Developing the vocal profile analysis scheme for forensic voice comparison

Eugenia San Segundo, Vincent Hughes,
Peter French, Paul Foulkes & Philip Harrison

University of York & J P French Associates

voice and identity

01011101ID1

BAAP Colloquium 2016
Lancaster University
30 March – 1 April

Arts & Humanities Research Council

Forensic Voice Comparison (FVC)

- **what:** analyse the speech of unknown offender and known suspect

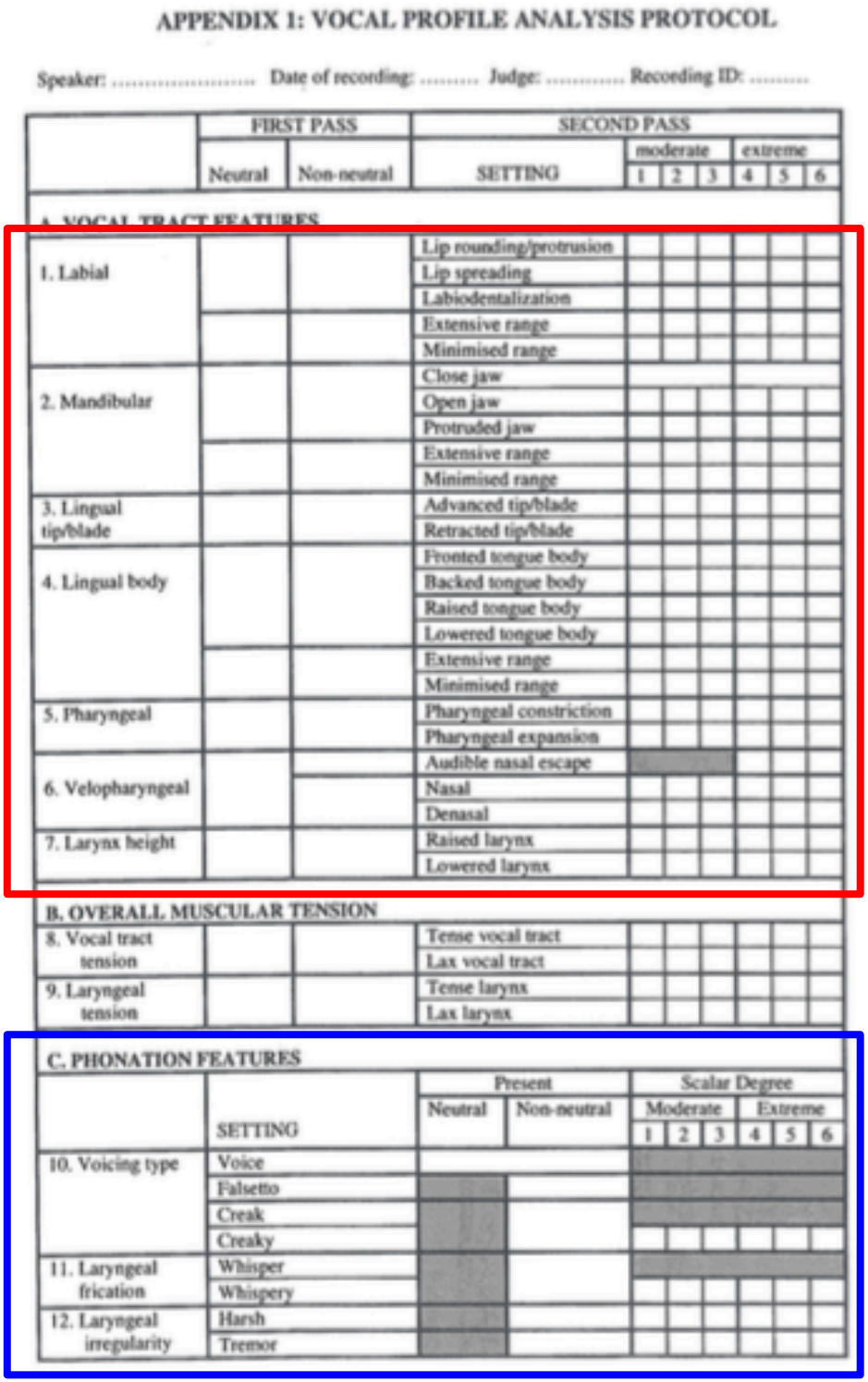
- **why:** aid the court in determining whether voices belong to the same or different speakers
- **how:** auditory phonetic *cum* acoustic phonetic analysis (Europe)

1. Introduction

- survey of practitioners (Gold & French 2011)
  - **voice quality (VQ):** one of most valuable features

    - 94% examine VQ
    - 68% of those do so ‘routinely’

    - 61% use recognised framework
    - 21% perform “auditory analysis and provide some form of a verbal description” (i.e. not following pre-established scheme)



APPENDIX 1: VOCAL PROFILE ANALYSIS PROTOCOL

Speaker: ..... Date of recording: ..... Judge: ..... Recording ID: .....

	FIRST PASS		SECOND PASS						
				moderate			extreme		
	Neutral	Non-neutral	SETTING	1	2	3	4	5	6
A. VOCAL TRACT FEATURES									
1. Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimised range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimised range						
3. Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
			Minimised range						
5. Pharyngeal			Pharyngeal constriction						
			Pharyngeal expansion						
6. Velopharyngeal			Audible nasal escape						
			Nasal						
			Denasal						
7. Larynx height			Raised larynx						
			Lowered larynx						
B. OVERALL MUSCULAR TENSION									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						

C. PHONATION FEATURES									
		Present		Scalar Degree					
				Moderate			Extreme		
	SETTING	Neutral	Non-neutral	1	2	3	4	5	6
10. Voicing type	Voice								
	Falsetto								
	Creak								
	Creaky								
11. Laryngeal frication	Whisper								
	Whispery								
12. Laryngeal irregularity	Harsh								
	Tremor								

Vocal Profile Analysis

- framework for systematic description of VQ
  - developed by Laver et al. (1981)
- modified by Beck (2007)
  - **25 supralaryngeal**
  - **7 laryngeal**
- comparison against ‘neutral setting’
  - clearly defined baseline with concrete acoustic and physiological correlates

1. Introduction

- Nolan (2005): first systematic discussion application of VQ analysis to FVC

issues with VPA for FVC

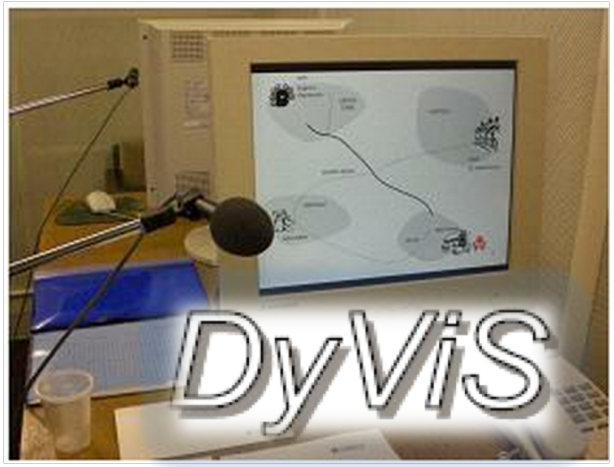
  - some dimensions never used in forensic analysis (redundancy)
  - forensic analysis needs to be based on independent features (avoid doubling evidence)
  - difficult to quantify
  - courts need to know reliability of the method (perception of subjectivity)

1. Introduction

research questions

1. what changes might we make to improve the useability of VPA for FVC?
2. how reliable are VPA scores across different analysts?
3. to what extent is a speaker's profile variable across recordings and how useful is VPA for speaker discrimination?

2. Methods



- **DyViS** (Nolan et al. 2009)
  - 100 male speakers
  - Standard Southern British English (SSBE)
  - 18-25 years old

Task 1

mock police interview

HQ, studio recording

(c. 20 mins)

Task 2

information exchange over telephone

HQ, near-end recording

(c. 10-15 mins)

3.1 VPA protocol

simplified version

- reduced scalar degrees
  - ‘present’ features (1-3)
- reduced N settings
  - mergers:
    - *fronted + raised*
    - *backed + lowered*
    - *creak + creaky*
    - *whisper + whispery*
  - deletion:
    - *audible nasal escape*
    - *protruded jaw*

	FIRST PASS		SECOND PASS				Notes
	Neutral	Non-Neutral	SETTING	Slight 1	Mark. 2	Extr. 3	
A. VOCAL TRACT FEATURES							
Labial			Lip rounding/protrusion				
			Lip spreading				
			Labiodentalisation				
			Extensive labial range				
			Minimised labial range				
Mandibular			Close jaw				
			Open jaw				
			Extensive mandibular range				
			Minimised mandibular range				
Lingual tip/blade			Advanced tongue tip/blade				
			Retracted tongue tip/blade				
Lingual body			Fronted tongue body				
			Backed tongue body				
			Extensive lingual range				
			Minimised lingual range				
Pharynx			Pharyngeal constriction				
			Pharyngeal expansion				
Velopharyngeal			Nasal				
			Denasal				
Larynx height			Raised larynx				
			Lowered larynx				
B. OVERALL MUSCULAR TENSION							
Vocal tract tension			Tense vocal tract				
			Lax vocal tract				
Laryngeal tension			Tense larynx				
			Lax larynx				

C. PHONATION FEATURES							
		Present		Scalar Degree			
	SETTING	Neutral	Non-neutral	Slight 1	Mark. 2	Extr. 3	
Voicing type	Voice						
	Falsetto						
	Creaky						
	Whispery						
	Breathy						
	Murmur						
	Harsh						
	Tremor						

3.1 VPA protocol

potential future simplifications

- further reduction of supralaryngeal settings:
  - extensive {mandibular | lingual | labial} range
    → **tense vocal tract**
- correlations between settings (e.g. lowered larynx ~ pharyngeal expansion)
  - sources of correlation (physiological, socioling…)
- dealing with polar opposites

lip rounding ⟷ ***lip spreading***

3.2. Interrater measures

pilot experiment:

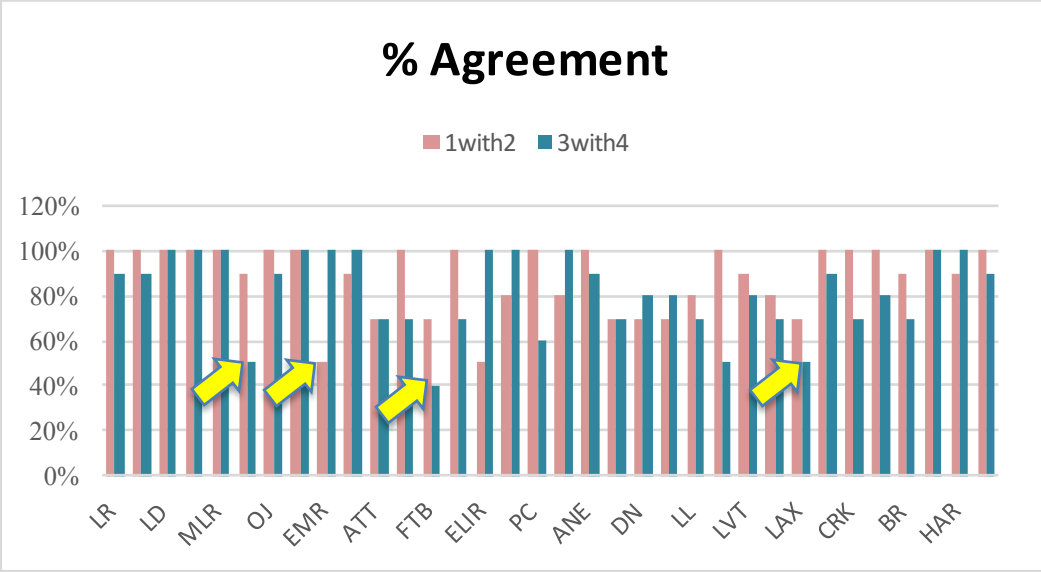
- 10 speakers randomly selected from Task 2
- 4 raters (blind perceptual analysis using SVPA)
- results based on absolute agreement

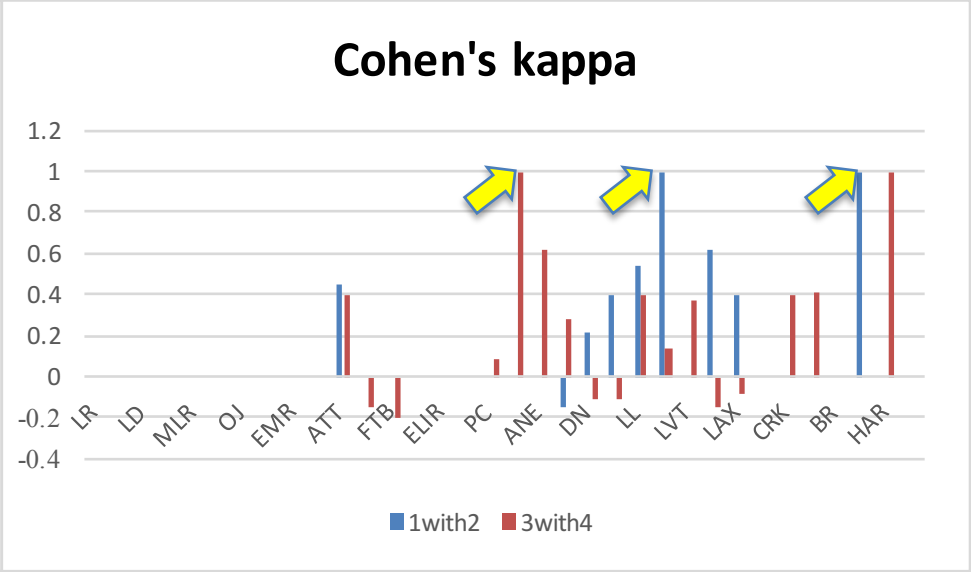
results

- high **percentage agreement** for most settings
  - only exceptionally < 70% (e.g. *fronted tongue body*)
- fair-moderate agreement with **Cohen's kappa**
  - $\kappa > 0.80$ (*pharyngeal expansion*, *harshness*)

3.2. Interrater measures

- Precalibration





- Good agreement for most settings
- Some exceptions (% < 0.70)
  - *close jaw*
  - *extensive mand./lingual range*
  - *fronted tongue body*
  - *lax larynx*
- But results not chance corrected!

- Very good agreement (κ > 0.80):
  - ✓ *pharyngeal expansion*
  - ✓ *harshness*
  - ✓ *tense vocal tract*
- Need to work on the rest of settings
- Problem with 'invariant values':
  - All coders attain 100% agreement
  - Only use one variable value

3.2. Interrater measures

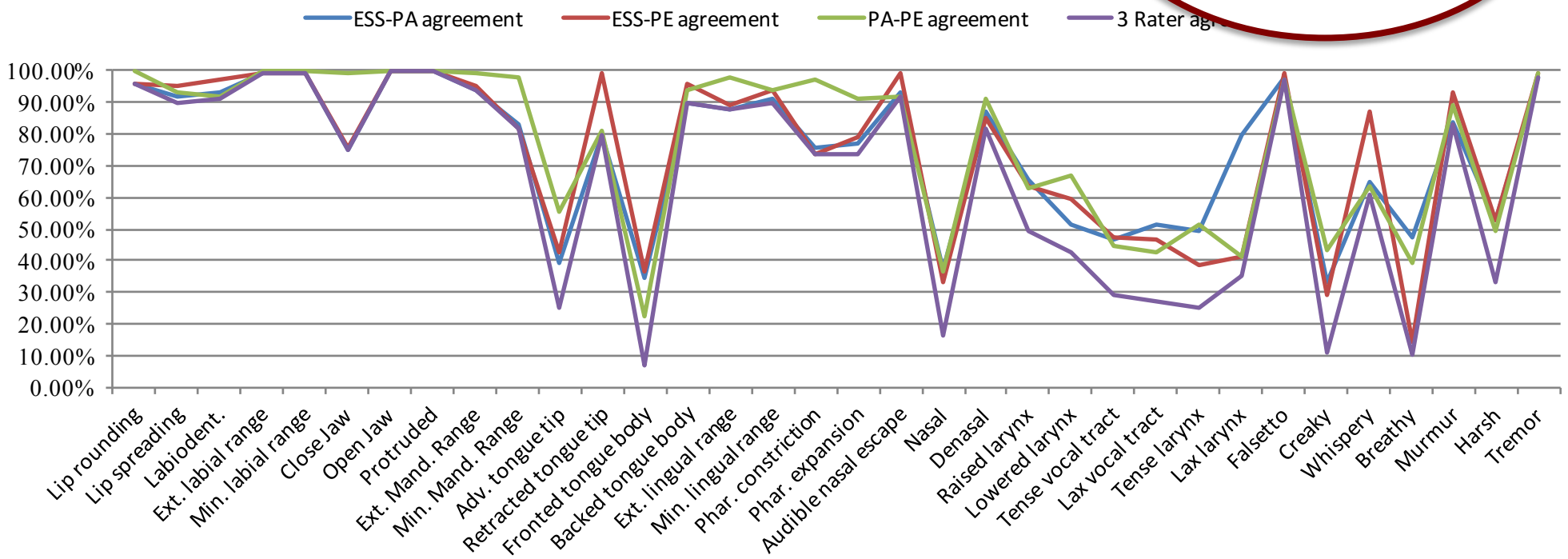
calibration procedure:

- joint listening
- redefining certain labels
- discussing the idea of ‘neutral’ voice as baseline for this population
- adjustment of the individual use of scalar degrees

3.2. Interrater measures

post-calibration interrater results

***REMEMBER**
Based on absolute agreement (same scalar degrees)!



- ✓ overall good agreement (80 – 100%) for most supralaryngal settings
- ✓ moderate agreement (< 60%) in phonatory settings (esp. *breathy*)
- ✓ Cohen's kappa confirms general patterns
  - but issues with use for invariant values
- ✓ exceptions to good agreement: e.g. *fronted tongue body* & *nasal*

3.2. Interrater measures

- more realistic definition of agreement/ disagreement:
  - disagreement about presence/ absence (0-1)
  - disagreement beyond 1 scalar degree

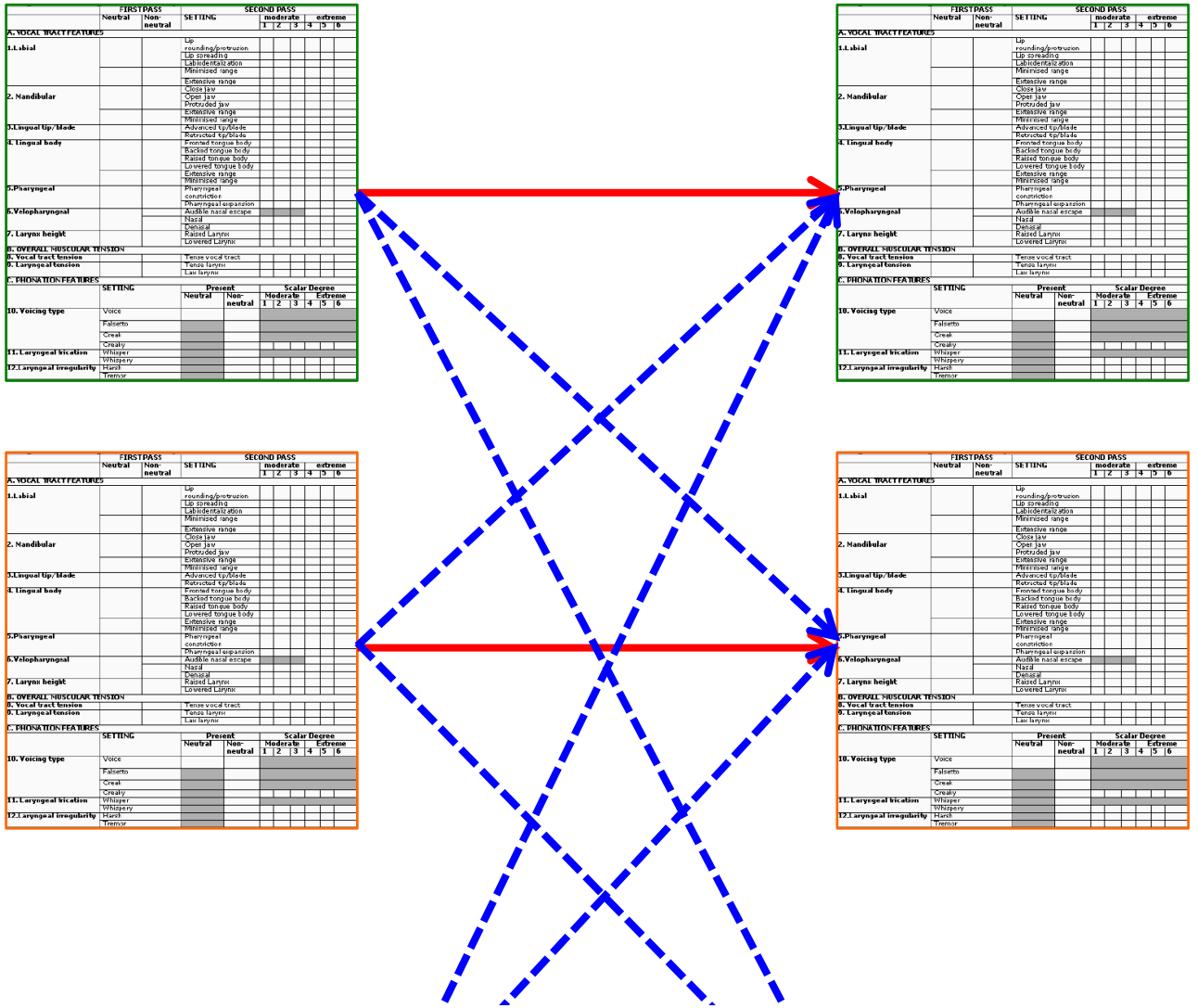
	% agreement
ESS ~ PE	88.1%
ESS ~ PA	87.3%
PA ~ PE	88.4%
mean	**87.9%**

3.3 Speaker discrimination

TASK 1 **TASK 2**

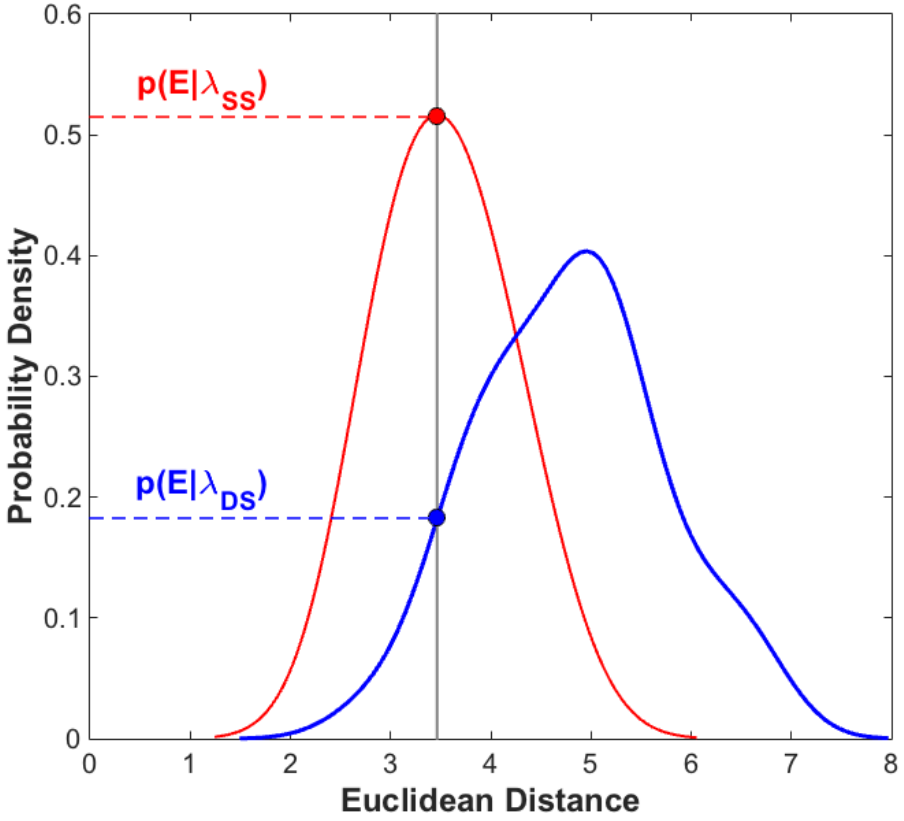
Speaker 1

Speaker 2



etc.

3.3 Speaker discrimination



$$\frac{p(\mathrm{ED}\,|\,\lambda_{SS})}{p(\mathrm{ED}\,|\,\lambda_{DS})} = \frac{0.515}{0.183}$$

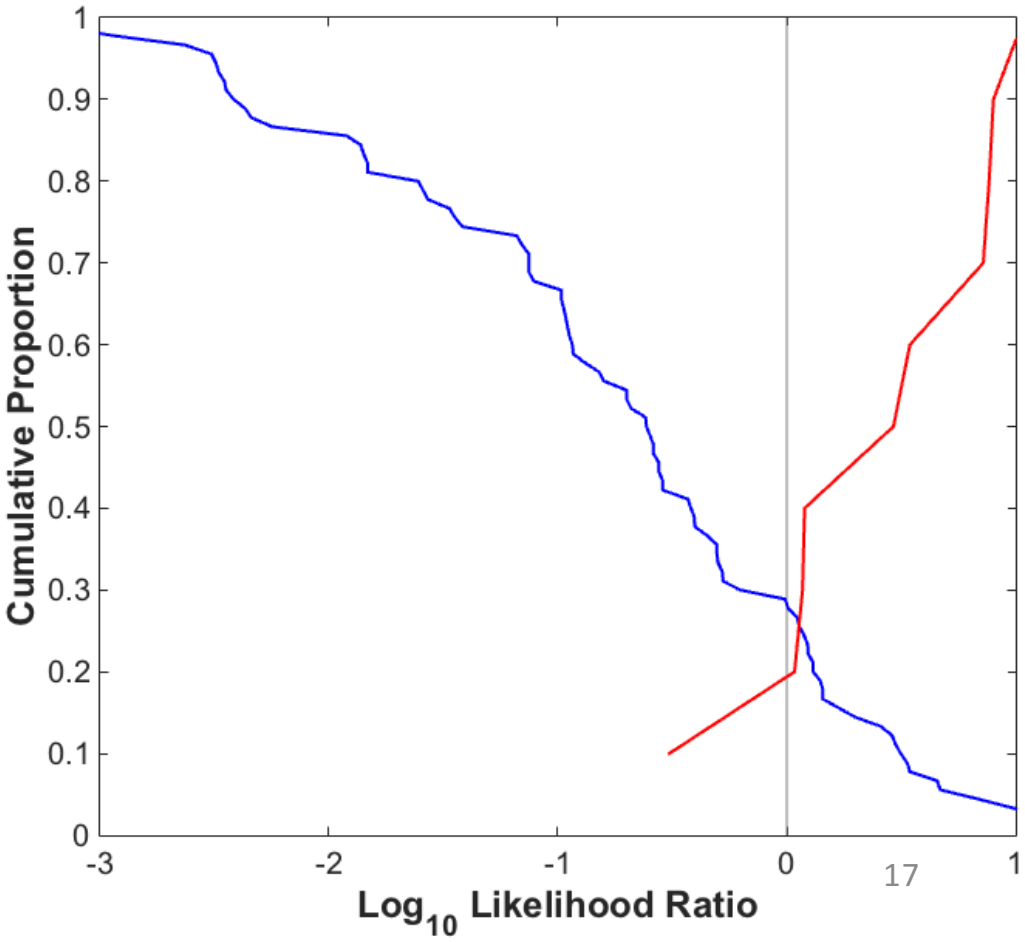
$$= 2.814$$

- performance measured using C_{llr}
  - magnitude of errors
  - < 1 is good

3.3 Speaker discrimination

best results: binary data

settings	C_{llr}
All	**0.6509**
Supralaryngeal	0.7137
Laryngeal	0.9984



4. Discussion: modifications

- first attempt at simplifying scheme
- issues with perceptual assessment of VQ
  - voice = highly multidimensional
  - difficulty isolating individual settings
  - reducing dimensionality/ maintaining value
- scope for further dimension reduction:
  - may be able to use interrater results

4. Discussion: interrater

- overall % agreement = good
  - some settings easier to agree upon? more salient?
  - ***labiodentalisation*** or ***harshness*** also high % agreement in previous studies (Beck 2005: **100%** and **84%** respectively)
- lower % agreement results may have simple explanations and solutions:
  - ***nasal-denasal*** co-occur → raters tick one label?
  - ***breathy-whisper*** → no clear perceptual boundary

4. Discussion: sp discrimination

- weak strength of evidence
  - small sample size
  - cross-validation
  - no calibration
  - issue with representation (quantification)
    - distances not weighted by auditorily marked features
    - averaging over settings
    - massive redundancy

5. Conclusion

- simplified VPA for forensic purposes
  - further modifications based on:
    - correlations/interrater results/speaker discrimination…
- overall good interrater agreement
  - systematic patterns (individuals/listening strategies)
  - other statistical measures
- promising speaker discriminatory value
  - more appropriate ways of quantifying VPA
  - optimising discriminatory value

Thanks! Questions?

THE UNIVERSITY of York

J P French λssociates

Forensic speech and acoustics laboratory