IAFPA 2021

IVIP

Phonetic correlates of listeners' judgements of voice similarity within and across accents

Kirsty McDougall, Alice Paver,

Francis Nolan, Nikolas Pautz,

Harriet M.J. Smith and Philip Harrison

UKRI Economic and Social Research Council

University of Cambridge

Nottingham Trent University

J P French Associates

The University of York

IVIP: Improving Voice Identification Procedures

IVIP

Project Team

University of Cambridge:	PI: Kirsty McDougall (phonetics)
	Francis Nolan (phonetics)
	Alice Paver (phonetics)
Nottingham Trent University:	Harriet Smith (psychology)
	Natalie Braber (sociolinguistics)
	David Wright (sociolinguistics)
	Nikolas Pautz (psychology)
De Montfort University:	Jeremy Robson (law)
University of Oxford:	Katrin Müller-Johnson (criminology, psychology)

Philip Harrison
University of York and
J.P. French Associates

UK ESRC Grant Ref: ES/S015965/1

UKRI Economic and Social Research Council

Improving Voice Identification Procedures (IVIP)

- Multi-disciplinary approach (psychology, linguistics, criminology & law)
- 4 different strands:

Strand 1: What are the optimal parameter values for voice parade procedures?

Strand 2: What are the psycho-phonetic underpinnings of voice distinctiveness?

Strand 3: How do social stereotypes affect voice identification?

Strand 4: How accurate are the normative assumptions of criminal justice practitioners in respect of voice identification procedures?

Outline

- Notion of Perceived Voice Similarity and previous research
- Experiment structure – stimuli, task, listeners
- Results – MDS and correlation analysis
- Main findings and discussion

Perceived Voice Similarity (PVS)

IVIP

- Principles for foil selection for voice parades still evolving
- Perceived voice similarity not well understood

- *What phonetic features contribute to certain speakers being judged as sounding more similar to or more different from each other?*
- *How do different accents affect judgements of voice similarity?*

Perceived Voice Similarity (PVS)

IVIP

- Within a group of speakers of same sex, age and accent background, listeners will perceive some speakers as more similar-sounding than others
- These similarity judgements are due to:
  - individual variation in vocal tract anatomy

    &

  - individual choices the speakers make in implementing their linguistic systems
- In the experiment to follow, we control the demographic characteristics (sex, age, accent) to enable us to examine this individual variation within each demographic profile

Previous research

IVIP

- Little previous (phonetically-informed) research (cf. Remez et al. 2007, Baumann and Belin 2010)
- Earlier study of PVS in Standard Southern British English (SSBE) (Nolan, McDougall and Hudson 2011, McDougall 2013)
    - developed from ESRC *VoiceSim* project with Francis Nolan and Toby Hudson
- Study by McDougall (2016) of SSBE versus York English

Summary of results - McDougall (2016)

- f0 important for SSBE, but less so York, possibly due to larger long-term f0 range for SSBE population
- Long-term formants playing key roles in PVS for both SSBE and York
- Limited role played by articulation rate, some significance for York
- Some agreement, but variation between SSBE/York listener groups on judgements of PVS

Cambridge (SSBE)

York

IVIP experiment on PVS

- 6 groups, 4 accents

Group	**Accent**	**Sex and age**	**No. speakers**
DyViS 1	SSBE (Standard Southern British English)	male, 18-25	15
DyViS 2	SSBE	male, 18-25	15
DyViS 3	SSBE	male, 18-25	15
YorViS	York English	male, 18-25	15
WYRED 1	Bradford English	male, 18-30	15
WYRED 2	Wakefield English	male, 18-30	15

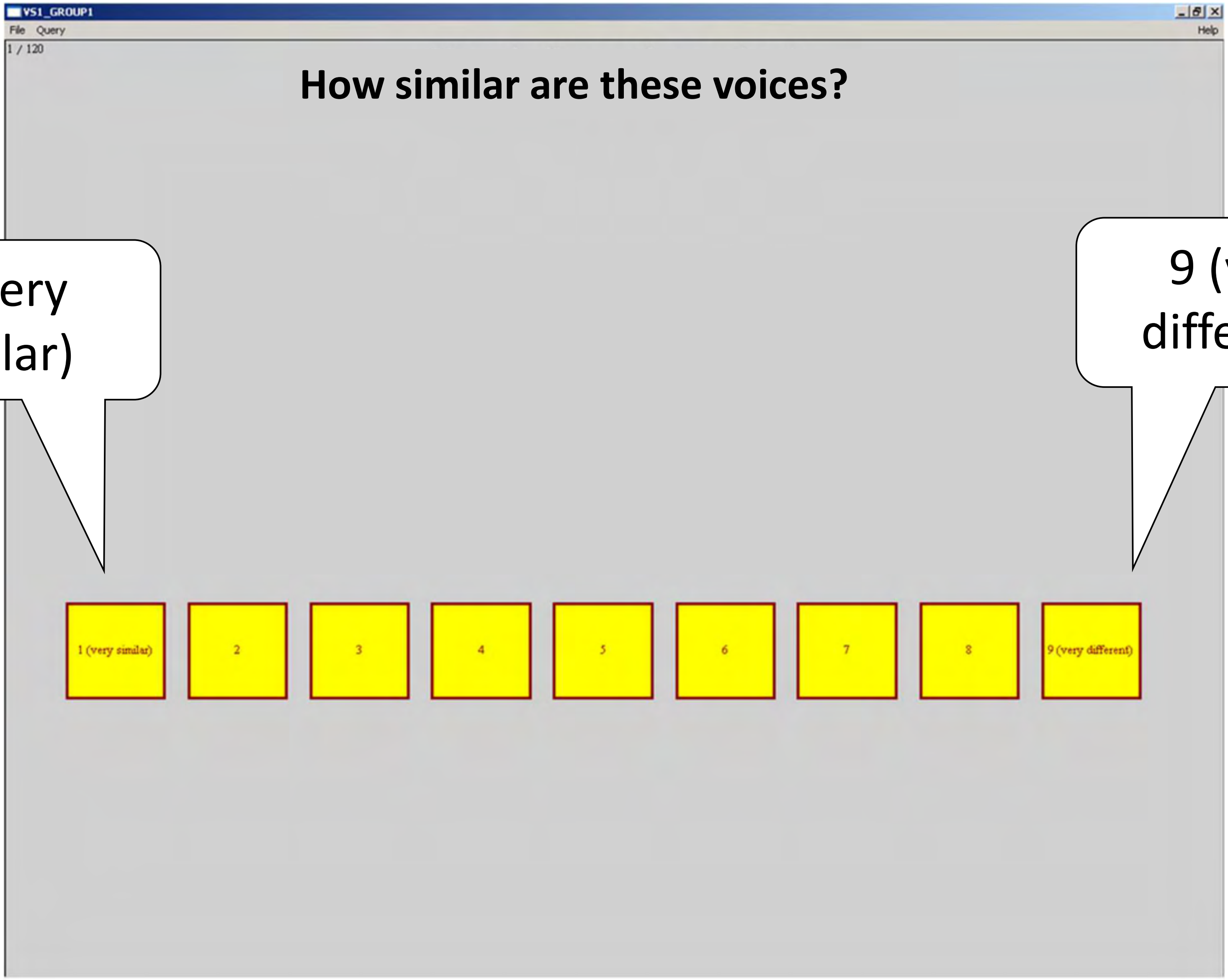
- *DyViS*, *YorViS* and *WYRED* databases – spontaneous speech, same elicitation technique

DyViS: Nolan et al. (2009);
YorViS: McDougall et al. (2015);
WYRED: Gold et al. (2018)

Stimuli

- two samples (approx. 3 secs) of spontaneous speech per speaker (telephone call, full bandwidth)
- within each 15-speaker group, samples paired, including same-speaker pairs (120 per group)
- task: to rate voices on 9-point distance scale from ‘very similar’ to ‘very different’

- DyViS 1, YorViS – in person (Praat)
- DyViS 2&3, WYRED 1&2 – online (Gorilla)



How similar are these voices?

1 (very similar)

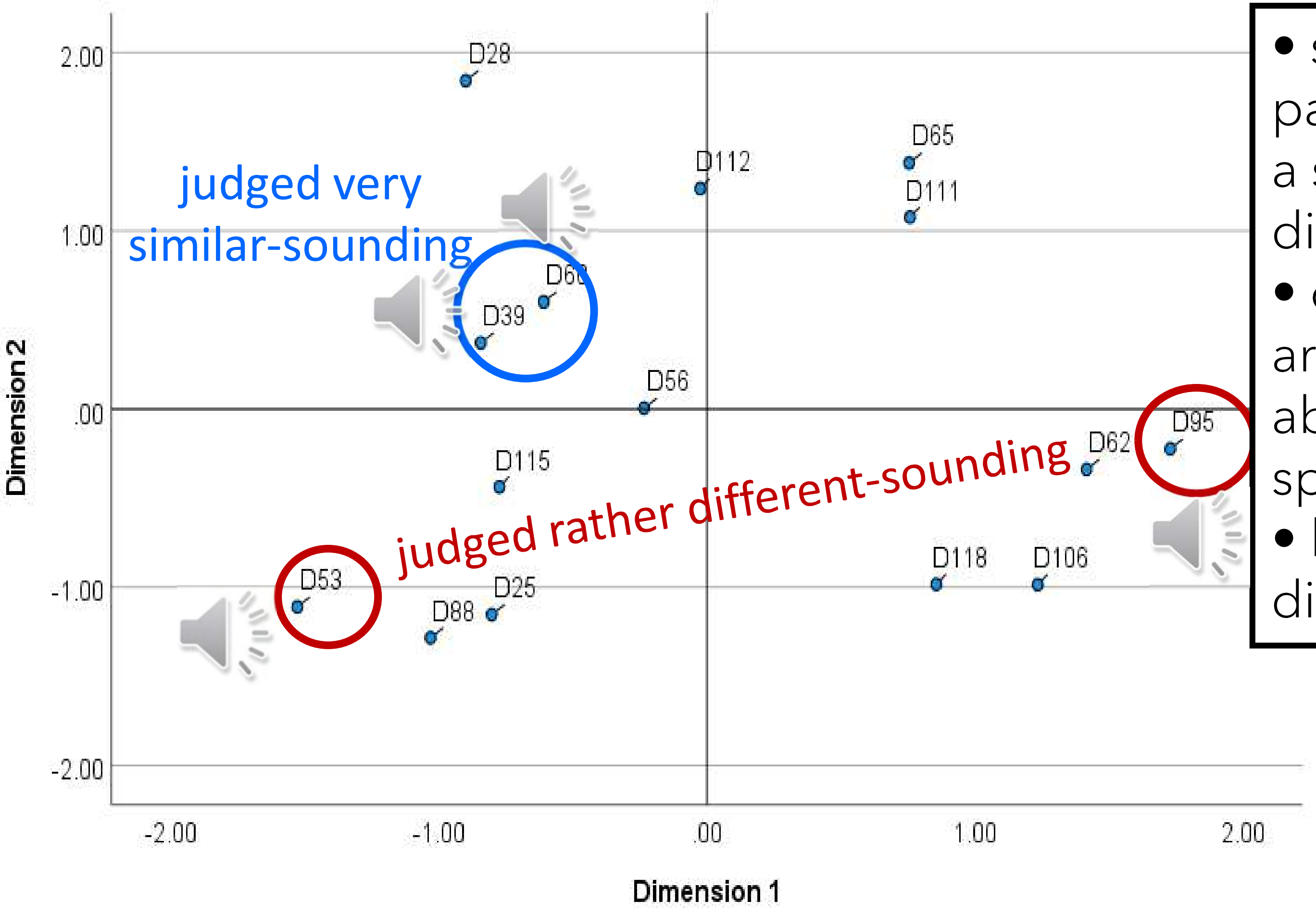
9 (very different)

Listeners

- N = 120 participants recruited at University of Cambridge, Nottingham Trent University and via Prolific (20 per group)
  - born in and lived most of their pre-18 lives in Great Britain
  - 1st language English
  - No hearing loss or hearing difficulties
  - Aged 18-40
  - Approx. half male, half female

Multidimensional scaling

SSBE group 1

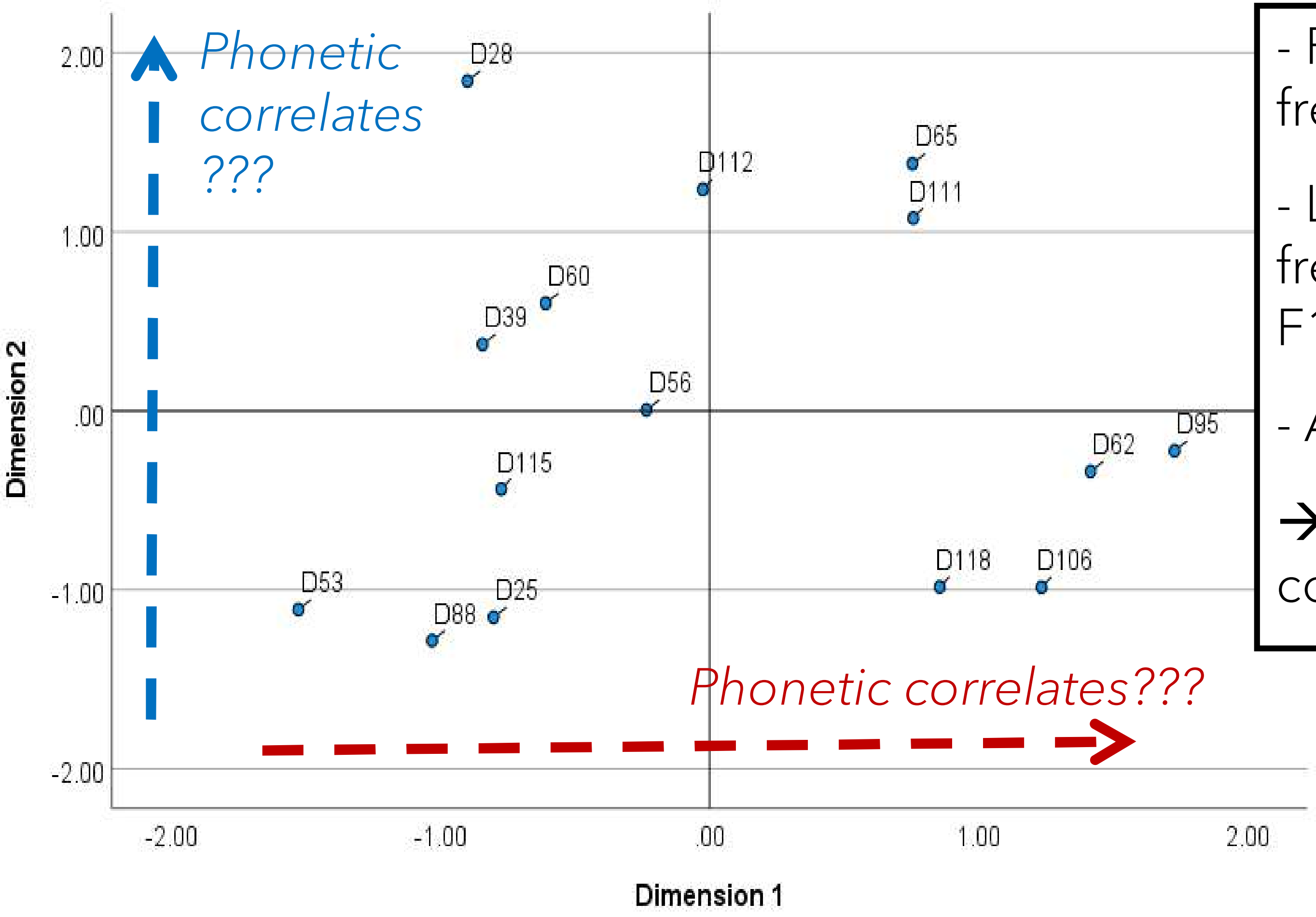


- summarises many pairwise distances in a smaller no. of dimensions
- objects (speakers) are placed in an abstract perceptual space
- here: 1st two of five dimensions shown

INSCAL, 5D solution: Stress = 0.17430, RSQ = 0.27713

Multidimensional scaling

SSBE group 1



- Fundamental frequency (f0)

- Long-term formant frequencies:
F1, F2, F3, F4

- Articulation rate

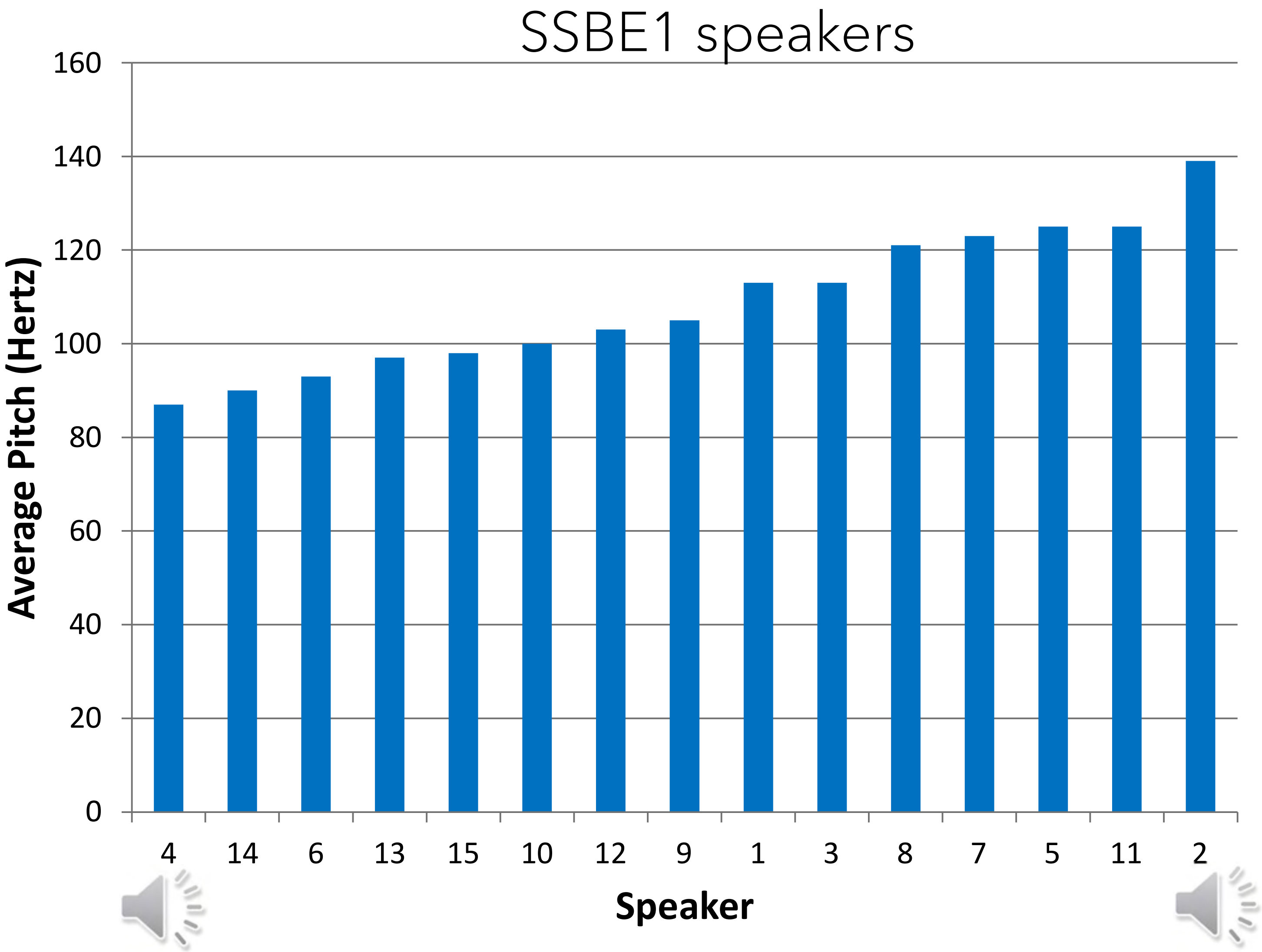
→ Test with Pearson correlation

INSCAL, 5D solution: Stress = 0.17430, RSQ = 0.27713

1. Long-term f0

- Long-term f0 calculated for each speaker using 6s speech from the 2 x 3s stimuli
- *Praat* script

1. Long-term f0

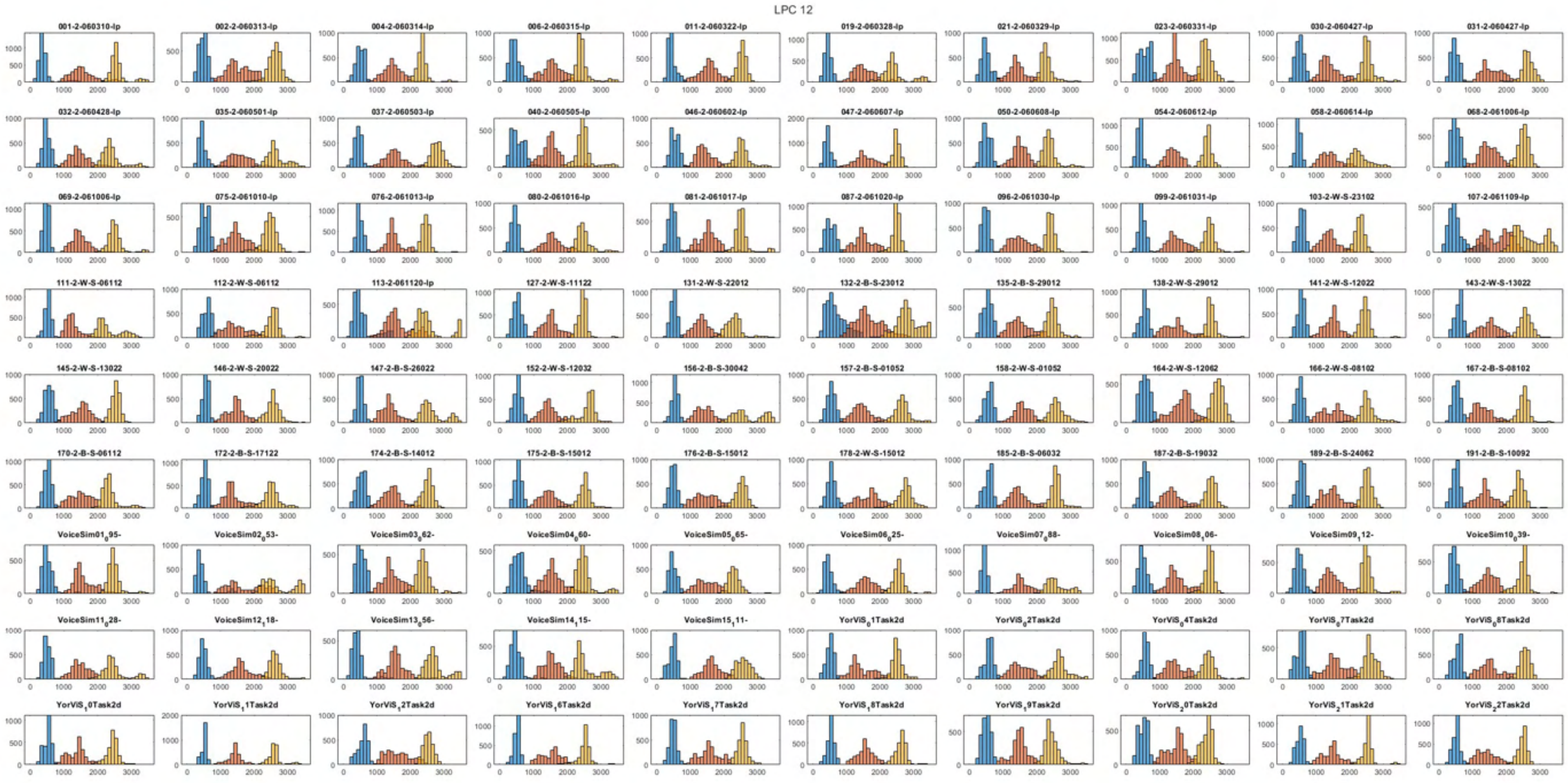


2. Long-term formants

IVIP

- Vowel/approximant material segmented from telephone task manually
  → 30s per speaker
- Snack Sound Toolkit (Sjölander, 1997)
- LTF for F1 to F4 - stable profiles achieved (except 5 speakers, excluded)
- Mean LTF values F1 to F4 calculated per speaker

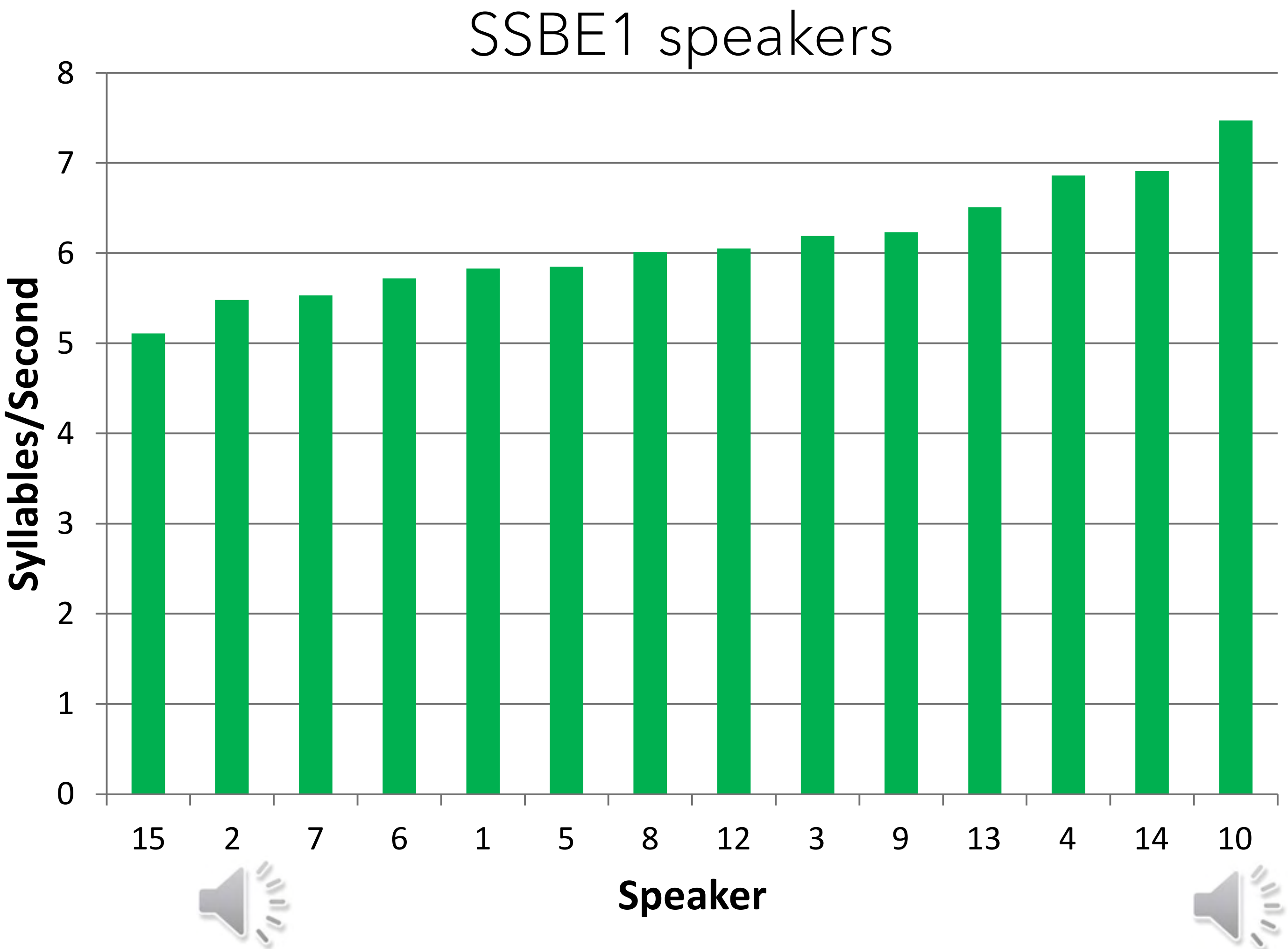
2. Long-term formants



3. Articulation rate

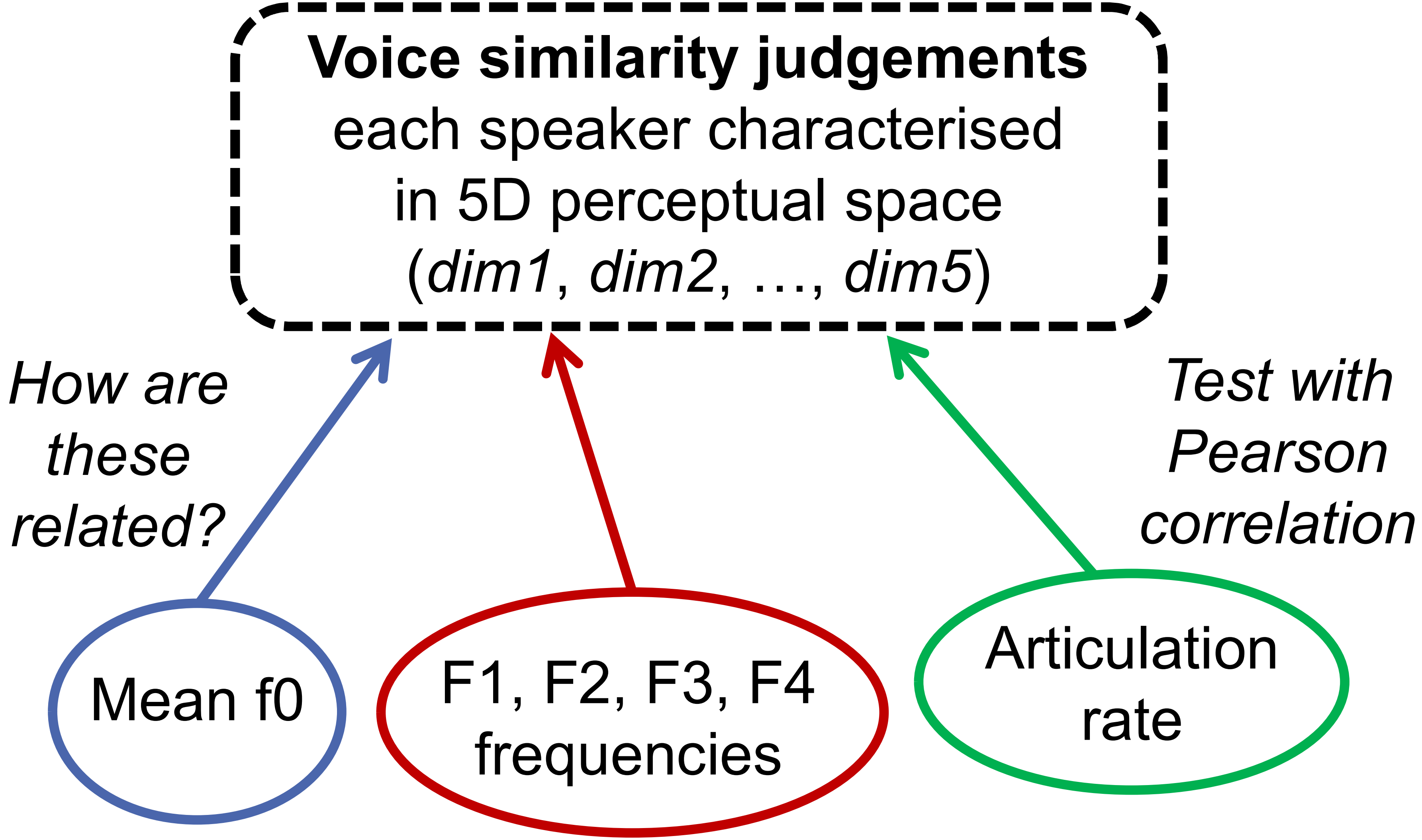
- Articulation rate (AR) calculated using telephone task recordings
- Jessen (2007) procedure for ‘global’ AR
- 30 ‘memory stretches’ of 5-20 phonetic syllables analysed, syllables determined auditorily

3. Articulation rate



Linking Voice Similarity to Phonetic Features

IVIP



Voice similarity judgements
each speaker characterised
in 5D perceptual space
(*dim1*, *dim2*, …, *dim5*)

How are these related?

Test with Pearson correlation

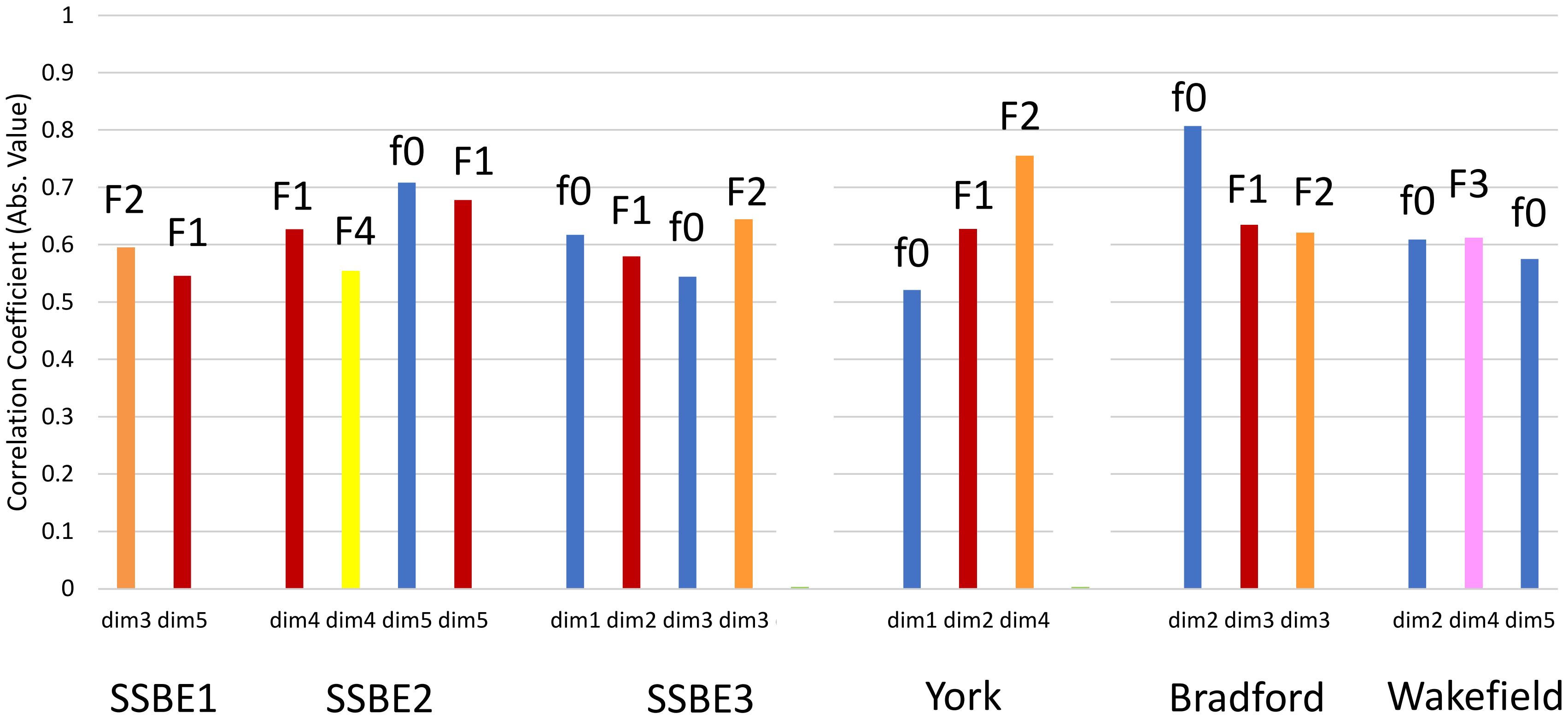
Mean f0

F1, F2, F3, F4 frequencies

Articulation rate

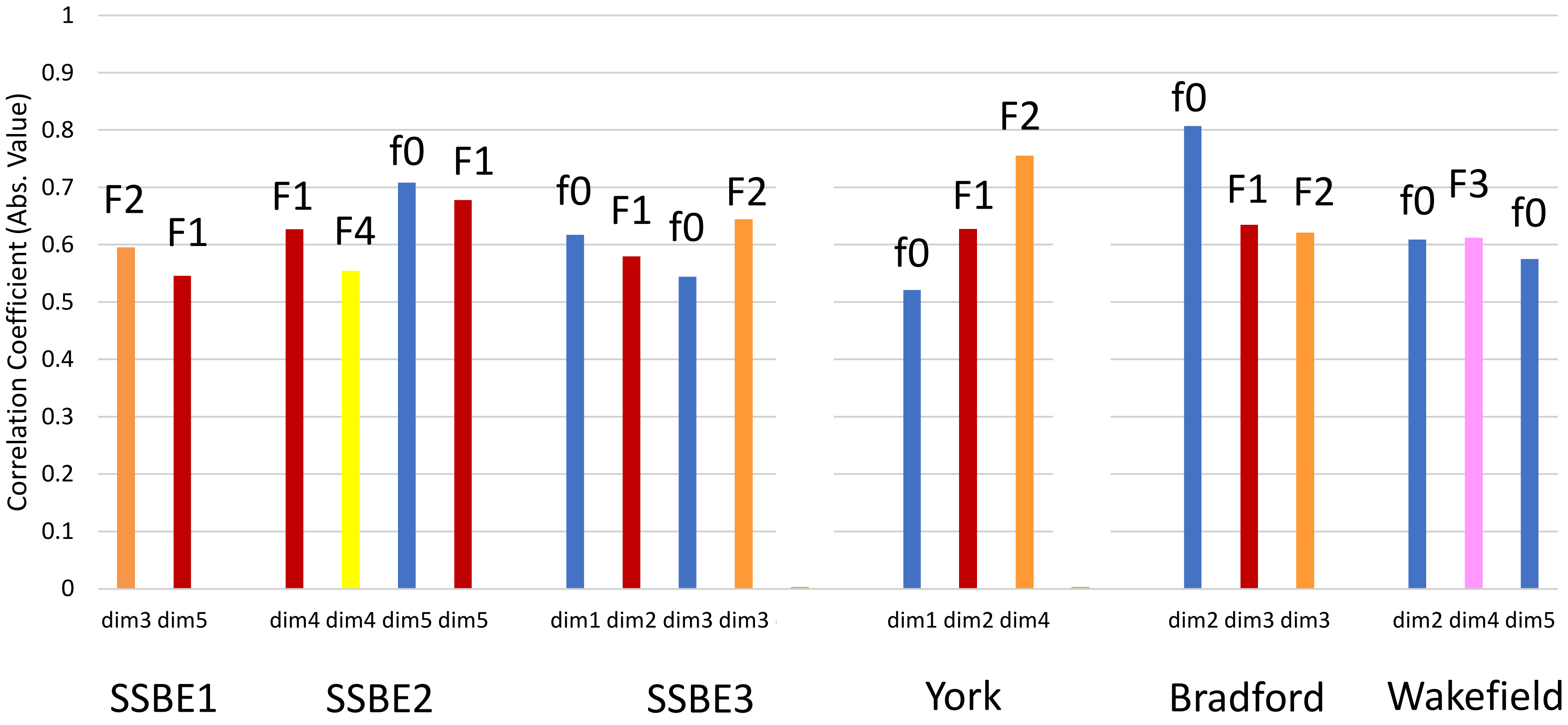
Correlation results – f0 & LTF

- Phonetic features yielding a significant correlation with a perceptual dimension are shown
- The lower the dimension number, the more important the feature



Correlation results – f0 & LTF

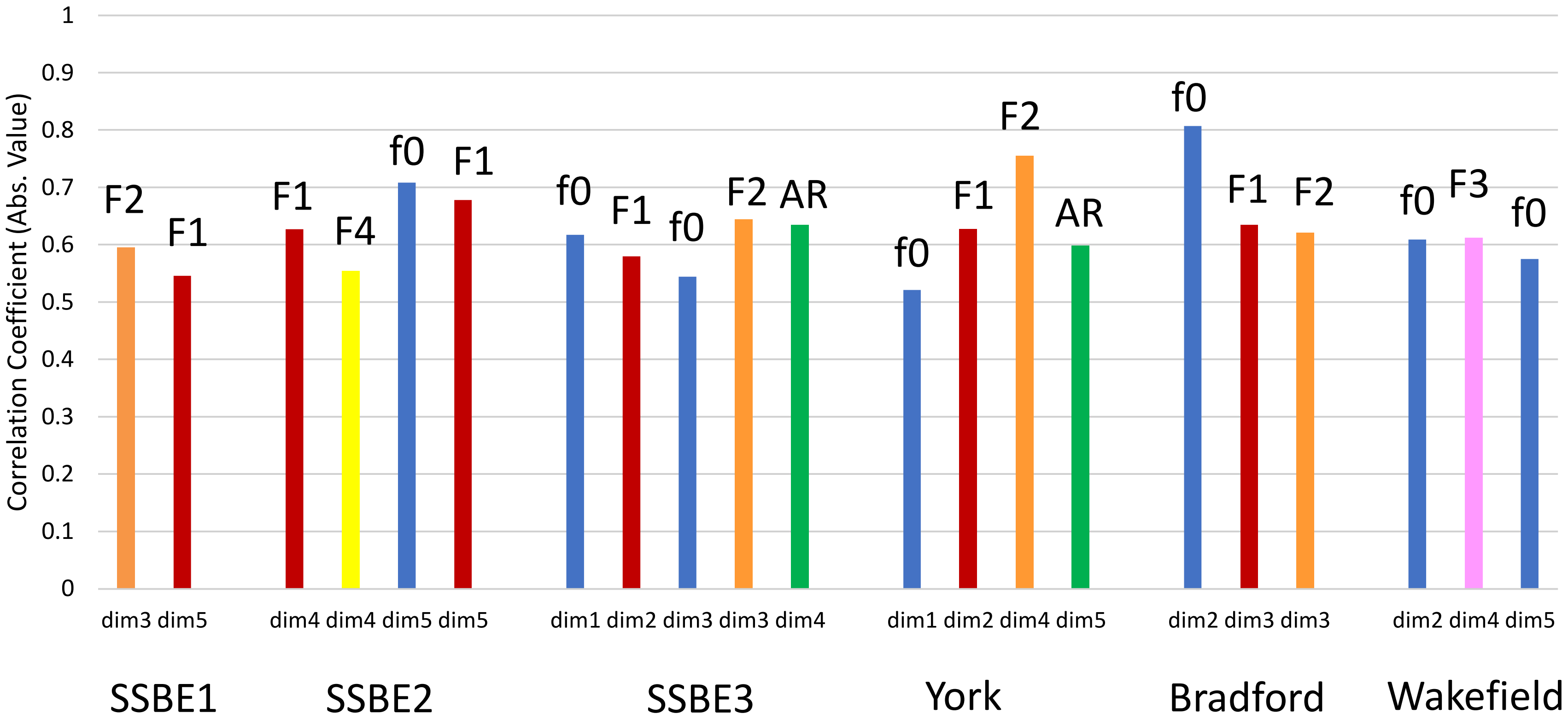
- f0 important for all groups except SSBE1
- Long-term formants playing differing roles in each accent
- Varying patterns within SSBE across the 3 groups



Correlation results – f0, LTF & AR

IVIP

- Articulation rate (AR) playing some role for SSBE3 and York
- Further experiments needed to investigate AR re sample duration (3 sec samples here)



Discussion and further work

IVIP

- f0 playing a key role
  - most important feature for each accent (except group 1 in SSBE)
- Long-term formants also playing a role, correlating with higher dimensions for each accent in different ways
- Some role for AR in SSBE and York – more data needed
- These results are for listeners from Britain broadly
- Also need to investigate judgements of mixed-accent groups
- Also Linda Gerlach's PhD research on the relationship between human-judged and ASR-assessed similarity of voices…. (15.30 today!)

See IVIP website for updates

https://www.phonetics.mmll.cam.ac.uk/ivip/

UKRI Economic and Social Research Council

References

- Baumann, O., and P. Belin. 2010. 'Perceptual scaling of voice identity: common dimensions for different vowels and speakers', *Psychological Research*, 74: 110–20.
- Gerlach, L., K. McDougall, F. Kelly, and A. Alexander. 2021. 'How do automatic speaker recognition systems 'perceive' voice similarity? Further exploration of the relationship between human and machine voice similarity ratings.' Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference (online), Marburg, 22-25 August 2021.
- Gold, E., S. Ross, and K. Earnshaw. 2018. 'The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework.' In *Proceedings of Interspeech 2018*, Hyderabad. 2748-52.
- Jessen, M. 2007. 'Forensic reference data on articulation rate in German', *Science and Justice*, 47: 50-67.
- McDougall, K. 2013. 'Earwitness evidence and the question of voice similarity', *British Academy Review*, 21: 18-21.
- McDougall, K., M. Duckworth, and T. Hudson. 2015. 'Individual and group variation in disfluency features: a cross-accent investigation.' In *Proceedings of the 18th International Congress of Phonetic Sciences*, edited by The Scottish Consortium for ICPhS 2015, Glasgow, Paper number 0308.1-5. http://www.icphs.info/pdfs/Papers/ICPHS08.pdf Glasgow: University of Glasgow.
- K. McDougall, T. Hudson and N. Atkinson. 2016. 'An investigation of the effect of listeners' accent background on judgments of voice similarity.' Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, York, 24-27 July 2016.
- Nolan, F., K. McDougall, G. de Jong, and T. Hudson. 2009. 'The *DyViS* database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *International Journal of Speech, Language and the Law*, 16: 31-57.
- Nolan, F., K. McDougall, G. de Jong, and T. Hudson. 2011. "Some acoustic correlates of perceived (dis)similarity between same-accent voices." In *Proceedings of the 17th International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 1506-09. http://www.icphs2011.hk/resources/OnlineProceedings/RegularSession/Nolan/Nolan.pdf. Hong Kong: City University of Hong Kong.
- Remez, R.E., J.M. Fellowes, and D.S. Nagel. 2007. 'On the perception of similarity among talkers', *Journal of the Acoustical Society of America*, 122: 3688-96.
- Sjölander, K. 1997. The Snack sound toolkit. [Computer program]. http://www.speech.kth.se/snack/.