SIMILAR AND DIFFERENT TONGUE SURFACE CONTOURS: INTRA-SPEAKER CONTROLS IN ULTRASOUND ANALYSIS

James M Scobbie

Queen Margaret University, Edinburgh, Scotland
jscobbie@qmu.ac.uk

ABSTRACT

Ultrasound studies of speech production analyse differences in dependent variables reflecting the tongue surface's location and shape. Inferential statistics distinguish theoretically-relevant from random effects, somewhat independently of the descriptive size of significant effects. Experimental designs induce measurable dependent changes by manipulating independent variables such as prosody, phonemic target, etc.

This paper presents descriptive statistics quantifying holistically all 15 pairwise differences between six monophthongal long vowel phonemes of one variety of English, comparing these to experimental noise differences attributable to the use of two identical blocks of data collection in sequence. Eight speakers were recorded, using two different ultrasound systems, and analysed in AAA using both edge-tracking and DeepLabCut pose estimation. The smallest phonemic contrast (~2mm) was greater than the experimental noise (~1mm), and was well-evidenced by AAA's t-test of radial difference.

Keywords: methods, ultrasound, articulation, experimental noise.

1. INTRODUCTION

Ultrasound scanning is a safe, accessible, non-intrusive and quick way of imaging the tongue during speech. Ultrasound (US) images are analysed manually or automatically to extract key features relevant to phonetic analysis, typically to derive a mid-sagittal tongue-surface contour. Experimental protocols enable analyses of the properties of these surface contours. In particular, comparisons of two or more groups of contours may, if the between-groups contours can be shown to be different, provide evidence for rejecting a null hypothesis.

There are a number of techniques for comparing two or more groups of tongue-surface contours sampled at single time points (or averaged from multiple time-points). These can be extended to the analysis of dynamic phenomena using time-series, but the focus here will be static configurations. In the analysis of untransformed surface contours, various statistical techniques can determine whether two groups of tongue surface contours are different or not. Popular approaches include SSANOVA e.g. [1] or [2], GAMMs e.g. [3] or [4], or the t-test difference function built into the data capture and analysis software package AAA e.g. [5], [6]. Global properties of contours (topological properties) can also be used, as well as location.

If a localised difference between two groups of contours reaches statistical significance, this does not imply that all aspects of the contours differ. So long as the confidence intervals flanking two mean contours are non-overlapping (for example), a "zone of difference" can be said to have been found. But how small can these zones be? While a threshold for confidence intervals at 95% is widely adopted, there is no general agreement on the minimum length of non-overlap needed to define such a zone of difference. One specific proposal [7] was to classify as significant only zones spanning at least six of AAA's 42 radii in a 135° field of view. This would be about 3cm of tongue surface, and appears rather an arbitrary threshold.

The distance between contours also matters, whether it be a single maximum value, averaged over a zone of statistical difference, or characteristic of some sub-part of the contour [7], [8]. Small distances can be critical in speech production, but these descriptive measures also inform statistical testing. Can a distance between two similar contours of just 1mm along 1cm of contour length be meaningful? Yes, if it is due to the speaker distinguishing between complete closure and close approximation. Probably not, if the difference is between two vowels. What if a mere 1mm difference was found in a midsagittal zone 5cm long?

When it comes to analysing the possible difference between two similar tongue contours, we should provide basic descriptive information, statistically-significant *or not*. We need to ask:

- How far apart are the contours (in a zone of interest)?
- What is the length of any zone of significant difference?

For such information to benefit future statistical modelling, we need descriptive figures on meaningless variation: i.e. *control* data. Experimental noise arises from issues such as instrumental accuracy and

precision, analytic variability, and speaker variation due to (task) effects such as fatigue, confidence, or implicit learning. These can be somewhat addressed by counter-balanced experimental designs, but not always, particularly when a baseline is required.

For concreteness, we have constructed a control study for a specific experiment [9], but the real goal is expository. We aim to exemplify the value of *a priori* meaningless (control) and meaningful descriptive findings, as a general principle. We will quantify: (a) a meaningless task non-difference between identical datasets presented in two blocks; (b) axiomatic contrasts between six vowel phonemes via holistic pairwise comparison; (c) two different ultrasound systems; and (d) two descriptive measures of difference, namely average radial distance difference and mean nearest neighbour distance difference.

2. METHOD

2.1. Protocol

Our protocol is intended to be compatible with a wide range of experimental studies and designs. It was based on studies in which several phonemes are elicited from a single speaker in two distinct prosodies, frames, languages, or affects, with a short rest between blocks. A specific model, [9], was chosen, so that this study is directly a control for [9]; but it also provides norms for Scottish English.

We used speaker-specific randomisations of our materials (see 2.2). There were no fillers or distractors. The materials were 12 words (or pseudowords) that were repeated three times in a block. Each block of 36 randomised words was presented to the speaker twice, with a short pause between blocks (one to two minutes). It was used as a brief rest, for small-talk, and to reset the computer. Speakers drank a little water, looked around, talked to the experimenter, and moved in their chair. They were aware this was the mid-point of the experiment.

2.2. Materials and speakers

The Scottish English materials were tightly structured, following [9]. In Scottish English, there are six "unchecked" or long monophthongal vowels /ieaɔoʉ/ with lexical set incidence: /i/ FLEECE, /e/ FACE, /a/ TRAP & BATH & PALM, /ɔ/ LOT & THOUGHT, /o/ GOAT, /ʉ/ GOOSE & FOOT. /ɔoʉ/ are phonologically rounded. Two C_C contexts were used, in which both C were labial (either /m/ or /p/) so as not to affect lingual behaviour. Thus the wordlist mostly included real words but also pseudowords. There was no carrier phrase.

- peep, pape, pap, pop, pope, poop
- meme, maim, mam, mom, moam, moom

The speakers were judged to have typical Scottish accents. Speakers c1-c4 were undergraduate students, some familiar with phonetics. Speaker c5 was the author, c6 was a colleague of the author, and c7 & c8 were family members of the author. Only c5 was fully aware of the purpose of the study.

2.3. Radial measurement

A typical vowel system for Scottish English is illustrated in Fig. 1, showing mean tongue surface contours labelled for the six vowels. Each mean was calculated from the radius crossing points of each of the 12 tokens, within AAA's fan-shaped analysis grid (not shown). Means are flanked by ± 1 standard deviation (along radii). Two key radii are shown, however: fanlines fn13 and fn33 are the anterior and posterior limits respectively for a common analysis sector of this speaker (c8), containing all contiguous radii in which edge-detection confidence in AAA was above a threshold of 80% for all vowels. For each of the 21 radii within this sector the 15 pairwise differences between each of the six vowels was measured (N=12 for each vowel). The *average* radial difference in distance from the origin is a holistic measure, irrespective of zones of statistical differences between vowels, but trimmed as described.

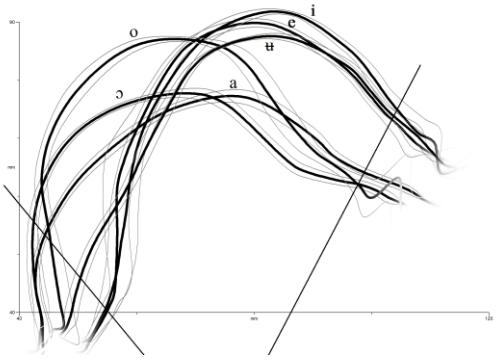

Figure 1: Typical Scottish English vowel space (c8), anterior to right, with a common analysis sector, showing the overall similarity of /i/, /e/ and rounded /ʉ/

Vowel-specific sectors (trimming off all surface contours <80% confidence) were used for within-vowel radial t-testing for Block 1 vs. Block 2 (N=6 for each vowel in each block). The number of contiguous radial fanlines with significant differences was noted.

2.4. Ultrasound system 1

The system used for speakers c1-c4 was similar to that used in [9],[10], comprising an Ultrasonix RP scanner, with an aluminium stabilisation headset [11], and a 5MHz 10mm radius Microconvex probe. AAA software [5] (version 2.19) was set to a 135°

field of view, at 121 frames per second, depth 80mm, with 63 scanlines. Tongue surface contours were created semi-automatically using AAA's edge-tracking procedures, on a fan-shaped grid (as above).

2.5. Ultrasound system 2

The system comprised a Telemed micro system scanner, and an aluminium stabilisation headset [11], with a 2-4MHz 20mm radius convex probe. AAA software [6] (version 220) was set to a 101° field of view, at 81 frames per second, depth 80mm, with 64 scanlines. It was used for speakers c5-c8.

AAA's innovative pose estimation [6], [12] using DeepLabCut (DLC) automatically created tongue surface contours, and AAA's nearest neighbour (NN) function was used to calculate both B1 vs. B2 differences and the pairwise vowel phoneme differences. We then converted the DLC contours to fan-based contours, enabling comparative radial differences to be calculated within AAA.

3. RESULTS, SYSTEM 1

The size of the sectors analysed (in fanlines) was minimally 12 and maximally 24 radii (mean 21, median 23). The locations of the analysis sectors varied, with the most retracted vowels /ɔ/ and /o/ also tending to have the smallest analysis sectors. Typically, the mean radial difference between blocks was below 1mm (Table 1). Only one vowel from one speaker (/i/, from c3) reached a threshold for having a zone of 6 contiguous radii where there was a significant difference in the radial distances.

Table 1: Size of noise (B1 vs. B2) as mean radial difference in distance from fan origin (mm). An asterisk indicates a zone of significant difference.

	c1	c2	c3	c4
i	0.7	0.9	*1.5	0.5
e	0.6	1.0	1.0	0.5
a	0.2	0.6	0.8	0.3
ɔ	0.3	0.7	0.9	0.4
o	1.1	0.9	0.4	1.1
ʉ	0.4	0.4	0.9	0.7

Speaker c3's /i/ vowel appeared to be higher and more anterior in the second block (Fig. 2). In Fig. 2, 12 thick radial lines projecting inwards from the arc indicate both the location of these radii with significant differences (p-value from 0.001 to 0.01) and their sizes (from 0.91mm, fn17, to 2.49mm, fn8). The most anterior radius used in the analysis was fn6 (the rightmost shown). The mean radial distance between these vowels in the zone of significance was 1.83mm, over a midsagittal tongue surface length of 42mm. The total length of confidently-traced tongue surface analysed, spanning from fn6 to fn29 (for c3's /i/), was 77mm. The average radial difference overall was just 1.5mm (Table 1).

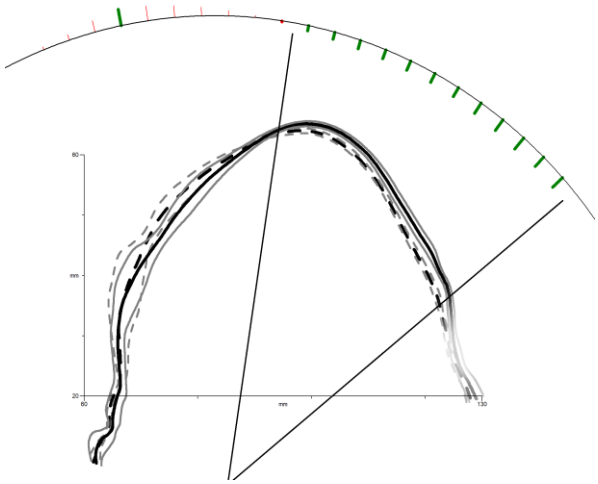

Figure 2: Two significantly different mean-radial tongue curves (speaker c3, /i/), each flanked by ± 1 standard deviation. Dashed /i/ from block 1, solid /i/ from block 2.

Table 2: Pairwise mean radial differences between phonemes (mm). All pairs had some zone of significance (length and width not reported).

	c1	c2	c3	c4	mean
ie	1.4	2.1	2.1	1.2	1.7
ia	12.2	13.2	10.6	11.4	11.8
iɔ	15.4	17.6	13.9	14.3	15.3
io	14.8	16.3	10.8	12.7	13.7
iʉ	5.1	7.3	4.7	3.6	5.2
ea	11.0	11.8	8.4	10.2	10.3
eɔ	14.1	16.2	11.7	13.1	13.8
eo	13.4	14.8	8.6	11.5	12.1
eʉ	3.8	5.6	2.6	2.4	3.6
aɔ	3.2	4.5	3.7	3.4	3.7
ao	4.1	4.5	4.0	5.3	4.5
aʉ	7.3	6.5	6.0	8.2	7.0
ɔo	2.8	2.8	4.0	3.4	3.3
ɔʉ	10.3	10.9	9.4	11.3	10.5
oʉ	9.7	9.3	7.2	10.0	9.0

As well as lacking statistical significance, most of the noise differences reported in Table 1 are tiny in absolute terms. Yet some are close to the smallest pairwise phoneme differences in Table 2, e.g. /i/ vs. /e/. Fronted /ʉ/ (GOOSE & FOOT) has a lingual shape and location very similar to /e/, while /ɔ/ was similar to /o/. All the pairwise differences had zones of significant difference, and though not reported

here, they were large (cf. Fig. 1). The maximum radial pairwise differences and mean difference just within zones of significance were greater than the overall mean pairwise differences in Table 2.

4. RESULTS, SYSTEM 2

The DLC pose estimation surface contours showed greater differences when quantified with AAA's nearest neighbour tool (Table 3 vs. Table 1).

Table 3: Size of noise (B1 vs. B2) as DLC mean nearest neighbour distance (mm).

	c5	c6	c7	c8
i	1.5	0.8	2.3	1.8
e	1.4	1.0	2.0	1.9
a	1.5	1.6	2.2	1.2
ɔ	2.1	2.4	1.7	1.3
o	2.0	1.7	2.6	1.9
ʉ	3.0	1.5	2.5	1.5

For a meaningful comparison, DLC tongue surface contours were therefore converted to contours on a fan grid. Significant difference could be tested. More zones of significance were found (Table 4) for inter-block noise for c5-c8 using system 2) than for c1-c4 using system 1.

Table 4: Size of noise (B1 vs. B2) as mean radial difference in distance from fan origin (mm).

	c5	c6	c7	c8
i	1.4	0.4	1.4	1.0
e	1.1	0.5	1.5	1.7
a	1.1	1.3	*1.9	*0.7
ɔ	*1.4	0.9	1.1	*1.0
o	1.3	1.1	1.4	*1.6
ʉ	2.6	0.9	1.1	1.0

Significant experimental noise was seen in a zone spanning 12 contiguous radii for c7's /a/ (mean 3mm difference). There were two zones of 6 radii for c5's /ɔ/ (together, 2mm). Significant zones for c8 spanned only 6 contiguous radii, and averaged just 1.2mm for /ɔ/, 1.8mm for /a/ and 2.1mm for /o/.

Pairwise matrix analysis (Tables 5 & 6) showed a similar rank order for radial and nearest neighbour measures. As seen in Table 2, the varied sizes of mid-sagittal contrasts are reflected intuitively well.

5. DISCUSSION AND CONCLUSION

Experimental (task) noise was occasionally statistically significant, and any inter-block differences could have been due to changes in speech production or instrumental instability. Combining all results, the small but significant phoneme radial difference of 1.7mm (/i/ vs. /e/) appears appreciably more than the rarely-significant noise in /i/ (0.9mm) and /e/ (1.0mm). This suggests small yet significant results in experiments lacking counter-balancing (e.g. Table 3 in [9]) should be treated with caution, given our current levels of statistical understanding. More control studies, involving a wide range of experimental protocols and systems, will provide a better context for inductive statistical analysis.

Table 5: Pairwise nearest neighbour differences between DLC curves (mm).

	c5	c6	c7	c8	mean
ie	1.8	2.6	2.1	2.3	2.2
ia	10.7	6.5	11.8	10.7	9.9
iɔ	16.3	13.4	13.7	13.5	14.2
io	16.6	12.5	11.3	12.4	13.2
iʉ	8.1	6.2	6.3	2.8	5.8
ea	9.5	4.7	11.1	9.9	8.8
eɔ	15.3	11.8	12.9	12.7	13.2
eo	15.6	10.4	10.5	11.0	11.9
eʉ	6.9	3.8	5.6	2.6	4.7
aɔ	6.5	7.6	3.2	3.7	5.3
ao	7.2	6.5	5.4	5.5	6.1
aʉ	4.8	3.7	6.8	8.8	6.0
ɔo	3.6	6.2	4.4	4.4	4.7
ɔʉ	9.8	9.2	8.4	11.9	9.8
oʉ	9.3	7.0	5.7	11.1	8.3

Table 6: Pairwise mean radial differences between phonemes (mm).

	c5	c6	c7	c8	mean
ie	1.2	2.5	1.1	2.1	1.7
ia	15.1	8.8	12.4	13.1	12.4
iɔ	22.3	17.3	15.3	17.1	18.0
io	22.8	16.3	13.5	16.8	17.4
iʉ	11.7	6.7	7.0	3.6	7.2
ea	14.0	7.2	11.5	12.3	11.3
eɔ	21.1	15.9	14.3	16.1	16.9
eo	21.6	14.0	12.5	15.2	15.8
eʉ	10.5	4.3	5.9	3.5	6.0
aɔ	7.4	8.7	3.4	4.6	6.0
ao	9.1	7.8	5.9	7.3	7.5
aʉ	6.4	4.9	7.1	9.8	7.1
ɔo	3.6	7.5	4.3	5.7	5.3
ɔʉ	11.7	12.8	8.9	14.1	11.9
oʉ	11.2	9.8	6.6	14.4	10.5

5. ACKNOWLEDGEMENTS

For funding support, thanks to ERC Award 101019847 “PlanArt: Planning the Articulation of Spoken Utterances” and Leverhulme Research Fellowship RF-2021-368 “Case studies in ultrasound tongue imaging”.

6. REFERENCES

[1] Davidson, L. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *J. Acoust. Soc. Am.* 120(1), 407–415.

[2] Mielke, J. 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *J. Acoust. Soc. Am.* 137(5), 2858–2869.

[3] Heyne, M. 2016. *The influence of First Language on playing brass instruments: An ultrasound study of Tongan and New Zealand trombonists.* PhD Thesis, University of Canterbury. New Zealand.

[4] Strycharczuk, P., Derrick, D., Shaw, J. 2020. Locating de-lateralization in the pathway of sound changes affecting coda /l/. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1), 21, 1–27.

[5] Articulate Instruments Ltd. 2012. *Articulate Assistant Advanced user guide: version 2.14.* Edinburgh, UK.

[6] Articulate Instruments Ltd. 2022. *Articulate Assistant Advanced, version 220.* http://www.articulateinstruments.com/downloads/

[7] Cleland, J. Scobbie, J.M., Wrench, A.A. 2015. Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics* 29(8–10), 575–597.

[8] Cleland, J. Scobbie, J.M. 2020. The dorsal differentiation of velar from alveolar stops in typically developing children and children with persistent velar fronting. *Journal of Speech, Language, and Hearing Research* 64(6S), 2347–2362.

[9] Scobbie, J.M., Ma, J. 2019. Say again? Individual articulatory strategies for producing a clearly-spoken minimal pair wordlist. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *19th International Congress of Phonetic Sciences, ICPhS 19, Melbourne,* Australia, August 5-9, 2019, *Proceedings.* Australasian Speech Science and Technology Association Inc., 358–362.

[10] Wrench, A.A., Scobbie, J.M. 2016. Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016). *CASL Working Papers*, WP-24. Queen Margaret University.

[11] Articulate Instruments Ltd. 2008. *Ultrasound Stabilisation Headset Users Manual.* Edinburgh, UK.

[12] Wrench, A., Balch-Tomes, J. 2022. Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut. *Sensors* 22(3), 1133.