

Speaker identification in whisper

India Evans* & **Paul Foulkes****

*University of York

**University of York & JP French Associates

pf11@york.ac.uk



Research questions

1. How well can listeners **identify familiar voices** disguised by **whisper**?
2. Does **length of sample** affect identification rates?
3. How well can they **recognise unknown** voices?

0. Outline

1. Introduction

- rationale for experiment

2. Experimental design

3. Results

4. Discussion & Conclusion

- comments on whisper production



1. Introduction

- many studies have assessed ability of lay listeners to identify familiar voices
 - to understand reliability of performance in forensic identification settings
- general conclusions:
 - identification is **not automatic**, and **not perfect**
 - wide **variability** in performance by **listeners**
 - wide **variability** in performance according to **voice**
 - numerous **potential effects** of performance...



1. Introduction

- ... identification rates tend to be lower where:
 - exposure to voice is **passive** (Hammersley & Read 1986)
 - sample is **short** (Ladefoged & Ladefoged 1980)
 - long **delay** between exposure and test (McGehee 1937)
 - listener has **poor hearing** ability (Bull & Clifford 1984)
 - speaker is **less familiar** (Hollien et al 1982)
 - sample is **degraded**, e.g.:
 - heard over telephone (Foulkes & Barron 2000)
 - shouted (Blatchford & Foulkes 2006)
 - speech is **disguised** (Hollien et al 1982)



1.1 disguise

- predictably, disguised voices generally prove harder to identify, e.g.:
 - Wagner and Köster (1999) identification rates:
 - undisguised: 97%
 - falsetto: 4%
- criticism of previous studies (Rodman 1998)
 - paucity of experimental work
 - lack of systematicity: unconstrained disguise types

1.2 whisper

- whisper is reportedly common in forensic cases
 - Clifford (1983), Masthoff (1996), Künzel (2000)
- but few previous speaker ID studies with whisper
 - Orchard & Yarmey (1995) & Yarmey et al (2001)
 - lay speaker ID
 - better with longer samples , normal speech & familiar voices
 - Zhang (2005): catastrophic for ASR (0% identification)

1.2 whisper

- predict high impact on identification rates
- cross-speaker differences severely reduced
 - changes to patterns of modal speech
 - removal of f_0
 - slower articulation rate, longer vowels etc (Tartter 1989)
- but:
 - effects of vocal tract transfer function remain (Nolan 1997)
 - 98% speaker sex identification (Schwartz & Rine 1968)
 - accent features should be robust



2. Experimental design

Network

- pre-existing work group: **'the group'**
 - 11 women (including IE)
 - ages 20 - 30
 - regular work-mates at cosmetics store
 - known each other between 8 months & 4 years
 - variety of accents from England and Scotland



2. Experimental design

Speakers

- 6 women from the group
- 3 foils
 - 2 female
 - 1 male
 - hypothesis: easy to reject male foil, even in whisper





Speech materials

- 3 x 100 word passages + extracts as word-list
 - read normally and whispered
 - whispered texts: c. 45 seconds to read
 - pilot study: 250 words
 - difficulty in maintaining whisper > 45 secs.
- recorded in quiet room at workplace or home
 - Zoom Handy Recorder H4
 - 16 bit, 44.1 kHz sampling rate
 - stereo .wav file

Listening tests

- extracts compiled into new .wav files
 - normalised for amplitude

	type	stimuli length	N stimuli
test 1	short whisper	4 sylls 	12 x 9 spkrs = 108
test 2	long whisper	16 sylls 	3 x 9 spkrs = 27
test 3	normal (control)	4 sylls	12 x 9 spkrs = 108

Listeners

- 10 women from the group (excluding IE)



Procedure

- tests presented through PowerPoint
 - automatic playing of labelled stimuli
 - 8 second delay between stimuli
 - conducted on laptop, Sennheiser headphones
 - short whisper > long whisper > normal speech
 - total test duration c. 45 minutes

- questionnaire: **closed set** identification
 - 6 familiar names + 'stranger'

Statistical analysis

- Anova via SPSS

3. Results

Overall correct identification: familiar voices

	type	length	familiar: % correct
test 1	short whisper	4 sylls	71
test 2	long whisper	16 sylls	87
test 3	normal (control)	4 sylls	93

Statistical significance markers:

- A bracket with an asterisk (*) spans the 71 and 87 values.
- A bracket with an asterisk (*) spans the 87 and 93 values.
- A bracket with a hash (#) spans the 93 and 71 values.

NB chance level = 14%

* p < .0001
p = .068



3. Results

Overall correct identification: foils

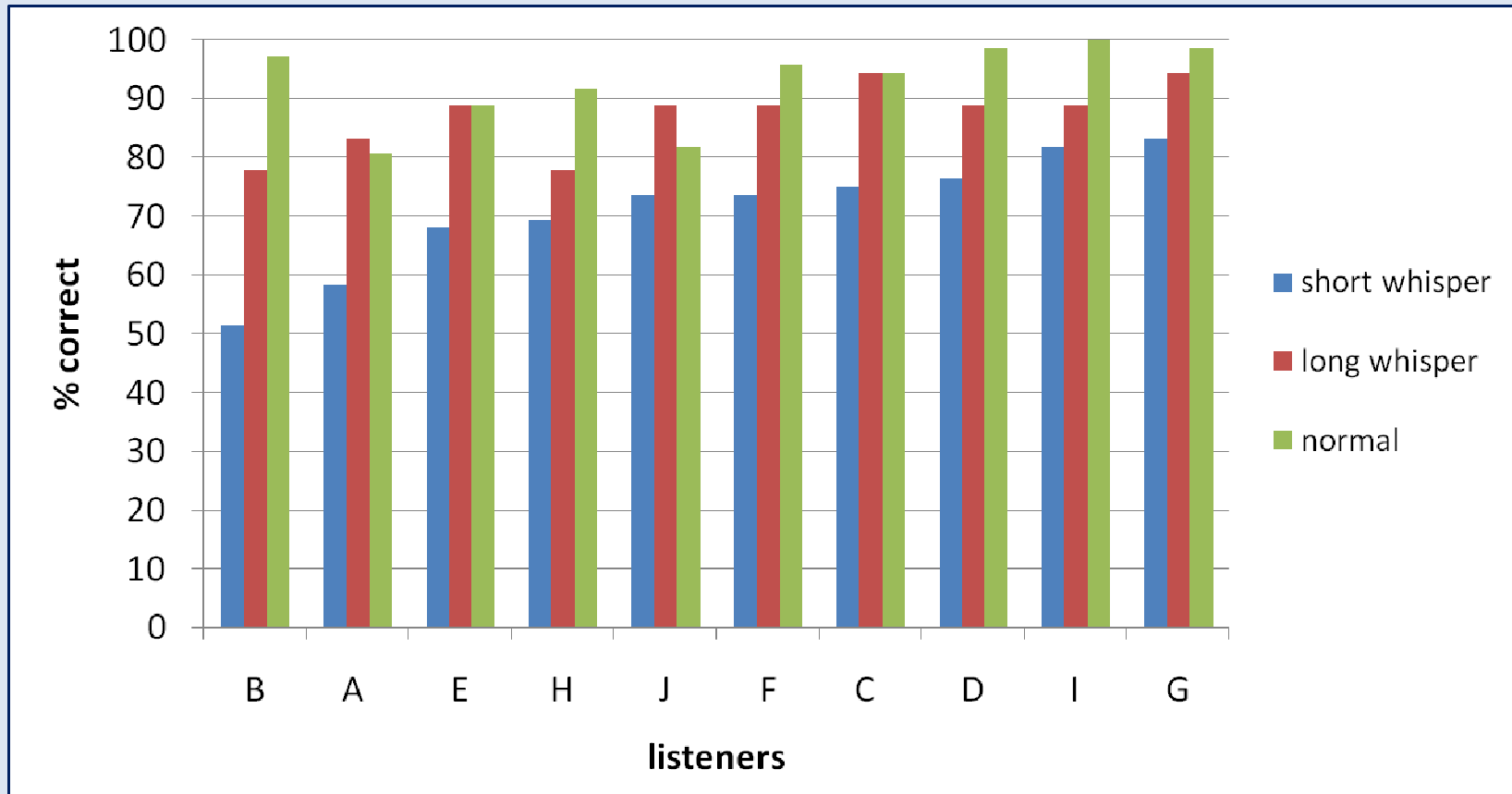
	type	females	male
test 1	short whisper	26%	98%
test 2	long whisper	33%	100%
test 3	normal (control)	36%	100%

NB chance level = 14%



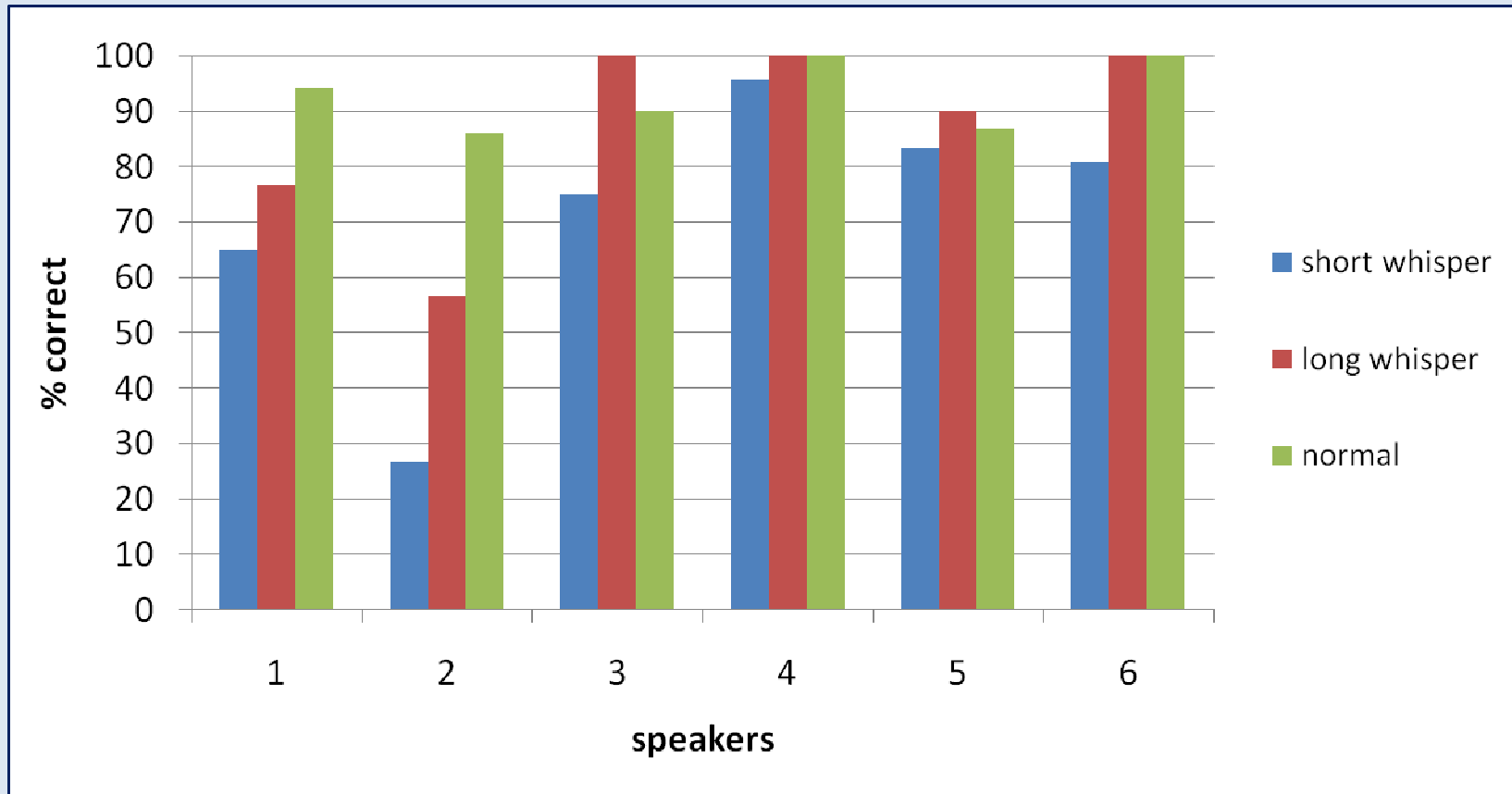
3.1 results by listener

Overall correct identification: familiar voices only



3.2 results by speaker

Overall correct identification: familiar voices only



4. Discussion

Research questions

1. How well can listeners **identify familiar voices** disguised by **whisper**?
2. Does **length of sample** affect identification rates?
3. How well can they **recognise unknown** voices?



4. Discussion

Research questions

1. How well can listeners **identify familiar voices** disguised by **whisper**?
 - surprisingly well: 71% with just 4 syllables
 - much higher than e.g. Yarmey et al (2001)
 - but significantly worse than with normal speech



4. Discussion

Research questions

1. How well can listeners **identify familiar voices** disguised by **whisper**?
 - NB significant variation by listener and by voice
 - thus statistics shouldn't be interpreted simplistically
 - in a forensic case, listener and speaker may affect ID
 - effects of situation also unknown



4. Discussion

Research questions

2. Does **length of sample** affect identification rates?

- yes: 71% → 87%
 - long samples less well identified than normal voices
 - but some listeners equally good with long whisper & normal samples



4. Discussion

Research questions

3. How well can listeners **recognise unknown** voices?

- male foil easily
 - no f_0 , but different vocal tract transfer function is clear
- female foils much less well
 - better than chance
 - better with longer sample

Final comments: production of whisper

- assumed whisper → reduction of cross-speaker differences
 - removal of f0
 - overall slower articulation rate & different CSPs
- supported to a large extent
- but...
 - some key distinguishing features **robust** to whisper
 - some **new sources** of cross-speaker variation

4. Discussion

Final comments: production of whisper

- accent features robust, as predicted
 - e.g. rhoticity in Scottish
 - northern STR[ʊ]T and B[a]TH
- help in constraining choices in closed set tests

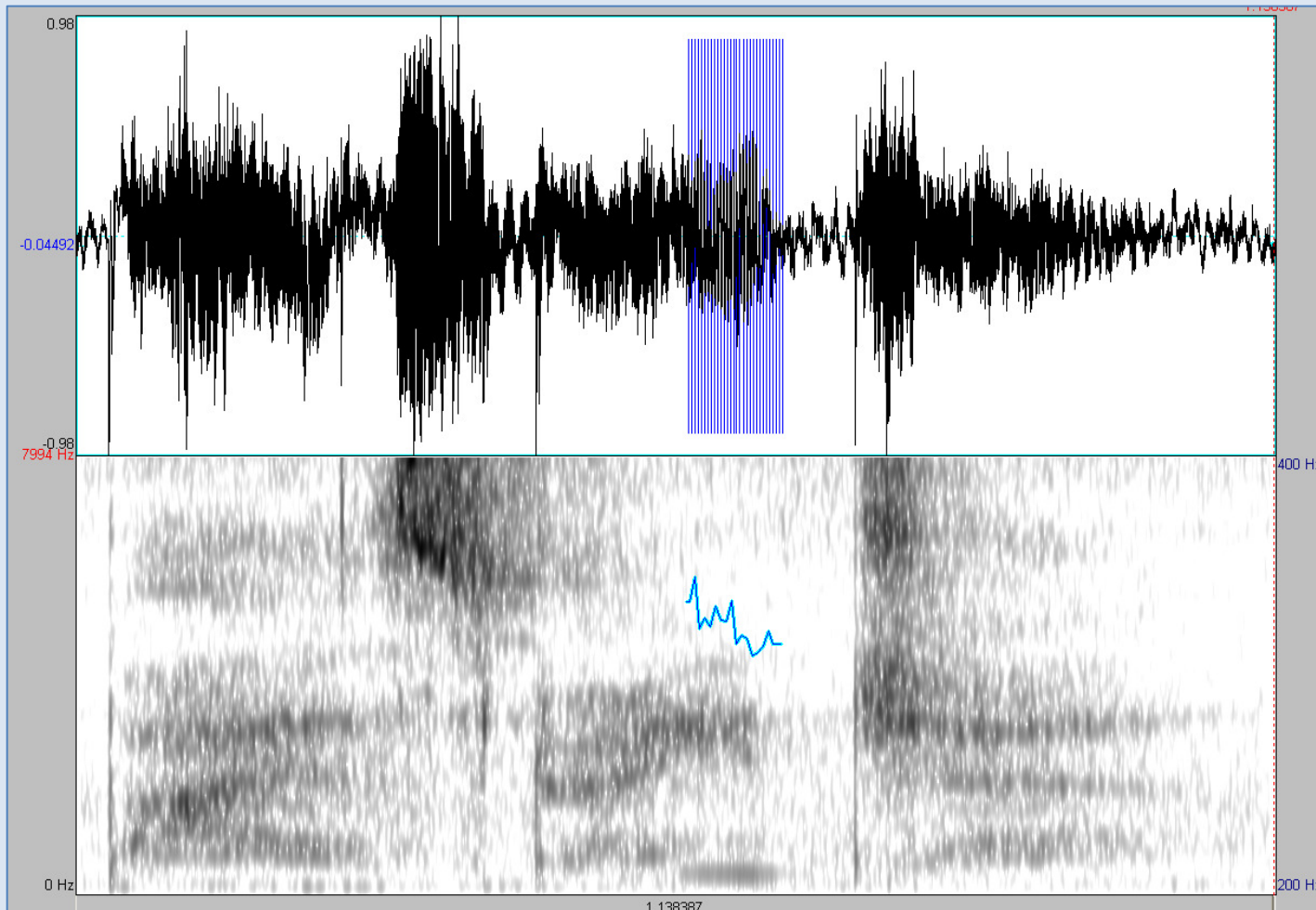
4. Discussion

Final comments: production of whisper

- ‘new’ sources of cross-speaker variation
 - **slower articulation**
 - but 1 speaker very rapid
 - **variation in amplitude**
 - loud ‘stage whisper’ or weak
 - differences in articulatory settings/airflow control (Laver 1980)
 - **audible breaths**, problems in controlling airflow

4. Discussion

- leakage of voicing
 - 8/9 speakers in long samples, 2 in short



4. Discussion

Final comments: production of whisper

- further work required to:
 - understand causes and distributions of variation
 - potential value for lay ears or acoustic analysis
- but all speakers struggled > 45 seconds
- in live forensic situations: *keep them talking*



thanks

questions?

