GETTING THE VOICE LINE-UP RIGHT: ANALYSIS OF A MULTIPLE AUDITORY CONFRONTATION

Andrew Butcher

Department of Speech Pathology
Flinders University of South Australia

ABSTRACT - The practice of confronting witnesses of a crime with a tape recorded 'voice line-up', where the voice of a suspect is included amongst a series of 'foils', is becoming more frequent as a forensic technique. A tape recently used in such a procedure was submitted to acoustic analysis and to auditory analysis by a panel of listeners. Two speakers were consistently identified as being different from the rest. One of these was the suspect. The voice line-up evidence was ruled inadmissible.

INTRODUCTION

The use of multiple auditory confrontations - more commonly known as 'voice line-ups' or 'voice parades' - by police forces in Europe, North America and Australia seems to be on the increase. Such a procedure may be used to obtain evidence of identification in cases where, in the course of committing a crime, the perpetrator has spoken in the presence of witnesses. In its usual form, the procedure involves putting together an audio tape which contains recordings of a number of speakers, including the suspect. This tape is played to the witness(es) and they are asked to state whether they can identify any of the voices as that of the perpetrator. In order to be entirely fair to the suspect, there are a number of criteria which need to be observed both in the construction of the tape and in the administration of the confrontation. As Broeders & Rietveld (1995:25) state: "Failure to observe proper procedure in the administration of an auditory confrontation will almost certainly render its evidential value null and void". As with visual identification parades, it is generally agreed that a general principle of fairness in the conducting of voice line-ups is that there should be no feature of the voices or the recordings which would cause non-witnesses to pick out a particular speaker (whether suspect or foil) as being different from the rest. One criterion is that the voices of the suspect and the foils should all broadly conform to the description of the voice of the offender provided by the witness(es). For example, in the visual context, if the perpetrator is described as bespectacled and clean shaven, one assumes the suspect conforms to this description also, otherwise the line-up is pointless. In this case, the inclusion of one or more bearded foils without glasses diminishes the effective number of speakers and increases the chances of the suspect being picked out. Another criterion is that all recordings should be carried out under similar conditions. Again in the visual context, if the suspect is wearing denims and all the others are in suits, the witness may be predisposed to select the odd man out.

In a recent armed robbery case in South Australia the prosecution submitted evidence obtained on the basis of a voice line-up. The defence asked for the tape to be analysed with regard to the above criteria. The recording consisted of a series of 9 phrases or sentences said to be similar to utterances made by the perpetrator in the course of carrying out one of the robberies. These were read out by 12 speakers.

AUDITORY ANALYSIS

Methodology

The analysis was carried out in the Speech Research Laboratory at the Department of Speech Pathology of the Flinders University of South Australia. The 12 recordings were digitised to hard disk into separate files at a sampling rate of 20.05 kHz and with 16-bit resolution, via a *Silicon Graphics Indy R4000 PC* workstation, using the *Sound Editor* software. A panel of five persons were asked to listen to the 12 recordings and report which, if any, they considered to stand out from the rest, in terms of recording conditions, voice quality, accent, or speech mannerisms. All five were members of staff of the Flinders University Department of Speech Pathology, with considerable experience in listening to and making judgements on samples of recorded speech. After an initial run through in the order as recorded on the tape, recordings were played in any order as requested by the panel member, and as often as required.

Results

The panel was equally divided between speaker no. 6 and speaker no. 8. Two members found speaker 6 the most exceptional, two members found speaker 8 the most exceptional, and the fifth member found both equally exceptional. All members were agreed that no other speakers stood out from the rest. Speaker 6 was judged to stand out on the grounds of his very broad (and possibly non-local) accent, together with his nasality. Speaker 8 was judged an exception on the grounds of his misarticulation of 'v' for 'th' and because of the 'echoey' sound of the recording.

ACOUSTIC ANALYSIS

Instrumentation

The acoustic analysis was carried out using *ESPS (version 5.0)* signal processing software, in conjunction with the *waves+* interactive graphics interface, running under *Unix* on a *Sun SPARC 10* workstation. The results of spectral analysis were not inconsistent with all recordings being made using the same equipment, as all had a very similar bandwidth, extending up to 7.5 - 8.0 kHz.

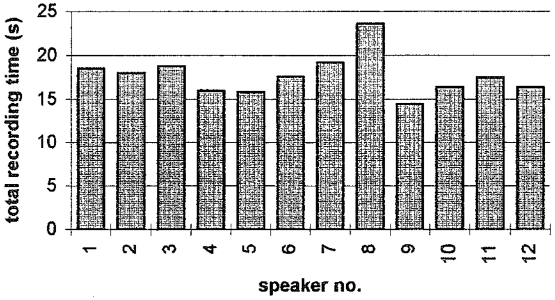

Figure 1. Total recording times for each speaker

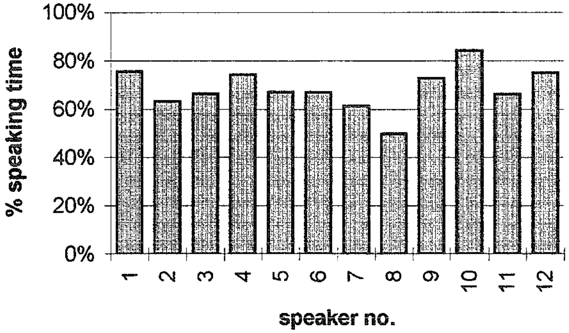

Figure 2. Proportion of speaking time to total recording time for each speaker

Durational Characteristics

The following measures were taken: total recording time, total speaking time (= total recording time minus pauses), and articulation rate (= number of syllables per second). The results are summarised in Figures 1 to 3.

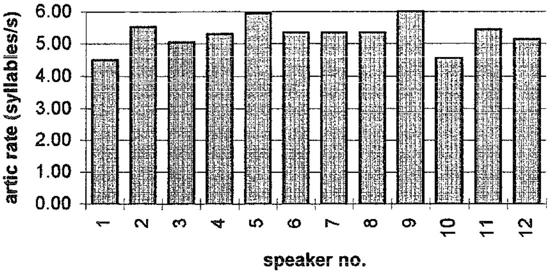

Figure 3. Articulation rates for each speaker

The overall mean length of recording is 17.7 s. Clearly the recording by speaker no. 8 is much longer than the rest. At 23.6 s, it is the only one longer than 20 s. From a comparison of Figure 2 and Figure 3, it is obvious that there are no great differences in articulation rate between speakers, and that the longer duration of the recording by speaker 8 is due to a longer time spent pausing between utterances. In fact speaker 8 only spends 50% of the total recording time actually speaking. No other speaker has a speaking time quotient of less than 60% (the mean for all speakers is 69%).

Fundamental Frequency
Obviously F0 variation is one of the main ways of conveying both grammatical and emotional meaning in speech. Nevertheless, each speaker has a particular range of fundamental frequency which s/he habitually uses and within which s/he feels most comfortable and this is a very important measure for forensic purposes, because it is one of the few measures for which we know the distribution amongst the adult European population. The average speaking fundamental frequency for an adult European male is 113 Hz (Künzel 1989) and 50% of the population lie somewhere between 100 to 130 Hz in spontaneous speech.

The average fundamental frequency of each recording was measured, by means of an algorithm which uses the normalised cross correlation function and dynamic programming. The means and standard deviations of these measures are shown in Figure 4.

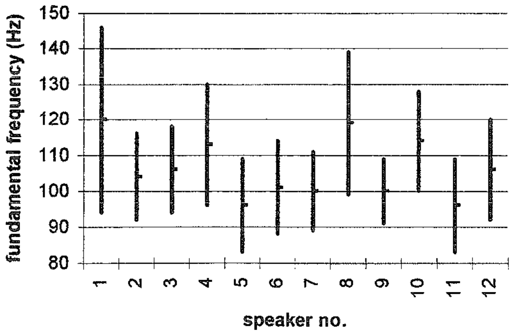

Figure 4. Means and standard deviations of fundamental frequency of all speakers

Note that the majority of speakers (8) have a mean fundamental frequency below the average for an adult male. Only four of the speakers are close to the average (1, 4, 8 and 10). If either Speaker 1 or Speaker 8 is the suspect speaker, then they would clearly be disadvantaged in that the other 10 speakers all have much lower fundamental frequencies than they do. If, on the other hand, the suspect is one of the 8 speakers with a fundamental frequency below 110 Hz, then the inclusion of speakers 1 and 8 as foils may well have decreased the effective line-up size.

Vowel Quality

Formant frequency values were obtained from the recordings using an LPC algorithm after visual location of the measurement point on a sound spectrogram display. For each speaker, three tokens of each of three vowels were measured: /i/ in three repetitions of 'money', /æ/ in three repetitions of 'bag', and /ʊ/ in two repetitions of 'put' and one of 'look'. Figure 5 shows the overall mean first and second formant frequencies of each of the three vowels from all twelve speakers plotted against one another in the standard way. Each vowel symbol is positioned at the intersection of the mean first and second formant frequencies. Ellipses are drawn at a distance of two standard deviations around the mean point for each vowel.

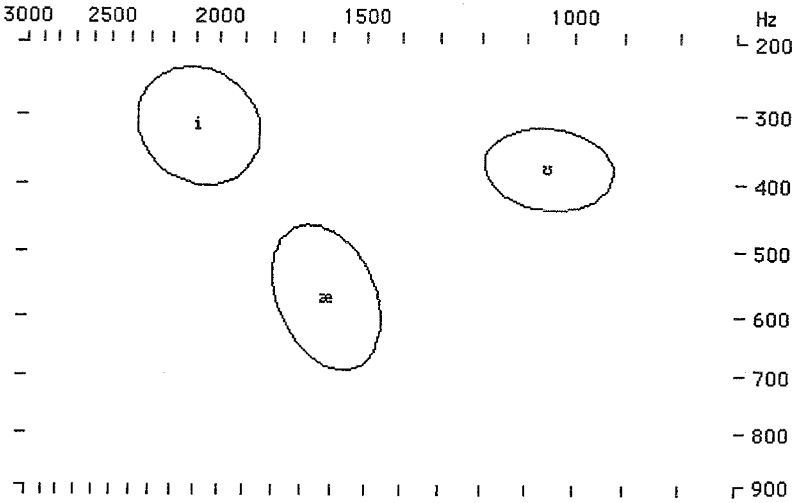

Figure 5. Plot of overall mean formant frequency values for stressed vowels in line-up recordings. Each letter represents the mean for that sound across all speakers. The ellipses represent two standard deviations around the mean.

There is, for the most part, a reasonable uniformity amongst the pronunciations of this group of speakers, as can be seen from the comparatively small standard deviation values. This is born out by an analysis of individual speakers' vowels, from which it can be seen that the means for almost every vowel of almost every speaker fall within or on the border of the two-standard-deviation area for the group as a whole. There is only one striking exception to this generalisation: the /æ/ of 'bag' as spoken by speaker no. 6 is well outside the area for the group, as can be seen from Figure 6. It is pronounced as a much closer vowel - i.e a sound that is closer to /ɛ/ - the vowel of 'beg' in many speakers.

Estimated vocal tract length

The 'neutral' vowel, characteristic of hesitation noises ('er', 'em', etc), but also occurring in unstressed syllables such as 'the' and 'a', is usually pronounced with a relatively unconstricted vocal tract. It may therefore be used to estimate the length of the tract from larynx to lips. Formant frequency values of neutral vowels in three examples of the word 'the' were measured in each recording. Not too much reliance can be placed on calculations made using such a small sample, but, using the standard formula, the vocal tract length of each speaker was calculated and these results are summarised in Figure 7.

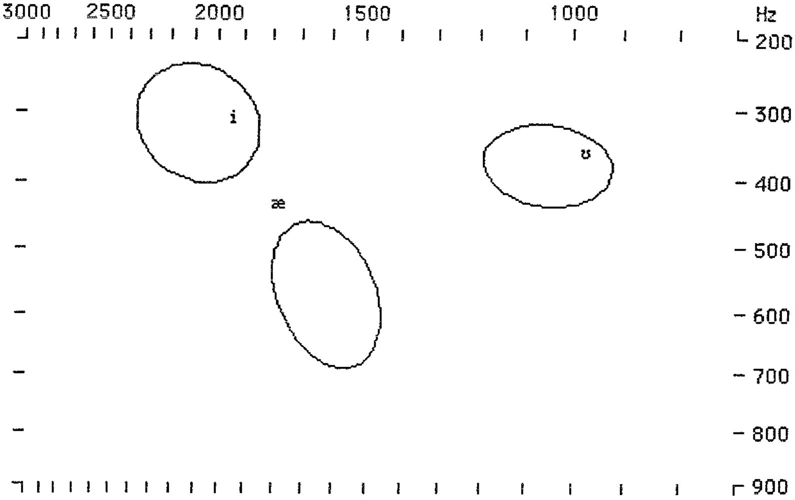

Figure 6. Plot of mean formant frequency values for stressed vowels of Speaker 6. Each letter represents his mean for that sound. The ellipses represent two standard deviations around the mean across all speakers, as in Figure 5.

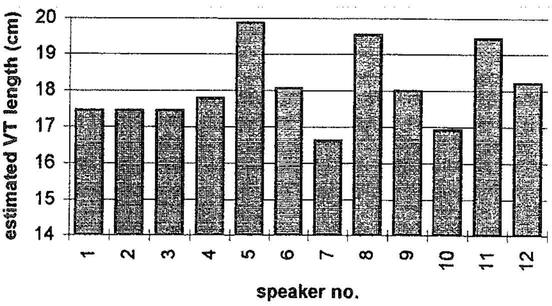

Figure 7. Estimated vocal tract lengths for all speakers

The average vocal tract length for an adult male is 17.5 cm. Variations in length are correlated with differences in higher resonances, which are important indicators of individual voice quality. Most of the speakers in these recordings are close to this average. There are three obvious exceptions. Speakers 5, 8, and 11 all have vocal tracts estimated to be substantially longer than the average (all greater than 19 cm). In the case of speakers 5 and 11 this is perhaps no surprise, as they also have the lowest fundamental frequency of all, indicating that they may be physically quite large individuals (there is by no means a reliable correlation - see Künzel 1989). On the other hand, speaker 8 has the second highest fundamental frequency in the group, combined with the second greatest estimated vocal tract length. I am not aware of any data on the distribution of these two measures in the population as a whole, but the combination is certainly unique to this group and may well correlate with unique vocal characteristics in speaker 8.

CONCLUSIONS

1. A listening test with a panel of five experts resulted in two speakers being singled out: speaker 8, on the grounds of idiosyncratic articulation and different recording conditions, and speaker 6 on the grounds of a possible regional accent difference and (variable) hypernasality.

2. The recording of speaker 8 is some 6 seconds longer than the average, as a result of much longer pauses between utterances.

3. Speakers 1 and 8 stand out as having a fundamental frequency slightly above average for the population as a whole. The majority of speakers in this group have a fundamental frequency well below the average.

4. The formant frequencies of the stressed vowels showed a high degree of similarity across the group. Speaker 6 stands out as having a characteristic (possibly regional) pronunciation of the vowel in 'bag'.

5. A highly tentative estimate of vocal tract length based on a small number of neutral vowels, suggests that the vocal tracts of speakers 5, 8, and 11 are longer than average. Only speaker 8 combines a relatively high fundamental frequency with a relatively long vocal tract.

It had to be concluded that if either speaker 6 or speaker 8 were the suspect, then the commonly accepted criteria that all foil speakers should have voices and accents broadly similar to those of the suspect and be recorded under the same conditions had not been met. In fact it turned out that speaker 8 was the suspect. In view of the this person's idiosyncratic pronunciation of "th" - and that this was a feature of the perpetrator's speech alluded to by the witness, it could be argued that it was incumbent upon the organisers of the line-up to find a majority of foils with a similar speech characteristic. The analogy would be that a fair visual line-up for a suspect with a squint would be expected to include a majority of foils with squints. At the very least, recordings should have been conducted under the same conditions. Recording a suspect in conditions which are audibly different from those of the foils could be likened to showing witnesses a photograph of a suspect with a black background, amongst a series of pictures of foils with a white background.

These findings were submitted by the defence at the *voire dire* hearing. The judge ruled that the evidence of identity based on the voice line-up was inadmissible in chief.

No line-up can be perfect, but all practical steps should be taken to make it as good as possible. A number of guidelines have recently been published which are useful in achieving this aim (e.g. Broeders & Rietveld 1995, Hollien, Künzel & Hollien 1995, Nolan & Grabe 1995)

REFERENCES

Broeders, A.P.A. & Rietveld, A.C.M. (1995) Speaker identification by earwitness. In A Braun and J-P Köster (eds), *Studies in Forensic Phonetics*, (Wissenschaftlicher Verlag Trier).

Hollien, H., Huntley, R.A., Künzel, H.J. & Hollien, P.A. (1995) Criteria for earwitness lineups. *Forensic Linguistics* 2, 143-153.

Künzel, H.J. (1989) How well does average fundamental frequency correlate with speaker height and weight? *Phonetica* 46, 117-125.

Nolan, F. & Grabe, E. (1995) Preparing a voice line-up. Unpublished ms, Dept of Linguistics, University of Cambridge, UK.