UNIVERSITY of York

Methodological issues in inter-rater agreement in voice quality analysis

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

1. Background of our research

- sociolinguistics, dialectology, general phonetics

- forensic speech analysis
  - comparison of general phonetic methods, acoustic measures & ASR approaches (AHRC grant, *Voice and Identity* 2015-19).
  - critical in forensic work for independent agreement on observations

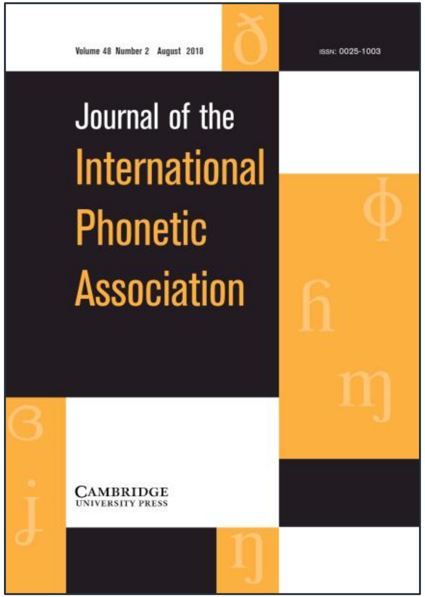
  → establishing inter-rater agreement in VQ analysis

- using modified Laver/Edinburgh VPA protocol within casework

Arts & Humanities Research Council

2. Outline

- establishing inter-rater agreement in VQ analysis (San Segundo et al, *JIPA* 2018)
- methods
- findings
  - issues with Edinburgh VPA
  - outcomes of inter-rater analysis
- outlook


Volume 48 Number 2 August 2018

ISSN: 0025-1003

Journal of the International Phonetic Association

CAMBRIDGE UNIVERSITY PRESS

3. Methods

- recordings: **DyViS** corpus (Nolan et al 2009)
  - forensic research
  - simulated police interview ca. 10 minutes

- 100 young men, Standard Southern British English (RP)
  - rather homogeneous, not typical of whole population

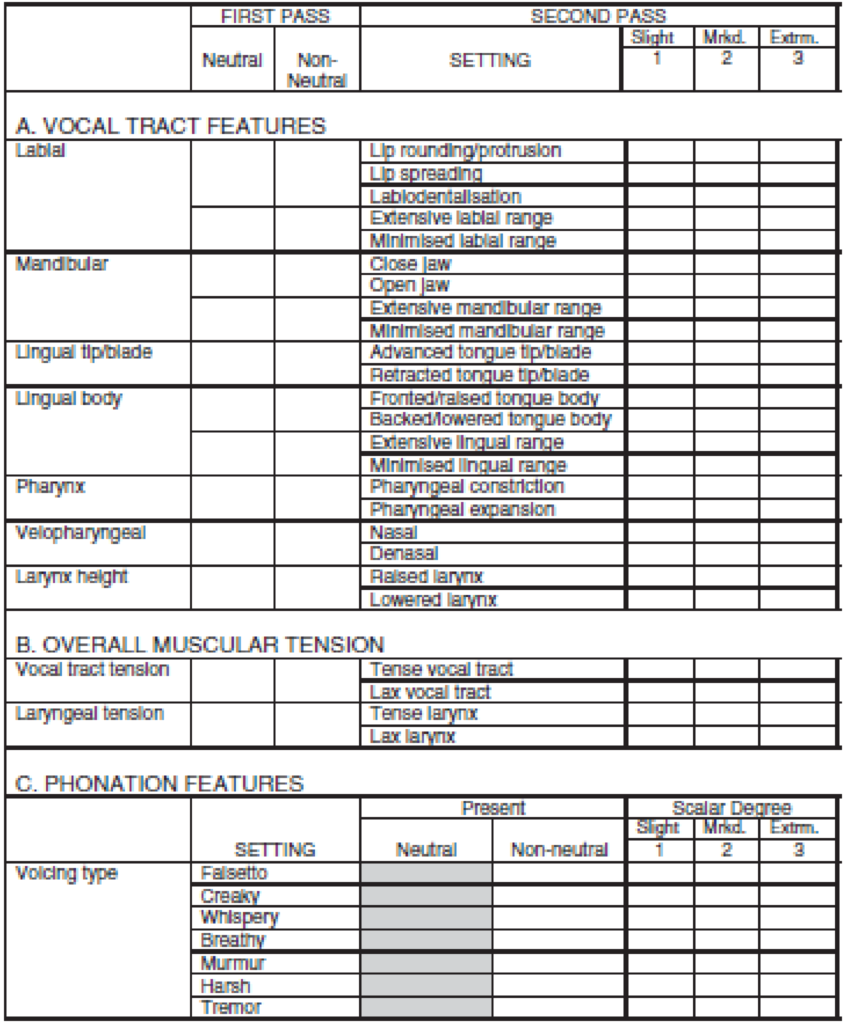
#067 🔊 🔊 #072

3. Methods

- 3 analysts – ESS, PF, JPF

- modified VPA used at J P French

- no pathological labels (4-6)
- grade 1 = slight (noticeable)
- grade 2 = marked
- grade 3 = extreme (not pathology)



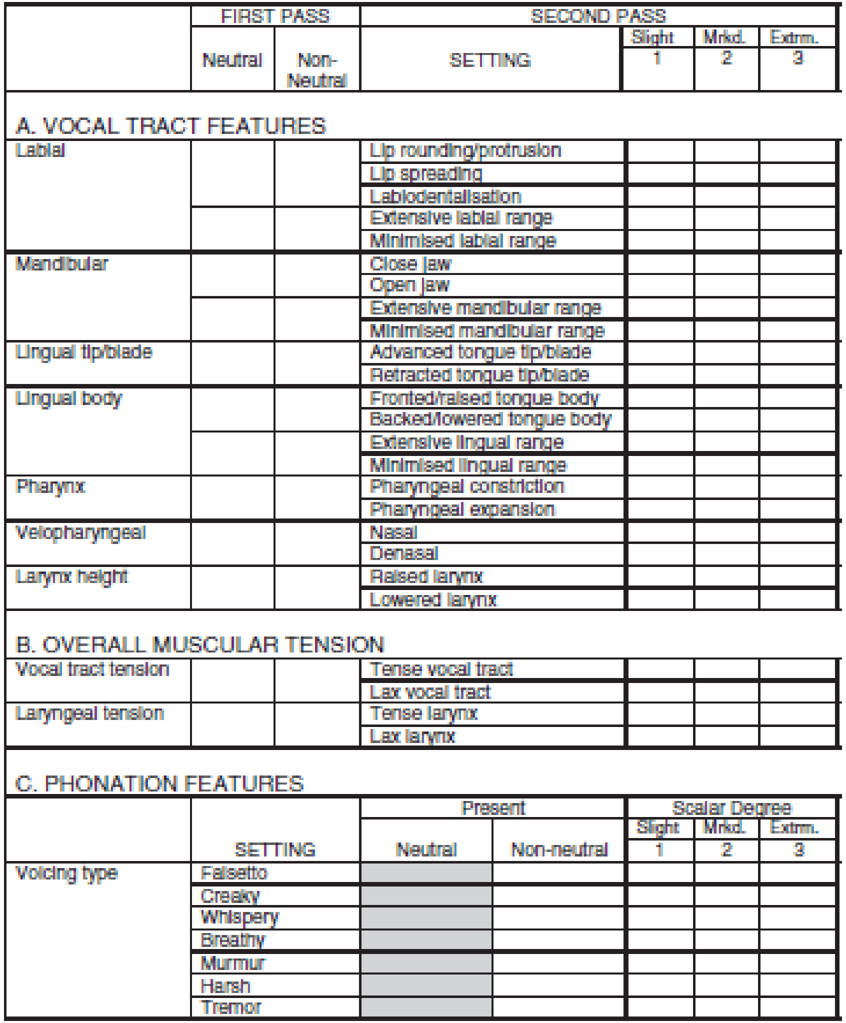
	FIRST PASS		SECOND PASS			
	Neutral	Non-Neutral	SETTING	Slight 1	Mrkd. 2	Extrm. 3
A. VOCAL TRACT FEATURES						
Labial			Lip rounding/protrusion			
			Lip spreading			
			Labiodentalisation			
			Extensive labial range			
			Minimised labial range			
Mandibular			Close jaw			
			Open jaw			
			Extensive mandibular range			
			Minimised mandibular range			
Lingual tip/blade			Advanced tongue tip/blade			
			Retracted tongue tip/blade			
Lingual body			Fronted/raised tongue body			
			Backed/lowered 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	Mrkd. 2	Extrm. 3
Voicing type	Falsetto					
	Creaky					
	Whispery					
	Breathy					
	Murmur					
	Harsh					
	Tremor					

3. Methods

- **stage 1**: 10 speakers
  - practice
- **stage 2**: calibration meeting
- **stage 3**: 99 speakers
  - first 10 redone blind
  - (1 technical problem)



	FIRST PASS		SECOND PASS			
	Neutral	Non-Neutral	SETTING	Slight 1	Mrkd. 2	Extrm. 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/raised tongue body			
			Backed/lowered 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		
				Slight	Mrkd.	Extrm.
	SETTING	Neutral	Non-neutral	1	2	3
Voicing type	Falsetto					
	Creaky					
	Whispery					
	Breathy					
	Murmur					
	Harsh					
	Tremor					

4. Issues with VPA

- our work raised various general issues with VPA conception & protocol (discussed also by others; summary in San Segundo et al 2018)

- **articulatory labels** but **perceptual** judgments
  - VQ as ‘an interaction between a listener and a signal’ (Kreiman & Sidtis 2011: 9)

- **neutral setting** as baseline
  - hypothetical, thus imaginary
  - difficult to avoid bias to dialect norms

    e.g. slight nasality, creak & tongue fronting for SSBE

4. Issues with VPA

- independence of 30-40 individual settings
  - how well can analysts focus on them separately?
  - physical linkages and perceptual correlations

  e.g. lowered larynx & expanded pharynx

4. Issues with VPA

- thresholds of permanence
  - how frequent/widespread must a setting be to count?

- VQ = long-term quasi-permanent setting/timbre
  - but any setting is also tied to key segments
  - thus by definition any setting is intermittent

- we attributed effects as segmental where possible
  - if limited to 1-2 segments e.g. labiodentalisation of /r/

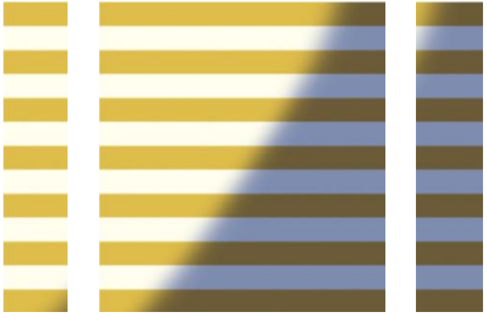
5.1 Outcomes: calibration

- calibration meeting: identified disagreement types & problems

- true error
  - analyst missed or mislabelled clear setting

- difficulty with distinctions
  - e.g. breathy~whisper

- systematic use of different labels for same percept
  - harsh phonation – tense larynx
  - retracted tongue body – constricted pharynx



5.1 Outcomes: calibration

- calibration meeting

- corrected the true errors

- established heuristics to
  - address systematic differences in scoring
  - combine perceptually equivalent labels
    - e.g. constricted pharynx & retracted tongue body
  - establish perceptual distinctions
    - e.g. whispery = higher friction, tension, poss. voicelessness
    - cf. breathy = lower friction, laxness

5.2 Outcomes: full analysis

- stage 3: full analysis of 99 speakers
- 3 analysts worked independently
- met to consider 3 versions
- agreed on mode rating if all within 1 scalar degree (1-2-2, 2-2-3…)
- re-listened collaboratively if:
  - difference in presence/absence (0-0-1, 0-1-1…)
  - wider disagreement (1-1-3, 1-3-3…)
  - apparent error

5.3 Outcomes: agreement

- inter-rater agreement

- no expectation of 100% agreement!
  - our VPA has 32 settings * 4 grades
  - logically $4^{32} = 1.84e^{19}$ combinations (> humans, < stars!)

- two classifications of results
  - absolute agreement
  - within 1 grade
  - Fleiss kappa statistic – quantifies agreement versus chance level

	absolute (%)	**± 1 grade (%)**		
Setting	**mean**	**mean**	**N**	**Fleiss kappa**
Overall agreement	**76**	**82**	99	

	absolute (%)	± 1 grade (%)			
Setting	**mean**	**mean**	**N**	**Fleiss kappa**	
Overall agreement	**76**	**82**	99	(> 0 is good)	
fronted tongue body	36	60	98	.01	slight
tense vocal tract	55	68	51	.22	fair
lax vocal tract	59	70	43	.29	fair
lax larynx	62	71	37	.31	fair
nasal	43	72	92	.13	slight
advanced tongue tip	59	73	56	.35	fair
lowered larynx	67	76	43	.41	moderate
tense larynx	67	76	47	.34	fair
breathy	52	78	73	.31	fair
creaky	46	81	83	.31	fair
raised larynx	74	82	34	.46	moderate
harsh	75	82	31	.43	moderate
whispery	91	96	10	.53	moderate

5.3 Outcomes: agreement

- all other settings 91-100% agreement
  - but N < 10 speakers
  - thus largely 0 scores

- NB: more frequent settings → lower agreement scores
  - easier to agree on absence than presence

5.3 Outcomes: agreement

- analyst pairwise ratings

- no striking differences between any pair of analysts

- we each acknowledged strengths, weaknesses, biases
  - e.g. PF: lax larynx, tense larynx, murmur

- team approach has clear benefit in addressing such issues

5.4 Outcomes: correlations

positively correlated VPA settings		Spearman's *r*	*C*
*raised larynx	tense larynx	.62	.58
*harsh	tense larynx	.36	.57
*lax larynx	lowered larynx	.57	.52
creaky	lax larynx	.46	.45
advanced tongue tip	fronted tongue body	.38	.41
creaky	lowered larynx	.35	.35

C = contingency coefficient, range 0-1

*noted by e.g. Beck (2007), but also predicted: lax lx ⇔ lowered lx ⇔ breathy/whispery

5.4 Outcomes: correlations

negatively correlated VPA settings		**Spearman's *r***	***C***
creaky	whispery	*-.36*	.37
lowered larynx	tense larynx	*-.47*	.46
creaky	raised larynx	*-.43*	.44
lax larynx	raised larynx	*-.51*	.47
lowered larynx	raised larynx	*-.55*	.51
lax larynx	tense larynx	*-.66*	.57
lax vocal tract	tense vocal tract	*-.73*	.61

C = contingency coefficient, range 0-1

NB opposites, but they do occur… forensically very valuable

6. Summary & outlook

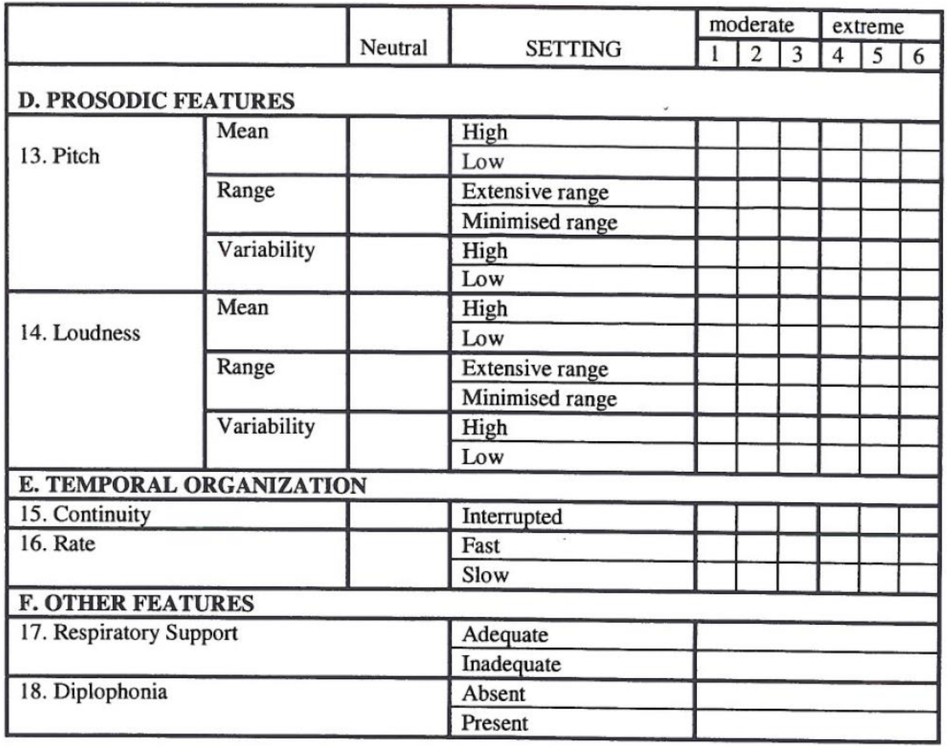
- team approach is not only possible but valuable
- agreement level overall is good, between each pair & all 3
- counters idiosyncrasies and biases
- calibration really helps
- focus on clearly notable features rather than exhaustive 32*4 grading

6. Summary & outlook

- supplementary settings in Beck (2007) potentially very helpful
  - not used here as ~acoustic or quantifiable

- holistic patterns
  - liveliness (wide f0 range + fast)
  - brightness, monotony, resonance
  - inconsistency in phonation

#009



		Neutral	SETTING	moderate			extreme		
				1	2	3	4	5	6
D. PROSODIC FEATURES									
13. Pitch	Mean		High						
			Low						
	Range		Extensive range						
			Minimised range						
	Variability		High						
			Low						
14. Loudness	Mean		High						
			Low						
	Range		Extensive range						
			Minimised range						
	Variability		High						
			Low						
E. TEMPORAL ORGANIZATION									
15. Continuity			Interrupted						
16. Rate			Fast						
			Slow						
F. OTHER FEATURES									
17. Respiratory Support			Adequate						
			Inadequate						
18. Diplophonia			Absent						
			Present						

UNIVERSITY of York

thank you, tack så mycket

questions?

voice and identity
0101101ID1

Arts & Humanities Research Council

Setting	**absolute agreement (%)**				**agreement within 1 scalar degree (%)**			
	ES-PF	**ES-JPF**	**JPF-PF**	**mean**	**ES-PF**	**ES-JPF**	**JPF-PF**	**mean**
Overall rate				**76**				**82**
nasal	43	36	49	**43**	66	75	75	**72**
denasal	90	87	92	90	91	88	93	91
raised larynx	78	73	71	74	85	84	79	82
lowered larynx	62	70	71	67	72	79	79	76
tense vocal tract	53	55	59	**55**	75	65	66	**68**
lax vocal tract	66	55	58	**59**	76	65	71	**70**
tense larynx	69	66	68	67	74	80	74	76
lax larynx	66	69	51	62	71	85	58	71
falsetto	100	100	100	100	100	100	100	100
creaky	42	37	59	**46**	80	79	85	**81**
whispery	90	94	88	91	95	98	95	96
breathy	49	42	64	**52**	72	77	85	**78**
murmur	99	100	99	99	100	100	100	100
harsh	75	74	76	75	84	80	84	82
tremor	100	100	100	100	100	100	100	100

Setting	absolute agreement (%)				agreement within 1 scalar degree (%)			
	ES-PF	ES-JPF	JPF-PF	mean	ES-PF	ES-JPF	JPF-PF	mean
lip rounding	96	96	100	97	96	96	100	97
lip spreading	94	95	95	95	94	95	95	95
labio-dentalisation	98	100	98	99	98	100	98	99
extensive labial range	100	100	100	100	100	100	100	100
minimised labial range	100	100	100	100	100	100	100	100
close jaw	96	96	100	97	96	96	100	97
open jaw	100	100	100	100	100	100	100	100
ext. mandibular range	99	99	100	99	99	99	100	99
min. mandibular range	96	96	98	97	98	98	98	98
advanced tongue tip	55	56	66	**59**	69	73	78	**73**
retracted tongue tip	92	99	92	94	93	99	92	95
fronted tongue body	33	43	31	**36**	51	69	62	**60**
backed tongue body	97	97	100	98	97	97	100	98
ext. lingual range	98	99	99	99	100	100	100	100
min. lingual range	98	98	100	99	99	99	100	99
pharyngeal constriction	97	95	98	97	98	97	99	98
pharyngeal expansion	97	98	97	97	99	100	99	99