Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity

Kirsty McDougall, Francis Nolan and Toby Hudson

Department of Theoretical and Applied Linguistics, University of Cambridge

kem37@cam.ac.uk, fjn1@cam.ac.uk, toh22@cam.ac.uk

Short title: Telephone transmission and earwitnesses

Address for correspondence: Dr Kirsty McDougall
Department of Theoretical and Applied Linguistics
University of Cambridge
Sidgwick Avenue
Cambridge CB3 9DA
United Kingdom

Telephone: +44 1223 760822
Fax: +44 1223 335062
Email: kem37@cam.ac.uk

Abstract

The effect of telephone transmission on a listener's ability to recognise a speaker in a voice parade is investigated. 100 listeners (25 per condition) heard one of five 'target' voices, then returned a week later for a voice parade. The four conditions were: target exposure and parade both at studio quality; exposure and parade both at telephone quality; studio exposure with telephone parade; and vice versa. Fewer correct identifications followed from telephone exposure and parade (64%) than from studio exposure and parade (76%). Fewer still resulted for studio exposure/telephone parade (60%), and, dramatically, only 32% for telephone exposure/studio parade. Certain speakers were identified more readily than others across all conditions. Confidence ratings reflected this effect of speaker, but not the effect of exposure/parade condition.

Keywords: earwitness evidence, telephone transmission, voice identification, voice line-ups, voice parades

Phonetics has taken as its central object of study the realisation of phonological and (to a lesser extent) paralinguistic contrasts. In this enterprise, the fact that the producers of the speech signal – the members of a speech community – are heterogeneous, provides a source of what is often regarded as unwanted noise. In another application of phonetics, however, those very speaker-specific characteristics and their perceptual processing by listeners are central. This is the application of phonetics to forensic investigations where speaker identity is at issue. Modelling the individuality of voices is still at a relatively early stage, as is our understanding of the factors which affect listeners' ability to respond to speaker-specific factors in the voice. In this paper, we focus on a question within the latter realm, and examine the effect on voice recall of information loss in the acoustic signal.

In certain crimes, a perpetrator may be heard but not seen. For example, the perpetrator may be masked, or the victim blindfolded, or it might have been too dark for a witness to see clearly, yet the voice of the criminal was heard. In such cases, 'earwitness' evidence may be obtained with the help of a 'voice parade'. The witness is asked whether he or she can pick out the voice heard at the time of the crime from a line-up of recordings which includes a suspect's voice and a number of 'foil' voices. As with visual parades, the main point of using a parade or 'line-up' rather than just confronting the witness with the suspect (visually or auditorily) is to safeguard an innocent suspect. Confronting the witness with a suspect inevitably risks biasing the witness toward identification. Whilst in the case of a parade it is still possible that a witness who is keen to help and feels pressure (despite instructions to the contrary) to pick someone will pick an innocent suspect, a fair parade – one in which guesses should be evenly distributed over the participants – at least affords the innocent suspect considerable statistical protection.

Voice parades are probably associated with more pitfalls than visual parades. The second author has three times evaluated parades constructed ‘in house’ by the police where it was perfectly possible to pick out the suspect with no knowledge of his voice. This was for the simple reason that the suspect’s sample was spontaneous speech, and the foils were reading the same words. Other pitfalls include overlooking important differences of regional or social accent between the suspect and the foils, and failing to match individual properties of the voice such as pitch or resonant frequencies. The multidimensional nature of speech makes achieving a fair parade a challenging task. For this reason, it has become common practice for phoneticians to be asked to assist in the creation of voice parades.

Forensic phoneticians, in consultation with psychologists and law enforcement officers, have put considerable effort into devising and refining procedures for constructing fair voice parades (e.g. Broeders and Rietveld, 1995; Hollien, 1996; Künzel, 1994; Nolan, 2003; Nolan and Grabe, 1996; Rietveld and Broeders, 1991). In the UK, the procedure set out in Home Office (2003) and Nolan (2003) still provides a template for the construction of voice parades, albeit modified in some details to accommodate technological advances. However, ensuring as fair as possible a parade is only one aspect of concern. Much of the reliability of earwitness evidence hinges on the reliability of the witness and the contingent factors affecting that reliability.

Many factors affect a listener’s ability to recognise a voice. These include the distinctiveness of the voice in question (e.g. Papçun, Kreiman, and Davis, 1989), the listener’s degree of prior familiarity with the voice (e.g. Hollien, Majewski and Doherty 1982; Schmidt-Nielsen and Stern 1985, Künzel 1994), the duration of the exposure to the voice, the listener’s emotional state and level of stress during the exposure, and the amount of time that has elapsed between the initial exposure to the voice and the presentation of the voice parade (Clifford, Rathborn and Bull, 1981; McGehee, 1937, 1944; Papçun. Kreiman and Davis, 1989; Wixted and Ebbssen, 1991). However, given the right combination of circumstances and conditions, earwitness evidence can be used and has been successfully used for a number of such cases in the UK (e.g. McDougall, 2013; Nolan, 2003; Nolan and Grabe, 1996).

A further potential confounding factor in earwitness evidence is the effect of transmission characteristics, of which the commonest everyday case will be the telephone. We are all aware that a voice sounds different over the phone, but also that we can often tell who the speaker is. It is not well understood, however, how differences in quality such as those imposed by the phone will interact with earwitness performance in voice parades. The present experiment investigates listeners’ accuracy in recognizing voices which have been heard over a landline telephone. Specifically, it aims to quantify the effect of different combinations of telephone speech and studio quality speech on identification performance in voice parades consisting of speakers closely matched for accent and personal voice quality.

A significant proportion of forensic phonetic cases involve speech affected by telephone transmission. Earwitness identification may be relevant in cases, for instance, of obscene or threatening calls, or fraud perpetrated on private individuals over the phone. Such calls may be made either over a landline, or, more commonly nowadays, over a mobile network (Öhman, Eriksson and Granhag 2010). Mobile telephone transmission compromises the speech signal more drastically than landline transmission (see e.g. Guillemin and Watson 2008); however, since mobile calls are subject to uncontrolled variation as a result of factors such as network congestion and packet loss, it was decided that the present study would, as a first step towards understanding the effects of acoustically degraded speech on earwitness performance, examine landline speech, the transmission characteristics of which are relatively stable.

Landline telephone transmission affects a speech signal by reducing the bandwidth of the signal to the range approximately 340 Hz – 3700 Hz, and by distorting the frequencies within the reduced bandwidth in a variety of ways (Foulkes and French, 2012). Phonetic research shows the effects of telephone transmission on linguistic features such as the formant frequencies of vowels (Byrne and Foulkes, 2004; Künzel, 2001; Nolan, 2002; Rose, 2003). However, research by Lawrence, Nolan and McDougall (2008) shows that a listener's perception of linguistic vowel quality may withstand acoustic distortion by telephone transmission. In this study, listeners who were trained phoneticians appeared to compensate for the telephone effect, judging a given vowel token recorded directly and over the telephone to have the same quality. However, considerable individual variation was exhibited in the phoneticians' responses such that further experimentation and analysis is required to substantiate this finding.

More recent work by the present authors (Nolan, McDougall and Hudson, 2013) investigates how a listener's perception of speaker-distinguishing properties of voices is affected by telephone transmission. Listeners rated the voice similarity of pairs of short spontaneous speech samples. The paired samples were either both of studio quality, both of telephone quality, or one of each (in either order). It emerged that, on average, a pair of speakers was rated more similar when the samples were presented in telephone quality compared to studio quality, presumably because potentially distinguishing spectral information was lost. This was also true on average when (by design, but unknown to the listeners) the two samples were from the same speaker. However, in the 'mixed' condition, with the samples in the pair in different transmission qualities, the effect of the telephone rather depended on how similar the samples were (as judged in the 'studio' condition): the difference between samples which were essentially similar (including same-speaker pairs) was increased, whereas the difference between samples that were less similar was decreased. These rather complex findings suggest that it is important to test how the effect of the telephone affects the performance of earwitnesses in the voice parade task.

A few studies of voice identification accuracy over the telephone by naïve listeners have been carried out, however results are conflicting, and the methodologies used contain a number of weaknesses and limitations. For example, Rathborn, Bull and Clifford (1981) compared listeners' ability to recognize a target voice to which they had been exposed via a full bandwidth recording or via a telephone-transmitted recording. Listeners were asked to pick the target voice from directly-recorded and telephone-recorded voice parades of six speakers (closed task), three male and three female. Significant effects of both target presentation mode (direct versus telephone) and recognition mode (direct versus telephone) were found, such that the direct-direct condition resulted in the highest correct identification scores (mean 3.886 out of a maximum possible of 6) while lower scores close in magnitude resulted for the direct-telephone (mean 2.486), telephone-direct (mean 2.667) and telephone-telephone (mean 2.657) conditions. However, the line-ups were played immediately after the target voice was heard, unlike a forensic situation where some period of time will have elapsed before the listener may have the opportunity to attempt to identify the voice. Further, while an earwitness's exposure to a voice would involve spontaneously produced speech, the recordings used in the experiment all involved read speech (see e.g. Laan, 1997 regarding differences between read and spontaneous speech). Furthermore, other than the fact that three speakers were female and three male, no mention is made of how the voices were selected and in particular whether they were auditorily similar-sounding. Using a mixed-sex parade is methodologically problematic since listeners can identify the sex of an adult speaker with a reasonable degree of accuracy, and so each parade effectively contained the target and just two foils.

The methodology used by Yarmey's (2003) study of earwitness identification in natural settings and over the telephone improves considerably on earlier work in terms of the selection of target and foil voices: the speakers were all female, of the same age, and shared the same geographical and educational background. Further, their selection was made on the basis of three judges' ratings of the similarity of the voice pairs of each potential foil and the target (but not comparing between potential foils). However, less forensically realistic aspects of this study are that the voice parades were constructed from read speech (a passage from a children's book) and that the parades were run within a few minutes of listeners being exposed to the target. This study used one-person show-ups and six-person voice parades (both target-present and target-absent), and found relatively poor levels of correct identification across both telephone and direct presentations of a target voice, with telephone presentations faring slightly better than direct.

A more recent study by Kerstholt, Jansen, van Amelsvoort and Broeders (2006) gives further conflicting results, with no difference in identification accuracy being found between telephone and directly-recorded presentations of a single target voice, for six-speaker target-present and target-absent voice line-ups played at intervals three and eight weeks after exposure to the target voice. This study controlled for age and accent, with speakers of regionally and strongly socially marked accents being used for the direct versus telephone part of the experiment. The parade was screened by naïve listeners for any unusual sounding samples and one speaker was removed due to a differing speech style, but there was no quantitative assessment of the voice similarity of the selection of voices used. Further, the speech material used was not forensically realistic, being taken from monologue descriptions of a picture.

Whereas the above studies used landline transmission, Öhman, Eriksson and Granhag (2010) investigated the effect of mobile telephone transmission on the identification accuracy of listeners. The target speaker exposure material was a 40 second recording of a scripted incriminating event being read out. Seven-speaker target-present voice parades were constructed using speech describing a walk prompted by pictures. The speakers in the parade were Swedish speakers from the Gothenburg area, aged 25-52 years (sex not specified). Some similarity testing was conducted, and the foils were chosen to include two “quite similar” to the “suspect”, two “rather dissimilar”, and two in between, on the basis of suggestions by Hollien (2002). This approach arguably reduces the selection to a smaller parade of three voices, at least on the more usual assumption that foils which are not similar to the suspect can be readily discounted by the witness. The present authors would argue against this approach, preferring to select foils which are approximately “equally” similar, but probably similar in different ways, to the suspect (see the method using multidimensional scaling below). The parades of Öhman *et al.* were carried out two weeks after exposure to the target voice. This study found no significant effect of presentation format or parade format, and very poor rates of recognition overall.

The present study improves on previous work by investigating listeners’ ability to distinguish among voices heard over the telephone compared with full bandwidth under relatively forensically realistic circumstances. In particular, spontaneous speech from simulated police interviews is used to construct the voice parades, and the parades are conducted at a separate session from the exposure. The parades are constructed using the recommended method for the United Kingdom (Home Office 2003), which involve nine-person line-ups, larger than the previous experiments described above. Crucially, the present study ensures that foil speakers are matched for age, sex, education, and accent, by selecting the samples from a database of speakers controlled for these four characteristics, and that foil speakers are selected on the basis of overall voice similarity, by pre-testing potential foils using pairwise naïve-listener judgments of similarity and multidimensional scaling. Further, while most studies use a single target speaker, five target speakers are used in the present experiment since it has been shown that some voices are more memorable than others (e.g. Foulkes and Barron, 2000; Papçun, Kreiman and Davis, 1989; Sørensen, 2012).

Method

The Source of the Speakers

The speakers were chosen from the *DyViS* database, a pre-existing database comprising recordings of 100 male speakers of Standard Southern British English (SSBE) aged 18-25, which exemplify a population of speakers of the same sex, age and accent group. It constitutes a corpus of voices where linguistic differences of accent or dialect are controlled, allowing the exploration of variation in personal voice quality untrammelled by such linguistic factors. The recordings were made between February 2006 and February 2007. Further details regarding the content of the database and elicitation techniques used are given in Nolan, McDougall, de Jong and Hudson (2009). The *DyViS* database of recordings and transcripts is available through the UK Data Service.[1]

Stimuli were constructed using data from *DyViS* Task 2, a telephone conversation recorded simultaneously in a recording studio and at the remote end of a telephone landline. The conversation involved the subject discussing with his 'accomplice' (an experimenter) his experiences in a previous task, a simulated police interview. The call was carried over the public telephone network to an office in the same building, and an intercept device picked up the subject's voice from the telephone line at the experimenter's end. Both the direct, studio recording of the subject and the telephone intercept were recorded at a sampling rate of 44.1 kHz. By using exactly the same speaking events for both the ‘studio’ and ‘telephone’ conditions in the experiment, effects of the telephone not related to telephone transmission, such as changes in conversational style and shifts in mean fundamental frequency, were controlled for.

Selection of the Speakers

The selection of similar speakers took as a starting point previous work mapping perceptual distances between speakers in the *DyViS* database. Nolan *et al.* (2013) selected fifteen speakers on a random basis (excluding any speakers whose voices sounded impressionistically relatively unusual, e.g. extremely high or low pitched). For each speaker, two short audio clips of approximately three seconds of spontaneous speech were selected from the recordings of *DyViS* Task 2 (described above). Both the studio quality recording and the telephone intercept recording of the same speech event were chosen. Each speaker was paired with all other speakers and with himself to form 120 pairings. The pairings were represented four times, once in studio quality, once in telephone quality, and twice in ‘mixed’ quality (the telephone or studio sample being heard first randomly). All 480 stimuli were randomised for each listener using the ‘ExperimentMFC’ facility in *Praat*. Twenty listeners (10 male, 10 female), all native speakers of British English aged 17-42 years, rated the similarity of the voices in each pair on a scale from 1 (very similar) to 9 (very different). The listeners were instructed to take into account voice quality and accent, but as far as possible to ignore the meaningful content of the speech.

The analysis from the similarity experiment which is relevant to the present study involved subjecting the similarity judgments on each pairing of studio-recorded voices to Multidimensional Scaling (MDS) (Schiffman, Lance Reynolds and Young, 1981). An analysis with five perceptual dimensions (stress = 0.18596, RSQ = 0.16006) was chosen (cf. Giguère's 2006 guideline thresholds for stress). Figure 1 gives a plot of the first two dimensions from this analysis showing the 15 speakers' locations along these dimensions. From this mapping of the listeners' perceptual space it can be seen, for instance, that speakers 1 and 11 are very similar, and speakers 8 and 13 are relatively less similar. Recall that these differences are within a tightly circumscribed part of 'speaker space', given that age, accent, and gender differences have been eliminated or minimised.

Figure 1 Plot of the 15 speakers' locations on the first two dimensions (of five) produced by multidimensional scaling using listeners' judgements of the pairings of speakers recorded in studio quality.[3]

The nine most similar-sounding of the speakers used in the similarity experiment were selected, to allow for voice parades with one target and eight foils, all the voices being similar enough to render the identification task sufficiently challenging that the effects of transmission quality could emerge. Using the studio condition data, each speaker had been characterised by a set of five coordinates on five perceptual dimensions of the form (dim1, dim2, dim3, dim4, dim5). These coordinates were used to calculate the Euclidean distances between all pairings of the speakers in the five-dimensional space, using the formula:

$$dist_{s1,s2} = \sqrt{(\dim 1_{s1} - \dim 1_{s2})^2 + (\dim 2_{s1} - \dim 2_{s2})^2 + \ldots + (\dim 5_{s1} - \dim 5_{s2})^2}$$

where $dist_{s1,s2}$ is the Euclidean distance between two speakers, *s*1 and *s*2, and $\mathrm{dim1}_{s1}$ represents the value of speaker *s*1 on perceptual dimension 1, etc.

In the case of a genuine voice parade case, one of the test speakers would have been the suspect and the eight speakers with the shortest Euclidean distances to the suspect would have been chosen to be the eight foils. However, for the present experiment there was no suspect as such; rather the aim was to choose the nine speakers judged to be the most similar-sounding of the fifteen. To this end, each of the fifteen speakers was in turn treated as a 'centre-speaker' (as if in the role of the suspect) and the Euclidean distances between the centre-speaker and each of the other fourteen speakers ranked. The centre speaker achieving the shortest sum of Euclidean distances between himself and the eight speakers closest to him was selected. This speaker, together with the group of eight speakers closest to him formed the nine speakers used for the voice parade. Based on the procedure mentioned here, the nine speakers with the shortest sum-of-Euclidean-distances were the ones called S1, S4, S5, S6, S9, S10, S11, S13, and S14[3] in Nolan *et al.* (2013); it is these nine speakers that were at the basis of the present experiment.

Voice Parade Construction

The parades were prepared in accordance with the Home Office guidelines (Home Office 2003; see also F. Nolan, 2003). The guidelines recommend constructing what might be termed a 'collage' of short audio clips totalling around a minute. This is so that no continuous narrative emerges that might give a clue to the speaker's status as foil or suspect. In the present experiment, two parallel target-present voice parades for each target were prepared containing identical speech material, one in studio quality, the other in telephone quality. The audio clips for each of the nine speakers in the parade were taken at random from his phonecall material in the *DyViS* database, and were no longer than six seconds in length. The segmentation, randomisation and parallel alignment of the two identical parades were achieved using a *Praat* (Boersma and Weenink, 1992-2014) script, with manual correction where necessary. Speakers S5, S9, S11, S13, S14 were selected randomly to serve as 'targets'. For each target speaker, a sample of around one minute of continuous speech (after excision of the interlocutor's voice) was taken from the end of the debriefing recording (both studio and telephone versions); this material did not overlap with the voice parade material.

Listeners

One hundred listeners (50 male, 50 female) undertook the experiment. Listeners were 17-42 years of age, were born and had lived mostly in the British Isles (with no strong Scottish, Welsh or Irish accent), and had no known hearing difficulties.

Procedure

The 100 listeners were divided into four groups of 25 listeners (roughly balanced for gender), one group per condition. The first group was exposed to a target voice at studio quality, and then returned to undertake the voice parade in studio quality; the second group initially heard the voice in telephone quality and then the parade in telephone quality; the third group heard the voice in studio quality and the parade in telephone quality; and the final group the reverse. Within each group of 25 listeners were 5 sub-groups, each presented with a different target voice. Thus the experiment is broken down into 20 (5×4) sub-experiments. Both familiarisation to the target and the parade were conducted in a sound-treated booth. All audio was played on a PC through a powered Yamaha speaker (Yamaha Monitor Speaker MS101 II). The parade was conducted using *PowerPoint*, which presented all voices in the parade (in a pre-determined random sequence), but also displayed the labels (A-J, excluding I) designating the voice samples.

Listeners undertook the experiment in groups of five. Each group of five listeners attended on a given day to hear the sample of the target voice. The group then returned exactly one week later (to the day) and listened to the parade to carry out the identification task.

At the first session, listeners were informed at the outset that the aim of the experiment was to see how reliably listeners can recognise voices with which they are not familiar. They were told that they would listen to a recording of one side of a section of a telephone conversation relating to a crime, one minute long, which would then be repeated once. Listeners were instructed to listen very carefully as this would be the only opportunity to hear the voice. Although this is unlike some earwitness cases where exposure to the voice of a perpetrator is unexpected and short, such as a bank raid by masked robbers, there are other situations where victims may have plenty of opportunity deliberately to memorise a voice, such as abduction or hostage holding by hooded perpetrators. The listeners were told that they need not be concerned with the details of what the speaker was saying: rather, that they should concentrate on the sound of his voice. Listeners were asked to let the experimenter know if they thought they knew the identity in real life of the speaker whose voice they heard at this exposure session, in which case they would not be able to continue to the second session of the experiment; no listeners recognised their target speaker upon exposure to his voice.

When they attended the second experimental session a week later, listeners were informed that they would be asked to attempt to select the voice they heard at the first session from a number of voices in a voice parade. They were told that they would hear nine speech samples, each a minute in duration, and each made of short extracts from a telephone call where the same events were being discussed. It was explained that the samples were labelled A, B, C, D, E, F, G, H, J, and that the appropriate letter would be displayed on the screen as the sample played. Listeners were told that the parade would be played straight through, that they must listen to the entire parade before making a selection, and that they would only hear the parade once. They were asked to indicate on a response sheet which voice sample matched the voice they had heard the previous week, and to rate their confidence in their identification on a scale from 0 to 10, where 0 = ‘not at all confident’ and 10 = ‘completely confident’. Listeners were allowed to take notes during the parade, if desired, and were instructed to keep their response sheets well hidden from other participants.

Results

The percentages of correct identifications made in the voice parades for each of the four conditions are shown in Figure 2.

Figure 2. Percentage of correct identifications made in the voice parade for each of the four combinations of studio/telephone quality exposure and studio/telephone quality voice parade (25 listeners in each condition).

Telephone exposure and parade led to fewer correct identifications (16 out of 25, i.e. 64%) than studio exposure and parade (19 out of 25 correct, i.e. 76%), but perhaps surprisingly only by a small margin. The two cross-modal conditions produced still fewer correct identifications: 15 out of 25 (60%) for studio exposure/telephone parade, and a much lower result of 8 out of 25 (32%) for telephone exposure/studio parade. A Pearson Chi-Square test showed a significant relationship between exposure/format condition and accuracy of identification ($\chi^2(3) = 10.673$, $p = 0.015$). Post hoc testing using Goodman's simultaneous confidence interval procedure (Jaccard and Becker, 1990) showed that the source of the significant relationship was indeed the comparison between the studio-studio and telephone-studio conditions.

Sex differences did not affect identification rates, with correct identification produced by 58% of female listeners and by 58% of male listeners overall. Considering the results for each exposure/parade condition separately, there were some differences in correct identification rates between the sexes, but two-tailed Fisher's Exact Tests showed that none of these were significant (studio-studio: 82% female versus 64% male, $p = 0.407$; studio-telephone: 67% female versus 70% male, $p = 1.0$; telephone-studio: 29% female versus 36% male, $p = 1.0$; telephone-telephone: 60% correct for both sexes).

If we break down the results by target speaker, we find that the overall picture is also affected by individual target speakers, as can be seen in Figure 3 which shows the number of correct identifications made within each of the four conditions broken down for the five target speakers. Target speaker S14 was more readily identified (18 correct identifications out of 20 across the four conditions) than the others, especially S5 and S9 (8/20 and 7/20), with S11 and S13 (13/20 each) being intermediate. For the telephone exposure-studio parade condition, S5 and S9 were not once correctly identified.

Figure 3. Number of correct identifications made in the voice parade for each of the four combinations of studio/telephone quality exposure and studio/telephone quality voice parade, for each target speaker (5 listeners per target speaker within each condition).

Listeners' assessments of how confident they were that they had made a correct identification are shown in Figure 4, with correct identifications shown in dark grey and incorrect identifications shown in light grey. These results are shown broken down by target speaker in Figure 5. As Figure 4 shows, 23 identifications were made with the maximum confidence ratings of 9 and 10, and these were correct in all but one case. 40 listeners chose confidence ratings of 6 or 7 and more than half of these identifications were inaccurate. This indicates that under different listening conditions confidence may be useful when the listener is at least 90% confident, but that confidence is not a good predictor of identification accuracy when the listener is less confident in his or her response.

Figure 4. Correct and incorrect identifications according to the level of confidence with which each listener made his or her identification.

There is little evidence that the confidence ratings reflect the distinct levels of accuracy in the four exposure/parade format conditions, the mean confidence ratings being 7.52 (studio-studio), 6.16 (telephone-telephone), 6.88 (studio-telephone), and 6.44 (telephone-studio, where the identification rate was less than half that of studio-studio.). This is confirmed by a univariate ANOVA with the factors Exposure/Parade Format (4) and Target Speaker (5), which showed no significant effect of Exposure/Parade Format (*F*(3, 80) = 2.050, *p* = 0.114), and neither was the interaction between Exposure/Parade Format and Target Speaker significant (*F*(12,80) = 1.272, *p* = 0.252).

Figure 5. Correct and incorrect identifications according to the level of confidence with which each listener made his or her identification, broken down by target speaker.

The confidence ratings broken down by target speaker in Figure 5 show that the target speaker who was most often correctly identified, S14, and one of the second-most readily identified speakers, S11, yielded the highest confidence ratings (mean = 7.6 for both). The speaker least often correctly identified, S5, was also associated with the lowest confidence ratings (mean = 5.5). The ANOVA described above showed a significant main effect of Target Speaker on confidence ratings (*F*(4, 80) = 3.746, *p* = 0.008). Pairwise comparisons with Bonferroni correction showed significant differences between the confidence ratings for target speakers S5 and S11, and between those of S5 and S14 (both *p* = 0.019), i.e. S5, the speaker yielding the lowest confidence ratings noted above, significantly differed from S11 and S14, the speakers with the equal highest confidence ratings. All other pairs of speakers showed no significant differences with respect to confidence ratings (*p* > 0.05).

Thus confidence ratings are not a reliable guide to performance accuary on the whole, consistent with the findings of previous research (e.g. Sørensen, 2012; Yarmey, 2004). However, the picture is a little complex: while the confidence ratings in the present experiment offered no predictive power of identification accuracy across different listening conditions, some listeners were accurate in judging certain voices as harder to identify than others.

Given that certain target speakers were more readily identified than others, and with higher confidence, it is instructive to look at the pattern of errors in the optimum listening condition (studio-studio) in the light of the perceptual distances between speakers as established in Nolan *et al.* (2013). Figure 6 shows the 36 different-speaker pairs arising from the nine speakers used in the parade rank ordered by the Euclidean distance between members of the pairs based on the five dimensions in the MDS analysis in the earlier study (using only ratings of studio-studio pairs).

Figure 6. Speaker pairs: Euclidean distances based on five MDS dimensions, ranked from shortest to longest. See text for explanation of the black and striped bars.

In the studio exposure-studio parade condition only six misidentifications were made, as shown by the black bar and striped bars in Figure 6. It can been seen from this figure that the misidentifications all correspond to speaker pairs with Euclidean distances towards the lower end of the scale. A glance back at Figure 2, plotting the first two (most important) MDS dimensions, shows that speakers S9 and S14, who are reciprocally misidentified, are indeed very close, as are S9 and S13 (the former being misidentified as the latter once). It is less clear why S5 should be misidentified as S11 twice over, or S11 as S14, though the distances involved here are still towards the lower end of the rank order (recall that Figure 2 shows only the first two MDS dimensions out of five, whereas the Euclidean distances in Figure 6 are calculated from all five, so the two displays do not entirely correspond). It is noteworthy that foil S14 suffered two false identifications, and foil S5 suffered none, though as a target he was twice falsely identified as S11; this underlines the point sometimes made in speaker verification studies that recognition errors are not distributed symmetrically within a set of voices (cf. Doddington *et al*., 1998).

Discussion

We now consider the implications of these findings, including their potential relevance to real-world voice parades. Whilst it is hard to compare absolute identification rates between speaker identification studies because each has different conditions, it could be argued that a 76% correct identification rate is reassuringly high in a parade where the speakers have been rigorously controlled to have the same accent and have been found by listeners rating them to be perceptually similar.

However, this relatively high identification rate was achieved by parade mock witnesses who were primed to remember the target voice, and hearing it consistently in good quality recordings. Performance in the telephone-telephone condition dropped, as expected, but (at 64%) was still around six times better than chance (11%). Likewise when the target was heard in studio quality, and a parade presented using samples recorded via the public telephone network, identification (at 60%) was still more than five times better than chance.

The unexpected, and disturbing result is the asymmetry in performance when the reverse order applies, namely a target heard in telephone quality and a parade constructed from studio quality samples. An identification rate of 32% is only three times better than chance. Even more disturbing is the fact that the mock witnesses' confidence ratings in this condition are not significantly lower than those in any other exposure/parade format condition. Many real world cases will involve a perpetrator being heard over the telephone, and this finding poses a serious challenge to using full-bandwidth speech in constructing a parade.

Why should this asymmetry in the 'mixed' conditions exist? The experiment was not designed to probe this phenomenon as it had not, to our knowledge, been reported. The best we can therefore do is to speculate about the processes which might be involved in comparing a memory of a voice with samples being heard. The explanation put forward here depends on two assumptions: that for voice samples to be compared they need to be commensurately represented (i.e. in this case bandwidth limited or full bandwidth); and that whilst we can cognitively process a memory of a voice heard full-bandwidth to model what it would sound like over the telephone, we cannot reliably reconstruct the missing information in our mental representation of a band-limited voice.

This means that the comparison has to be made in the domain of band-limited speech. As shown schematically in Figure 7, if the target voice stored in memory was heard as a studio sample, only one transformation need be made on the basis of knowledge of the telephone effect. The parade samples can then be compared without further transformation. If, on the other hand, the target has been heard in telephone quality, each of the parade samples will have to be transformed to be compatible. Since this cognitive transformation is always open to error, the opportunities for error are multiplied nine-fold (in a nine sample parade), and the cognitive load during the parade substantially increased.

Figure 7. Schematic representation of the process of voice comparison when parades are presented with transmission characteristics different from target exposure. When the target was heard in studio quality, one transformation ('modelling') will make the cognitive representation of the voice compatible with all samples in the parade. When a phone target was heard, all parade samples will have to be (cognitively) filtered, multiplying the possibilities for error.

Whatever the mechanism that accounts for the discrepancy in the two 'mixed' conditions, if it is substantiated by further research, there is a clear practical message for the conduct of voice parades: if the target (perpetrator's) voice has been heard over the telephone, the parade should be conducted with telephone quality samples (either recorded over the telephone, or with the effects of the telephone simulated by filtering of the samples).

The results of this experiment also highlight the importance of the role of the individual target speaker in earwitness research. Identification accuracy varied considerably for the different target speakers in the experiment across the various format conditions. The identifiability of different target speakers appears to vary greatly, yet most previous telephone studies use only one target speaker. Further, while confidence ratings did not correspond to the pattern of correct and incorrect identifications across format conditions, the target speakers who were identified correctly most often were also associated with higher confidence ratings. An important area for further research is that of speaker characteristics of target speakers (e.g. Sørensen (2012) on the role of mean fundamental frequency in earwitness identification), but it is clear that much more research is needed into the ways in which a variety of individual properties of a voice can affect the accuracy with which it is identified. That research will need not only to carry out empirical testing of the kind reported here, but also to build a phonetically-informed and perceptually-relevant model of speaker-identity. Such a model would aim, for instance, to assign relative weights to acoustic dimensions (e.g. fundamental frequency, formant frequencies, and articulatory timing patterns) in keeping with their role in perceived similarity between speakers, and to predict the importance of linguistic-phonetic differences between speakers relative to differences of personal voice quality. Work towards this kind of model is already underway (e.g. Nolan *et al.* 2011, McDougall *et al.* in prep.).

Conclusion

The present study investigated the effect of the telephone on the identification accuracy of earwitnesses using voice parades with foils matched for accent and personal voice similarity, the latter quantified by multidimensional scaling of perceptual similarity ratings. Exposure to a voice recorded at studio quality followed by a studio quality voice parade led to correct identifications in 76% of cases, while telephone quality exposure and voice parade produced correct identifications 64% of the time. The cross-modal conditions of studio-telephone and telephone-studio exposure/parade gave 60% and the markedly lower 32% respectively. Whilst no controlled experiment can achieve complete ‘ecological validity’, the present experiment does replicate a sufficient number of the characteristics of real-world voice parades that its results potentially have practical implications for the preparation of voice parades in cases where the perpetrator’s voice has been witnessed over the telephone. Full bandwidth speech samples should not be used for such voice parades, rather, speech recorded over the telephone or speech samples filtered to resemble telephone speech should be used.

Notes

1. http://www.ukdataservice.ac.uk

2. The equivalent *DyViS* speaker numbers are 95, 60, 65, 25, 112, 39, 28, 56 and 115, in that order.

3. This figure also appears in McDougall, K. (2013) ‘Earwitness evidence and the question of voice similarity’ *British Academy Review, 21*, (18-21).

References

Boersma P, Weenink D (1992-2015): Praat: doing phonetics by computer. http://www.praat.org/.

Broeders APA, Rietveld ACM (1995): Speaker identification by earwitnesses; in Braun A, Köster J-P, Studies in forensic phonetics: Beiträge zur Phonetik und Linguistik 64: 24-40.

Byrne C, Foulkes P (2004): The 'mobile phone effect' on vowel formants. Int J Speech, Language and Law 11.1: 83-102.

Clifford B, Rathborn H, Bull R (1981): The effects of delay on voice recognition accuracy. Law and Human Behaviour 5: 201-208.

Doddington G, Liggett W, Martin A, Przybocki M, Reynolds D (1998): Sheep, goats, lambs and wolves: a statistical analysis of speaker performance. Proc IC-SLD'98, NIST 1998 Speaker Recognition Evaluation, Sydney, Australia, pp 1351–1354.

Foulkes P, Barron A (2000): Telephone speaker recognition amongst members of a close social network. Forensic Linguistics 7.2: 180-198.

Foulkes P, French P (2012): Forensic speaker comparison: a linguistic-acoustic perspective, in Tiersma P, Solan L, Oxford Handbook of Language and Law. Oxford, Oxford University Press, pp 557-572.

Giguère G (2006): Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. Tutorial in Quantitative Methods for Psychology 2.1: 27-38.

Guillemin B, Watson C (2008): Impact of the GSM mobile phone network on the speech signal: some preliminary findings. Int. J. Speech, Language and the Law. 15.2: 193-218.

Hollien H (1996): Consideration of guidelines for earwitness lineups. Forensic Linguistics 3.1: 14-23.

Hollien H (2002): Forensic Voice Identification. London, Academic Press.

Hollien H, Majewski W, Doherty ET (1982): Perceptual identification of voices under normal, stress and disguise speaking conditions. J Phon. 10: 139-148.

Home Office (2003): Advice on the use of voice identification parades. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership and Powers Unit. https://www.gov.uk/government/publications/advice-on-the-use-of-voice-identification-parades

Jaccard J, Becker MA (1990): Statistics for the behavioural sciences (2nd ed.). Wadsworth, Belmont.

Kerstholt JH, Jansen NJM, van Amelsvoort AG, Broeders APA (2006): Earwitnesses: effects of accent, retention and telephone. App Cog Psych 20.2: 187-197.

Künzel HJ (1994): On the problem of speaker identification by victims and witnesses. Forensic Linguistics 1.1: 45-57.

Künzel HJ (2001): Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics 8.1: 80-99.

Laan GPM (1997): The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. Speech Commun 22: 43-65.

Lawrence S, Nolan F, McDougall K (2008): Acoustic and perceptual effects of telephone transmission on vowel quality. Int J Speech, Language and the Law 15.2: 161-192.

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

McDougall K, Nolan F, French P, Stevens L, Hudson T (in prep): Phonetic correlates of the perception of voice similarity: an experiment on Standard Southern British English spoken by males.

McGehee F (1937): The reliability of the identification of the human voice. J Gen Psych 17: 249-271.

McGehee F (1944): An experimental study in voice recognition. J Gen Psych 31: 53-65.

Nolan F (2002): The 'telephone effect' on formants: a response. Forensic Linguistics, 9.1: 74-82.

Nolan F (2003): A recent voice parade. Int J Speech, Language and the Law, 10.2: 277-291.

Nolan F, Grabe E (1996): Preparing a voice lineup. Forensic Linguistics 3.1: 74-94.

Nolan F, McDougall K, de Jong G, Hudson T (2009): The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. Int J Speech, Language and the Law, 16.1: 31-57.

Nolan F, McDougall K, Hudson T (2011): Some acoustic correlates of perceived (dis)similarity between same-accent voices. Proc 17th ICPhS, Hong Kong, pp 1506-1509.

Nolan F, McDougall K, Hudson T (2013): Effects of the telephone on perceived voice similarity: implications for voice line-ups. Int J Speech, Language and the Law 20.2: 229-246.

Öhman L, Eriksson A, Granhag PA (2010): Mobile phone quality vs. direct quality: how the presentation format affects earwitness identification accuracy. Eur J Psych Applied to Legal Context, 2.2: 161-182.

Papçun G, Kreiman J, Davis A (1989): Long-term memory for unfamiliar voices. J Acoust Soc Am 85: 913-925.

Rathborn HA, Bull RH, Clifford BR (1981): Voice recognition over the telephone. J Police Sci and Admin. 9: 280-284.

Rietveld ACM, Broeders APA (1991): Testing the fairness of voice identity parades: the similarity criterion. Proc 12th ICPhS, Aix-en-Provence, pp 46-49.

Rose P (2003): The technical comparison of forensic voice samples; in Freckleton; Selby, Expert evidence (Chapter 99). Sydney, Lawbook Co.

Schiffman SS, Lance Reynolds M, Young FW (1981): Introduction to multidimensional scaling: theory, methods, and applications. Academic Press, New York.

Schmidt-Nielsen A, Stern KR (1985): Identification of known voices as a function of familiarity and narrow-band coding. J Acoust Soc Am 77: 658–663

Sørensen MH (2012): Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. Int J Speech, Language and the Law 19.2: 145-158.

Wixted J, Ebbssen E (1991): On the form of forgetting. Psych Sci 2: 409-415.

Yarmey AD (2003): Earwitness identification over the telephone and in field settings. Forensic Linguistics 10.1: 62-74.

Yarmey D (2004): Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. Int J Speech, Language and the Law, 11.2: 267-277.

List of figures

the target was heard in studio quality, as in (a), one transformation ('modelling') will make the cognitive representation of the voice compatible with all samples in the parade. When a phone target was heard, as in (b), all parade samples will have to be (cognitively) filtered, multiplying the possibilities for error.

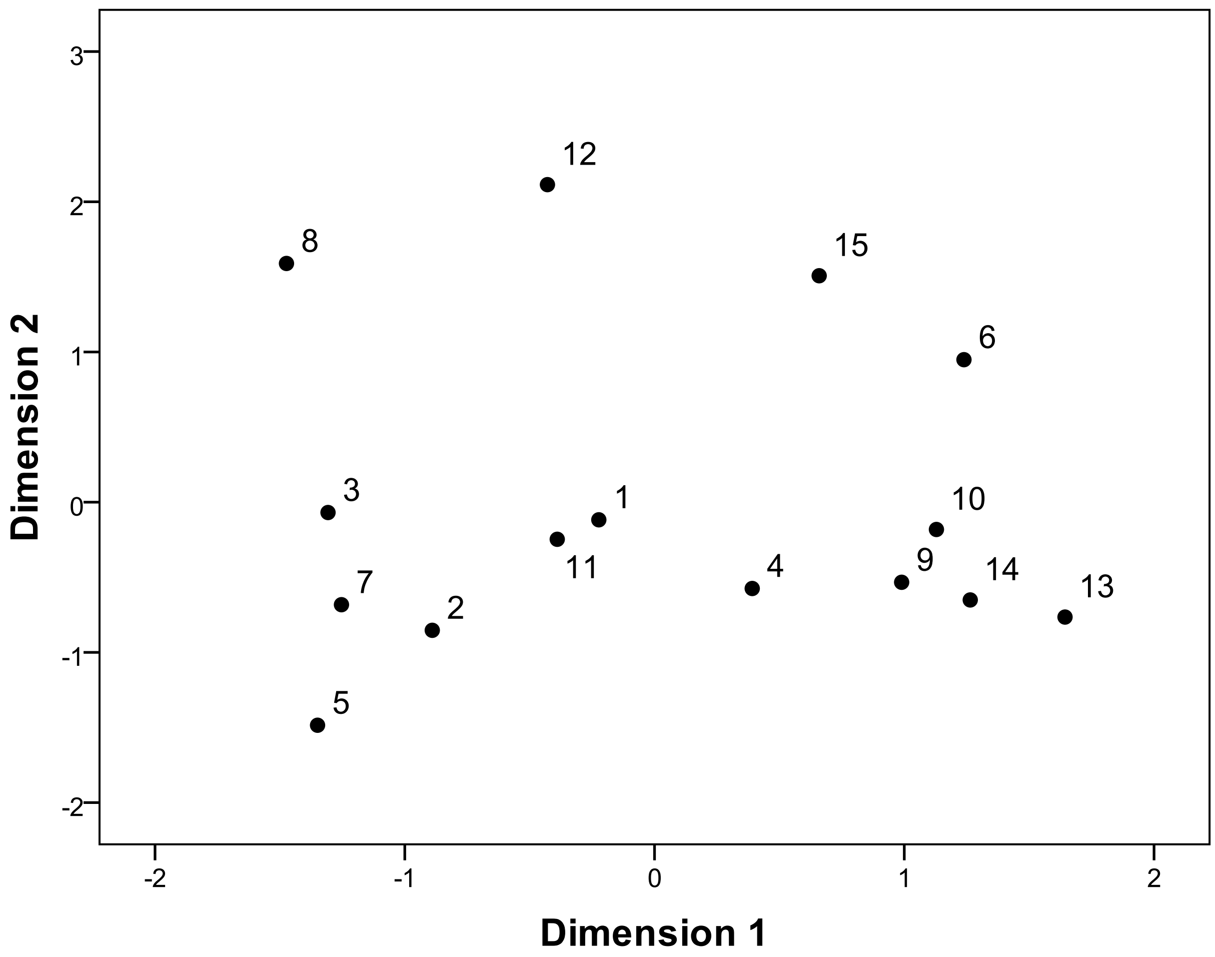


Figure 1. Plot of the 15 speakers' locations on the first two dimensions (of five) produced by multi-dimensional scaling using listeners' judgements of the pairings of speakers recorded in studio quality.[3]

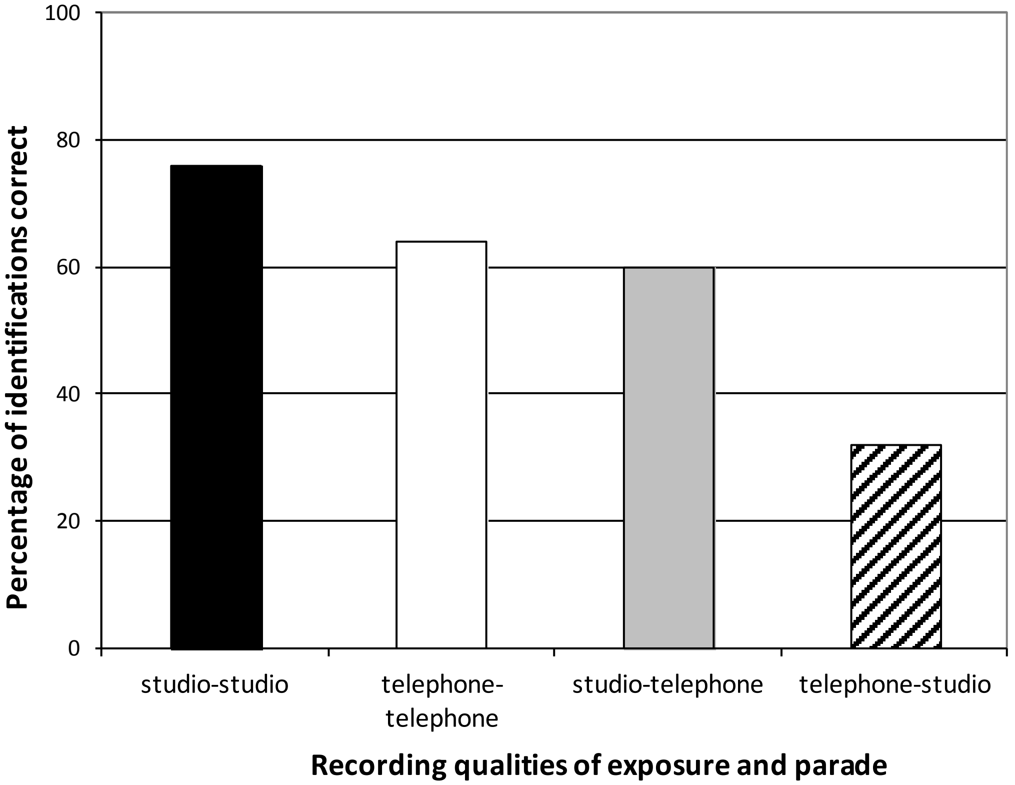


Figure 2. *Percentage of correct identifications made in the voice parade for each of the four combinations of studio/telephone quality exposure and studio/telephone quality voice parade (25 listeners in each condition).*

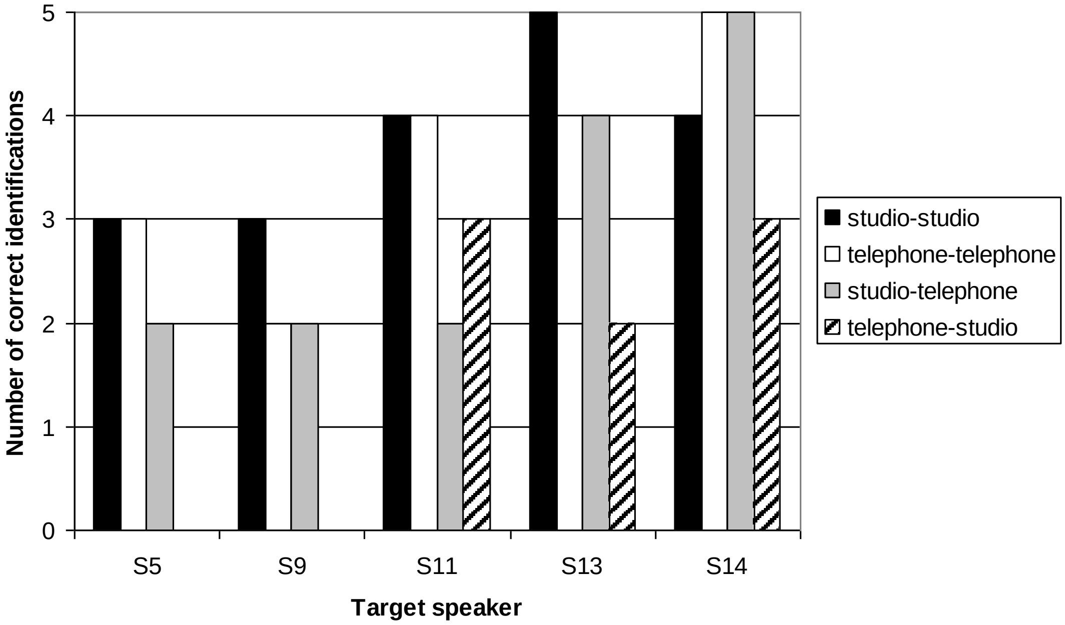


Figure 3. *Number of correct identifications made in the voice parade for each of the four combinations of studio/telephone quality exposure and studio/telephone quality voice parade, for each target speaker (5 listeners per target speaker within each condition).*

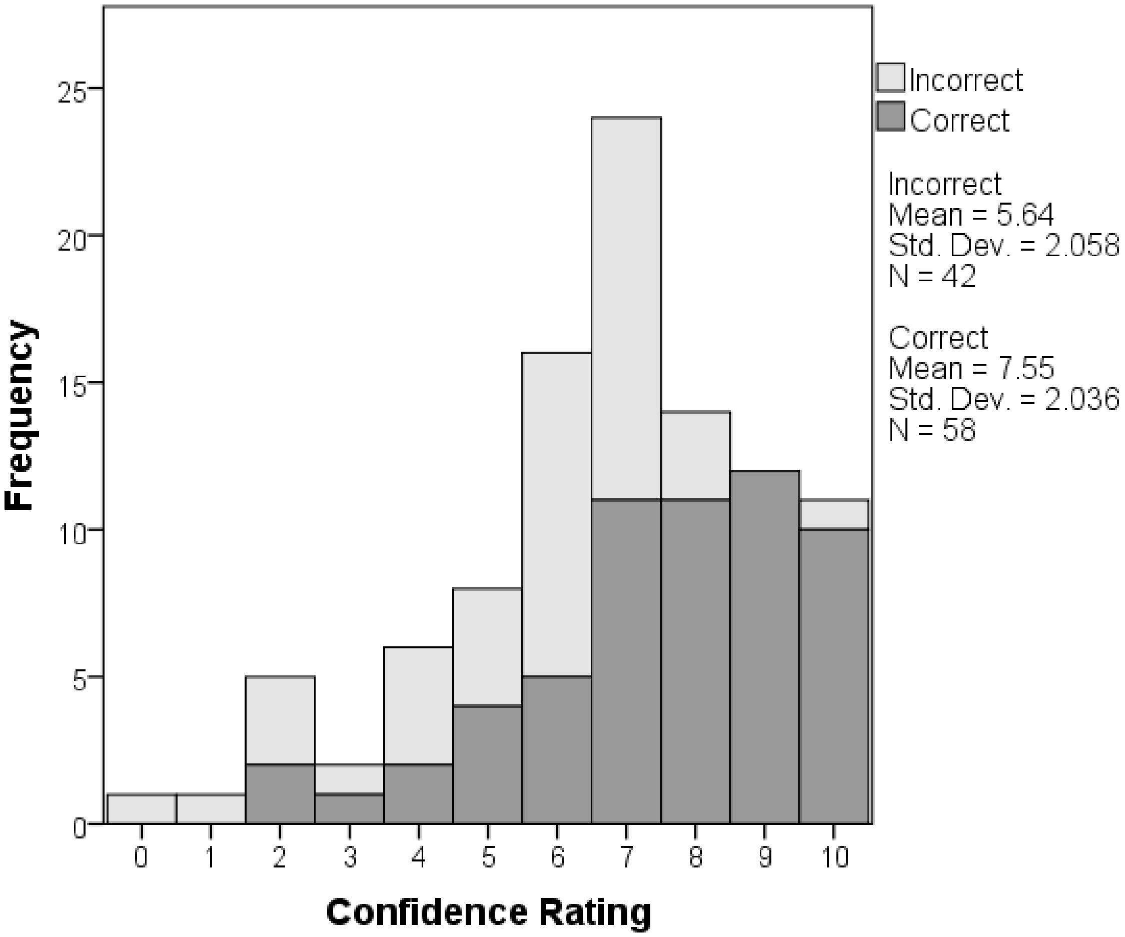


Figure 4. *Correct and incorrect identifications according to the level of confidence with which each listener made his or her identification.*

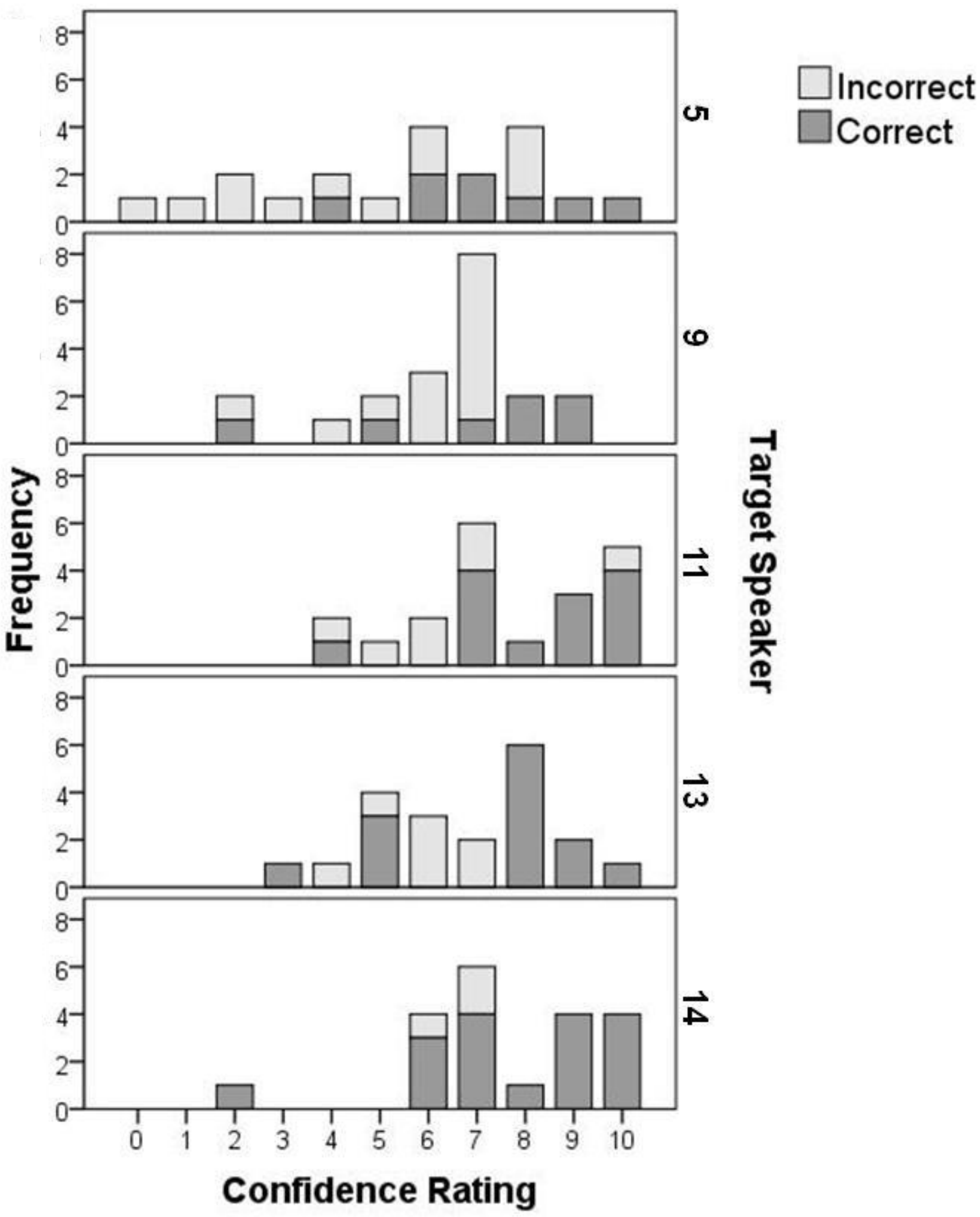


Figure 5. *Correct and incorrect identifications according to the level of confidence with which each listener made his or her identification, broken down by target speaker.*

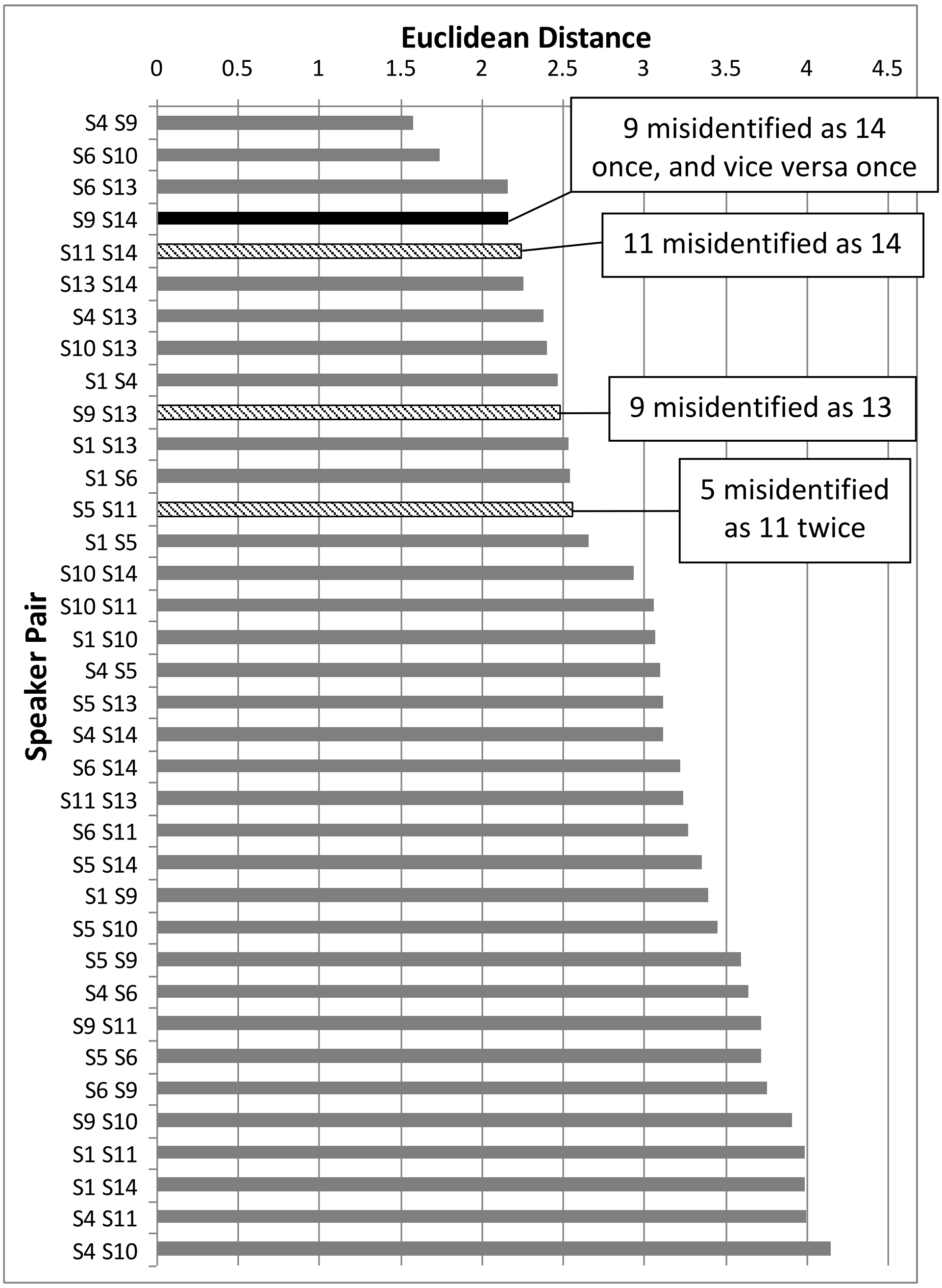
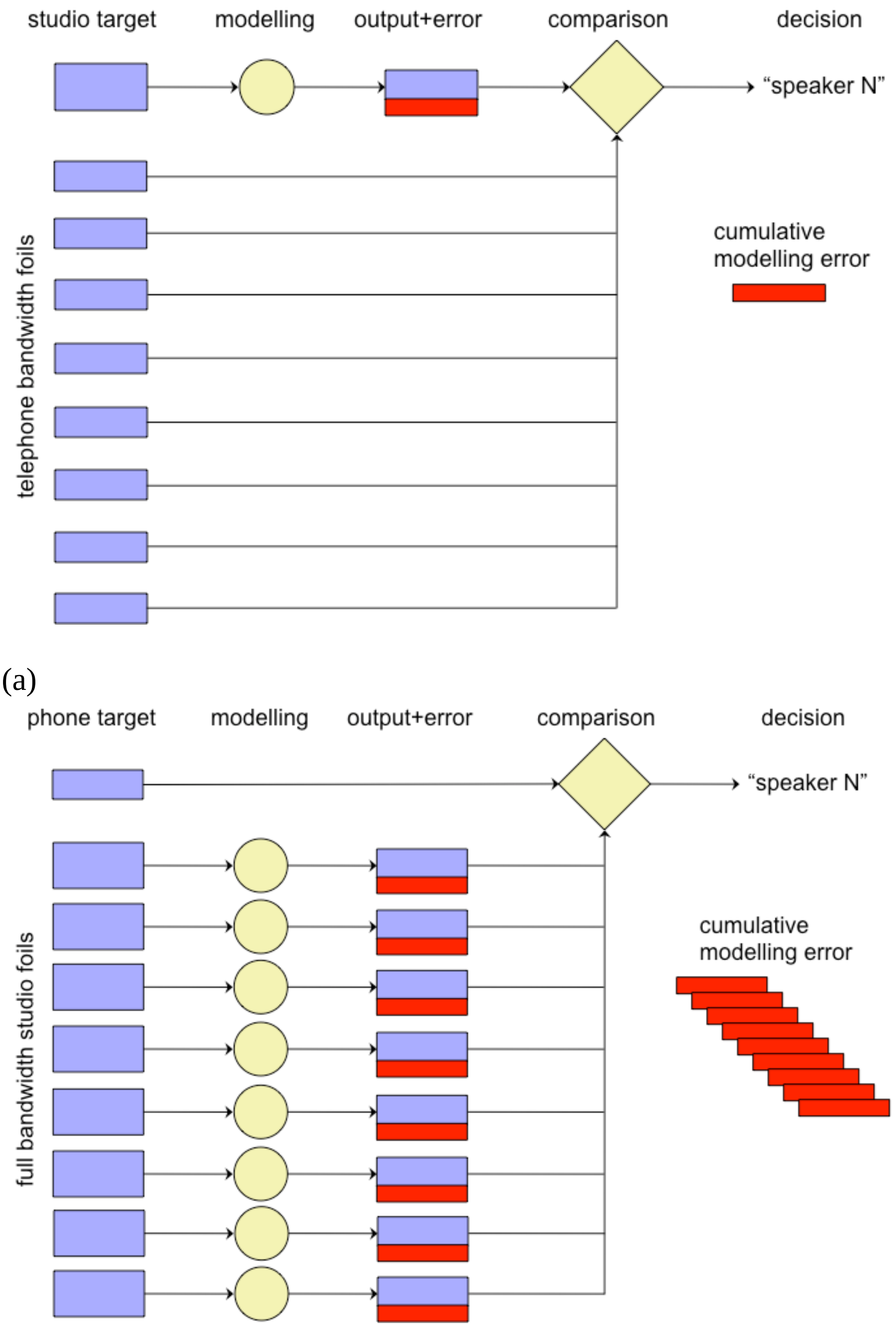


Figure 6. Speaker pairs: Euclidean distances based on five MDS dimensions, ranked from shortest to longest. See text for explanation of the black and striped bars.



(a)

(b)

Figure 7. *Schematic representation of the process of voice comparison when parades are presented with transmission characteristics different from target exposure. When the target was heard in studio quality, as in (a), one transformation ('modelling') will make the cognitive representation of the voice compatible with all samples in the*

parade. When a phone target was heard, as in (b), all parade samples will have to be (cognitively) filtered, multiplying the possibilities for error.