Long term measures of the resonating vocal tract: establishing correlation and complementarity

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

University of York & J P French Associates

voice and identity

0101101ID1

IAFPA Annual Conference 2015
Universiteit Leiden
10th – 13th July

Arts & Humanities Research Council

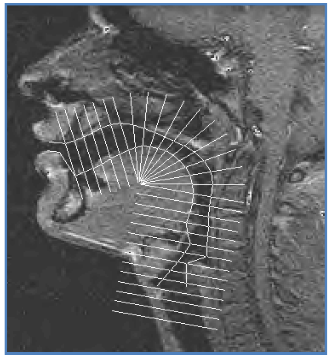
Voice and identity: source, filter, biometric

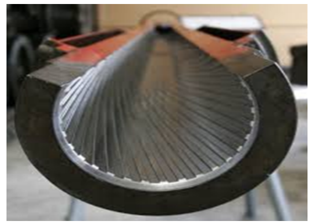
- **aims**
  - individual speaker characterisation: properties of the voice that are specific to the individual
    - focus on (1) **filter (vocal tract)** and (2) **source (larynx)**
  - combination of linguistic/phonetic and ASR methods (cf. Gonzalez-Rodriguez et al. 2014)
  - improve the performance of forensic voice comparison systems

AH/M003396/1
Arts & Humanities Research Council

1. Introduction

- **stage 1:** focus on the vocal tract (filter)
- underlying assumption: physiology of vocal tract = unique to individuals
  - differences between individuals should be manifested in vocal tract output
- **but…**
  - no direct access to physiological measures in FVC
  - limited to indirect output measures



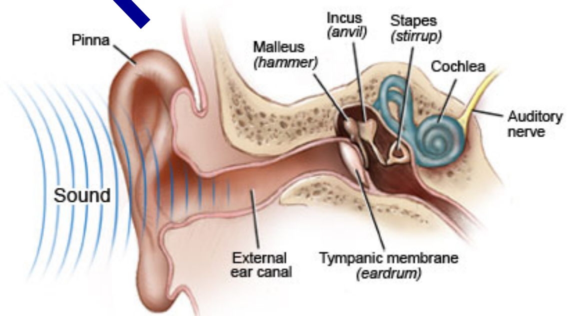


1. Introduction

- to better understand the speaker-specifics of the vocal tract…

biology/ physiology

acoustic output



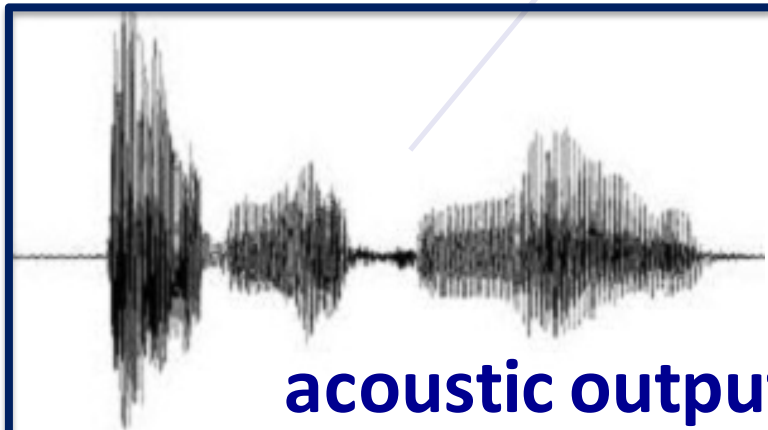
auditory percept

1. Introduction

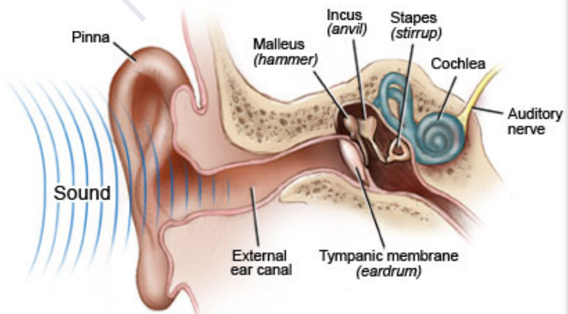
- to better understand the speaker-specifics of the vocal tract...

biology/
physiology

first step: important to understand the relationships between vocal tract output measures



acoustic output



auditory percept

1. Introduction

- auditory features
  - **Vocal Profile Analysis** (VPA; Laver et al. 1981)
  - 27 supralaryngeal features (linguistic-phonetic)
    - labial, mandibular, lingual, pharyngeal, vocal tract tension features
- acoustic features
  - semi-automatic: **long-term formant distributions** (LTFDs) (Jessen, Heeren et al., Krebs & Braun, Meuwly et al. @ IAFPA 2015) (linguistic-phonetic)
  - automatic: **MFCCs/ LPCCs** (ASR)

1. Introduction

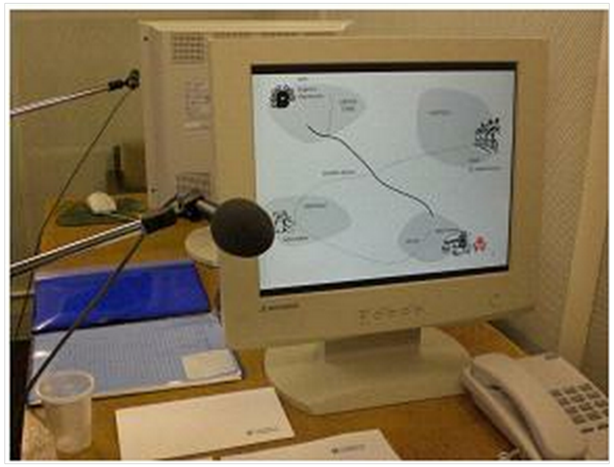
- why these features?
  - long-term features = more likely to capture broad individual differences in vocal tract physiology
    - cf. segmental variables: more susceptible to systematic within-sp variability (empirical question?)
    - easier to extract data automatically
  - combination of features from linguistics/ phonetics and ASR
    - general move towards the integration of analytic approach from different sub-fields

1. Introduction

research questions

1. to what extent are long-term vocal tract output measures related?
2. to what extent do these long-term vocal tract measures provide complementary information?

2. Methods



DyViS

- corpus = DyViS (Nolan et al. 2009)
  - 100 male speakers
  - Standard Southern British English (SSBE)
  - 18-25 years old
- Task 2 studio (near-end) recordings
  - information exchange task with ‘accomplice’ over landline telephone
  - 44.1kHz/ 16-bit depth audio
  - 10-15 minutes in duration

2. Methods

preparation of sound files

- manual editing to remove overlapping speech, overlapping background noise and non-linguistic sounds (e.g. clicks, audible breath)
- silences > 100 ms removed
- clipping detected and sections removed
- samples reduced to 4 minutes

2.1 VPA analysis

- in-house JPFA version of Laver (1981) VPA scheme
  - used 7 (incl. 0) scalar degrees
  - representing deviations from 'neutral' setting

VOCAL PROFILE ANALYSIS PROTOCOL

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

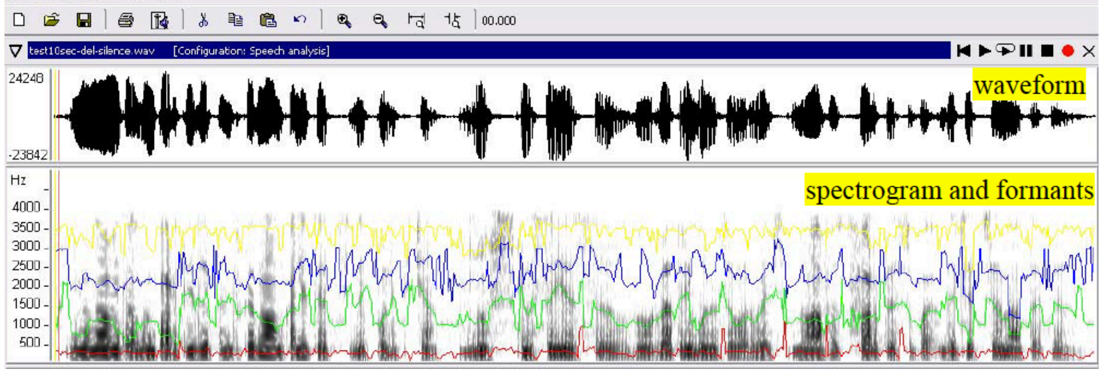
	FIRST PASS		SECOND PASS						
				moderate			extreme		
	Neutral	Non-neutral	SETTING	1	2	3	4	5	6
A. VOCAL TRACT FEATURES									
1. Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimised range						
2. Mandibular			Close jaw						
			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						

- auditory analysis performed by LS
  - only 27 supralaryngeal features analysed here

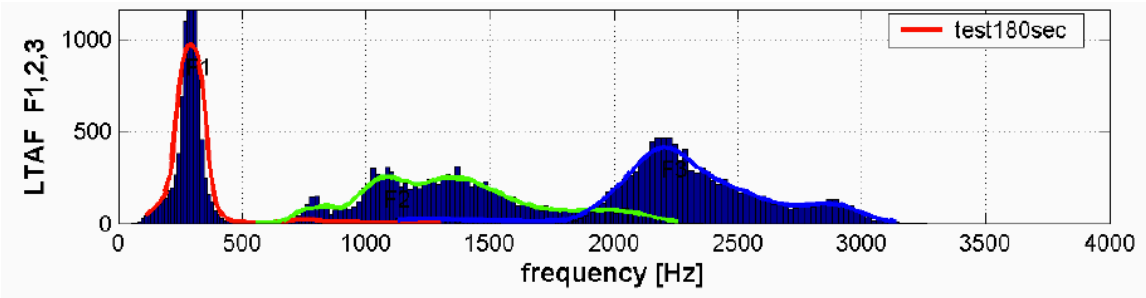
2.2 MFCC/LPCC analyses

- pre-emphasis filter applied (value = 0.97)
- entire signal divided into a series of overlapping *frames*
  - 20 ms hamming window shifted at 10 ms intervals
  - 50% overlap between adjacent frames
- 16 MFCCs/16 LPCCs extracted from each frame using RASTAMAT toolkit (Ellis 2005) in MATLAB

2.3 LTFDs



- automatic separation into C and V using StkCV (Andre-Obrecht 1988)
- vowel-only samples
  - 25 ms Gaussian window shifted at 5 ms
- F1~F4 values extracted from each frame
  - iCAbS tracker (Harrison & Clermont 2012)



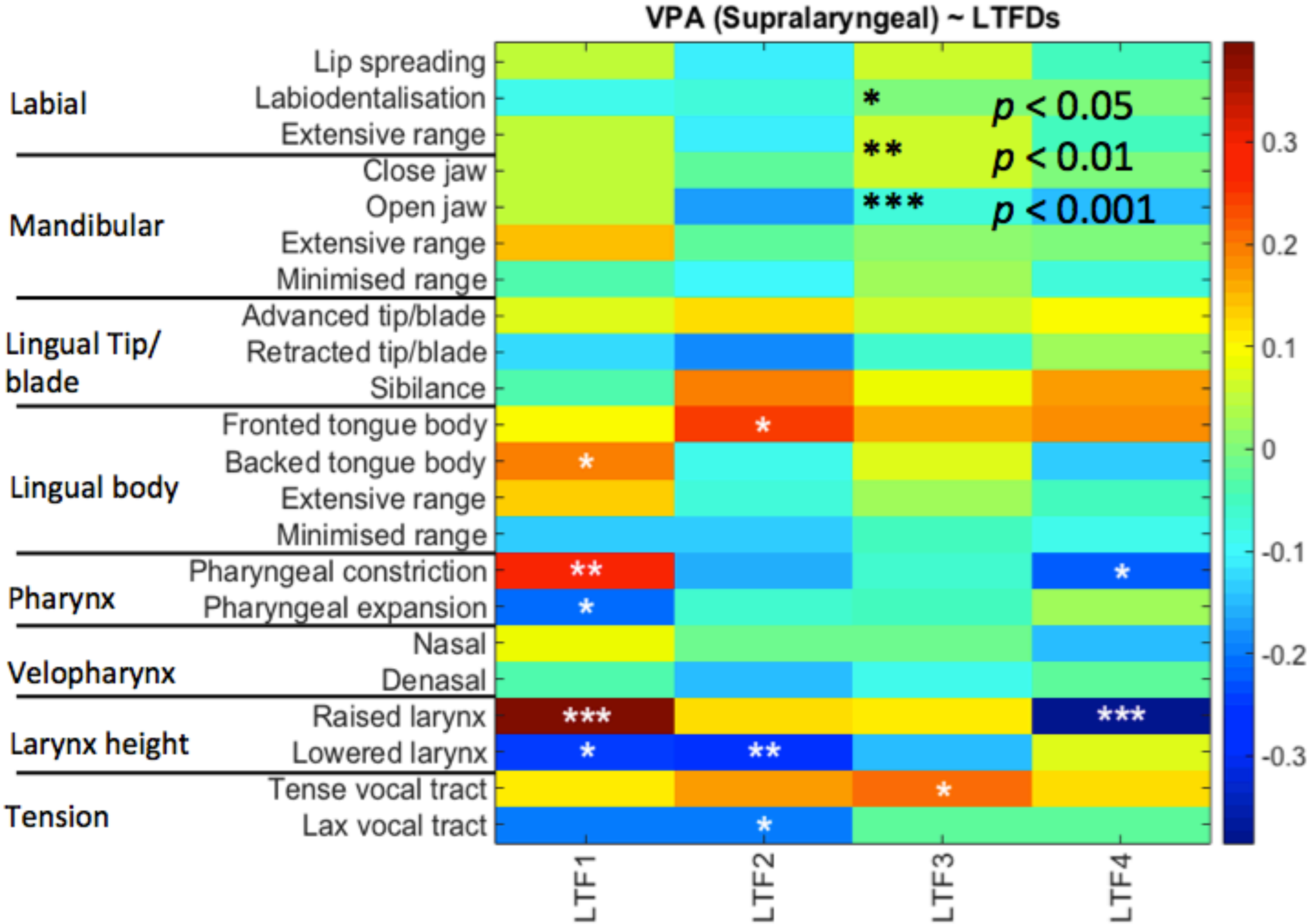
3. Experiment (1): correlations

method

- by-speaker means calculated for LTF1~LTF4 (LTFM = long term formant mean)
- Spearman correlations (non-parametric) matrix generated for LTFDs and VPA scores
- plotted as heatmaps based on rho value:
  - dark colours = stronger correlation
  - **red** = positive correlation
  - **blue** = negative correlation

3. Experiment (1): correlations

VPA (supralaryngeal) ~ LTFD


VPA (Supralaryngeal) ~ LTFDs
Labial
Lip spreading
Labiodentalisation
Extensive range
Mandibular
Close jaw
Open jaw
Extensive range
Minimised range
Lingual Tip/ blade
Advanced tip/blade
Retracted tip/blade
Sibilance
Lingual body
Fronted tongue body
Backed tongue body
Extensive range
Minimised range
Pharynx
Pharyngeal constriction
Pharyngeal expansion
Velopharynx
Nasal
Denasal
Larynx height
Raised larynx
Lowered larynx
Tension
Tense vocal tract
Lax vocal tract
* p < 0.05
** p < 0.01
*** p < 0.001
LTF1
LTF2
LTF3
LTF4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3

3. Experiment (1): correlations

LTFD 1

• backed tongue body	rho = 0.200	$p = 0.045$*
• pharyngeal constriction	rho = 0.298	$p = 0.0026$**
• pharyngeal expansion	rho = -0.213	$p = 0.034$*
• raised larynx	rho = 0.397	$p < 0.0001$***
• lowered larynx	rho = -0.248	$p = 0.013$*

LTFD 2

• fronted tongue body	rho = 0.239	$p = 0.0164$*
• lowered larynx	rho = -0.257	$p = 0.0097$**
• lax vocal tract	rho = -0.197	$p = 0.049$*

LTFD 3

• tense vocal tract	rho = 0.242	$p = 0.041$*

LTFD 4

• pharyngeal constriction	rho = -0.220	$p = 0.028$*
• raised larynx	rho = -0.385	$p < 0.0001$***

4. Experiment (2): clustering

method

- 1024 Gaussian GMMs generated in MATLAB (ISP toolkit) for each speaker for the MFCCs/LPCCs
- Kullback-Leibler (KL) divergences between speaker models
  - measure of distance (similarity) between speakers
  - **near** = similar/ **far** = dissimilar
- speakers plotted in 2D KL divergence space using multidimensional scaling

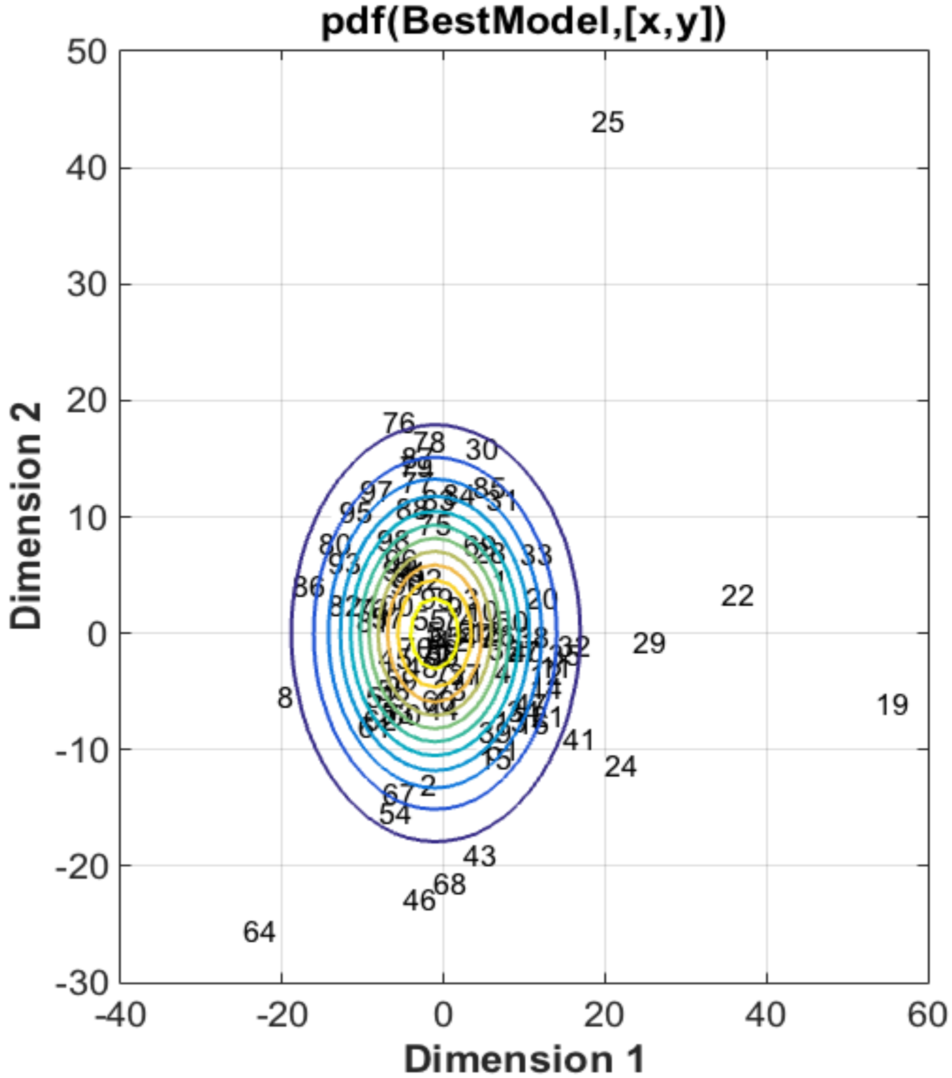
4. Experiment (2): clustering

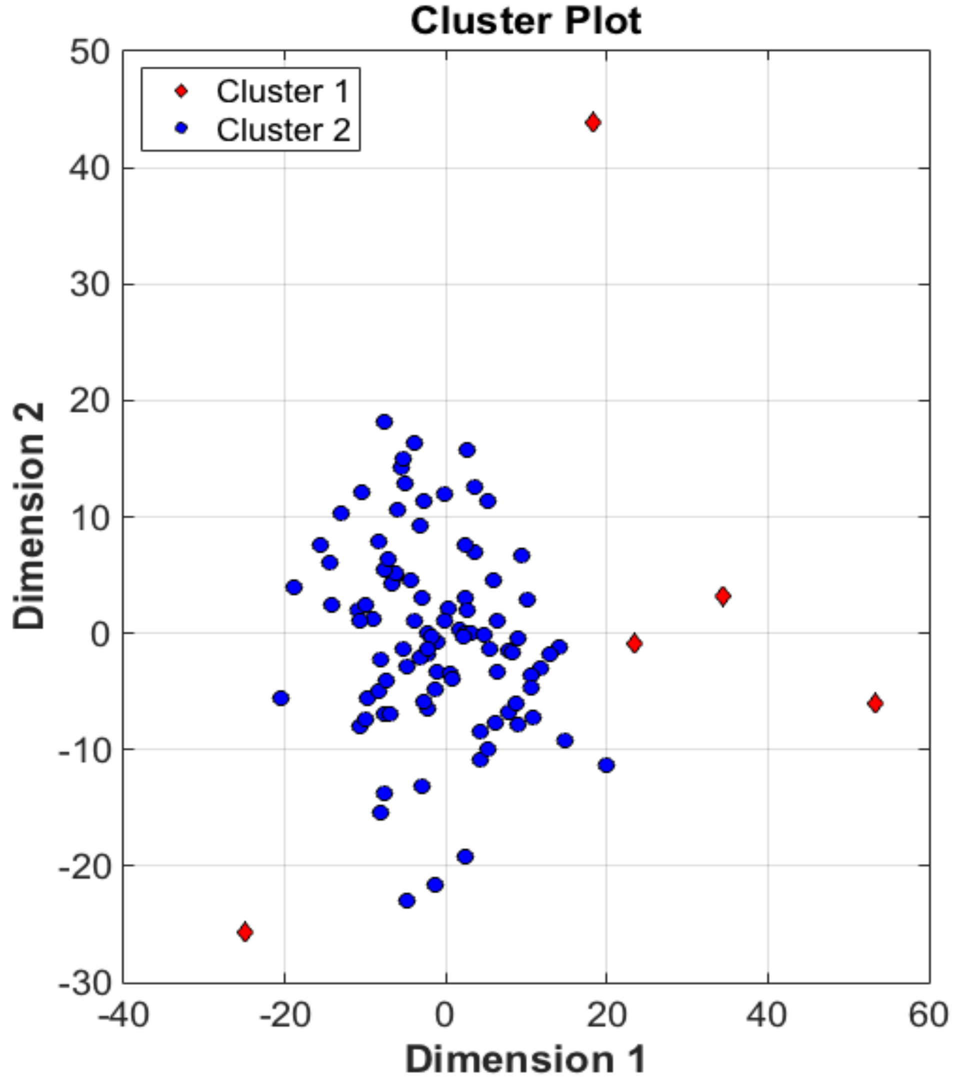
method

- cluster analysis (using GMMs) performed using co-ordinates in KL divergence space to identify speaker groups
  - N clusters determined by AIC fit statistic
- speaker clusters analysed relative to VPA profiles
- outlying speakers identified and analysed
  - supralaryngeal VPA scores
  - any other features (e.g. segmental, temporal, technical) which might separate these speakers from the clusters

4. Experiment (2): clustering

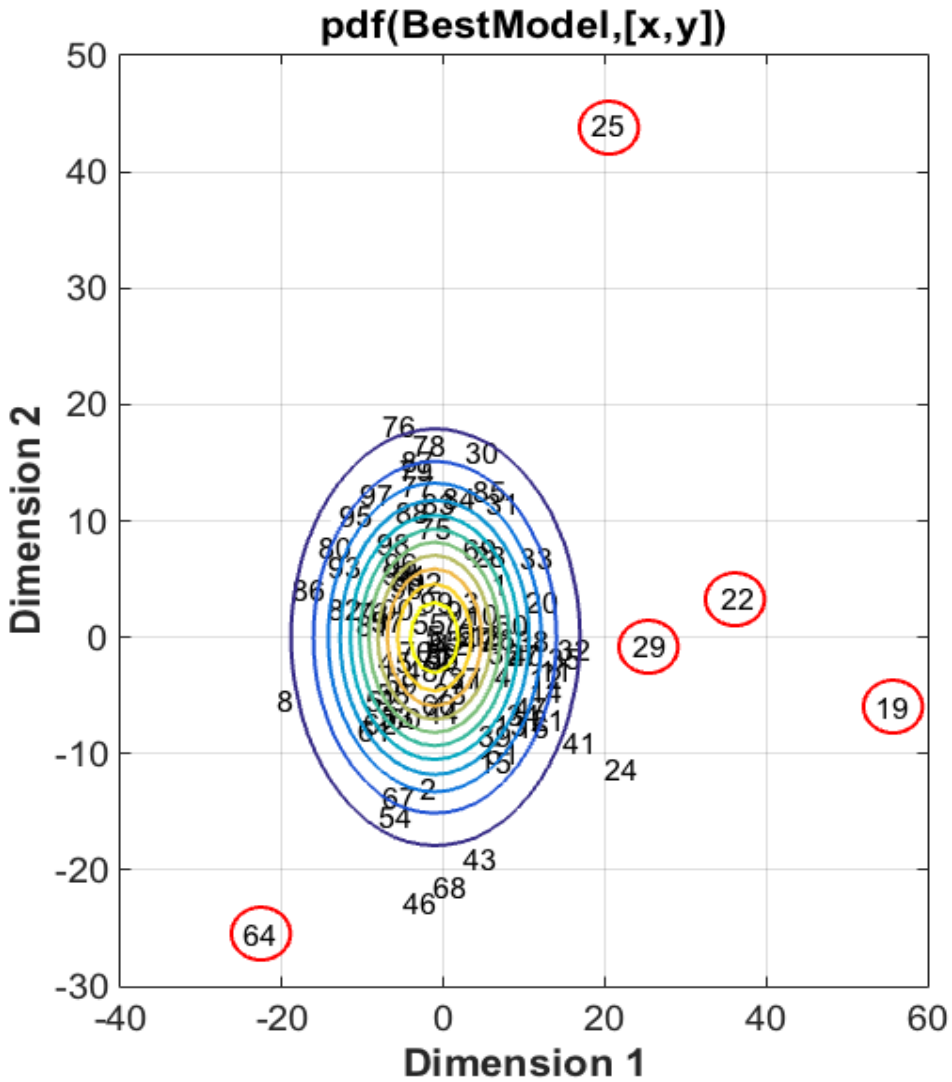
16 MFCCs

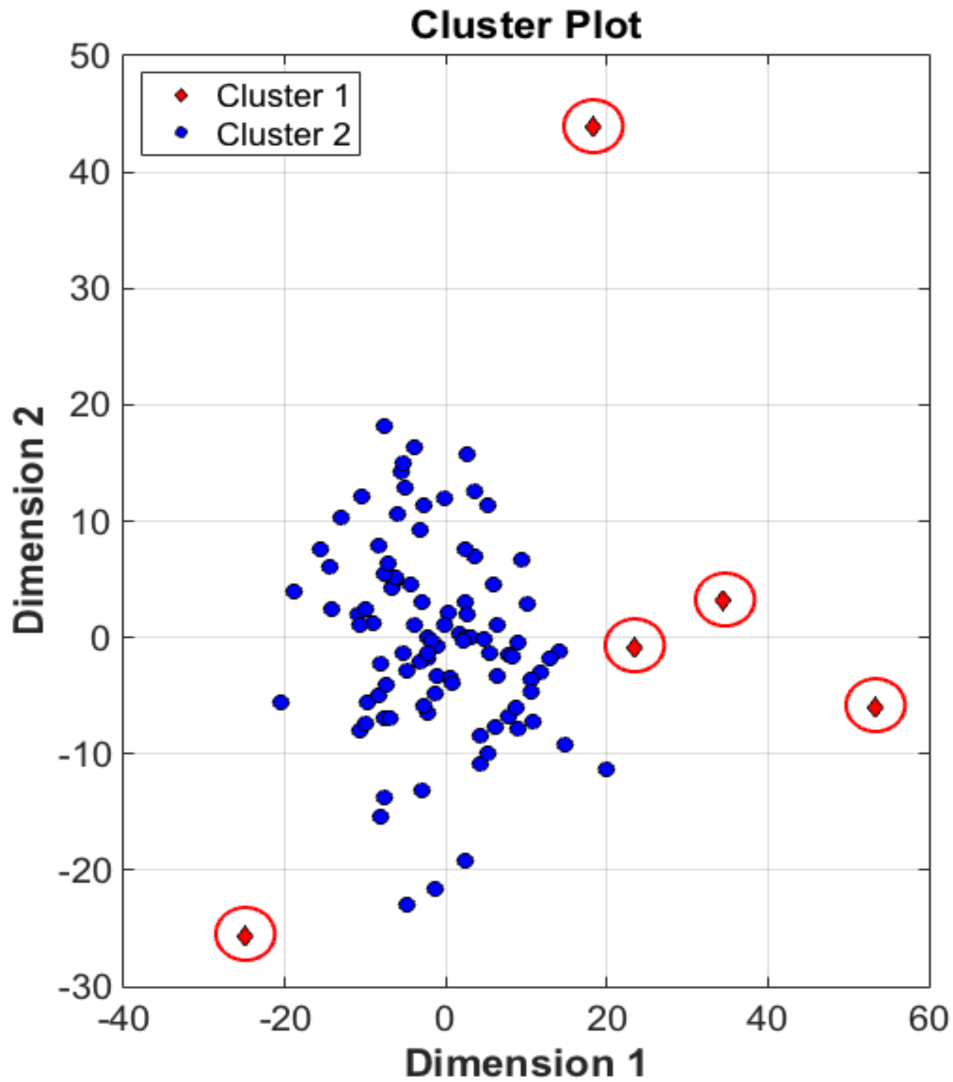




4. Experiment (2): clustering

16 MFCCs





4. Experiment (2): clustering

16 MFCCs

- outliers (as identified by the clustering):
  - 19 (022-2-060330)
  - 22 (025-2-060425)
  - 35 (028-2-060426)
  - 29 (032-2-060428)
  - 64 (072-2-061009)
- **are these speakers unusual in terms of overall supralaryngeal VPA profiles?**

4. Experiment (2): clustering

yes…

Sp 19	Sp 22	Sp 25	Sp 29	Sp 64
Advanced tongue tip	Low larynx	Audible nasal escape	Lax larynx	Advance tongue tip
Tense vocal tract	Lax larynx			Lax larynx
Nasal				Whispery

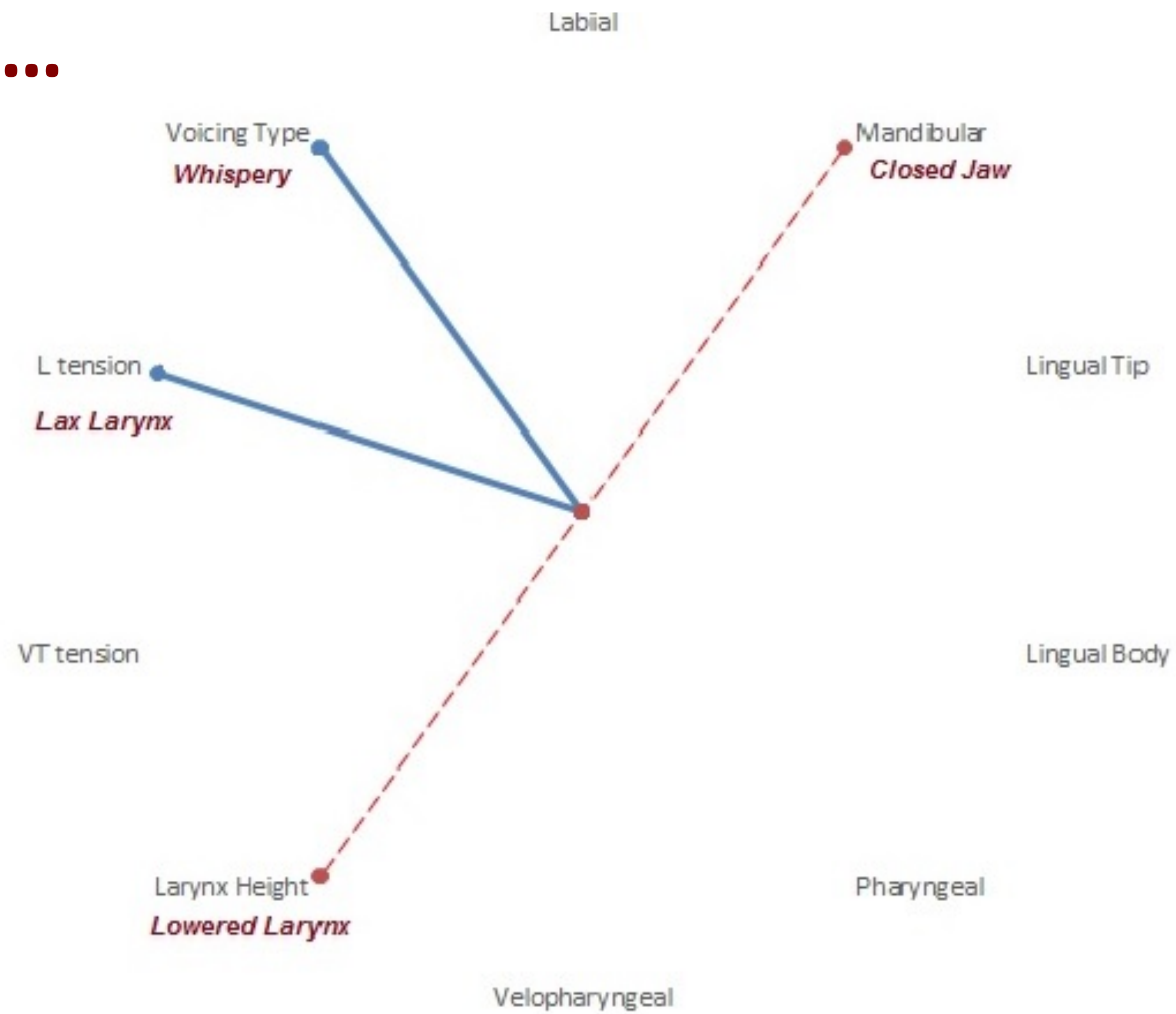
**Agreement reached between two independent phoneticians. Procedure: blind evaluation; two passes each expert.

4. Experiment (2): clustering

is there systematicity in the clustering of speakers? yes...

Speakers clustered in the middle:

Sp 05
Sp 07
Sp 42
Sp 56
Sp 89

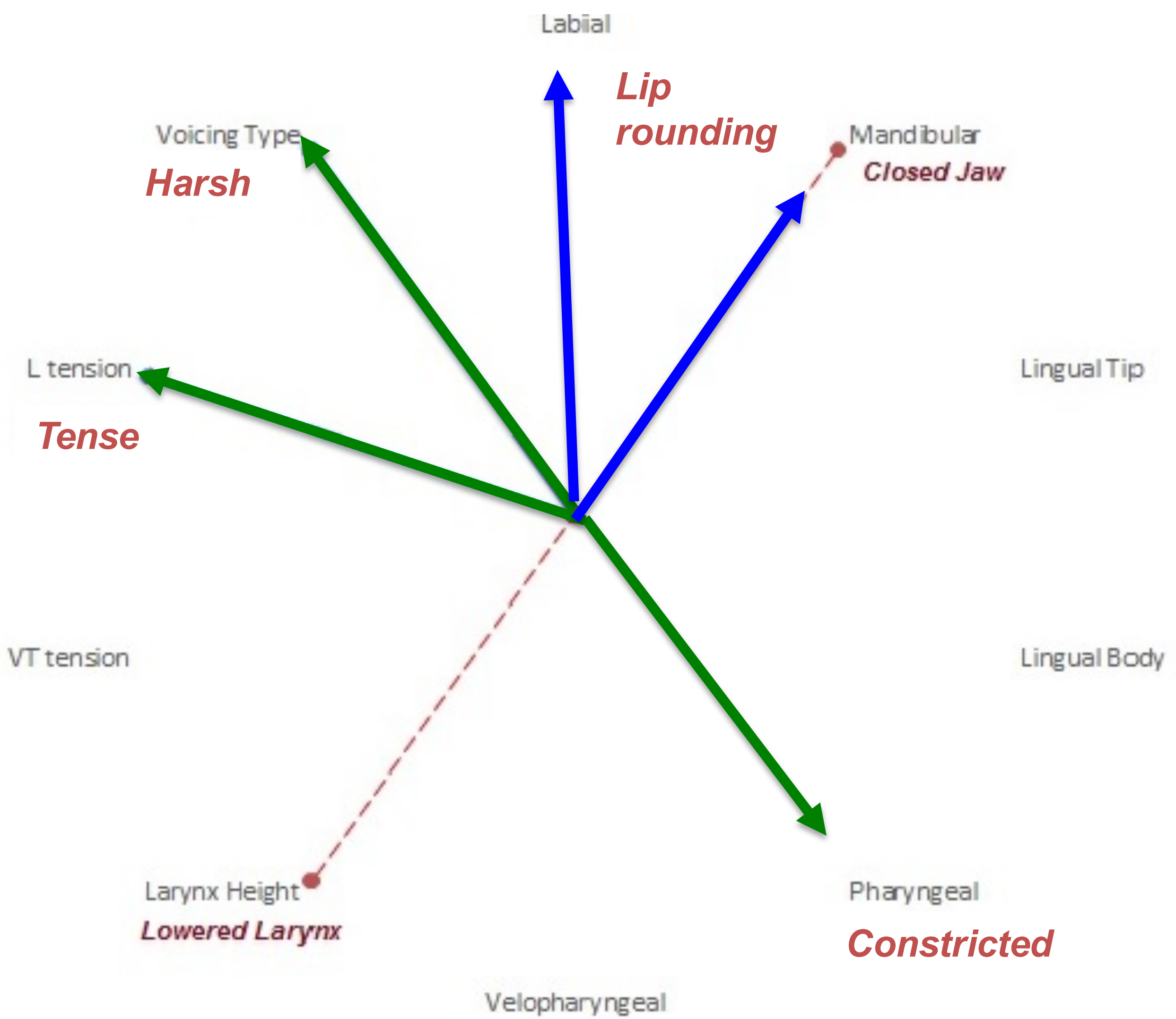


4. Experiment (2): clustering

…and no

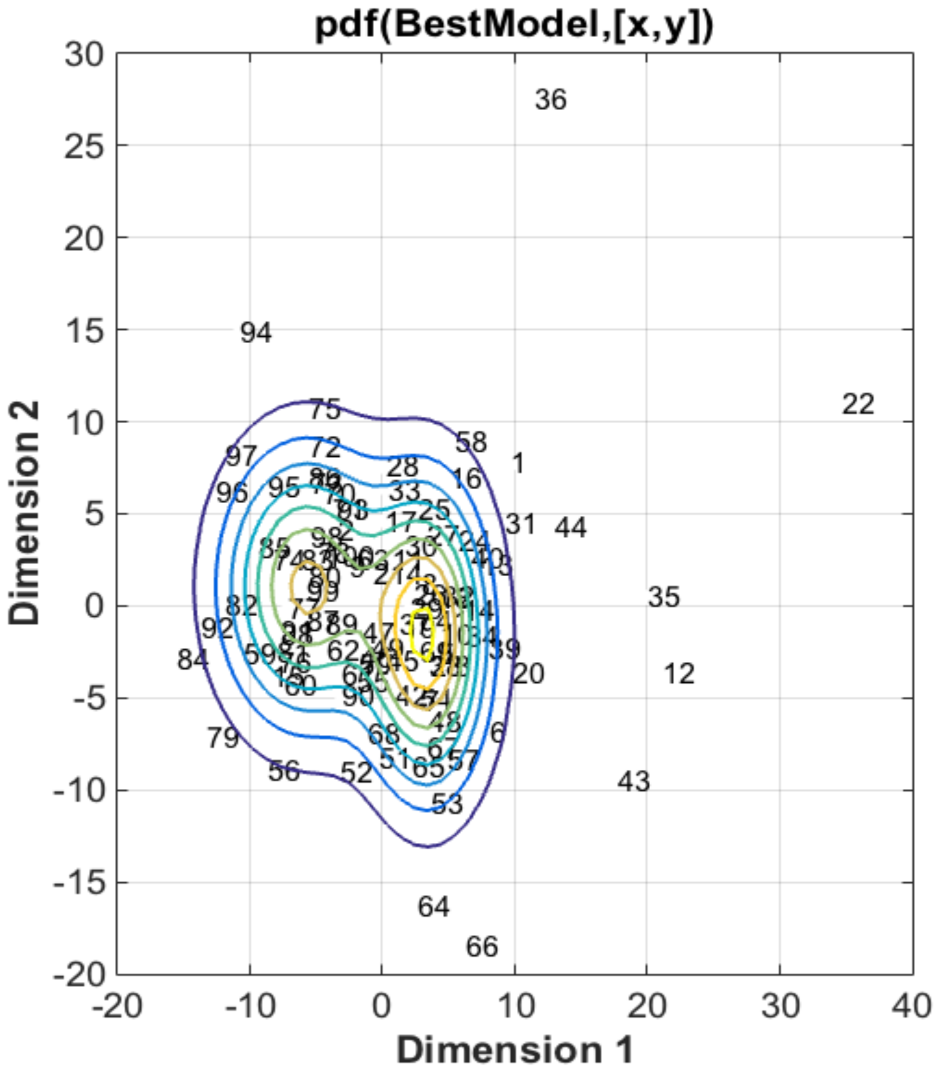
Speakers clustered in the middle:

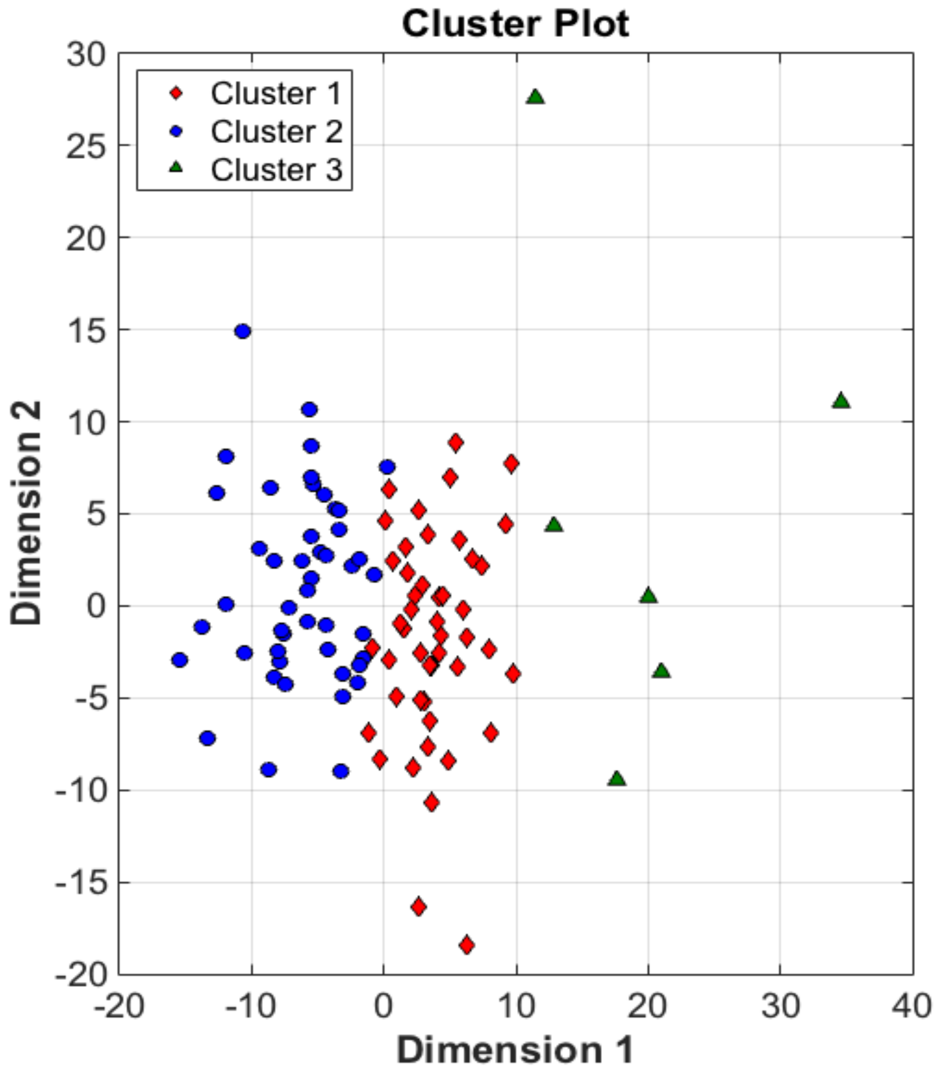
Sp 05
Sp 07
Sp 42
Sp 56
Sp 89



4. Experiment (2): clustering

16 LPCCs





4. Experiment (2): clustering

16 LPCCs

- which speakers are grouped together?
  - no clear explanation for the groupings of speakers in the two main clusters
  - general supralaryngeal VPA profiles = very similar (accent features)
    - advanced tongue tip
    - sibilance
    - fronted tongue body

4. Experiment (2): clustering

16 LPCCs

- outliers (as identified by the clustering):
  - 12 (015-2-060324)
  - 22 (025-2-060425)
  - 35 (038-2-060504)
  - 36 (039-2-060504)
  - 43 (047-2-060607)
  - 44 (048-2-060608)
- **are these speakers unusual in terms of overall supralaryngeal VPA profiles?**

4. Experiment (2): clustering

yes…

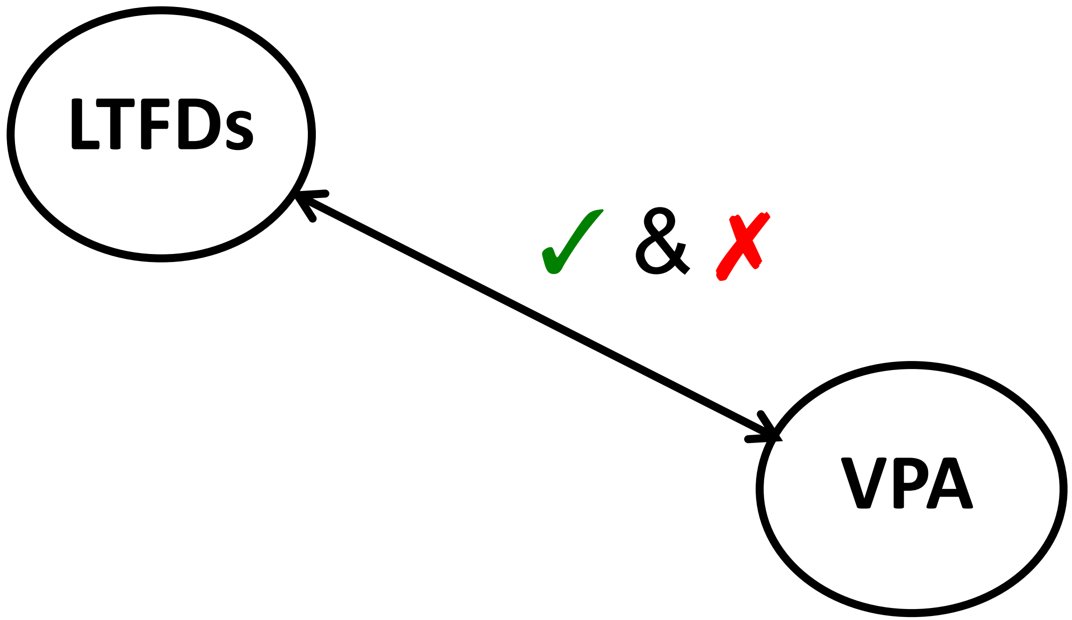
- possible to find dimensions on which speakers differ

… and no

- but these speakers aren’t especially distinctive relative to the group
- greater between-speaker VPA differences for speaker pairs in the centre of the clusters

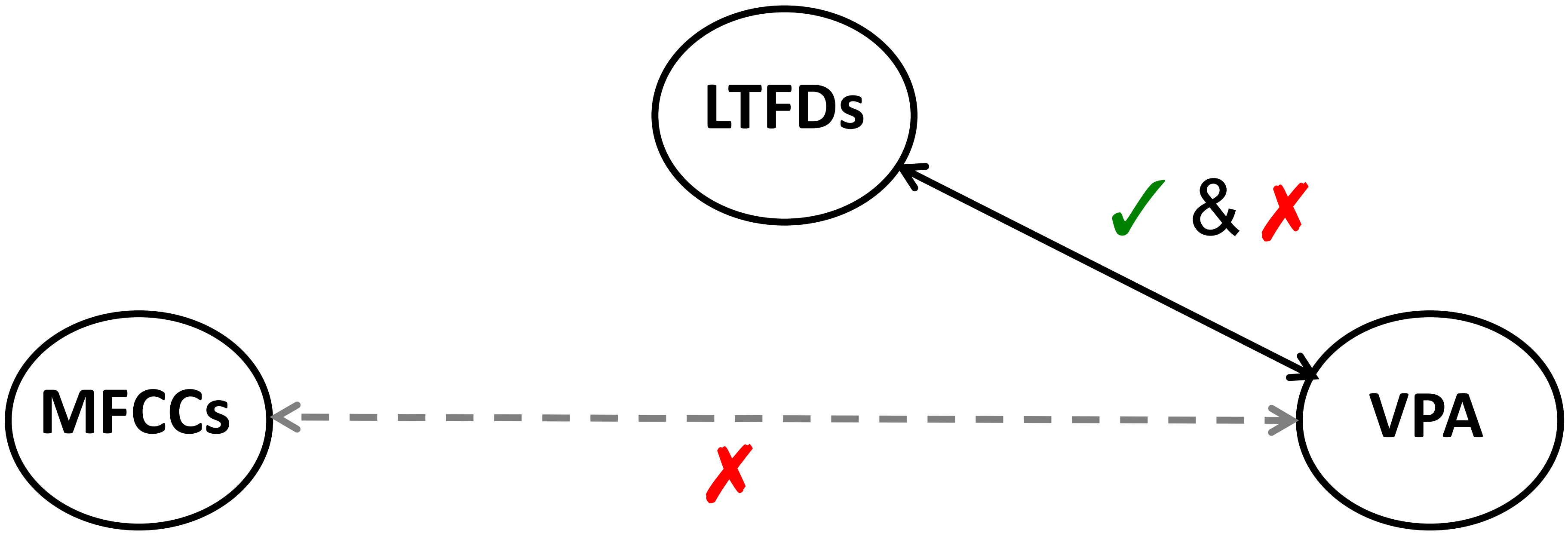
5. Discussion

- interrelationships between long-term measures of vocal tract output…



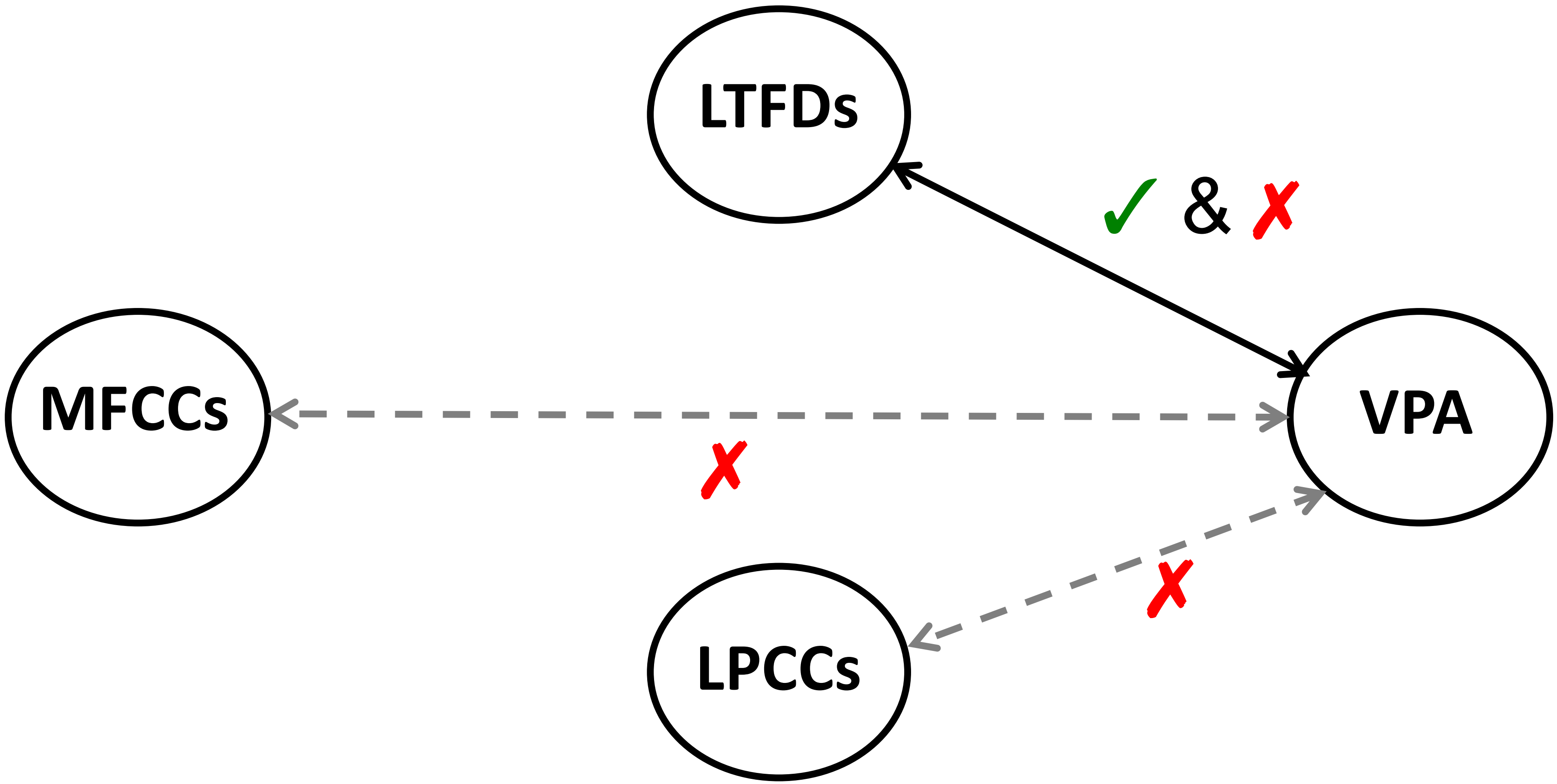
5. Discussion

- interrelationships between long-term measures of vocal tract output…



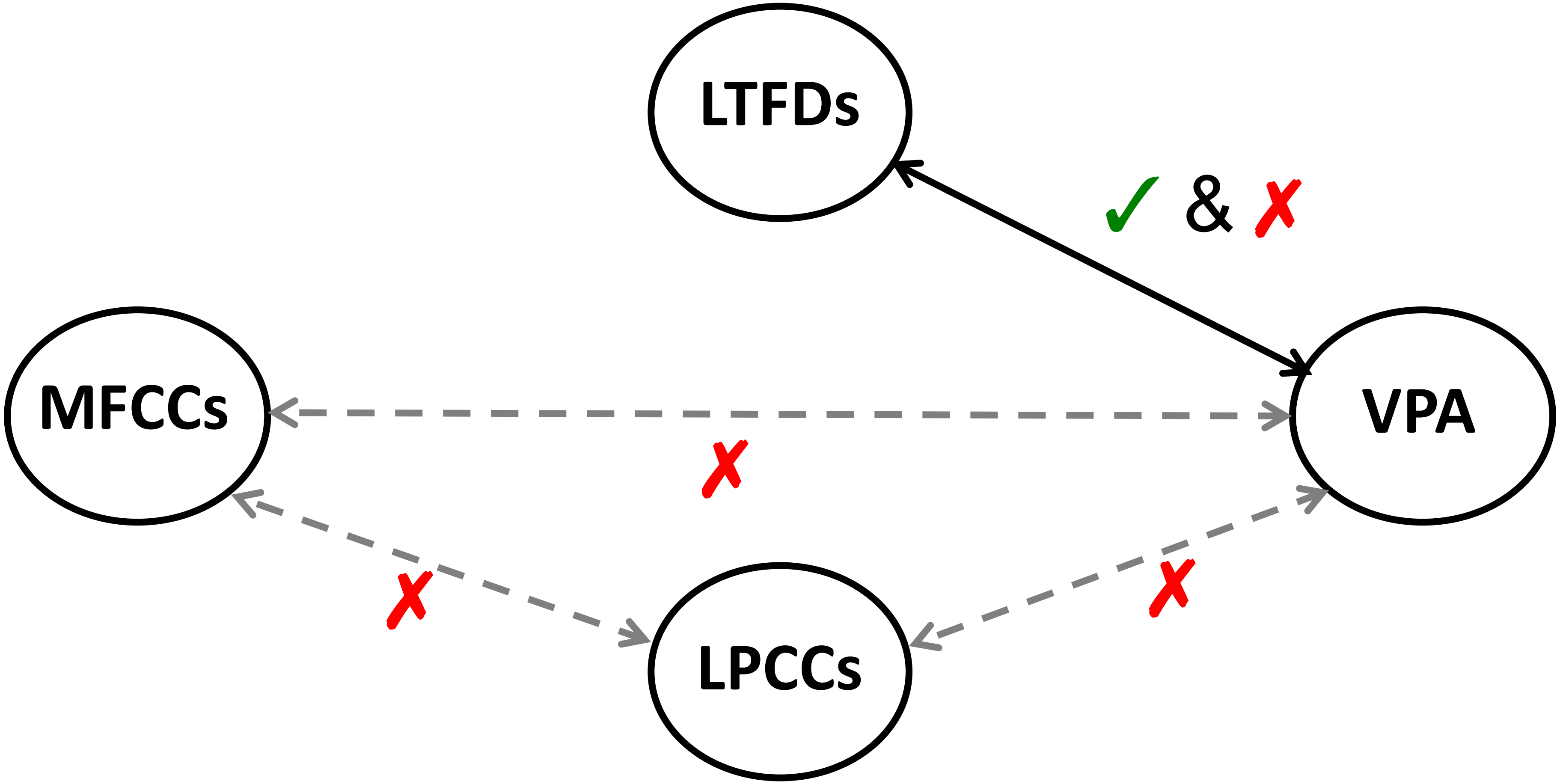
5. Discussion

- interrelationships between long-term measures of vocal tract output…



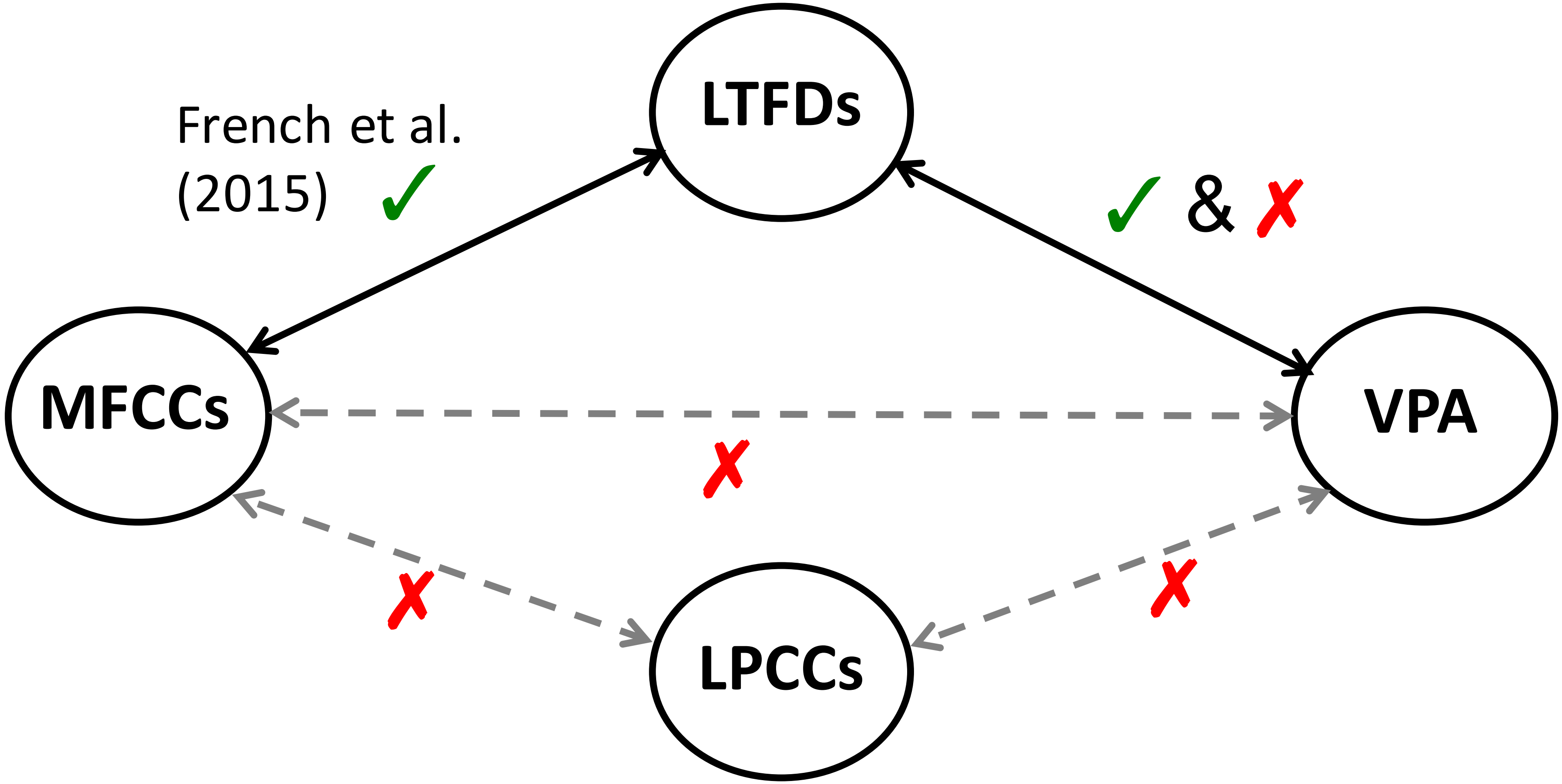
5. Discussion

- interrelationships between long-term measures of vocal tract output...



5. Discussion

- interrelationships between long-term measures of vocal tract output…



6. Conclusion

- complementary VT information provided by auditory (supralaryngeal VPA) and acoustic (LTFDs to some extent and CCs) analyses
  - potential for improving the performance of ASRs by including independent VPA information
- further complementary information provided by laryngeal VPA (Gonzalez-Rodriguez et al. 2014) and segmental features

Thanks! Questions?