www.phonetics.mmll.cam.ac.uk/ivip
@ivipproject
kem37|aep58|fjn1@cam.ac.uk

The impact of duration of speech sample on listeners' judgements of voice similarity

Kirsty McDougall, Alice Paver and Francis Nolan

Theoretical and Applied Linguistics Section, University of Cambridge

UKRI Economic and Social Research Council

IVIP

1. Background

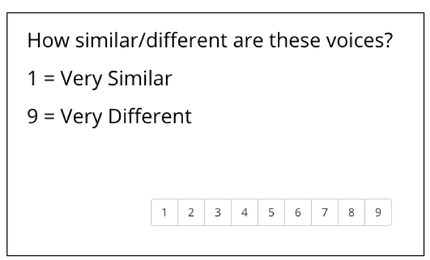
- Listeners perceive some speakers as sounding more similar than others, yet the phonetic underpinnings of perceived voice similarity (PVS) are not well understood
- Improved understanding of PVS will aid foil selection for voice identification parades
- Which phonetic features contribute to voice similarity judgements and does this vary across accent?

2. Previous Study: McDougall (in press) [1]

- Listeners (N = 120) judged the similarity of 120 pairs of voices on a 9-point Likert scale
- 6 groups of 15 speakers (male, 18-30 years):

3	SSBE	*DyViS* [2]
1	York English	*YorViS* [3]
1	Bradford English	*WYRED* [4]
1	Wakefield English	*WYRED* [4]

- 3s stimuli, spont. speech from phonecall
- Data collection online through Gorilla



- 20 listeners per speaker group
- Multidimensional scaling (MDS) applied to judgements to characterize speakers in each group in a pseudo-perceptual space
- Correlations between MDS dimensions and mean f0, long-term formants (LTF) F1-F4 and articulation rate (AR) calculated

➢ Variable results across the 6 speaker groups, even for 3 same-accent SSBE groups:

Feature	Significant for
f0	York, Bradford, Wakefield, SSBE x 2
LTF	All groups but variable
AR	York, SSBE x 1 (higher dims only)

- Were the 3-second stimuli too brief to establish long-term phonetic properties for listeners making judgements?

3. Present Study: Follow-up Experiment

- Experiment repeated on SSBE group 2 with 10s stimuli, to compare against 3s results
- Longer task so 120 stimuli pairs divided into 4 blocks (80 listeners, 20 per block)
- Listener judgements normalized and averaged to form combined sets
- MDS applied to new 3s & 10s matrices

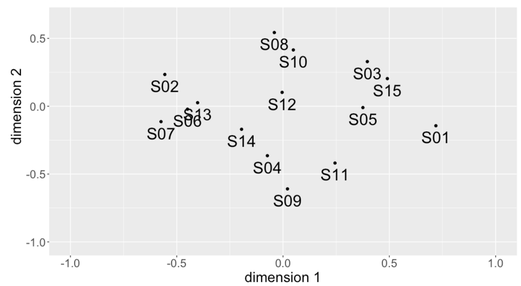


Figure 1. Scatterplot of first 2 MDS dimensions (of 5) for 10s stimuli showing the 15 speakers' locations in pseudo-perceptual space

4. Acoustic Analysis

- Mean f0 (semitones)
- LTF F1-F4 in ERB, *Praat* tracker
- F2-F1
- F2′ [5]
- Articulation Rate (AR): number of phonetic syllables per second [6]

5. Results

Figure 2. Scatterplot of judgements of 3s vs 10s stimulus pairs. Each datapoint represents the mean across the listeners' judgements for one stimulus pair (Likert scale normalised within each listener's responses).

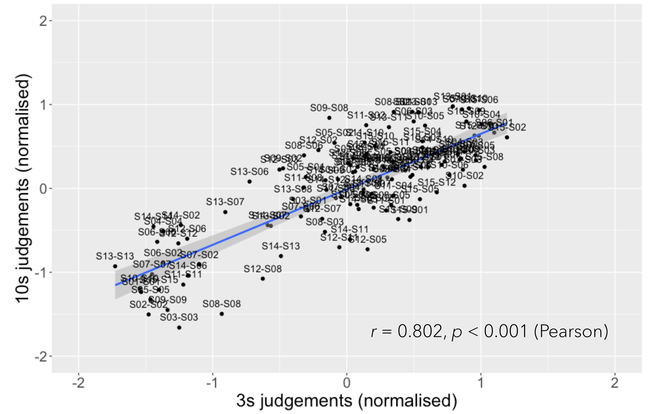
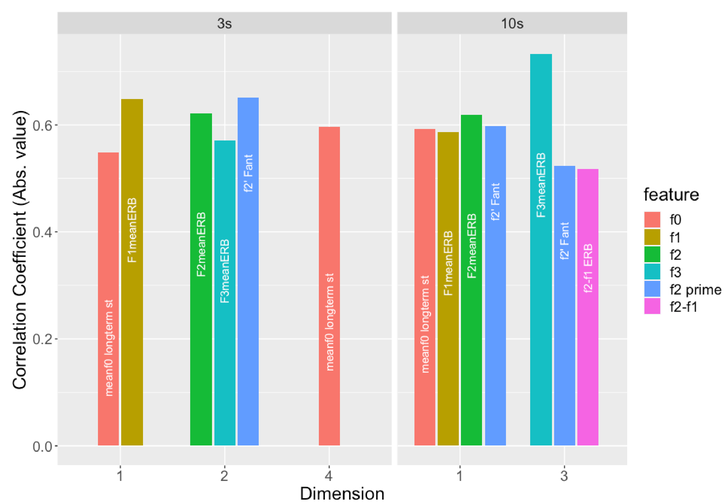


Figure 3. Significant correlations (absolute value) between the acoustic features tested and the five pseudo-perceptual dimensions generated by the listeners' voice similarity judgements by the MDS analysis for 3s stimuli (left panel) and 10s stimuli (right panel).



6. Discussion

- f0: key role for PVS in 3s & 10s
- LTF relatively consistent correl. patterns in higher dimensions
- AR: no sig. correl. for 3s or 10s
- PVS quite variable; new analysis does not highlight importance of different phonetic features when longer (10s) stimuli used
- Short samples (3s) seem to be sufficient for listeners to get hold of speakers' characteristics, at least within the same accent – further investigation needed re different-accent
- Proceed with 3s samples for MDS 'fairness' check in voice parade construction

7. References

1. McDougall, K. (in press) Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. *Proc. XVII AISV (Associazione Italiana Scienze della Voce) Conference: 'Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications'*, 4-5 February 2021, University of Zürich.
2. Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009). The *DyViS* database: Style-controlled recordings of 100 homogeneous speakers for forensic research. *International Journal of Speech, Language, and the Law*, 16(1), 31-57.
3. McDougall, K., Duckworth, M. & Hudson, T. (2015). Individual and group variation in disfluency features: A cross-accent investigation. *ICPhS 2015. Proc. 18th International Congress of Phonetic Sciences*. Paper number 0308.1-5.
4. Gold, E., Ross, S. & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, September 2018*, 2748-2752.
5. Bladon, R.A.W. & Fant, G. (1978) A two-formant model and the cardinal vowels, *Speech Transmission Laboratory Quarterly Progress Report*, Stockholm, KTH, Jan. 1-8.
6. Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice*, 47, 50-67.

ESRC Grant Reference: ES/S015965/1

UNIVERSITY OF CAMBRIDGE