VOICE LINEUPS: A PRACTICAL GUIDE

Gea de Jong-Lendle[1], Francis Nolan[2], Kirsty McDougall[2], Toby Hudson[2]

[1]Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany
[2]Department of Theoretical and Applied Linguistics, University of Cambridge, United Kingdom
gea.dejong@staff.uni-marburg.de, fjn1@cam.ac.uk, kem37@cam.ac.uk, toh22@cam.ac.uk

ABSTRACT

In this article the authors aim to offer some practical advice for phoneticians being confronted with a voice parade request for the first time, by providing a detailed description of a parade carried out successfully in the UK. The methodology used in this parade is based on a number of studies on earwitnesses carried out since the 1990s. However, it primarily builds on the work by Nolan [5] for general advice on the construction of parades and on Rietveld and Broeders [7] for measuring the similarity of voices. In addition, useful ways are suggested to instruct and train identification officers involved. Special care is taken to ensure an efficient and smooth execution of the parade, keeping the risk of errors to a minimum. Finally, general issues important to voice parades are raised and discussed.

Keywords: earwitness identification, voice lineup, voice parade, forensic speaker analysis.

1. BACKGROUND

In the village P in the north of Britain an incident occurred several years ago, where two men were involved in an assault upon two teenagers V1 and V2. One of the attackers was wearing a mask. The victim V1 however claimed that he recognised the voice of the masked man as the voice of a local police officer (hereafter referred to as SUS). Some six months later, Detective Inspector DI contacted us and requested our assistance with the construction of a voice parade.

2. CONSTRUCTION OF THE PARADE

As the construction of a voice parade is costly and time consuming, it was first checked whether the circumstances of this crime satisfied the conditions for carrying out a parade. The following issues were considered: 1) is the parade really necessary? Is there enough evidence for the trial to go ahead without the lineup or is there only a small amount so that a positive identification could just provide that missing piece of evidence in a trial? As there is always the risk of a negative outcome despite the suspect being present in the parade (i.e. incorrect identification/no selection), it is important to discuss the effect this could have on an otherwise strong case. 2) Is the voice heard at the time of the crime familiar or unfamiliar to the victim, requiring a type I or type II parade respectively? 3) How much time has passed since the time of the crime? In the case of an unfamiliar voice this time delay is a crucial factor, as recognition accuracy decays with time [2, 4, 7]. How long a time delay can be obviously also depends on factors like the extent and nature of exposure to the target voice [6, 7]. However, it should normally not exceed a couple of weeks. In the case of a familiar voice, the time issue is less relevant. 4) How much speech was heard at the time of the crime? 5) Do we have access to recordings of at least 7 or 8 voices which are compatible with the voice of the suspect, to act as foil voices, or is the suspect voice too unusual or distinctive? 6) Does the severity of the crime justify the costs?

In this particular case, it was decided that a voice parade should be carried out, a type II parade serving the purpose of testing the victim´s familiarity with the voice (and only that).

2.1. Constructing the samples

In order to create speech compilations for 8 foils, interview tapes from other unrelated crimes in the local area were requested. The first set of 20 interview tapes involved, white British males, aged 30 to 36 years, and born around the P-area, to match the characteristics of the suspect. The interviews selected were associated with violent crimes.

Auditory analysis revealed that whilst the suspect's recording would provide an adequate edited sample of one minute for the voice parade, only five of the potential 'foil' interviews were satisfactory. Most of the rejected tapes contained too little speech of a useable kind (i.e. speech not immediately identifiable with a particular crime), or contained speech too dissimilar from the suspect's in terms of accent (including educational background) and/or voice quality, were too poor in recording quality, or involved the interviewee suffering from extreme tiredness, a cold, or alcohol-intoxication. The last of these was mainly a consequence of the fact that a large number of the speakers had been arrested after a violent episode late evening and

were often still under the influence of alcohol. The speech of the intoxicated suspects was extremely different from the clean sample produced by a totally sober (and educated) police officer. The DI selected a further 33 interview tapes, focusing this time on suspects with manual or professional jobs and extending the type of offences to damage, public order and theft.

During the auditory analysis it was noticed that some police interviewers also matched the linguistic profile of the suspect quite well, since in a typical police interview the officer also spends a considerable amount of time describing the incident in order to confirm particular details with the suspect., Permission to use officers' speech as possible foils was requested and subsequently granted. The resulting set consisted of 11 foils (9 interviewees and 2 officers).

The selection of speech extracts to be used in the lineup and the pre-tests was carried out as follows: First, short, coherent chunks which could be segmented from each 'foil' interview without unnaturalness were identified. These ranged in length from single words, usually less than a second in length, to short sentences, up to about 3 seconds.

The reason for choosing rather short utterances is that long chunks are hard to find in some interviews, and, more importantly, frequently reveal the nature of the crime for which the suspect is being interviewed. As far as possible the chunks were chosen to be comparable in content across samples, including for instance brief answers to questions about times and whereabouts (avoiding, however, identifiable local landmarks in the area of the crime in the present case or references to a particular type of crime).

2.2. Testing the fairness of the parade

A major criterion for the reliability of the parade is that it should be fair, one of the implications being that the voices in the parade should be similar. To assess the fairness of the lineup two pre-parade experiments were conducted: the Paired Comparison Test and the Mock Parade.

2.2.1. The Paired Comparison test

To test the similarity of the selected voices statistically, the Paired Comparison Test was conducted. Rietveld and Broeders [7] suggested implementing this test in voice parades; a full outline of the methodology can be found in [3]. In short, the Paired Comparison Test consists of pairs of voices that are presented to listeners who are asked to assess the similarity on a numerical rating scale. Multidimensional Scaling (MDS) is subsequently applied to show the relationships of perceived similarity within the whole group of foils and suspect. It is used here 1) to make sure that the voice of the suspect is not too different from the foils, and 2) to enable us to select the best 8 foils from the available set of 11 possible foils.

Each experimental stimulus (i.e. pair half) consisted of a single sound file of approximately 5 seconds of one or two chunks of speech per speaker. Each foil sample was paired with every other foil sample and the suspect's voice, with one second of silence in between resulting in 66 pairs in total. Before the test, every listener was familiarised with the entire set of 12 voices as part of the practice run. For each listener all pairings were presented in a different randomised order. *Praat ExperimentMFC* [1] was used to present the voice samples and to record the listeners' answers. Listeners were given the instructions shown in Fig. 1:

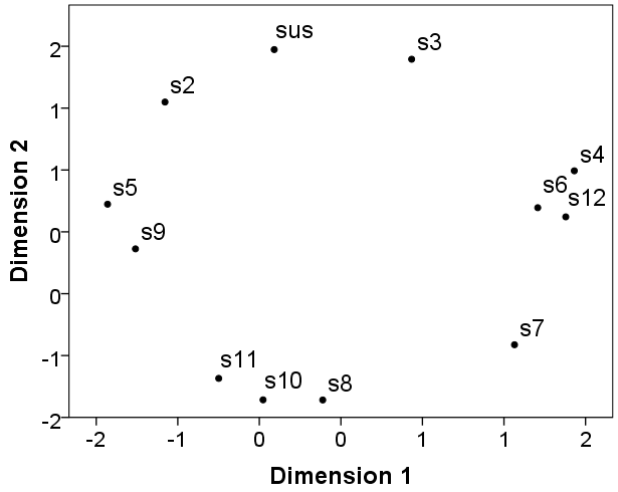
Fig. 1: Paired Comparison Test: Instructions

> "In this session you are asked to compare a number of voices and assess their degree of similarity. For each comparison you will hear 2 short speech recordings (approximately 5 seconds each) separated by a brief pause. Taking into account voice quality and accent, but as far as possible ignoring the meaningful content of the speech and background noise, please rate the similarity of each pair of voices on a scale from 1 (very similar) to 9 (very different). Record your judgment as a rating of the difference between the voices by left-clicking on a displayed number. You will then hear the next pair of voices.
> Some voices may be very similar, but don't worry about whether you are hearing the same speaker: just rate the similarity. The session consists of 66 audio pairs for comparison and will take around 20 minutes in total. The speed at which you proceed through the comparisons is in your power but since a snap reaction is what is required, do not stop to agonise over any single comparison. Before the session you will be given a practice run of 6 comparisons for you to practise making judgments. In this pre-test you will be exposed to the full range of voices to be compared during the experiment."

Altogether 12 subjects, all students and staff at the University of Cambridge (aged 23-61y), took part. Subjects listened to the samples on a PC using Sennheiser-HD 570 headphones in a quiet room. The listeners' judgements were subjected MDS and the first two dimensions of the resulting output plotted on a scatterplot, enabling the perceived distance/similarity among the twelve speakers on those two dimensions to be visualised, as shown in Fig. 2. The figure confirms that no individual speaker stands out as sounding markedly different from the other 11 speakers. Overall the data points are relatively evenly spaced, with no single speaker's data point being located a long way from the group. The suspect's data point was not an outlier but appropriately spaced among the foil

speakers and close enough to all other speakers. Rietveld and Broeders [7] suggested that it should be situated no further from the centroid than the average distance + 1 SD and this was indeed the case. In the ideal case SUS would be in the centre.

Figure 2: Two-dimensional spatial configuration derived by MDS for the 12 listeners' judgments of the similarity between the suspect "SUS" and the 11 potential foil speakers (S2-S12).



In order to select the 8 foils to be included in the voice parade, the Euclidean distance between the suspect and each of speakers S2-S12 on the two-dimensional map was calculated. The 8 speakers with the shortest distances to the suspect were selected for inclusion in the voice parade.

Table 1: Euclidean distances between the suspect and each potential foil speaker, and relative ranking of each speaker (1 = judged most similar to suspect, 11 = judged most dissimilar). The speakers selected for inclusion (ranking 1-8) are indicated in bold.

Potential foil	Euclidean distance to suspect	Rank
s2	0.7928	1
s3	0.8460	2
s4	2.0842	6
s5	1.6137	3
s6	2.0598	5
s7 (officer)	2.8038	9
s8	2.8475	11
s9	1.8217	4
s10	2.8314	10
s11 (officer)	2.6794	8
s12	2.2412	7

2.2.1. The mock witness test

Once the final set of 8 foils was selected, speech samples for the mock witness test (and the final voice parade) were prepared. Here, selections of speech chunks characteristic of the given speaker were concatenated into a single sound file of approximately 45 seconds.

The suspect's sample was constructed last to ensure that its style and content fell within that established by the other samples. Again, any reference even remotely connectable to the case under investigation or the suspect was avoided.

Next the mock witness test was carried out. 12 subjects, all students or staff members at the University of Cambridge (aged 21-55y), took part. None of the listeners had participated in the earlier Paired Comparison Test. Order effects were controlled for.

These listeners ('witnesses') were asked to read the following instructions, and any questions they had were then answered.

Fig. 3: Mock witness test: Instructions

"Your task is to listen to nine speech samples, each made up of extracts from a real police interview with a suspect and lasting about 45 seconds. The interviews range over a number of crimes. For each sample, you must estimate how likely or unlikely it would be for the speaker: to risk his life for a stranger / to commit a violent assault / to be a police officer / to be a drug addict.

This you will do on the nine-point scales below. You may well feel that such judgments are difficult or impossible to make, but it is important that you respond to every question. The speech samples are labelled A, B, C, D, E, F, G, H, J. After hearing each one, please circle the point on the nine-point scale which best reflects the likelihood."

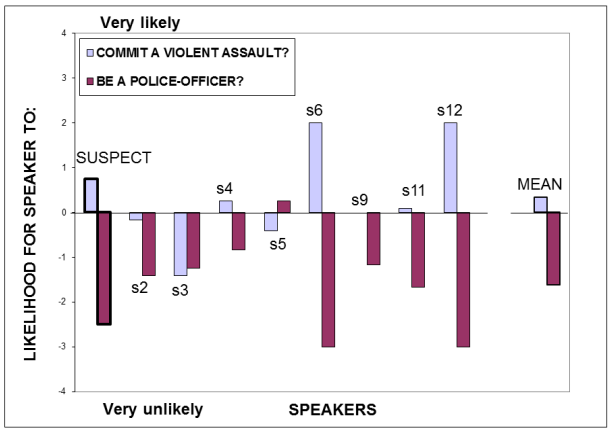
They were then given rating scales (very unlikely -4 to +4 very likely) and listened sequentially to each sample, rating at the end of the sample.

Under normal circumstances one would have briefly described the crime in question and then asked the listener how likely or unlikely it is that the suspect is being interviewed about crime X. In this case, however, where a police officer was suspected of an assault, it was judged that a set-up like that might be too confusing for the listeners: police officers are usually not associated with committing crime in general. It was therefore decided to ask more than one question and to separate the two issues: the suspect being a police officer and the suspect being suspected of an assault. Question 1 and 4 were inserted as distractor questions.

It can be seen from Fig. 4 that none of the speakers was given an extreme rating as either being very likely to commit a violent assault or to be a police officer; even if a witness were unable to

recognise any voice and resorted to guessing on the basis of content, the suspect's sample should be unlikely to be chosen on either account.

Figure 4: Average likelihood judgment scores for the suspect and 8 foils to: 1. Commit a violent assault (purple), or 2. Be a police officer (red). The suspect's scores are the two leftmost bars. Rightmost is the average for all listeners (*n*=12).



2.3. Constructing the final parade

Three randomised orders (order 1-3) within three PowerPoint files were constructed to present the parade. To ensure that the volume was appropriately set, a slide was inserted before each parade, which contained an unrelated sound file (with the same average amplitude as the sound files of the parade). The officer could play this sound file while adjusting the volume until the witness indicated that the setting was loud enough and comfortable.

A second slide was inserted with the heading 'Instructions to the witness'. This slide was left empty to allow the Police to outline their set of instructions to the witness. One important issue was discussed in detail, however: the question to be asked of the witness. The identification of the kind normally made at a parade, that is, a decision of the kind 'that's the voice I heard on the day in question' would in this case be problematic; the witness had already decided that he knew the voice of the perpetrator. In our case, the witness would in fact try to pick the voice of the police officer he claimed to know and not the voice of the person who assaulted him in on the day in question. It was therefore recommended that the Prosecution ask the witness to select the voice of this police officer.

The samples were copied into the PowerPoint file while making sure that they remained uncompressed when played. The samples in each set were labelled sequentially in the order A-J (leaving out the I to avoid confusion with the number 1). Since full control over each slide and sound file was requested, the presentation was created in such a way that each slide would appear with a large letter and a large play button to play the associated voice sample. The point was stressed that the witness should hear the entire set of samples at least once before making a decision. After all samples had appeared and been heard sequentially, they then appeared all together on one slide, each one again with its own play button and identifying letter. At this point, and after the witness has listened to the entire set of samples, he was given the opportunity to listen to any sample again.

A demonstration version of the parade using unrelated speakers from an experimental research database (and excluding the voice of the suspect or any other sound file received from the Police) was sent beforehand in order to familiarise the officers with the set-up of our parade, to allow them to request changes if necessary, and to enable them to test their equipment.

After completing preparation of the three actual parade PowerPoint files, each file plus the entire set of sound files was burnt onto a non-rewritable CD. All speech files (in .wav format, sample rate 44.1K, resolution 16-bit) had been renamed to ensure that the order could not be directly concluded from the file listing on the CD. The three CDs were checked for any problems, tested on a variety of computers and systems, and then separately sealed in a police evidence bag.

2.4. Outcome

The voice lineup was carried out straightforwardly by an independent police officer (though familiar with the demonstration parade) in the presence of the defence lawyer. None of the lineup consultants was present nor was the investigating detective.

The lineup resulted in the victim V1 identifying the suspect from the voice parade and demonstrating in this way his familiarity with the voice of the suspect. The defence lawyer confirmed that the lineup itself and its execution had been fair.

3. CONCLUSION

A detailed account has been given of the construction of a voice parade, where different methods and ideas have been combined and implemented that may contribute to the fairness of a parade. In addition, new strategies have been suggested that may help to make the execution of a parade more efficient and less prone to errors.

4. REFERENCES

[1] Boersma, P., Weenink, D. 2008 Praat: doing phonetics by computer [Computer program]. Version 5.0.23 retrieved May 2008 from http://www.praat.org/.

[2] Clifford, B.R., Rathborn, H., Bull, R. 1981. The effect of delay on voice recognition accuracy. *Law and Human Behavior*, 4: 373-94.

[3] McDougall, K. 2013. Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language and the Law* 20(2), 163-172.

[4] McGehee, F. 1937 The reliability of the identification of the human voice. *Journal of General Psychology*, 17:249-71.

[5] Nolan, F. 2003 A recent voice parade. *Forensic Linguistics* 10(2), 2003.277-291.

[6] Nolan, F., Grabe 1996 Preparing a voice lineup. *Forensic Linguistics*, 3(1): 74-94.

[7] Rietveld, A.C.M., Broeders, A.P.A. 1991. Testing the fairness of voice parades: the similarity criterion. *Proc. of the 12th International Congress of Phonetic Science*s. Aix-en-Provence, Université de Provence, Service des Publications. 5: 46-49.