and from what is known about the relationship between Harris and Beckford it is likely that our network members had had considerably longer and richer experience of each others' voices than Harris had of Beckford's.

Secondly, our results indicated that there was substantial variation in how well individual voices were identified. Some voices were identified more accurately and more consistently than others, which in part may be due to the specific acoustic, phonetic and linguistic characteristics of the voice concerned. It remains a moot point whether the salience of certain characteristics is a relative phenomenon – i.e. some features may make a voice stand out relative to others in a particular community of speakers – or whether there are features which have universal salience to listeners. Either way, we cannot be sure whether Harris would find Beckford's voice more or less easily identifiable than any other. (Harris did in fact claim that the assailant's voice was distinctive, and described a number of features of the voice which underpinned his identification of Beckford, but the investigation by French and Harrison (2005) raised questions about the distinctiveness of these features relative to Beckford's peer group.)

Thirdly, our experiments also showed considerable variation in the ability of listeners to perform the task. Some did very well, even with the stimulus *get him!* On the other hand, even with the longer sample in Test 2 one participant still identified fewer than half of the voices.

Previous research on voice identification has highlighted the range of factors that can affect performance. Although we also set out to investigate some of these factors, in terms of varied duration and the use of shouting, we consider that the most important finding from the forensic perspective is the variation in the results. Listeners displayed highly divergent abilities to perform the identification tasks. Furthermore, some voices proved easier to identify than others. Such variation means that the ability of a witness to identify a voice cannot be taken on trust.

Naturally, caution must be exercised when extrapolating the results of a controlled experiment using a closed test in ideal listening conditions to a real forensic situation. Harris heard his assailant's voice in a street and under what must have been highly stressful circumstances. The chances of his correctly identifying his assailant under those conditions cannot be predicted directly on the basis of experimental data. The variation inherent in our results would allow us only to predict a listener's ability in an identification task within a relatively broad range. This emphasises the need to subject a witness's ability to identify a voice to formal testing whenever it is feasible to do so, particularly when the sample at issue was short or shouted.

Acknowledgements

Versions of this material were presented at the 2005 IAFPA conference in Marrakech, the 2006 BAAP colloquium in Edinburgh, and the 2006 Sociolinguistics Symposium in Limerick. We are grateful to colleagues at those meetings, as well as two reviewers, for their comments. We also record our thanks to Peter French and Philip Harrison for sharing their data and report concerning the Beckford case, and to Huw Llewellyn-Jones for technical assistance.

References

- Bull, R. and Clifford, B. R. (1984) 'Earwitness voice recognition accuracy', in G. L. Wells and E. F. Loftus (eds) *Eyewitness Testimony: Psychological Perspectives*. Cambridge: Cambridge University Press, 92–123.
- Brungart, D. S., Scott, K. R. and Simpson, B. D. (2001) 'The influence of vocal effort on human speaker identification', *Eurospeech*, 2001: 747–50.
- Foulkes, P. and Barron, A. (2000) 'Telephone speaker recognition amongst members of a close social network', *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 7: 181–98.
- Foulkes, P. and Docherty, G. J. (2000) 'Another chapter in the story of /r/: 'labiodental' variants in British English', *Journal of Sociolinguistics*, 4: 30–59.
- French, P. and Harrison, P. T. (2005) 'Lay-witness voice description and identification: consideration of shouting, pitch and accent', paper presented at paper presented at the IAFPA conference, Marrakech, Morocco.
- Hollien, H., Majewski, W. and Doherty, E. T. (1982) 'Perceptual identification of voices under normal, stress, and disguise speaking conditions', *Journal of Phonetics*, 10: 139–48.
- Künzel, H. (1994) 'On the problem of speaker identification by victims and witnesses', *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 1: 45–58.

- Ladefoged, P. and Ladefoged, J. (1980) 'The ability of listeners to identify voices,' *UCLA Working Papers in Phonetics*, 49: 43–51.
- Rathborn, H., Bull, R. and Clifford, B. R. (1981) 'Voice recognition over the telephone,' *Journal of Police Science and Administration*, 9: 280–84.
- Rose, P. and Duncan, S. (1995) 'Naive auditory identification and discrimination of similar voices by familiar listeners,' *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 2: 1–17.
- Rostolland, D. (1982) 'Acoustic features of shouted voice,' *Acustica*, 50: 118–25.
- Rostolland, D. (1985) 'Intelligibility of shouted voice,' *Acustica*, 57: 103–21.
- Traunmüller, H. and Erickson, A. (2000) 'Acoustic effects of variation in vocal effort by men, women, and children,' *Journal of the Acoustical Society of America*, 107: 3348–51.
- Yarmey, A. D. (2004) 'Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations,' *International Journal of Speech, Language and the Law*, 11: 267–77.
- Yarmey, A. D., Yarmey, A. L., Yarmey, M. J. and Parliament, L. (2001) 'Commonsense beliefs and the identification of familiar voices,' *Applied Cognitive Psychology*, 15: 283–99.
- Wells, J. C. (1982) *Accents of English*. (3 vols.) Cambridge: Cambridge University Press.