TESTING THE FAIRNESS OF VOICE IDENTITY PARADES: THE SIMILARITY CRITERION

A.C.M. Rietveld* & A.P.A. Broeders**

*Dept. of Language & Speech, University of Nijmegen, Netherlands
**National Forensic Science Laboratory, Rijswijk, Netherlands

ABSTRACT

Several factors may adversely affect the reliability of a voice identity parade. This paper concentrates on one of these, the degree of similarity of the voices used in the line-up. It describes two techniques which may be used to measure voice similarity - pairwise comparison and the use of semantic scales - and compares the results with the scores obtained in a recognition experiment using a six-speaker voice line-up. It is suggested that a modified version of the paired comparison technique could usefully be applied either to select voices for inclusion in a line-up or to interpret the results of a voice line-up.

1. INTRODUCTION

In a recent survey of the literature on voice identification research , Deffenbacher et al. [2] observe that voice recognition studies are few in number and widely scattered. Studies dealing with voice identity parades are fewer still. This is somewhat surprising in view of the fact that the voice identity parade is a procedure which, potentially, has a wide range of application in police investigations and legal proceedings. It is frequently the case that the voice of a person involved in the commission of a crime is one of the few clues to the identity of that person. Of course, once a suspect is available, it is possible to ask an earwitness if the voice of the suspect is the same as that heard at the time of the crime. But there is obviously a real danger here that the earwitness may 'recognize' the voice, simply because the suspect's voice sounds similar to that of the criminal. There are many other types of bias which may render the results of an aural confrontation virtually meaningless. Some of these are similar to those that apply to visual identity parades (Clifford [1]). Hammersley and Read [3] briefly discuss some of the precautions that should be taken in conducting a voice line-up. A major criterion for the reliability of the voice line-up is that it should be 'fair'. An important implication of this is that the voices should be similar. It would therefore be useful if some objective measure existed by means of which voice similarity could be determined. Ideally, this would be used in the selection of voices for inclusion in the line-up but if, for some reason, this has not happened, it could also be a useful tool to interpret the results of a voice parade.

2. AIM OF THIS STUDY

The main objective of this study is to explore the possibility of measuring voice similarity for the purposes of a voice identity parade. Two methods, the use of semantic scales and pairwise comparison, are examined, and the results are compared with those of a voice recognition experiment involving a six-speaker line-up.

3. STIMULUS MATERIAL

Various precautions were taken to avoid bias in the stimulus material. Five educated male speakers of Dutch were recruited to produce material similar to that contained in an authentic recording of a 'target' speaker. They were selected on the basis of close similarity to the target speaker in terms of age, educational background and accentedness, and on the basis of their credibility in terms of the role they were asked to play. The six speakers selected were between 35 and 50 years of age and had a mild to very mild The Hague accent. Special care was taken to avoid bias due to speech content and to speech style as a function of language use. The five foils all took part in two approximately 5-minute telephone conversations with a third party, as part of a (fictitious) campaign to recruit representatives for a new company. Prior to the recording of the telephone conversations, the five foils were given two sheets, one describing the aims and organisation of the company and the future activities of the representatives, the other containing a list of keywords and phrases to be used as a prompt during the telephone conversations. The third party was provided with similar information to enable him to pose as a prospective representative. Although the material produced by the foils was therefore neither strictly unrehearsed nor unmonitored, none of the listeners turned out to be aware of this. The telephone conversations were edited to remove the voice of the third party. Four stimulus tapes were produced: The first, Tape 1, consisted of 150 sec. samples of nett speech from each of the 6 speakers (i.e. the five foils and the target). The second, Tape 2, consisted of 60 sec. samples of nett speech different from that used for the compilation of the first tape. Tapes 1 and 2 were used in the voice recognition experiment. The third tape, Tape 3, consisted of two sets of 10 sec. samples taken from the 60 sec. samples on Tape 2, randomly arranged to form two series of 15 pairs. Tape 4 consisted of a single listing of the six 10 sec. samples used for Tape 3.

4. EXPERIMENTS CONDUCTED

4.1 Semantic Scales

In the first experiment, 10 listeners were played Tape 2, consisting of 150 sec. samples of each of the six stimulus voices, and asked to rate the speakers on 16 scales, eight of them referring to speech characteristics and eight to speaker personality characteristics.

4.2 Pairwise Comparisons

In the second experiment, two groups of 5 subjects first listened to Tape 3. They were asked to familiarize themselves with the range of voices on this tape before moving on to Tape 4, containing the 15 randomized pairs. For each pair, they were instructed to express the degree of difference between the voices they heard on a scale from 1 to 10, with 1 standing for 'the same' and 10 for 'very different'.

4.3 Voice Recognition Parade

The third experiment was a voice identity parade. The 150 sec. samples of speakers number 1, 3 and 4 respectively, on Tape 1 were played to three groups of second-year Business Communications students. They were told to pay special attention to the voice they heard rather than to what was being said, as they would be questioned about this voice in a week's time. Exactly one week later they listened to Tape 2, containing the 60 sec. fragments. Prior to this, they were told that they were going to hear six speakers, one of whom might be identical to the one they had heard a week earlier. Contrary to the truth, they were also told that none of the speakers might be identical to the target voice, since the experiment involved several groups and that in some of these the target voice was not on the tape. They were asked to circle the number on their answer sheet corresponding to the number preceding the speaker of their choice and to circle 0 if they judged none of the speakers identical. They were also asked to indicate the degree of confidence in their decision on a 5- point scale.

4.4 Mean F0

The 10 sec. samples of Tape 4, used for the Pairwise Comparison task were also used to arrive at a mean F0-value + standard deviation for each of the 6 speakers, using the SIFT algorithm. The following values were found:

Table 1 Mean F0-values and standard deviations

Speaker	Mean F0 (Hz)	S.D. (Hz)
1	103	16
2	139	24
3	82	18
4	123	19
5	110	34
6	104	15

5. RESULTS

5.1 Semantic Scales and Paired Comparisons

The scale values obtained on the 16 scales were used to calculate interspeaker distances by means of the common squared Euclidean metric. Three sets of distances were calculated: a) distances based on the complete set of 16 scales, b) distances based on the 8 speech scales, and c) distances reflecting the scores on the 8 personality scales. The calculation of distances on the basis of scale values is a somewhat hazardous affair, as the number of scales covering a specific aspect of the object under investigation may have a substantial effect on the distance obtained. For that reason, factor scores should be used, as these scores are not correlated. In actual forensic practice, however, only a limited number of subjects will normally be available, so that the use of factor analysis is not possible, as the number of variables largely exceeds the number of cases.

The following table shows the correlations between *Dall* (distances based on all 16 scales), *Dsp* (distances based on speech scales), *Dper* (personality scales) and *Diss* (the overall dissimilarities, obtained in the paired comparison test), for all 15 pairs of the 6 speakers involved.

Table 2 Correlations between 3 types of distances and the overall dissimilarities; N= 15 (see text). Significant correlations (p = 0.05) are marked *.

	Dall	Dsp	Dper	Diss
Dall				
Dsp	.98*			
Dper	.95*	.97*		
Diss	.45	.46	.51	

The correlations given above show that the scale-based distances and the overall dissimilarities are not equivalent. We also applied cluster analyses to the distances and dissimilarities. The results provided confirmation for the difference found between the two approaches: paired comparisons vs. the use of semantic scales. So we have reason to believe that the two methods of assessing the homogeneity of a group of subjects are not equivalent.

5.2. Voice Recognition Parade

The results of this experiment were as follows:

Table 3 Identification results (Gr= Group; C.I. = correct identification, F.I. = false identification, F.E. = false elimination).

Gr	Target	C.I.	F.I.	F.E.	N
A	1	91.7%	8.3%	-	12
B	3	90.1%	-	9.9%	11
C	4	69.2%	30.8%	-	13

An analysis of the False Identifications reveals that, while one listener mistook Speaker 4 for Speaker 1, 4 listeners wrongly identified Speaker 1 as Speaker 4. The bias towards Speaker 1 may be due to the order of presentation. Group A, whose target speaker was Speaker 1, heard their target before they could be confused by Speaker 4, while Group C, whose target was Speaker 4, first heard the apparently rather similar Speaker 1.

6. FURTHER CONSIDERATIONS

Given the fact that the two methods do not produce equivalent results, it would obviously be desirable to assess their validity by means of some independent test. One way of doing this would be to correlate the results of the two techniques with the confusion scores of a large number of voice recognition tests involving all six speakers in turn as targets, conducted at various time intervals. So far, only the results presented in 5.1 are available. Unfortunately, it appears that with a one-week interval between presentation and recognition sessions, recognition scores are very high so that a ceiling effect is produced. It is expected that longer delays between presentation and recognition sessions will produce the type of scores that are required to calculate correlations. On present information, we would be inclined to prefer the paired comparison test. It has two distinct advantages over the semantic scale test, one theoretical, the other practical. As we have already observed, the use of semantic scales inevitably involves a certain amount of overlap between the scales used, which may seriously affect the distance indices obtained. The distances obtained with the paired comparison test should provide a more accurate reflection of the dissimilarity of the voices. From a practical point of view too, the paired comparison test is preferable since it is considerably less time-consuming and labour-intensive. However, it should be noted that the dissimilarity measures obtained in the paired comparison test need to be converted to metrical distances by means of a Multidimensional Scaling Technique. This conversion presupposes a small 'Stress-value', associated with a relatively small number of underlying dimensions.

7. A PRACTICAL PROPOSAL

As discussed in the introduction, the voices used in a voice parade should not constitute too heterogeneous a set. The target voice in particular, should not occupy an outlying position in the perceptual space. Presumably, an ideal situation for a voice parade would be for all voices to be located in equidistant positions on a concentric circle around the centroid of the perceptual space. A rule of thumb might be: the target voice should not be situated at a distance from the centroid greater than the average distance + its standard deviation. If this principle is applied to the distances obtained in the paired comparison test, the result is that illustrated in Fig.1

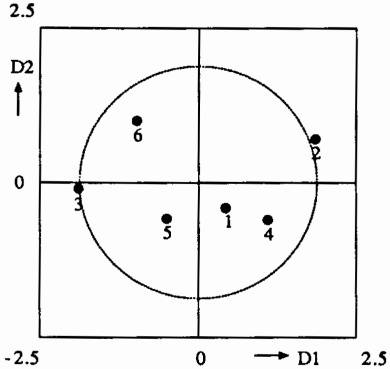

Figure 1 The locations of the 6 speakers in the Perceptual Space

The SPSS-procedure ALSCAL was used; with two dimensions Kruskal's Stress formula 1 was .063 (RSQ = .972). Here, the sum of the average distance plus the s.d. (=1.882) is taken as the radius of a circle around that point. The value of the target voice should not exceed that value. It appears that our target (no 6) is located within the desired distance to the centroid.

8. CONCLUSION

Of the two procedures which may be used to assess the similarity of voices used in a voice line-up, the paired comparison test appears the more promising. Further research to obtain independent support for its validity is in progress.

A final observation concerns the proposed sphere of application of the similarity test. It is clearly not intended as a foolproof procedure which can simply be applied to any random set of voices. It is only when every conceivable effort has been made to avoid bias of any kind in the selection of the voices that the results of the test will be meaningful.

REFERENCES

[1] CLIFFORD, B.R. (1980), "Voice Identification by Human Listeners: On Earwitness Reliability", *Law and Human Behavior*, 4, 373-394.

[2] DEFFENBACHER, K.A. et al. (1989). "Relevance of Voice Identification Research to Criteria for Evaluating Reliability of an Identification". *Journal of Psychology*, 123(2), 109-119.

[3] HAMMERSLEY, R.H. & J.D. READ (1983). "Testing Witnesses' Voice Recognition: Some Practical Recommendations". *Journal of the Forensic Science Society*, 23, 203-208.