Earwitness Line-ups: Effects of Speech Duration, Retention Interval and Acoustic Environment on Identification Accuracy

J.H. Kerstholt & E.J.M. Jansen*, A.G. van Amelsvoort** and A.P.A. Broeders***

*both at TNO Human Factors, Soesterberg; **LSOP Police Knowledge and Expertise Centre, Apeldoorn; ***Netherlands Forensic Institute, Rijswijk; all in the Netherlands
t.broeders@nfi.minjus.nl

Abstract

An experiment was conducted to investigate the effects of retention interval, exposure duration and acoustic environment on speaker identification accuracy in voice line-ups. In addition, the relation between confidence assessments by participants and test assistant and identification accuracy was explored. A total of 361 participants heard a single target voice in one of four exposure conditions (short or long speech sample, recorded only indoors or indoors and outdoors). Half the participants were tested immediately after exposure to the target voice and half one week later. The results show that the target was correctly identified in 42% of cases. In the target-absent condition there were 51% false alarms. Acoustic environment did not affect identification accuracy. There was an interaction between speech duration and retention interval in the target-absent condition: after a one-week interval, listeners made fewer false identifications if the speech sample was long. No effects were found when participants were tested immediately. Only the confidence scores of the test assistant had predictive value. Taking the confidence score of the test assistant into account therefore increases the diagnostic value of the line-up.

1. Introduction

In the last decade there has been a growing interest in questions raised by the identification of persons by witnesses based on their voices. In the absence of clear criteria for the design of earwitness identification procedures, in the Netherlands a set of guidelines has been developed [1] which is partly based on a tried and trusted method developed by the Dutch police and judiciary for the visual identification of persons by eyewitnesses [2]. Procedures based on these guidelines are currently used in several European countries.

Witnesses frequently constitute an important source of evidence in criminal cases. But although the reliability of eyewitness identifications has been extensively investigated, comparatively little research has been carried out on earwitness identifications [3, 4]. Still, there are several situations in which speaker identifications by witnesses, in the form of voice line-ups or otherwise, could conceivably make a significant contribution to the judicial fact-finding process. Lighting conditions or disguise may have prevented visual perception, the witness may have been blindfolded, or the perpetrator may have been heard over the telephone.

The quality of an identification procedure for witnesses is determined by two sets of variables: system variables and estimator variables [5]. System variables may be controlled by the investigator through the use of a proper line-up procedure. Estimator variables are beyond the control of the investigator but they will also affect the success of the line-up. The present experiment looks into the effect of some of these estimator variables, i.e., exposure duration, retention interval and acoustic environment, may have on the reliability of voice line-ups. In addition, the relation between confidence assessments by both listeners/subjects and test assistant, and identification accuracy was explored.

Percentages of correct identifications by earwitnesses vary widely in studies conducted to date. In [6] a percentage of 75% correct identifications is reported for a short retention interval and of 38% correct for a longer interval, whereas participants performed at chance level in one condition in [7]. Although the underlying causes of these differences in accuracy are not entirely clear, several factors have been identified that affect speaker identification [8]. The present study examines the effects of three of these factors: retention interval, exposure duration and acoustic environment.

In line with general theories on decay of memory traces, it might be expected that the more time elapses between exposure to witnesses are presented with a line-up, the poorer will be their performance. However, the extent of the delay beyond which a voice line-up is unlikely to produce a positive result is unknown. In general, exposure duration seems to affect identification accuracy positively [3]. A significant improvement in performance was found when exposure duration increased from 6 to 60 seconds [9]. In all, previous findings indicate that the (combined) effects of exposure and retention interval may be quite complex. The most straightforward prediction for the present experiment is that a longer retention interval decreases performance, but less so as exposure duration goes up.

Voice identification may be difficult due to within speaker variation: people’s voices are known to vary widely from occasion to occasion [8]. As a result, a witness might not be able to extract sufficient information from a brief exposure to identify the speaker from a reference speech sample. Apart from the effect of factors such as disguise, emotions or the acoustic environment, performance will generally decrease when the perpetrator has been heard only briefly and there is a change of tone between initial exposure and that in the speech samples used in the line-up. Performance dropped when whispered voices had to be recognized [10, 11]. When the speaker’s tone changed from angry during exposure to normal during test, performance almost dropped to chance level [7]. At present, it is not known how differentiation in acoustic environment affects identification performance. In the present experiment, the target’s exposure and test samples were not only recorded in different rooms but also at different points in time (about two weeks apart). In addition, the differentiation of the acoustic environment was varied during exposure. The speech sample was either recorded in one acoustic environment (inside a building) or in two different environments, that is, both indoors and outdoors.
There relation between (self-)reported confidence and accuracy seems to be poor. Even though a significant correlation between recognition performance and confidence ratings was found [7, 12], the absolute values were small. The accuracy-confidence relationship seems to be improved by an increase in speech sample duration [13]. In an earlier experiment on eyewitness identifications, the test assistant was also asked to assess his confidence in the accuracy of the witness. This assessment proved to be the best predictor of accuracy (37% correct prediction after chance correction). The present experiment examined whether this conclusion also holds for the auditory domain.

To summarize, the present experiment explored the influence of retention interval, exposure duration and acoustic environment on the accuracy of earwitness identifications. Based on previous findings we predict that performance deteriorates as retention interval (immediate test versus one-week delay) increases, but less so when the speech samples are longer (30 seconds versus 70 seconds). Furthermore, in one condition we recorded speech samples in two different acoustic environments: indoors and outside a building. In view of the reduced quality of the outdoor recordings we predict that performance will decreases compared to that for indoor only recordings. Lastly, we predict that the confidence ratings of the test assistant are more predictive of accuracy than are the confidence ratings of the witnesses themselves.

2. Method

2.1 Participants
A total of 361 persons took part in the experiment. Eight participants failed to turn up after the one-week delay. Of the remaining 353 persons, 184 were men and 169 women. Mean age of the participants was 37 years (sd=11). Three educational levels were distinguished: vocational training for 12-16 year olds (N=52), vocational training for 16-18 year olds (N=163) and vocational training or higher education for 18+ year olds (N=138). Care was taken to balance participants on all characteristics across experimental conditions (approximately 22 persons per condition).

2.2 Material
Potential candidates for the line-up were selected on gender (men only), accent and age. Candidates were asked to give a detailed verbal description of a picture and to use a normal conversational speaking style. The interviews were taped with a DAT-recorder. The fourth author selected 8 voices from a total set of 25 taped voices. One of these eight voices was selected as the target voice. The line-up selection was based on expert judgments; the extent to which a particular voice matched the voice of the target was taken as a criterion in selecting the remaining voices for the line-up. Selected voices matched the target voice on all phonetic levels, that is, the same (slight) regional accent, similar voice quality, pitch, perceived age etc.

Fairness of the line-up was assessed by five independent listeners who were not familiar with the target voice. One voice was removed from the set as the test panel thought his way of speaking differed too much from the other voices.

The target, a man aged 20, was interviewed again after about two weeks. He was asked questions about his education, his hobbies, family, etc. Two speech samples were made, one inside and the other outside the building. When the target speaker was interviewed outside the building, questions were aimed to elicit emotional statements from him as well, in an attempt to increase differences in acoustic features of the two speech samples.

Four exposure conditions were created, varying in exposure duration and acoustic environment. Exposure duration was either 30 or 70 seconds and the voice recordings were made only inside the building or both inside and outside. The indoor voice samples consisted of a single, uninterrupted speech sample. The samples consisting of material recorded both inside and outside had a pause between the two parts in which a female voice announced the beginning of the second part. Both parts lasted either 15 seconds or 35 seconds each.

Seven audio-samples were made for the line-up. Each line-up consisted of six voices and each sample lasted 20 seconds. In the target-absent condition, the target voice was replaced by a foil. Care was taken to have the same sound intensity for each voice. For both sets six orders were constructed according to a balanced diagram design [14].

2.3 Design
16 conditions could be distinguished (retention interval (2) x exposure duration (2) x acoustic environment (2) x line-up (2)). Retention interval was manipulated by presenting participants with the voice line-up immediately or after a week. Exposure duration was either 30 or 70 seconds. Acoustic environment was defined as a single speech sample recorded inside the building or a speech sample that was recorded half inside and half outside the building. For each condition, half the participants received a target-present line-up and half a target-absent line-up. A between participants’ design was used. Participants were randomly assigned to one of the experimental conditions.

2.4 Procedure
Participants came in groups of 8 persons. They knew that the experiment focused on voice identification. Each group of 8 participants was exposed to the voice of the target. Half the participants were allowed to leave after exposure and were asked to return after a week. The other half received a written instruction and were given the opportunity to ask further questions. The instruction stressed the following aspects:

- There is a possibility that the target voice is not in the line-up.
- The test assistant who is operating the recorder does not know whether the line-up contains the target.
- If you have any doubts, do not mention anyone.

After they had indicated that they understood the instructions, participants were individually presented with a voice line-up. They made a single decision, after hearing all the speakers in the line-up, and were not allowed to hear the line-up more than once. They were asked to indicate whether or not the target was in the line-up. If they answered that they did not know, they were subsequently encouraged to make a (forced) choice between either option. If participants thought the target voice was present, they were asked to indicate the voice number. After this response, the research assistant indicated on a 7-point scale how confident he was about the correctness of the participant’s answer. Care was taken to prevent the participant from seeing this score. After the research assistant had given his confidence judgment, the participants were asked to indi-
icate how confident they were about their judgment, also on a 7-point scale. In all, the identification test took about 5 minutes.

3. Results

3.1 Overall performance
Table 1 shows the answers that were given by participants across all conditions. Twenty-three percent of participants indicated that they ‘didn’t know’. In the target-absent condition, 51% of participants erroneously believed they correctly recognized the target’s voice (false alarms). In the target-present condition, the target was identified correctly by 42% of participants (hits), while 24% of participants picked the wrong voice and 34% did not select a voice or responded that they did not know.

<table>
<thead>
<tr>
<th>Speaker selected</th>
<th>No selection</th>
<th>?</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>-T</td>
<td>90 (51%)</td>
<td>41 (23%)</td>
<td>47 (26%)</td>
</tr>
<tr>
<td>+T</td>
<td>115 (73 correct) (66% (42%)</td>
<td>26 (15%)</td>
<td>34 (19%)</td>
</tr>
<tr>
<td>N</td>
<td>205 (58%)</td>
<td>67 (19%)</td>
<td>81 (23%)</td>
</tr>
</tbody>
</table>

Table 1. Number of participants that selected a voice, did not select a voice, or indicated that they did not know, for target-absent (-T) and target-present (+T) conditions. False Positives (51%) and Correct Identifications (42%) are in bold type.

3.2 Diagnostic value of the test results
On the basis of these data, the diagnostic value of the test may be calculated. Strictly speaking, a separate diagnostic value can be calculated for each result (identification, elimination, ‘don’t know’), to indicate the extent to which the result in question makes a particular hypothesis more or less probable. For an identification, for example, the percentage of hits (hits divided by total number of judgments in the target-present condition) is divided by the percentage of false alarms in the target-absent condition. As there is only one suspect in the line-up, the percentage of false alarms in the target-absent condition should be divided by the number of speakers in the line-up. In the present case, the number of false alarms is therefore first divided by 6 and then by the total number of judgments in the target-absent condition). The resulting value indicates how diagnostic the test is for the specific case. If the resulting value is ‘1’, the test is not diagnostic at all; if the diagnostic value is smaller than 1, the test result provides evidence in support of the opposite of the hypothesis. In the present experiment the diagnostic value of a target identification is 5 (73/175 : (90/6/178)). For an elimination, i.e., the diagnostic value of a response to the effect that the target speaker is not in the lineup, it is (41/178 : 26/175 =) 1.55 for non-identity, and 0.7 (26/175 : 41/178) for identity. Finally, the diagnostic value of ‘don’t know’ for identity is (35/175 : 47/178 =) 0.75.

Participants who indicated they did not know were asked to make a forced choice. As appears from Table 2, most of the participants who were encouraged to make a forced choice did so (84%). In the target-absent condition this tendency resulted in a large amount of false alarms (85%). In the target-present condition, 82% of participants thought they recognized the target, but only half of them correctly identified the target’s voice.

Table 2: Number of times speaker was selected or not selected by participants who gave judgment after initially indicating that they did not know. False Positives (85%) and Correct Identifications (38%) are in bold type.

<table>
<thead>
<tr>
<th>Speaker selected</th>
<th>No speaker selected</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>-T</td>
<td>40 (85%)</td>
<td>7 (15%)</td>
</tr>
<tr>
<td>+T</td>
<td>28 (82%) (13 correct, 38%)</td>
<td>6 (18%)</td>
</tr>
<tr>
<td>N</td>
<td>68 (84%)</td>
<td>13 (16%)</td>
</tr>
</tbody>
</table>

Figure 1: Percentage of correct responses as a function of line-up (left: target-absent (-T), right: target-present (+T)), retention interval (left bars: immediate (0), right bars: after a week (1 w)), and exposure duration (shaded bars: 30 seconds, white bars: 70 seconds).

A distinction was made between target-absent and target-present conditions. No significant differences were observed in the tar-get-present condition (mean percentage correct is 53%). In the target-absent condition, an interaction was found between expo-sure duration and retention interval ($\chi^2 (1)=4.63, p<0.05$). No effects of retention interval were found when exposure duration was short (33% versus 23%, ($\chi^2 (1)=1$). However, when exposure duration was longer, participants performed better after one week than when they were immediately tested (23% versus 50%, ($\chi^2 (1)=4.67, p<.05$).

The diagnostic value of identification and elimination for each experimental condition is shown in Table 3.
Exposure Retention Identification Elimination
30 sec. direct 4.2 1.4
30 sec. week 4.1 .18
70 sec. direct 4.4 .95
70 sec. week 8.8 2.5

Table 3: Diagnostic values of identification and elimination by listener as a function of exposure duration and retention interval.

It appears that the diagnostic value of the test is highest (for both identification and elimination) when both exposure duration and retention interval are long.

3.4 Predictive variables
In order to investigate whether accuracy of the judgments could be predicted by specific variables, a logistic regression analysis was carried out with correctness of the judgments as dependent variable and age, sex, education, confidence judgment of the research assistant, and confidence judgment of the participants as predictors. Only the confidence score of the research assistant proved to have predictive value. Overall, with only this variable 64% of judgments could be predicted correctly. After chance correction (Cohen’s Kappa), 23% of judgments were correctly predicted. If all variables were included in the comparison, again 23% of judgments were correctly predicted. Age, sex, education and confidence level of the participant therefore did not appear to have additional predictive value.

If only the judgments were taken into consideration for which the research assistant felt confident, the diagnostic value of identifications increased from 5 to 6.8. The diagnostic value of elimination increased from 1.55 to 2.2.

4. Discussion
The results show that 42% of participants correctly identified the target voice, while there were 51% false alarms in the target-absent condition. Thus, under the present circumstances, the probability that an innocent suspect is incorrectly identified is 9% (51/6). These findings correspond with previous results: [13], for example, found a mean hit rate of about .40 and a false alarm rate of .52. As the diagnostic value of the test is quite low, the conclusion must be that earwitness identifications have to be treated with extreme caution [5, 8]. Two reasons can be mentioned that may have affected the large number of incorrect identifications. First, the target voice could be described as reasonably average. In general, it is harder to recognize such a voice than a voice with more characteristic features [5]. Second, the voices in the line-up may have been too similar. Even though the target voice might be well encoded in memory, discrimination may be too difficult a task if the other voices in the line-up are very similar to the target voice [7].

The most striking finding was the relationship between exposure duration and retention interval for the target-absent condition. No effects of retention interval were found when exposure duration was short (30 seconds), but with the longer exposure duration there were fewer false alarms after a one-week delay than with an immediate test. This finding is not in agreement with previous findings, [9], for example, did not find any effects of retention interval while [13] found an increase in false alarm rate for the target-present condition.

In agreement with previous research, the confidence score of the participants had no predictive value over the accuracy of his or her judgment [13]. Confidence of the witness should therefore not be used to assess the reliability of a judgment. The only value that could partly explain accuracy was the confidence score of the research assistant. In 23% of cases he was able to correctly predict accuracy of a witness’s judgment. Evidently, an observer is better able to evaluate the reliability of a judgement than is the listener him- or herself. Possible explanations are that observers confine themselves to more diagnostic cues such as facial expressions, or that the listener is influenced by psychological mechanisms such as uncertainty reduction. In any case, taking into account the research assistant’s confidence score increases the diagnostic value of the line-up.

5. References