Behavioural Speaker Identification for Forensic Applications*

S. J. Aiken, D. G. Jamieson, & V. Parsa

Hearing Health Care Research Unit, School of Communication Sciences and Disorders
The University of Western Ontario, London, Ontario, CANADA

Background

Behavioural speaker identification refers to the process of identifying an individual as the speaker of a given utterance, based only on auditory information. This type of identification is periodically used in forensic applications [1,3], although such usage is not without controversy [2,3]. Given that successful speaker identification is not always achieved under the most ideal conditions [3], it is of dubious value in forensic environments, where conditions are usually far less than ideal [1]. Nevertheless, when a recorded or remembered voice constitutes the strongest evidence in a case, behavioral speaker identification may be an invaluable resource.

This situation occurred in a recent criminal investigation, where the police hypothesized that a specific person was the speaker in a number of potentially incriminating telephone calls. These calls had been made on another person's telephone and recorded under court order. The suspect acknowledged speaking on some calls, but denied being a participant in most of the calls. He cooperated by permitting recording of his voice for comparison.

Objectives

The objectives of the present study were to test the hypothesis of the police, and to specify which calls were likely made by the suspected speaker.

Method

Two listeners rated 40 samples of the word "okay" as same or different in a paired comparison task. Twenty-six of the samples were obtained from police wire-tap (where the identity of the speaker was not known), while the remaining 16 samples were obtained directly from the suspect (also via telephone wire-tap).

The samples were digitized at a frequency of 22 kHz, low-pass filtered at 10 kHz, and edited using CSRE [4], to isolate the word "okay" from the surrounding acoustic information. All samples were presented to the subjects monaurally, via an ER-3A insert earphone, using a listening test generated in ECoS/Win [5].

All possible pairs of different samples were used in the task. Thus, each subject rated 1560 pairs of samples, divided equally into 20 blocks. The pairs were randomized across the 20 blocks, and were further randomized for each subject. Due to the length of the task, raters completed the experiment in two sessions, with 10 blocks in each session. Immediately after hearing each pair of samples, raters indicated whether the speaker of the samples was the same or different, and whether they were certain or uncertain of this decision. Raters could replay the samples as many times as they wished.

Results

Different samples from within telephone calls were not differentiated, as they could not represent different speakers. Thus, there were 143 unique pairs of calls. In order to test the hypothesis of the police, only those voice samples that were obtained by the police were used for the analysis. Thus, 129 unique pairs, along with presentation order and rater, were subjected to an ANOVA, with the assigned rating as the dependent variable. While there was no significant effect of presentation order, there was a significant effect of rater ($p < .005$). The effect of sample pair was also significant ($p < .0001$). Therefore, the hypothesis generated by the police (ie. that the voice samples were produced by a single speaker), was not supported.

A rough estimate of accuracy was generated by comparing the hit and miss rates for voice samples that were known to have been generated by the same speaker (ie. samples obtained from a single telephone call). The wire-tap samples obtained by the laboratory were not included in this comparison, however, because the superior quality of these samples oversimplified the same-different task. The average hit rate was 0.89, and the average miss rate was 0.11. This estimate of accuracy should be interpreted with caution, however, because samples from within single telephone calls share acoustic information apart from the voice spectra (such as specific telephone noise), which could have simplified the same-different decision, and inflated the accuracy rate.

The significant effect of rater indicates differences in rater judgement. Unfortunately, this lack of agreement adds difficulty to the task of speaker identification. If raters do not consistently agree, there is no way to know which rater is correct, and the rating task provides little useful information. The potential for accurate speaker identification diminishes in accordance with lack of inter-rater agreement. Nevertheless, in accordance with the second objective, the data were subjected to a cluster analysis. Two very distinct clusters emerged, with eight calls in each cluster. Only one call did not fit into a cluster. Interestingly, the calls in one cluster were those to which the suspect admitted participating, with only one exception. The suspect denied participating in one call in the cluster, and admitted to participating in one call from the other cluster.

Summary and Conclusions

The present experiment tested the hypothesis that a particular individual was the only speaker in a large set of calls. Results of the experiment failed to support this hypothesis, indicating that there was more than one speaker in the set of calls. Moreover, a cluster analysis revealed two distinct clusters, suggesting the participation of a second speaker. Interestingly, although the police hypothesis was not supported, the cluster analysis clearly attributed to the suspect one call in which the suspect denied participating.

References

[1]Koenig, B. E. (1988). Enhancement of forensic audio recordings. *Journal of the Audio Engineering Society, 36,* pp. 884- 894.

[2]Nolan, F. (1984). *The Phonetic Bases of Speaker Recognition.* New York: Cambridge University Press.

[3]Yarmey, A. D. (1994). Earwitness evidence: memory for a perpetrator's voice. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult Eyewitness Testimony: Current Trends and Developments*(pp. 101-124). New York: Cambridge University Press.

[4]*Computerized Speech Research Environment (CSRE),* Version 4.5 (1996), Avaaz Innovations Inc., London, Ontario, CANADA.

[5]*ECoS/Win: Experiment Generator and Controller (1996),* Avaaz Innovations, Inc., London, Ontario, CANADA.

**Supported by a grant from NSERC to D.G.J.*