AN INTERNATIONAL INVESTIGATION OF FORENSIC SPEAKER COMPARISON PRACTICES

Erica Gold[a] *& Peter French*[a,b]

[a]The University of York, UK; [b]J P French Associates, UK
erica.gold@york.ac.uk

ABSTRACT

The results of the first international survey on forensic speaker comparison practices are presented in this paper. Thirty-four experts from 13 countries and 5 continents responded to a series of questions concerning their practices in casework and which features they found to be useful speaker discriminants. Despite the responses revealing some prominent trends, there is wide variation in methodology, importance assigned to particular speech features, and choice of framework for expressing conclusions.

Keywords: forensic speaker comparison, survey, methodology, features, conclusion frameworks

1. INTRODUCTION

The purpose of this paper is to present the results of the first international survey on forensic speaker comparison (FSC) practices. The motivation for the survey was twofold:

- To make available for the first time to the wider forensic, legal, and speech science communities basic information concerning the working practices of FSC experts across the globe.
- To draw upon the very considerable collective expertise of FSC experts worldwide in order to identify effective working methods and features of speech that have the greatest potential for discriminating between individuals.

2. PARTICIPANTS

Potential participants were contacted through their professional and research organizations [1] and invited to take part in an online survey. 34 responses were collected. Respondents were given the freedom to respond to all or some of the questions.

2.1. Countries

Respondents (21 male; 13 female) were from the following 13 countries: Australia, Austria, Brazil, China, Germany, Italy, the Netherlands, South Korea, Spain, Sweden, Turkey, UK, and USA**.**

2.2. Place of work

Respondents identified their place of work [2] or affiliation. 18 participants represented universities or research institutes followed by 11 employed in government laboratories/agencies. 9 of the experts are affiliated with private laboratories, and 7 work as individuals.

2.3. Experience

The total number of cases from respondents' estimates collectively was 17,951, ranging from 4 to 6,000, with a mean of 528. The respondents had a range of 2 to 50 years of experience in FSC analysis, with a mean of 15.7.

3. METHODS OF ANALYSIS

There is at present no consensus of opinion in the scientific community as to how FSC analysis should be carried out. Rather, a wide range of methods is employed. Methods may be grouped under the following headings:

Auditory Phonetic Analysis Only (AuPA):
The expert listens analytically to the speech samples and attends to aspects of speech at the segmental and suprasegmental levels [7].

Acoustic Phonetic Analysis Only (AcPA):
The expert analyzes and quantifies physical parameters of the speech signal using computer software. As with AuPA, this is labor intensive, involving a high degree of human input and judgment [7].

Auditory Phonetic cum Acoustic Phonetic Analysis (AuPA+AcPA):
This combines the preceding two methods [6].

Analysis by Automatic Speaker Recognition System (ASR):
This requires the use of specialist software designed to identify speakers automatically, with

minimal human input. There are three sequential stages employed by ASR: parameter extraction, parameter modeling, and the calculation of distances [7].

Analysis by Automatic Speaker Recognition System with Human Analysis (HASR):
This involves the use of an automatic system in conjunction with analysis of the auditory and/or acoustic phonetic kind [6].

The distribution of these methods across the 13 countries is provided in Table 1.

Table 1: Methods of analysis employed by countries.

Method	Countries
AuPA	Netherlands, USA
AcPA	Italy
AuPA+AcPA	Australia, Austria, Brazil, China, Germany, Netherlands, Spain, Turkey, UK, USA
HASR	Germany, South Korea, Sweden, USA

The specific features of speech that are analyzed and considered important vary from analyst to analyst within each of the method categories. The data relating to this variation are presented in Section 5.

4. CONCLUSION FRAMEWORKS

As with method of analysis, there is no consensus within the forensic speech science community as to how conclusions are and should be expressed. Currently, there is much debate in the field on the 'logical' and 'legally correct' frameworks for conclusions [4, 5, 7, 8, 9].

A variety of frameworks for expressing conclusions is currently utilized across the world. The conclusion frameworks may be grouped under the following headings:

Binary Decision:
A two way choice – either the criminal and suspect are the same person or different people [2].

Classical Probability Scale (CPS):
The probability or likelihood of identity between the criminal and suspect is stated [7]. Typically, the assessment is a verbal rather than a mathematical one and it may use such terms as "likely/ very likely to be the same or different speakers."

Likelihood Ratio (LR):
This expresses the results as the likelihood of finding the degree of correspondence or mismatch between the samples on the basis of the prosecution hypothesis that they come from the same speaker, against the defense hypothesis that they come from different speakers [9]. Some analysts express the likelihood ratio as a number; others do so verbally [3].

UK Position Statement:
This potentially involves a two-part decision. The first part concerns the assessment of whether the samples are compatible or consistent with having come from the same person. The second part, which only comes into play if there is a positive decision concerning consistency, involves an evaluation of how unusual or distinctive the features common to the samples may be [5].

Some methods of analysis lend themselves more readily than others to the adoption of certain conclusion frameworks. For example, some automatic systems express the results of the comparison as a numerical LR as the default option. A breakdown of methods against conclusion frameworks appears in Table 2.

Table 2: Methods used for analysis in forensic speaker comparisons against conclusion frameworks.

	Binary Decision	CPS	Numerical LR	Verbal LR	UK Position Statement	Other
AuPA		1				1
AcPA			1			
AuPA+AcPA	2	9	1	2	10	
HASR		3	1	1	1	

As seen in Table 2, there is a tendency for participants using AuPA+AcPA to adopt the classical probability scale and UK Position Statement conclusion frameworks.

Table 3 breaks down conclusion frameworks by country. Some countries appear more than once, as there were multiple respondents from the same country, with individual experts implementing different conclusion frameworks.

Table 3: Conclusion frameworks used by countries.

Conclusion Framework	Countries
Binary Decision	Brazil, China
Classical Probability Scale	Australia, Austria, Brazil, Germany, Netherlands, South Korea, Sweden, UK, USA
Numerical LR	Australia, Germany, Italy
Verbal LR	Netherlands, USA
UK Position Statement	Spain, Turkey, UK, USA
Other	USA

A Likert Scale was used to measure the level of satisfaction with a respondent's conclusion method. Likert ratings were averaged across respondents. The scale ranged from 1 (extremely dissatisfied) to 6 (extremely satisfied). Table 4

reports the mean scores of satisfaction by conclusion frameworks.

Table 4: Satisfaction with conclusion framework.

Conclusion Framework	Mean Likert Rating
Numerical LR	5.00
UK Position Statement	4.27
Verbal LR	4.00
Classical Probability Scale	3.67
Binary Decision	3.50

4.1. Population statistics

22 of 31 respondents reported that they use some form of population statistics in arriving at their conclusions. 18 of 31 stated that they had personally collected population statistics for the incidence of occurrence of one or more phonetic or acoustic features.

5. FEATURES EXAMINED IN DETAIL

This section reports on the aspects of recorded speech respondents take into account or consider important in FSC cases. Since respondents were not required to answer every question, responses are given in percentages for those responding to a given question.

5.1. Phonetic features

Respondents were asked whether and with what frequency they examined the following features.

5.1.1 Segmental features

All respondents analyze vowel and consonant sounds in the course of their examinations. In regards to **vowels**, 83% invariably carried out some form of analysis and 17% routinely did so. 94% evaluated the auditory quality of vowels, 97% carried out some form of formant examinations and 55% measured durations.

Of those undertaking **formant** examinations, all measure the second resonance (F2). 86% of respondents reported measuring F1 and an equal percentage reported measuring F3. 18% of respondents stated that they measure F4. In respect of which **aspects of formants** are examined, 93% reported measuring center frequencies of formants of monophthongs, 69% reported measuring formant trajectories of diphthongs and 41% examined vowel-consonant or consonant-vowel formant transitions. 38% stated that they examine formant bandwidth and 14% reported examining formant densities.

In relation to **consonants,** all respondents reported subjecting them to some form of examination; 55% invariably did so. 90% of respondents reported evaluating auditory quality. 81% stated that they examined aspects of timing and 48% reported measuring the frequencies of energy loci.

Table 5 reports the frequency with which consonants, broken down by manner of articulation, are analyzed in FSC cases. Respondents reported using a 6-point Likert Scale ranging from 1 (never) to 6 (always). Mean Likert ratings are represented in Table 5 for those respondents who are native English speakers only.

Table 5: Frequency of consonant analysis in English.

Manner of Articulation	Mean Likert Rating
Fricatives	4.85
Plosives	4.73
Approximants	4.50
Laterals	4.46
Nasals	4.08
Affricates	3.82
Taps/Flaps	3.70
Trills	3.18

5.1.2 Suprasegmental features

All respondents routinely measure **fundamental frequency** in their comparisons. 94% of respondents stated that they examine **voice quality** as part of their overall procedure, although only 68% of these invariably or routinely examine it. Further to this, only 61% of those who examine voice quality do so using a recognized scheme, or modified variant of such a scheme, for its description. 84% of respondents stated that they examine **intonation** with one or another level of frequency. However, of these only 23% do so invariably.

93% of respondents stated that they analyze **tempo** with varying degrees of frequency. Of those analyzing tempo, 80% apply a formal measure (e.g. speaking rate or articulation rate). 71% stated that they examine speech **rhythm** with varying regularity.

5.2. Non-phonetic features

5.2.1 Higher order linguistic features

In addition to examining phonetic features, 79% of respondents reported examining **discourse features** and/or **conversational behaviors** (discourse markers, aspects of turn-taking, telephone opening and closing behaviors, patterns of code switching). 86% stated that they examine **lexico-grammatical** usage. Lexical features were examined most frequently, followed by syntax and morphology.

5.2.2 Non-linguistic features

For the respondents who answered this question,

97% reported examining non-linguistic features at least some of the time. In descending frequency order, specific features were as follows: filled pauses, tongue-clicking, audible breathing, throat clearing, and laughter.

6. WHAT IS CONSIDERED DISCRIMINANT

"The whole is greater than the sum of the parts [1]*."*

In addition to being asked about features within linguistic, phonetic and acoustic domains, participants were given the opportunity to identify which feature from *any* domain they found most useful. Voice quality was reported most often (33%), followed by dialect/accent variants and vowel formants (both 29%). 21% reported speaking tempo as a useful parameter. This was followed by rhythm and F0 (both 17%). Lexical and grammatical choices, vowel and consonant realizations, phonological processes (e.g. connected speech processes) and fluency were all reported by 13% of the respondents. And one respondent went as far as stating that vowel formant analysis "is rarely insightful."

Interestingly, though perhaps not surprisingly, the vast majority of participants alluded to the fact that despite some individual parameters holding significant weight, it is the overall *combination* of features that they consider crucial in discriminating between speakers.

7. DISCUSSION

The purpose of this article has not been to advance any argument or to develop theoretical propositions. Rather, its objective has been the much more mundane one of laying out basic factual information concerning the practice of FSC internationally in the present day.

Those not directly involved in this specialist field but working in related areas, e.g., phoneticians with non-forensic interests and forensic scientists from other disciplines, may well be surprised at the lack of consensus over such fundamental matters as how speech samples are to be analyzed and compared, which aspects of the samples are to be assigned greatest importance during the analytic process, and how conclusions are to be expressed at the end of it. Indeed, it will be apparent that there was hardly a single issue explored in the survey with which anything approaching a consensus of practice or opinion was found. Whilst other areas of forensic science would undoubtedly show some degree of variation across individual practitioners, the wide disparities reported here must surely call for greater consultation, debate, and co-operation across experts, institutions and nations.

The prerequisite for a resolution of the differences is, of course, knowledge of their existence. Insofar as the present study lays bare that information, it may be considered as making a modest first step towards methodological unity.

8. ACKNOWLEDGEMENTS

This project is supported by the Marie Curie Initial Training Network, *Bayesian Biometrics for Forensics.* (*http://bbfor2.net*).

9. REFERENCES

[1] Aristotle. *Metaphysica* 10f-1045a.
[2] Broeders, A.P.A. 2001. Forensic speech and audio analysis. Forensic linguistics. 1998 to 2001. A Review. *Proc. 13th INTERPOL Forensic Sciences Symposium* Lyon, France.
[3] Champod, C., Evett, I.W. 2000. Commentary on Broeders 1999. *Forensic Linguistics* 7(2), 238-243.
[4] French, J.P., et al. 2010. The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *Int'l. Journal of Speech, Language and the Law*, 17(1), 143-152.
[5] French, J.P., Harrison, P. 2007. Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *Int'l. Journal of Speech, Language and the Law* 14, 137-144.
[6] Greenberg, C.S., et al., 2010. Human assisted speaker recognition in SRE10, *Proc. Odyssey 2010*, Brno, Czech Republic.
[7] Jessen, M. 2008. Forensic phonetics. *Language and Linguistics Compass* 2(4), 671-711.
[8] Morrison, G. 2009. Forensic voice comparison and the paradigm shift. *Science and Justice* 49, 298-308.
[9] Rose, P., Morrison, G. 2009. A response to the UK position statement on forensic speaker comparison. *Int'l. Journal of Speech Language and the Law* 16(1), 139-163.

[1] Emails were sent to the European Network of Forensic Science Institutes, the National Institute for Standards and Technology for those who participate in the NIST Speaker Recognition Evaluations, and the International Association for Forensic Phonetics and Acoustics. A number of individuals working at government laboratories/agencies were also contacted to participate in the survey.

[2] Some respondents are associated with multiple places of work.