Beyond 2D in articulatory data acquisition and analysis

Philip Hoole[1], Andreas Zierdt[1,2] and Christian Geng[2]

[1]Institut für Phonetik und Sprachliche Kommunikation, Munich
[2]Zentrum für Allgemeine Sprachwissenschaft, Berlin

Email: hoole|andi@phonetik.uni-muenchen.de, geng@zas.gwz-berlin.de

Work supported by German Research Council and NTT Basic Research Labs
MRI acquisition in collaboration with Dr. Axel Wismüller, Dept. Radiology, Klinikum Innenstadt, Ludwig-Maximilians-Universität, Munich

This document contains the oral presentation given at ICPhS Barcelona, 2003. The section on 5D EMA has been supplemented with a large number of annotations to hopefully make it easier to follow as a stand-alone document (click on the yellow note symbols).
Links to QuickTime animations, sounds and other documents are highlighted in red. The QuickTime animations can be downloaded all at once in the following ZIP file, and should be stored in the same subdirectory as this document:
www.phonetik.uni-muenchen.de/~hoole/icphs03_anim.zip
(click on adjacent red box to follow this link)

Topics

1. Three-dimensional electromagnetic articulography
    Potential advantages for monitoring speech movements

2. Three-dimensional tongue shape from multi-speaker, multi-volume MRI
    Analysis with 3-way statistical techniques

Linking 1 and 2?

1. Three-dimensional electromagnetic articulography

The magic number seven (plus or) minus two = **five**

Better thought of as a **five**-dimensional system

(Or: The last **five** percent of its development have been a real pain ....)

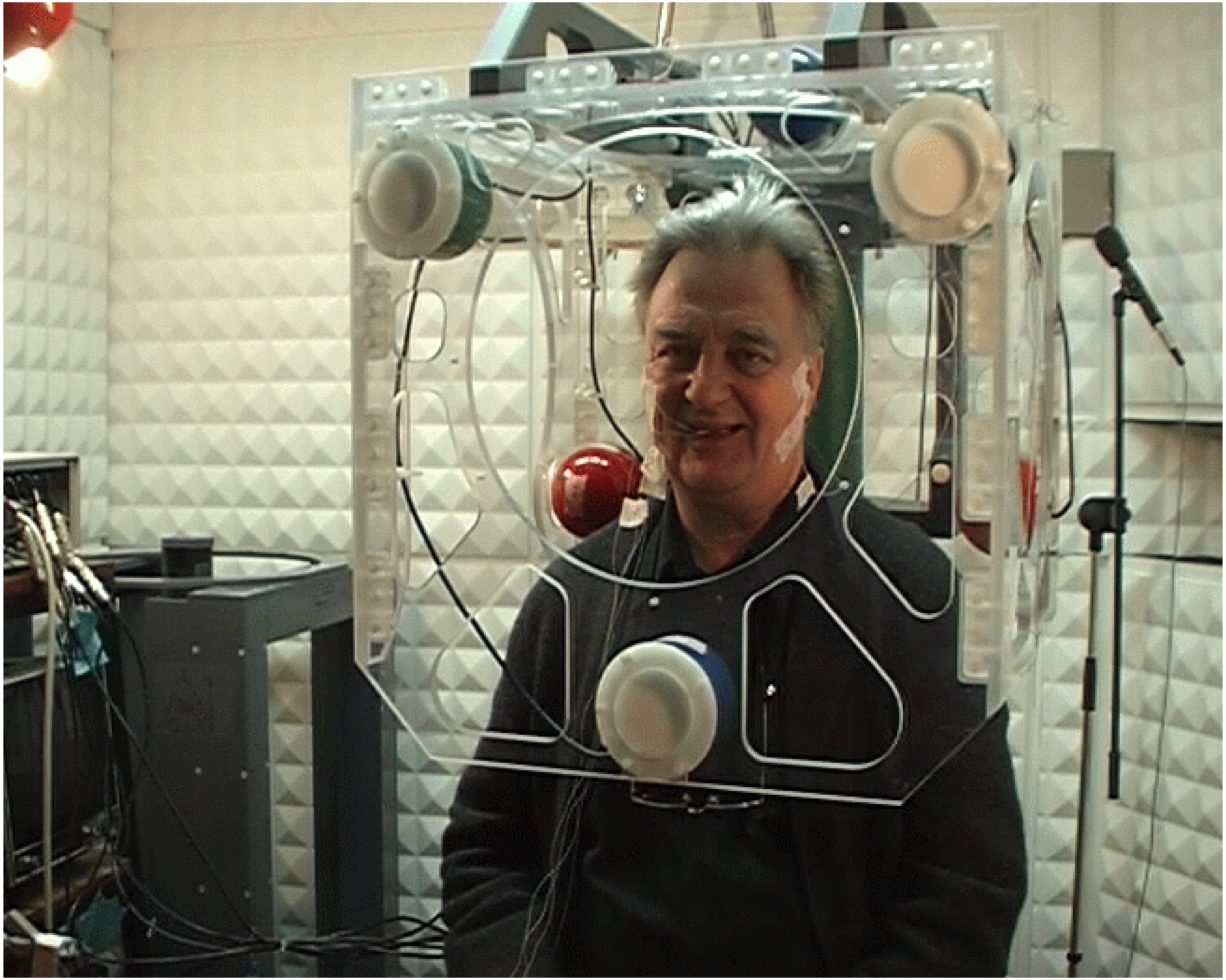
Same sensors as for traditional 2D EMMA

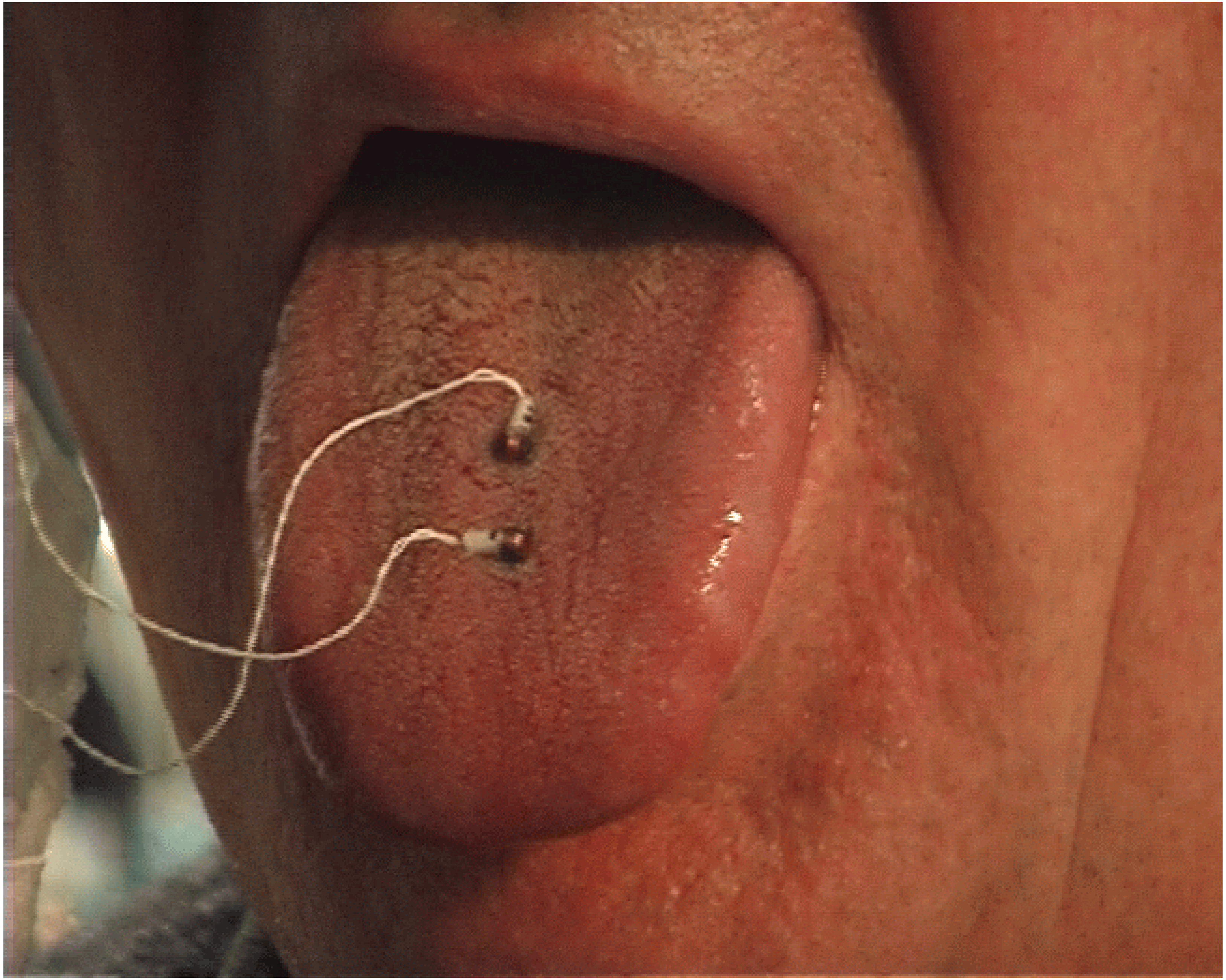
But 6 transmitters

➔ 3 Cartesian (x, y, z) and 2 angular (azimuth, elevation) coordinates for each sensor (1 rotational degree of freedom not accounted for)

The five dimensions: both the benefit and the bane of the new system

The benefit:	Very high information density per sensor
The bane:	Very difficult to demonstrate reliability over the full 5-dimensional space





Assessment of performance under realistic conditions

Do the benefits outweigh the problems?

See Proceedings for comparison of a set of speech tasks recorded consecutively with the 2D and 5D systems.

Examples from more recent recordings:

(1) Head Movement

(2) Reconstruction of tongue shape

Examples (1)
Head Movement

Good example of high information density per sensor:

Two sensors (upper incisors, bridge of nose) sufficient to determine the six rigid-body degrees of freedom of the head ➔ 3 translations, 3 rotations.

The same two sensors in 2D EMMA ➔ 2 translations, 1 rotation

Potential benefits:

- Freedom of head movement for subject ➔ more natural speaking situation
- Communicative relevance of head movement (prosody)

Head movement: Animations

1. With synchronized video

2. Systematic variation of 3 translations and 3 rotations

3. Tongue movement after correction for head movement

   Utterance: **ʃviːʁɪk**

Estimating positional errors during actual movement tasks

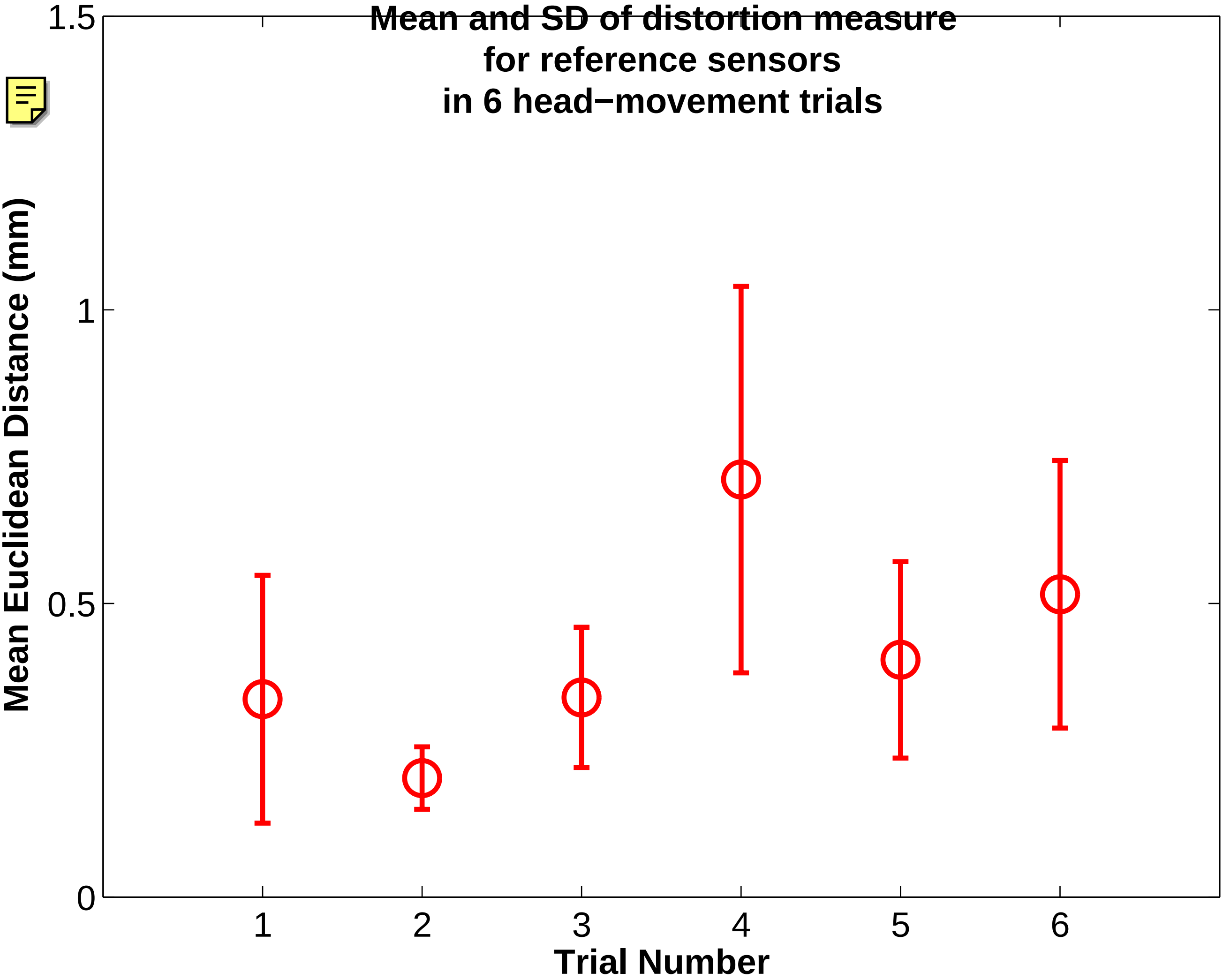
The two reference sensors with their orientations define a rigid structure.

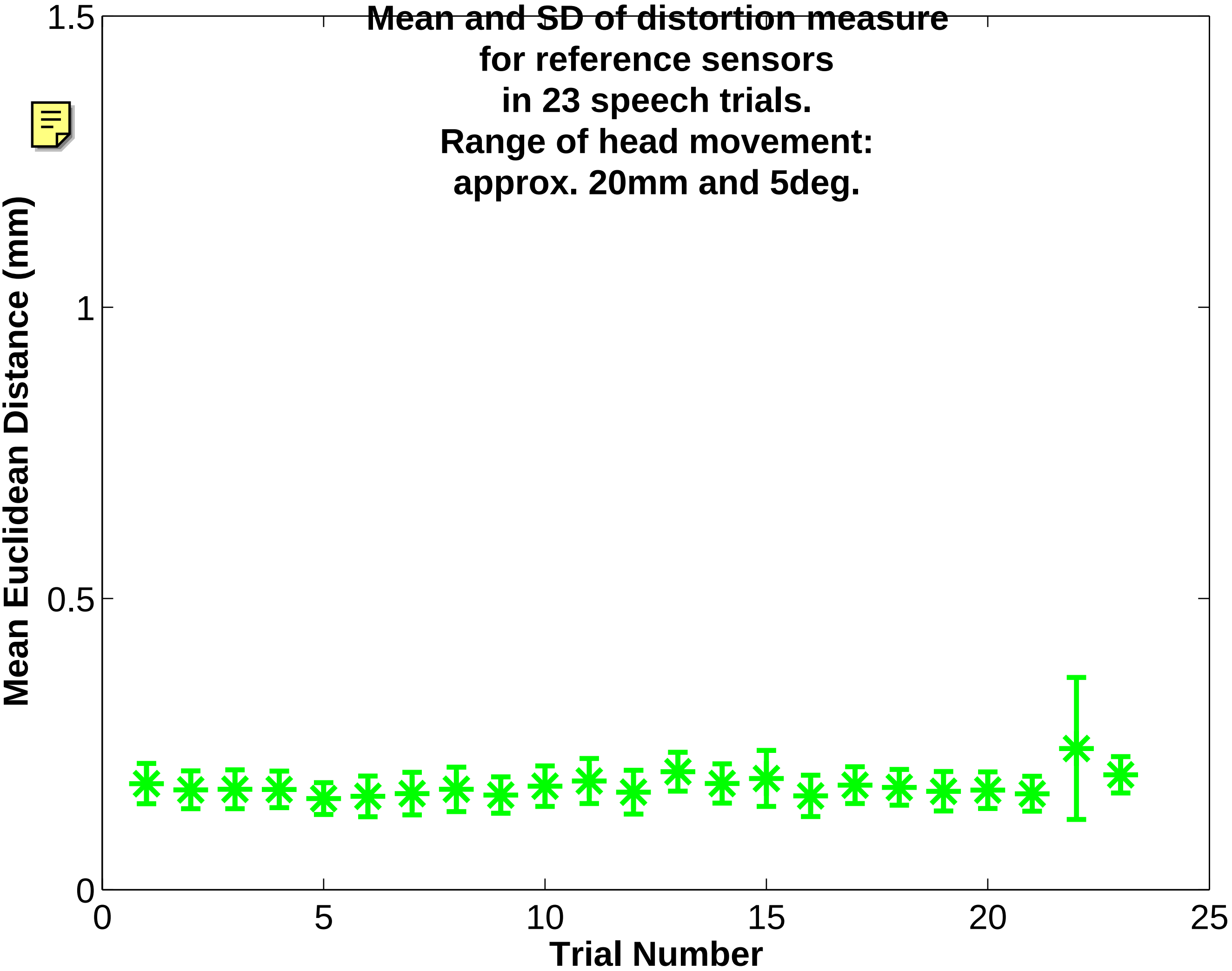
How much is this rigid structure distorted when mapping all head positions to the desired reference position?

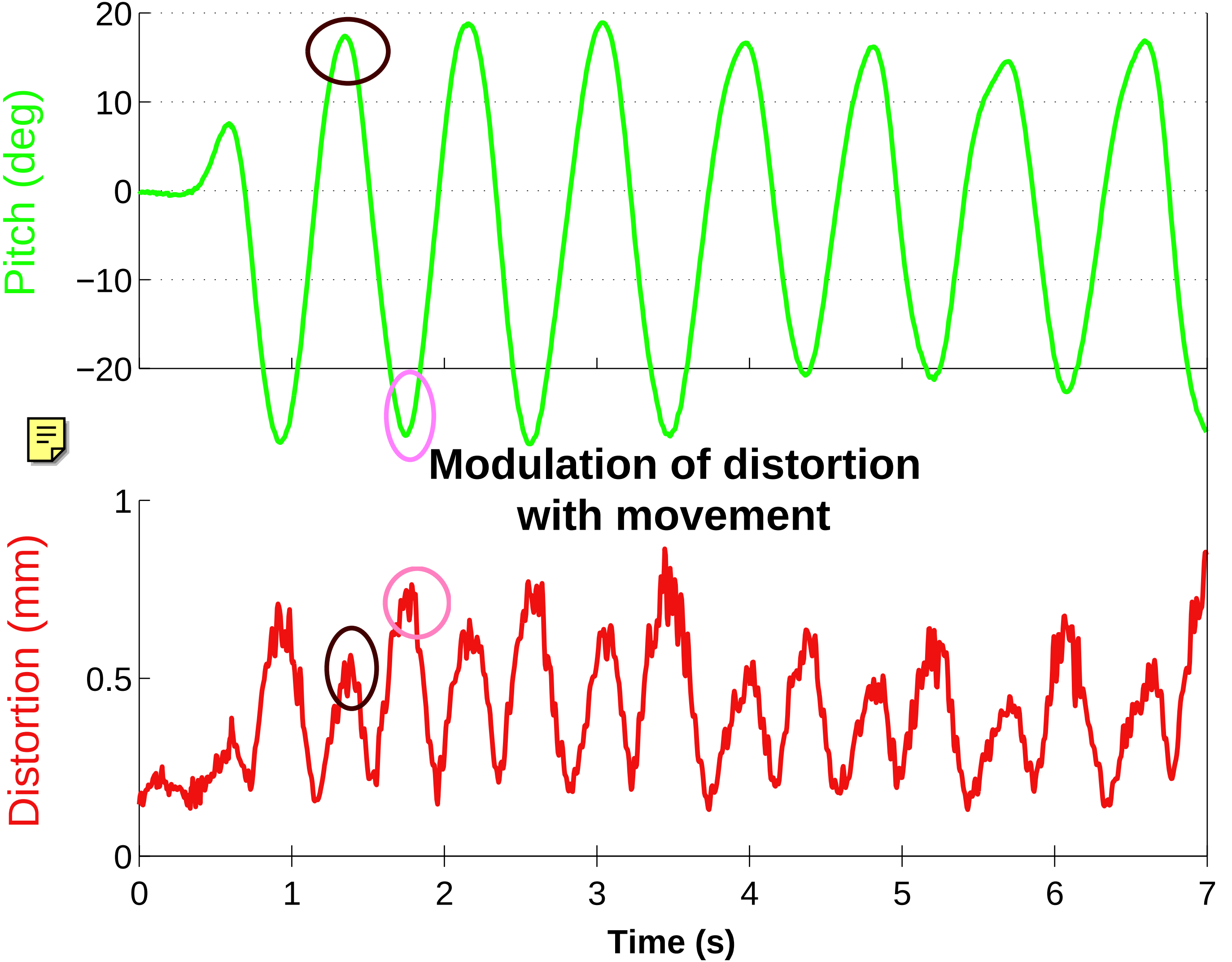
Calculate average Euclidean distance between transformed head position and reference head position.

➔ Distortion measure. Good estimate of relative accuracy.

Corresponds to stability of distance between upper incisor and nose sensor in 2D EMMA


Mean and SD of distortion measure
for reference sensors
in 6 head−movement trials
Mean Euclidean Distance (mm)
1.5
1
0.5
0
1
2
3
4
5
6
Trial Number


Mean and SD of distortion measure
for reference sensors
in 23 speech trials.
Range of head movement:
approx. 20mm and 5deg.
Mean Euclidean Distance (mm)
1.5
1
0.5
0
0
5
10
15
20
25
Trial Number



Modulation of distortion with movement

Examples (2)
Reconstruction of tongue shape

Information on sensor orientation improves estimate of tongue position in “difficult” regions:

tongue tip

tongue root

Tongue movement: Animations

/isi/ and /usu/ slowmotion movie

Static tongue configurations

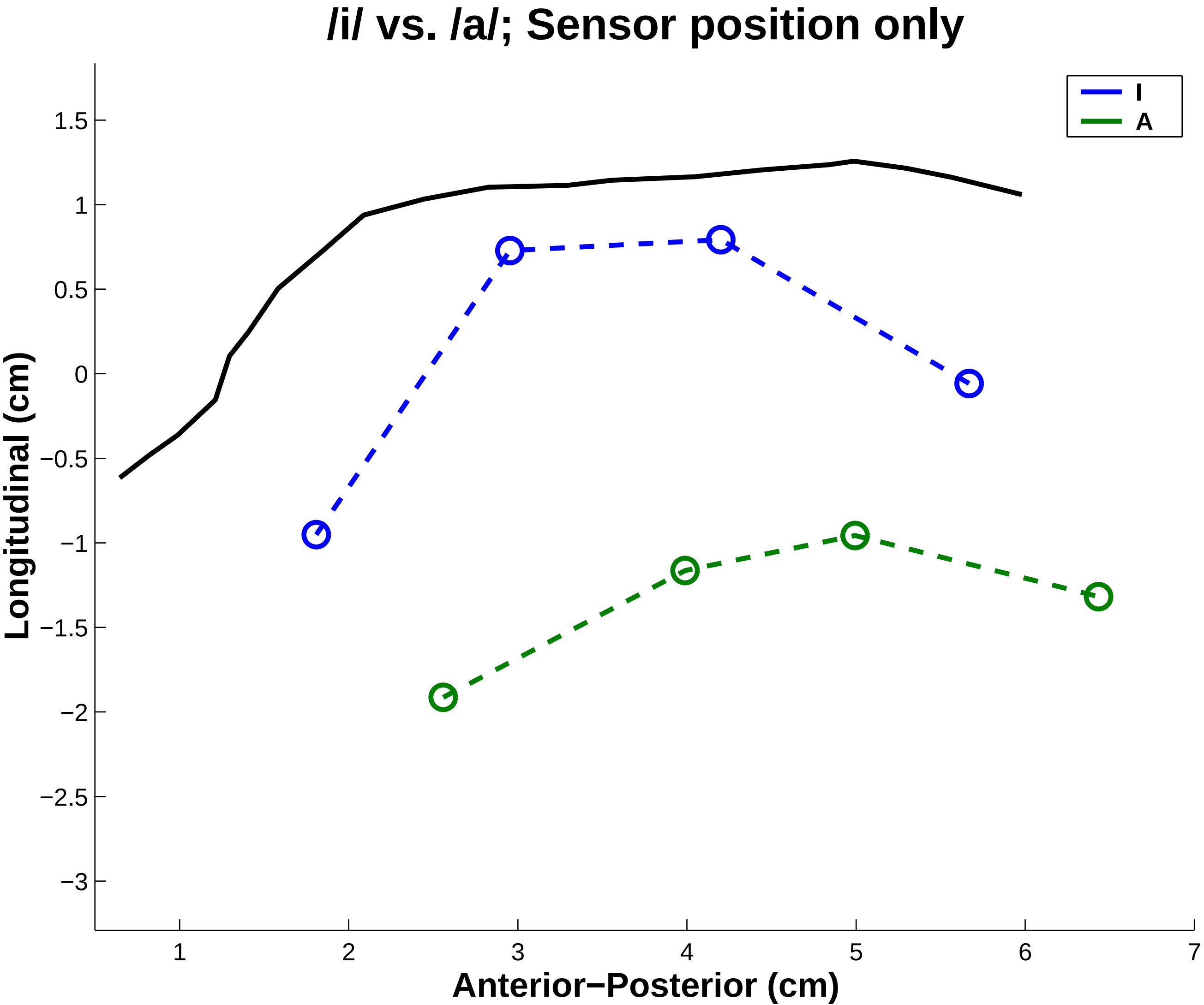
Contrasting

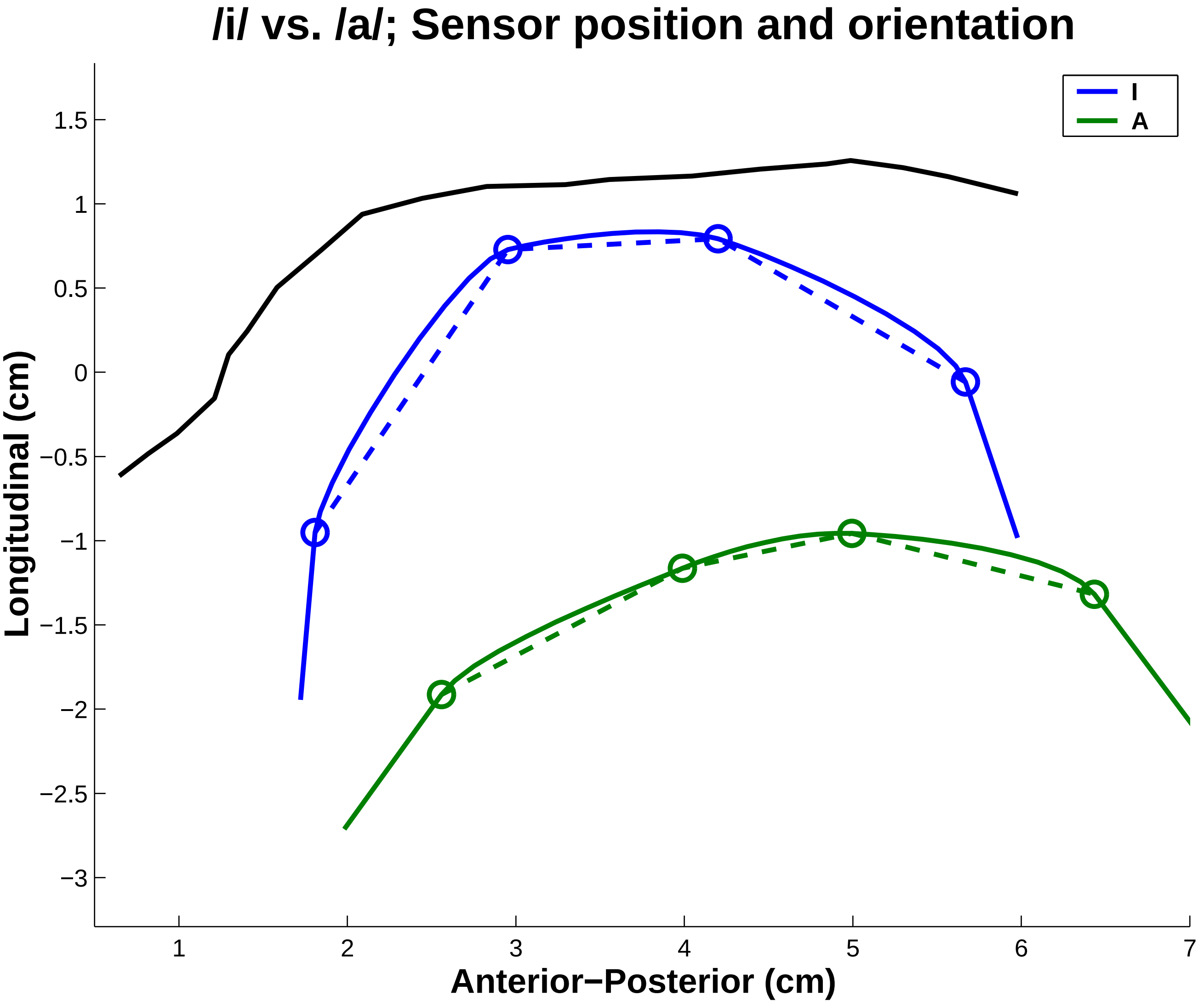
use of purely positional information

vs.

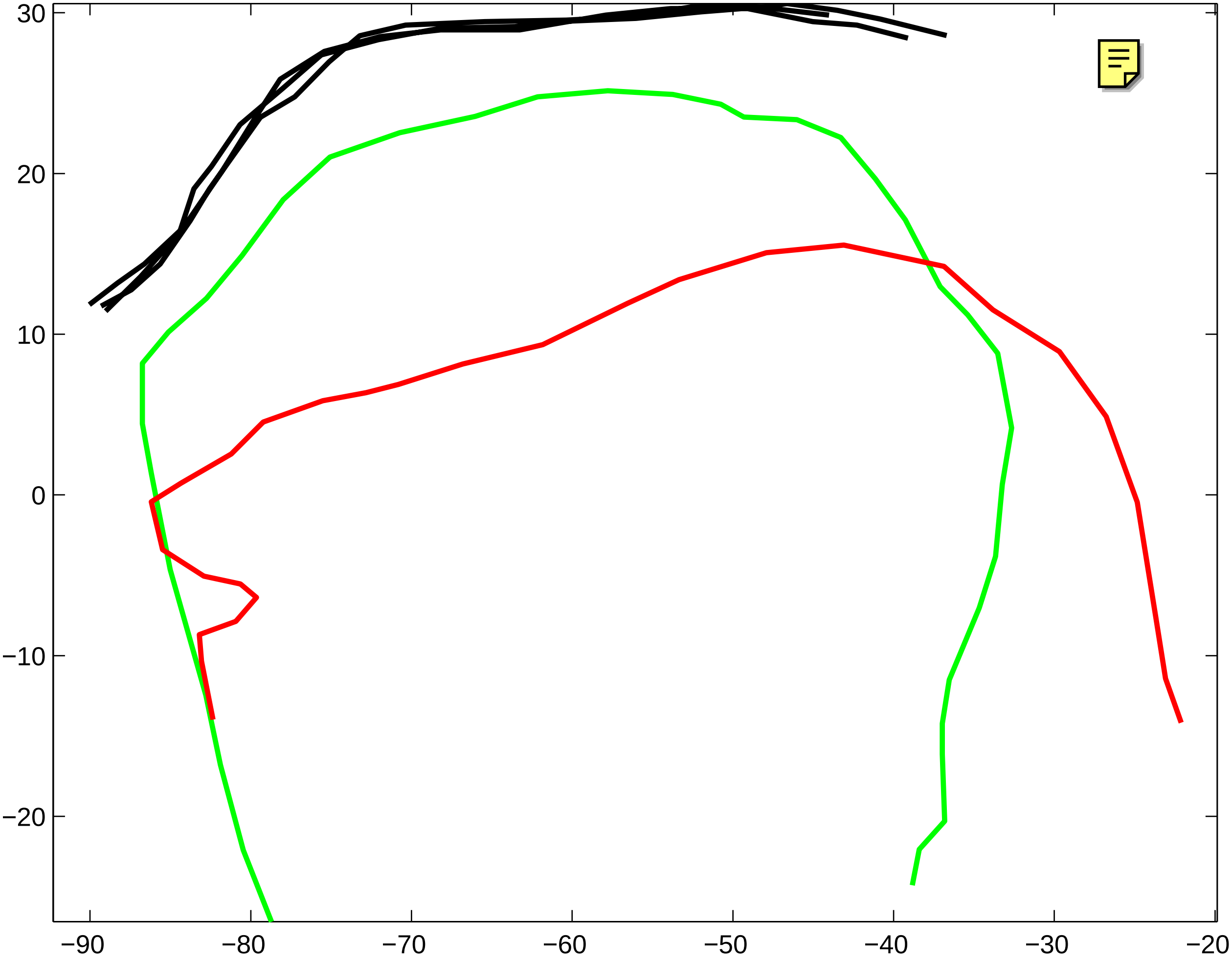
use of both position and orientation information (spline interpolation between sensors and extension of tongue 1cm beyond frontmost and rearmost sensor)

(1) Vowels


/i/ vs. /a/; Sensor position only
Longitudinal (cm)
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)
I
A


/i/ vs. /a/; Sensor position and orientation
I
A
Longitudinal (cm)
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)

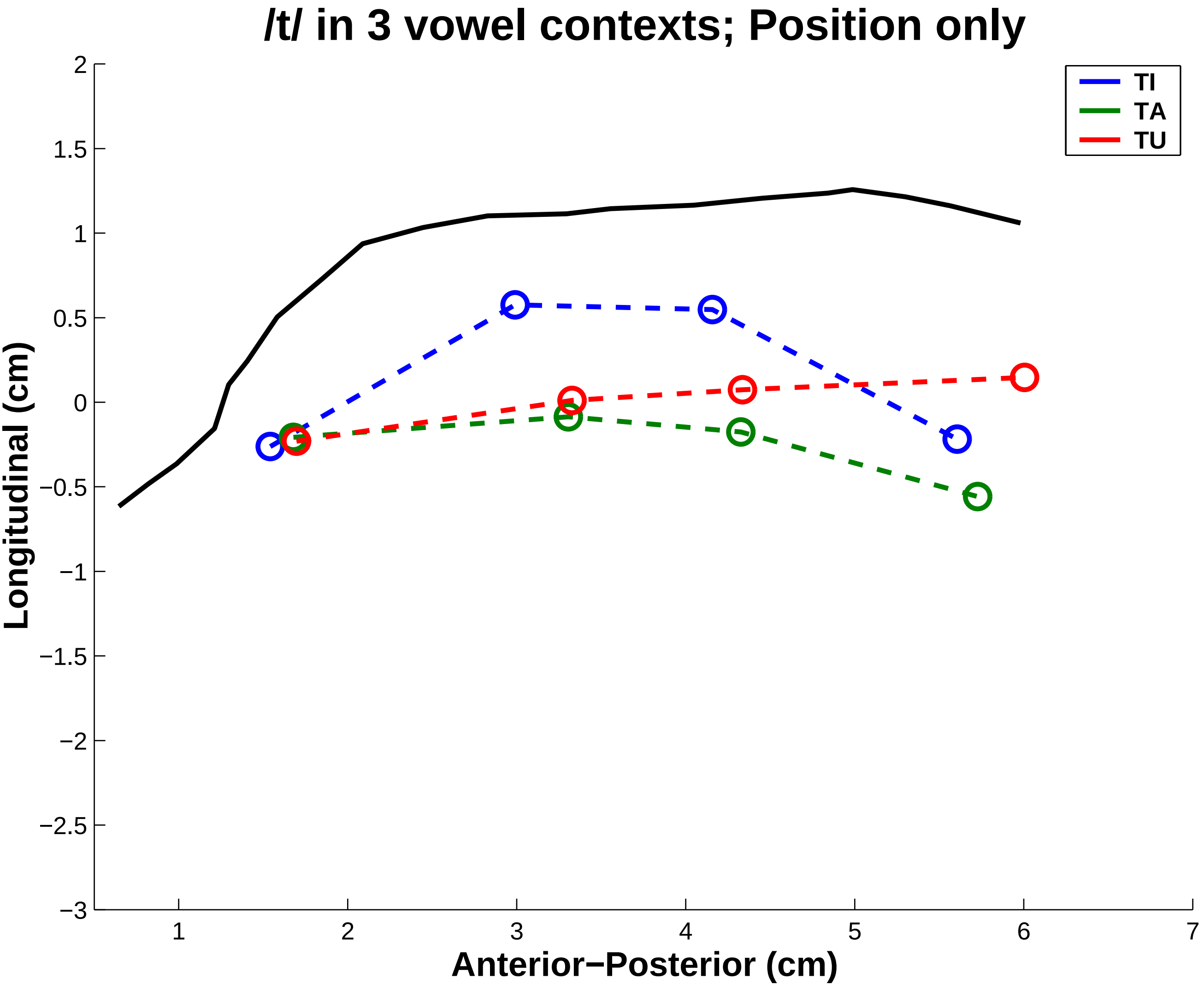
Speaker HP, Midsagittal /i, a/ from MRI



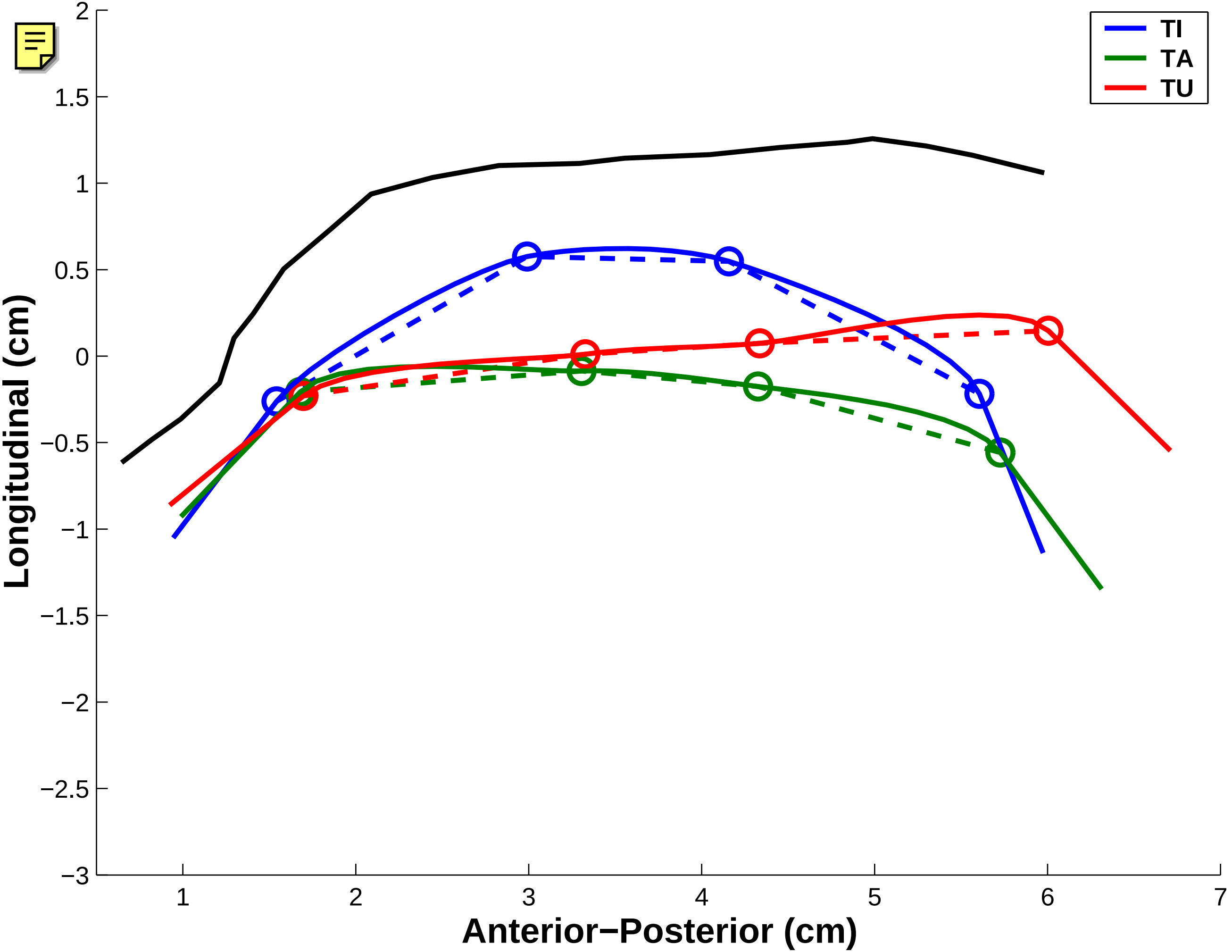
Static tongue configurations

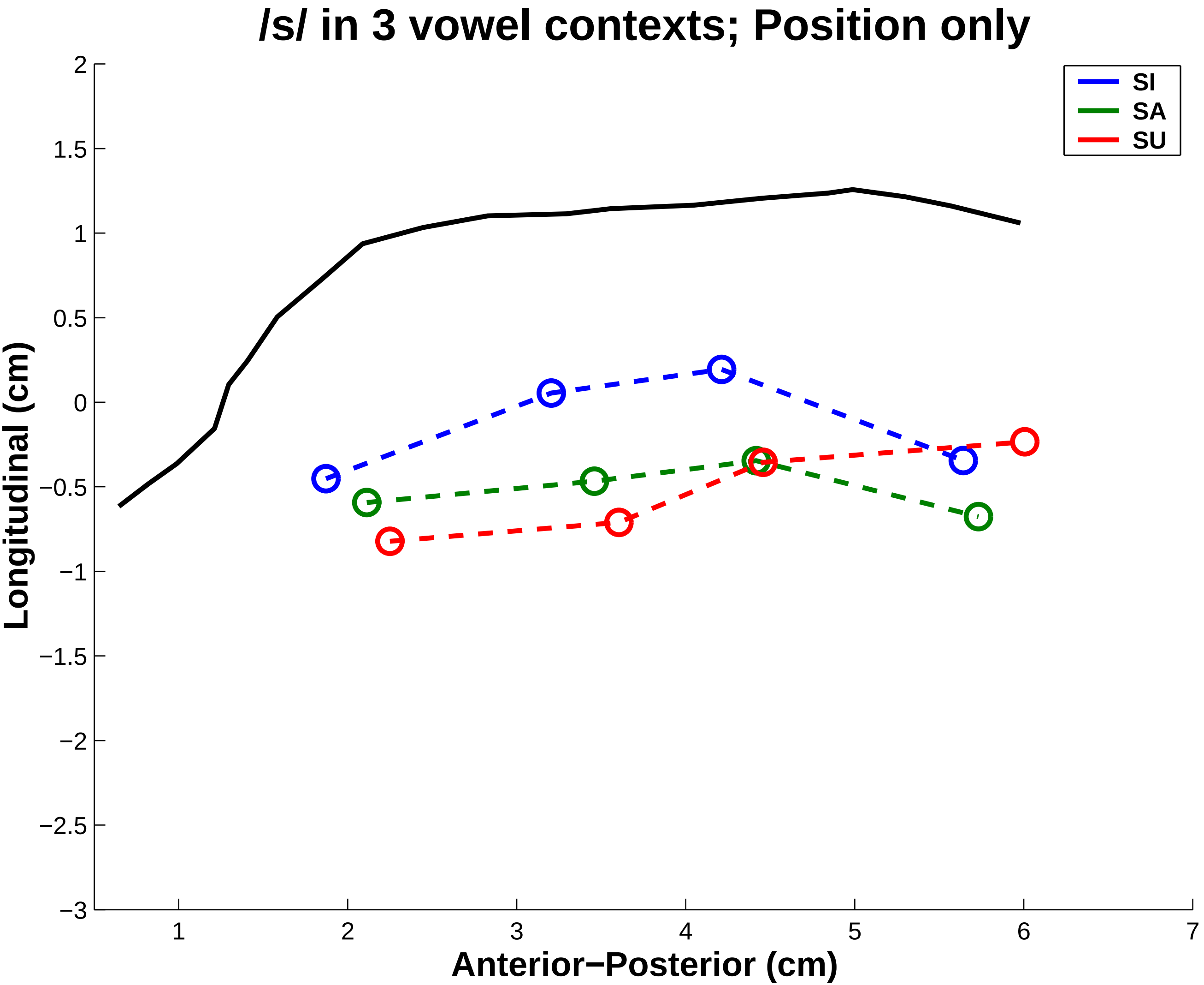
(2) Consonants

(Showing some classic coarticulation effects along the way)

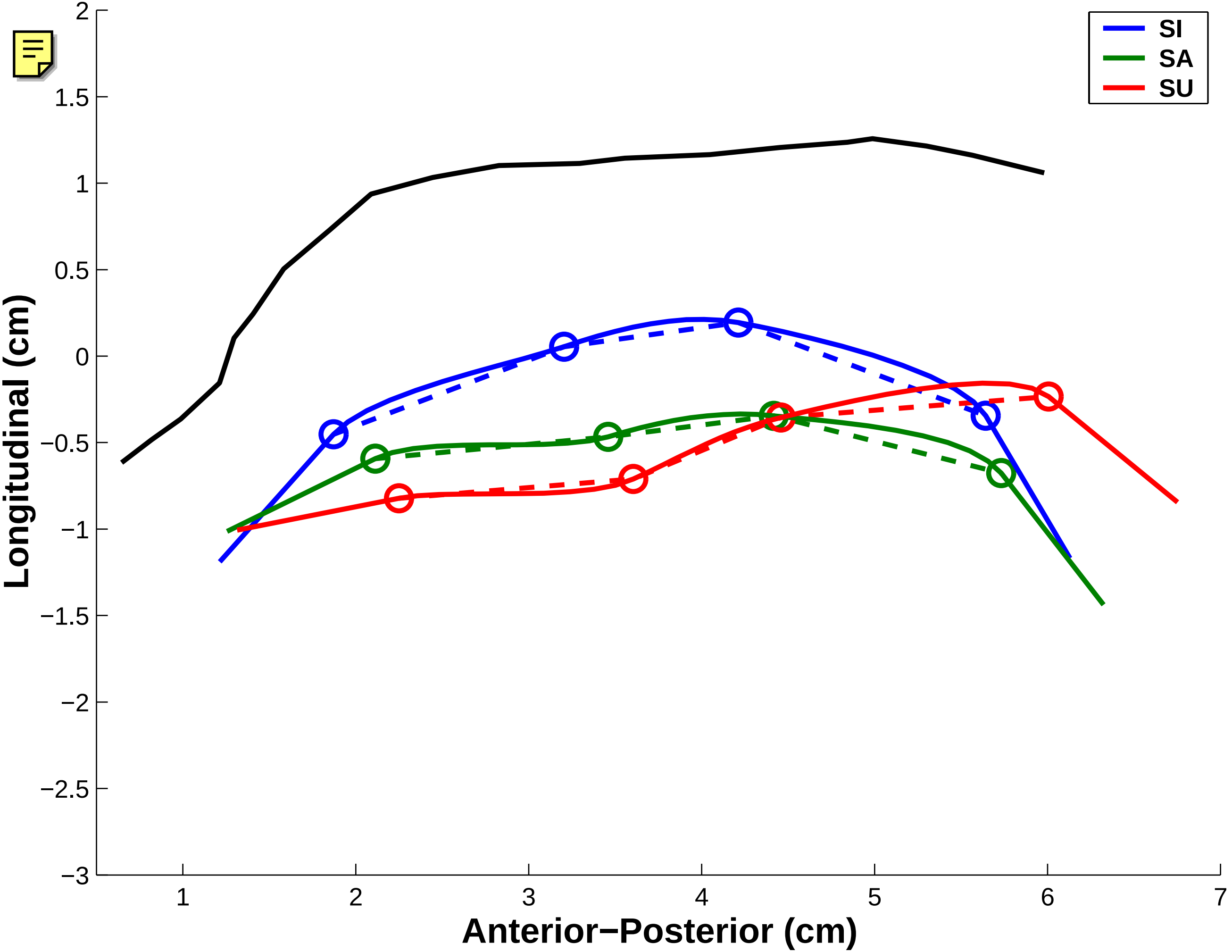

/t/ in 3 vowel contexts; Position only
TI
TA
TU
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)

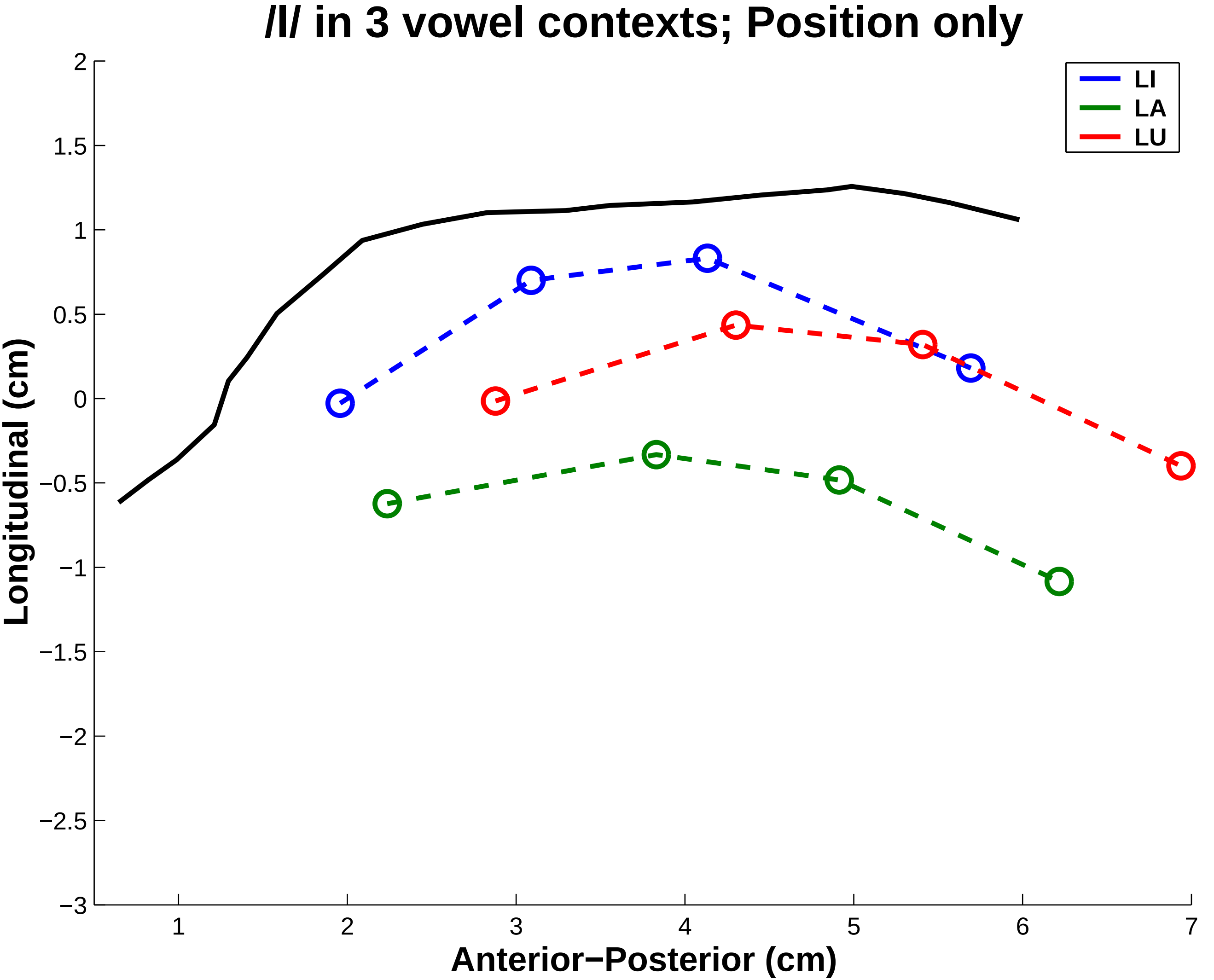
/t/ in 3 vowel contexts; Position and orientation

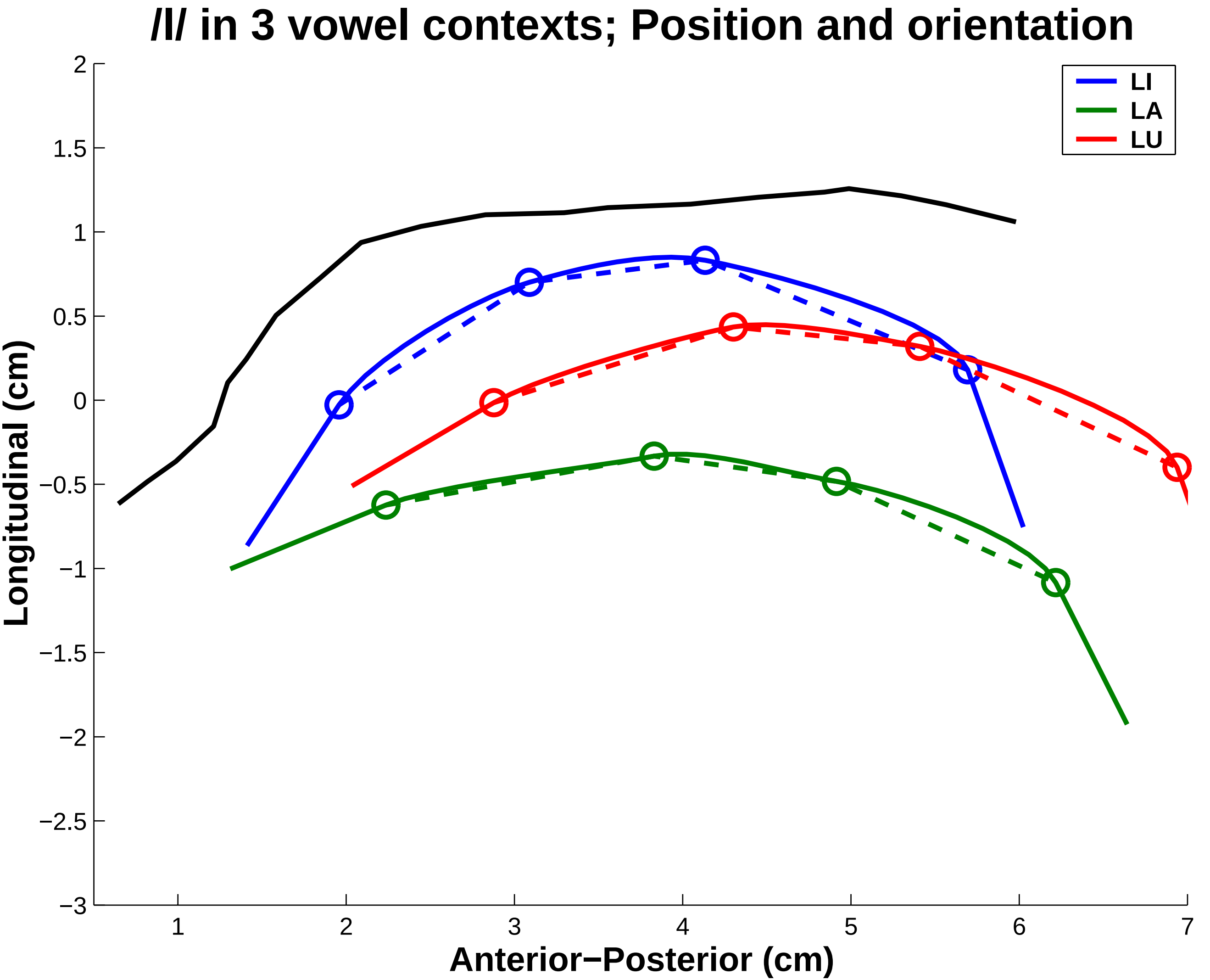


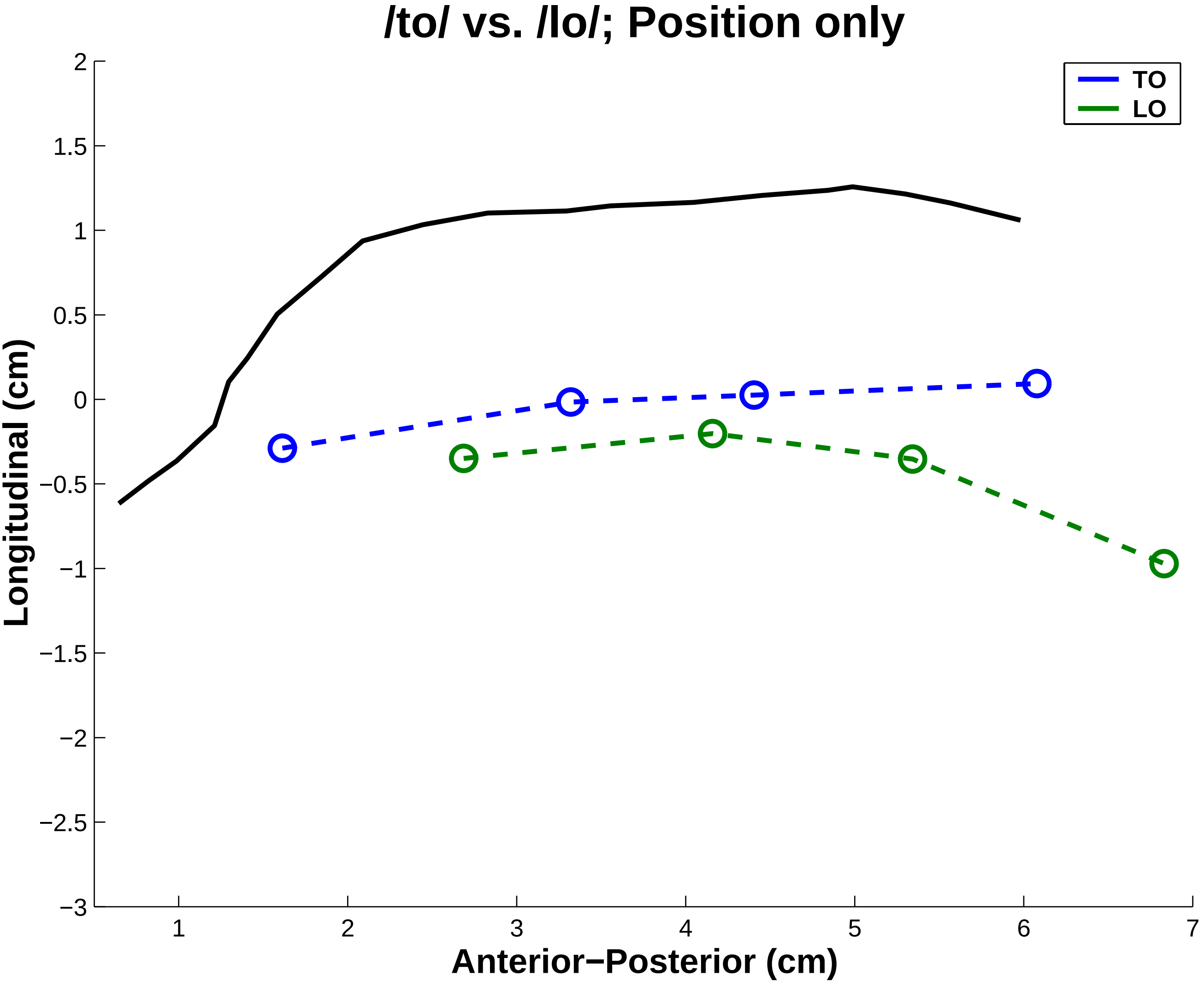

/s/ in 3 vowel contexts; Position only
SI
SA
SU
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)

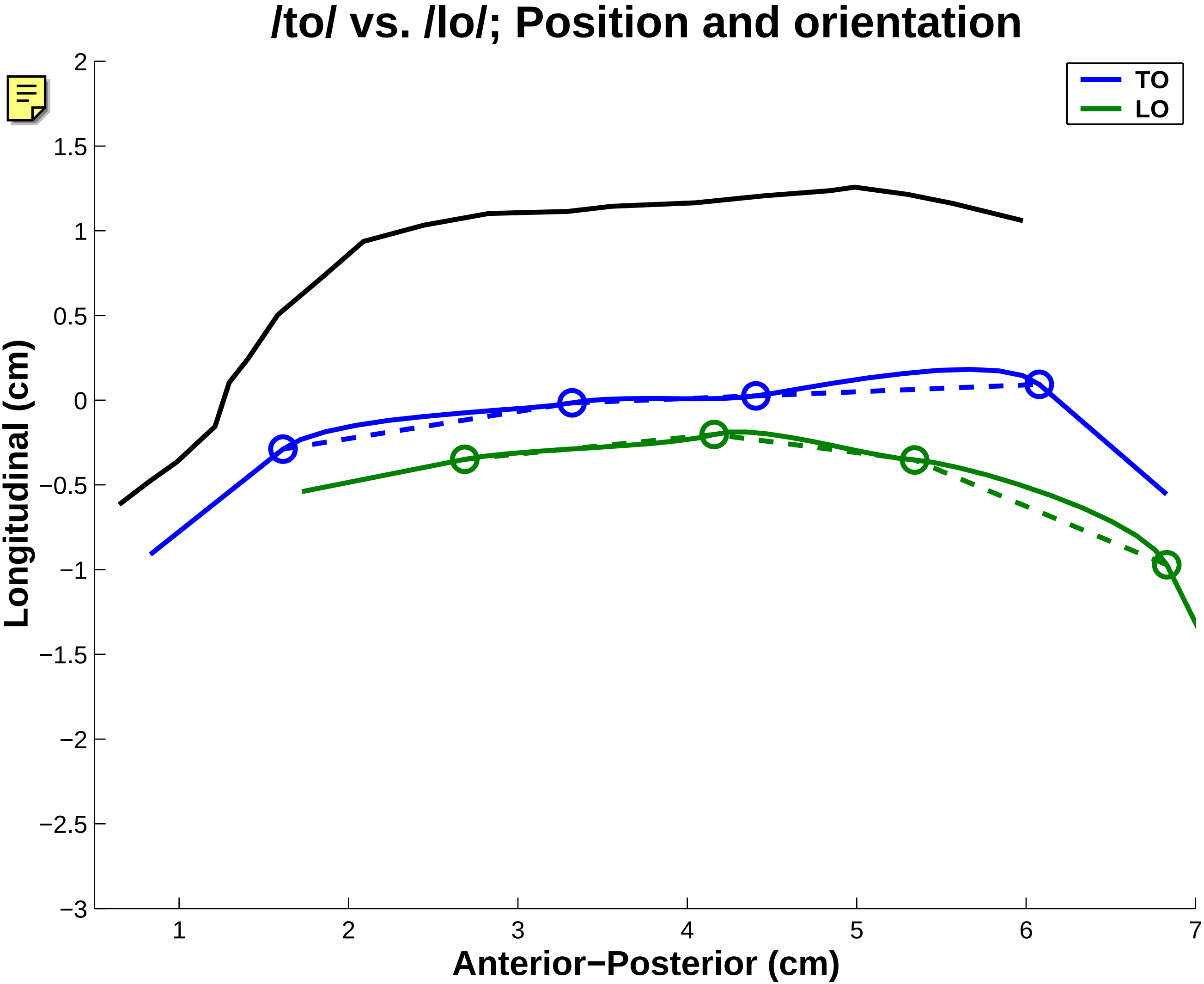
/s/ in 3 vowel contexts; Position and orientation




/l/ in 3 vowel contexts; Position only
LI
LA
LU
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)


/l/ in 3 vowel contexts; Position and orientation
LI
LA
LU
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)


/to/ vs. /lo/; Position only
TO
LO
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)


/to/ vs. /lo/; Position and orientation
TO
LO
Longitudinal (cm)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
1
2
3
4
5
6
7
Anterior−Posterior (cm)

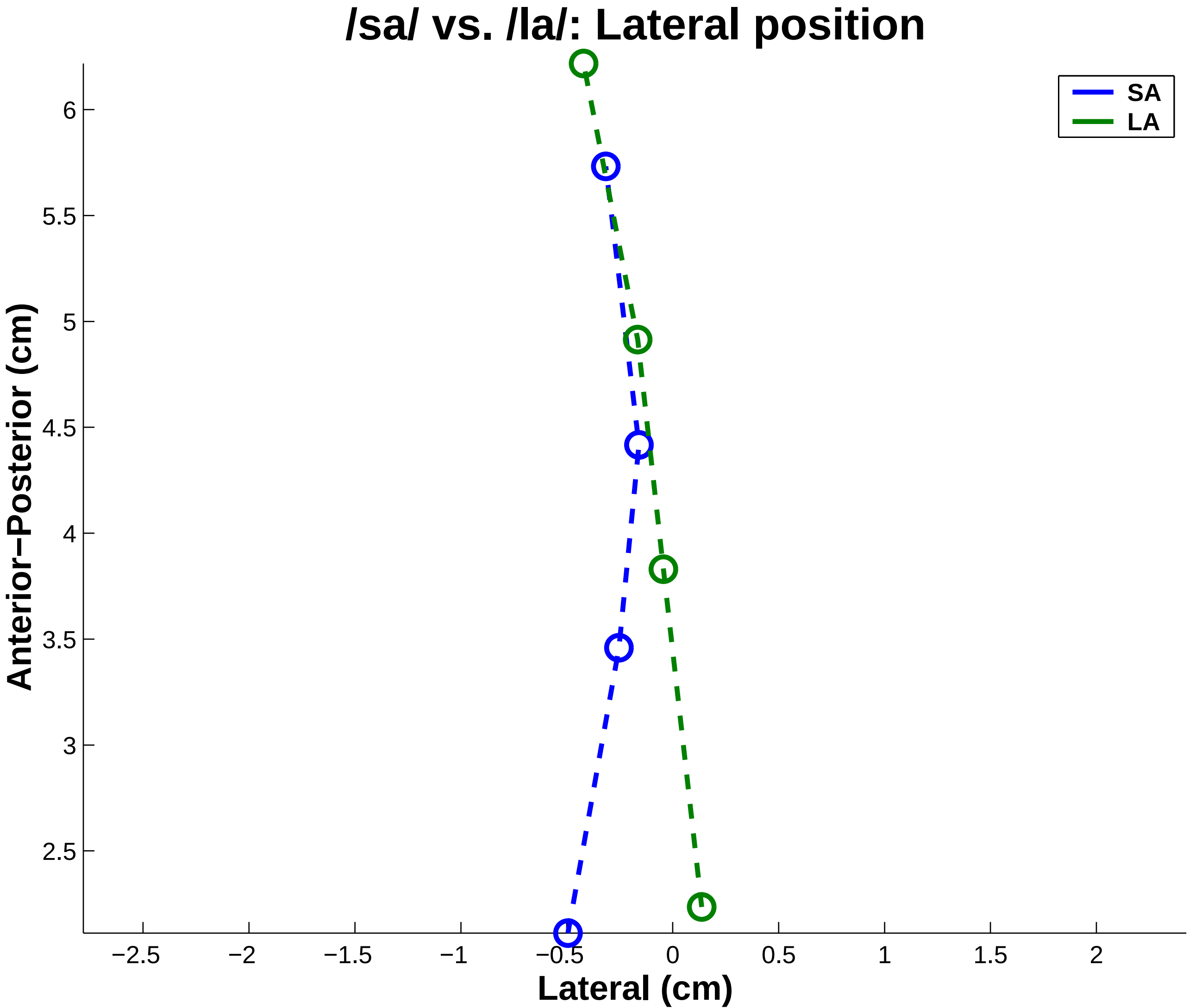
Further tongue animations

/oto/ vs. /olo/	slowmotion movie
retroflex	slowmotion movie

Further Topics

Lateral movement of the articulators

(more examples in Proceedings and on web)


/sa/ vs. /la/: Lateral position
SA
LA
Anterior–Posterior (cm)
6
5.5
5
4.5
4
3.5
3
2.5
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
Lateral (cm)

Bad data happens

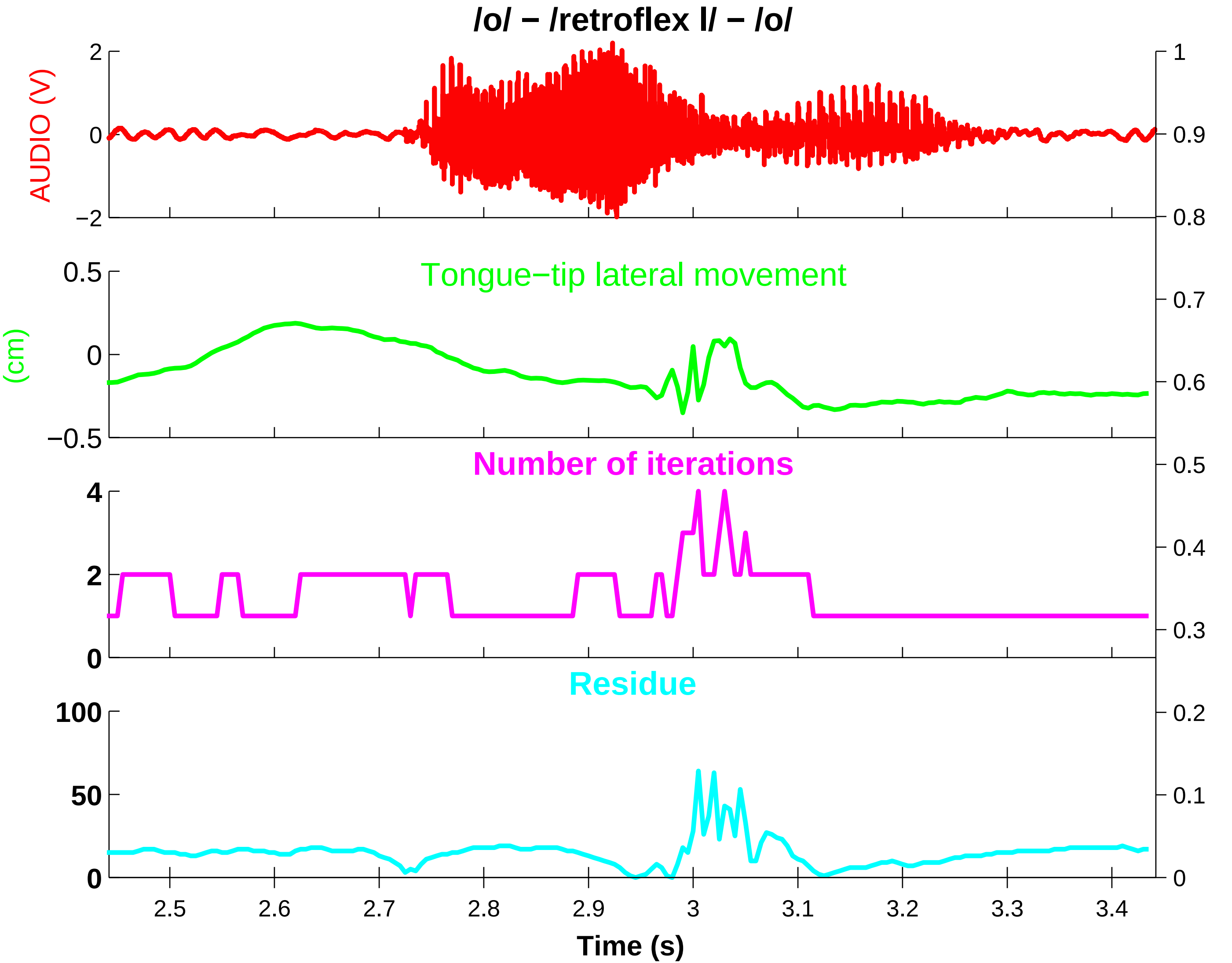
But it can usually be identified

Background:

Unlike 2D EMMA, no closed-form solution to the non-linear equations relating magnetic field strength from the six transmitters to sensor position and orientation.

The iterative algorithm may get caught in a local minimum

➔ mistracking


/o/ − /retroflex l/ − /o/
AUDIO (V)
2
0
−2
Tongue−tip lateral movement
(cm)
0.5
0
−0.5
Number of iterations
4
2
0
Residue
100
50
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
3.4
Time (s)

5D EMA: Summary

(1) Rigid structures

- Tracking of head translations and rotations with only two sensors appears feasible

  (may not yet be accurate enough to recover speech movement with sub-millimetre accuracy when superimposed on large head movements, but probably sufficient to give the subject freedom of head-movement within limits adequate for many kinds of experiments)
- Extension to jaw tracking should be straightforward

(2) Soft structures

Compared to 2D EMMA:

- More information from the same number of sensors (or similar information from fewer sensors).
- Even for traditional midsagittal measurements less need to worry about deviation from midline or sensor misalignment.

Potential for:

- Lateral movements, lateral tongue shaping
- More comprehensive tracking in labial region

(3) The to-do list

- Accuracy and robustness of the calibration and tracking still need improving

2.

Three-dimensional tongue shape from multi-speaker, multi-volume MRI

Analysis with 3-way statistical techniques

Aims

In general

- What are the nature and the number of the underlying patterns of tongue shapes used in speech (here: vowels)?

In detail

- In earlier work we derived a two-factor PARAFAC model of tongue shapes for vowels based on EMMA data (Hoole, 1999a), and on midsagittal NMRI data (Hoole et al., 2000). The PARAFAC model provides a strong motivation for working from a multi-speaker perspective.

  BUT this earlier work indicated that the PARAFAC constraints on speaker-specific behaviour may be too strong.

  PARAFAC is just one member of a large family of n-way methods.

  What methods are most promising for modelling multi-speaker, three-dimensional tongue data?

Three-Mode Analysis (PARAFAC)

(e.g Harshman et al., 1977)

Systematic exploitation of a third dimension to solve the problem of rotational indeterminacy in the factor axes. The speakers represent this third dimension here.

Model prediction for speaker *k*:

$$\mathbf{Y}_k=\mathbf{A}\mathbf{S}_k\mathbf{V}^{\mathrm{T}}$$

where **V**, **A** and **S** are 3 loading matrices (for vowels, articulators and speakers, respectively)

and where $\mathbf{S}_k$ is a matrix with the *k*th row of **S** on the main diagonal and zero elsewhere

Hence very strong assumptions on possible speaker-specific behaviour

If assumptions are met

Very parsimonious representation

Close relationship of factors to the underlying behavioural dimensions

Material and Procedures

- 9 speakers
- 7 long German vowels /i, e, y, ø, a, o, u/
  (7 of the 9 speakers also recorded the consonants /t, s, n, l, ʃ/; these are not discussed further here)
- Complete sagittal, coronal and axial scans. All scans encompassed the complete vocal tract
- Pixel resolution 1.17mm
- Slice thickness 4mm; interslice interval 5mm (4mm for recent sagittal scans)

To date, 5 speakers ready for use in the statistical modelling

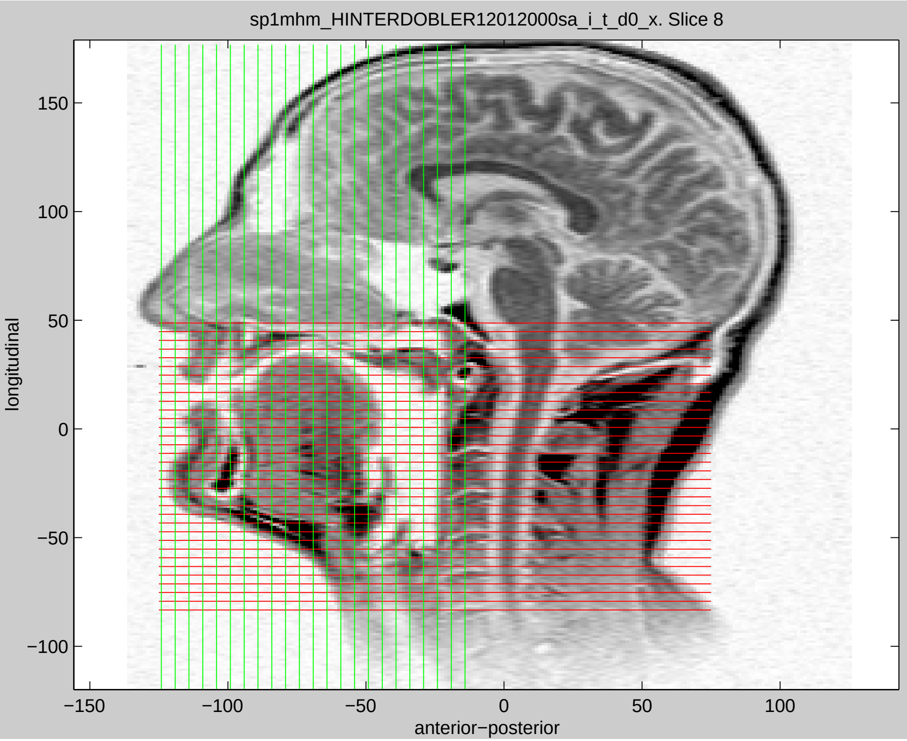


Fig. 1: Coronal and axial slice locations shown on sagittal image

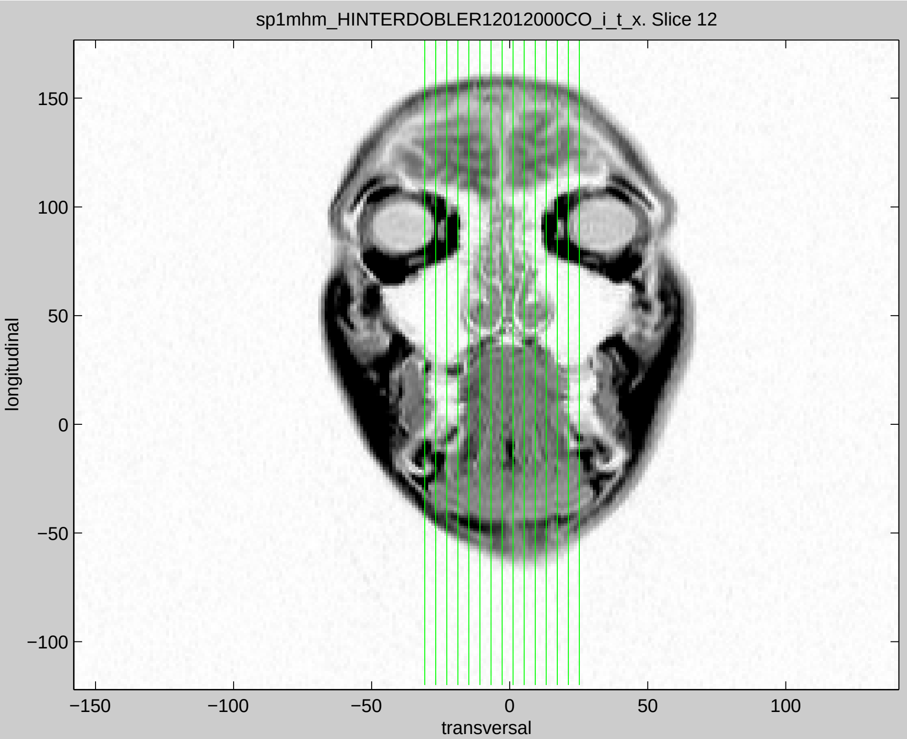
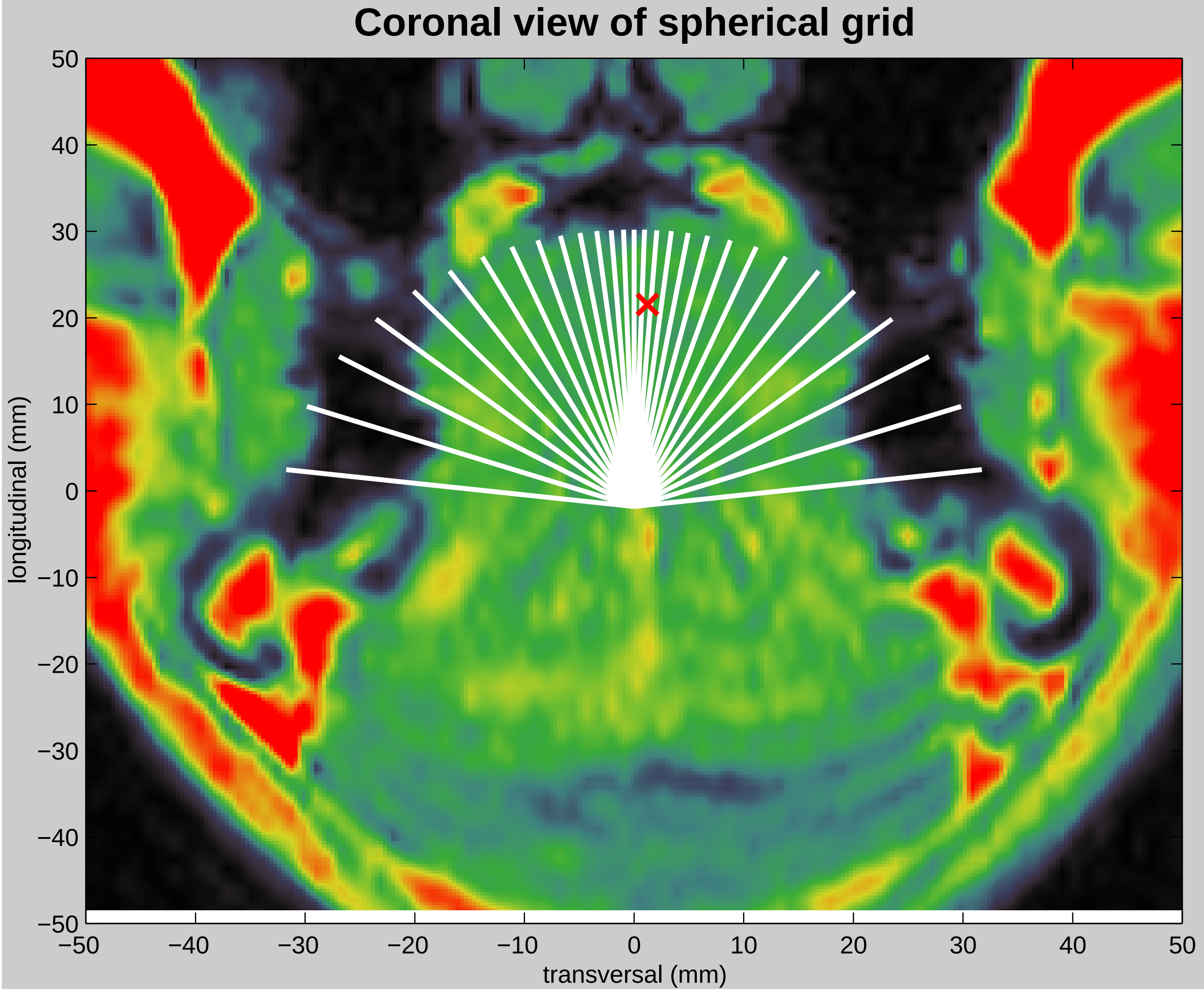
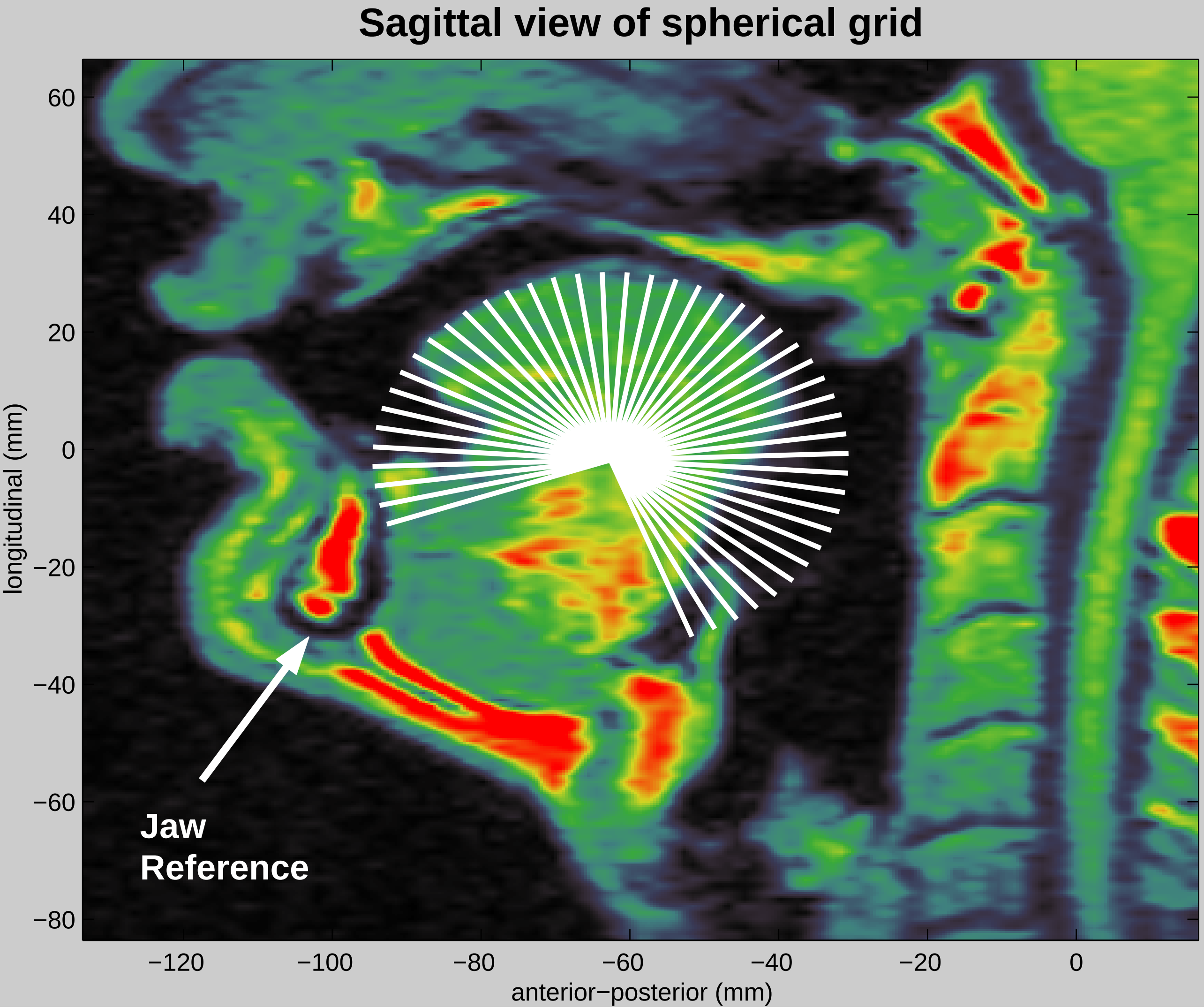


Fig. 2: Sagittal slice locations on coronal image (for earlier subjects more slices were acquired, with correspondingly increased lateral range)

Generation of tongue surfaces for statistical analysis

- Extract raw contours of tongue from each coronal and axial slice
- Subtract jaw position
- Align midline plane of tongue at zero on the lateral axis
- Calculate centre of tongue in midsagittal plane
- Define a grid in spherical coordinates with tongue centre as origin
- Separately for axial and coronal volumes:
  - Convert raw contour data to spherical coordinates and use to predict position of tongue surface at the grid locations
- Merge tongue surfaces derived from the axial and coronal data.

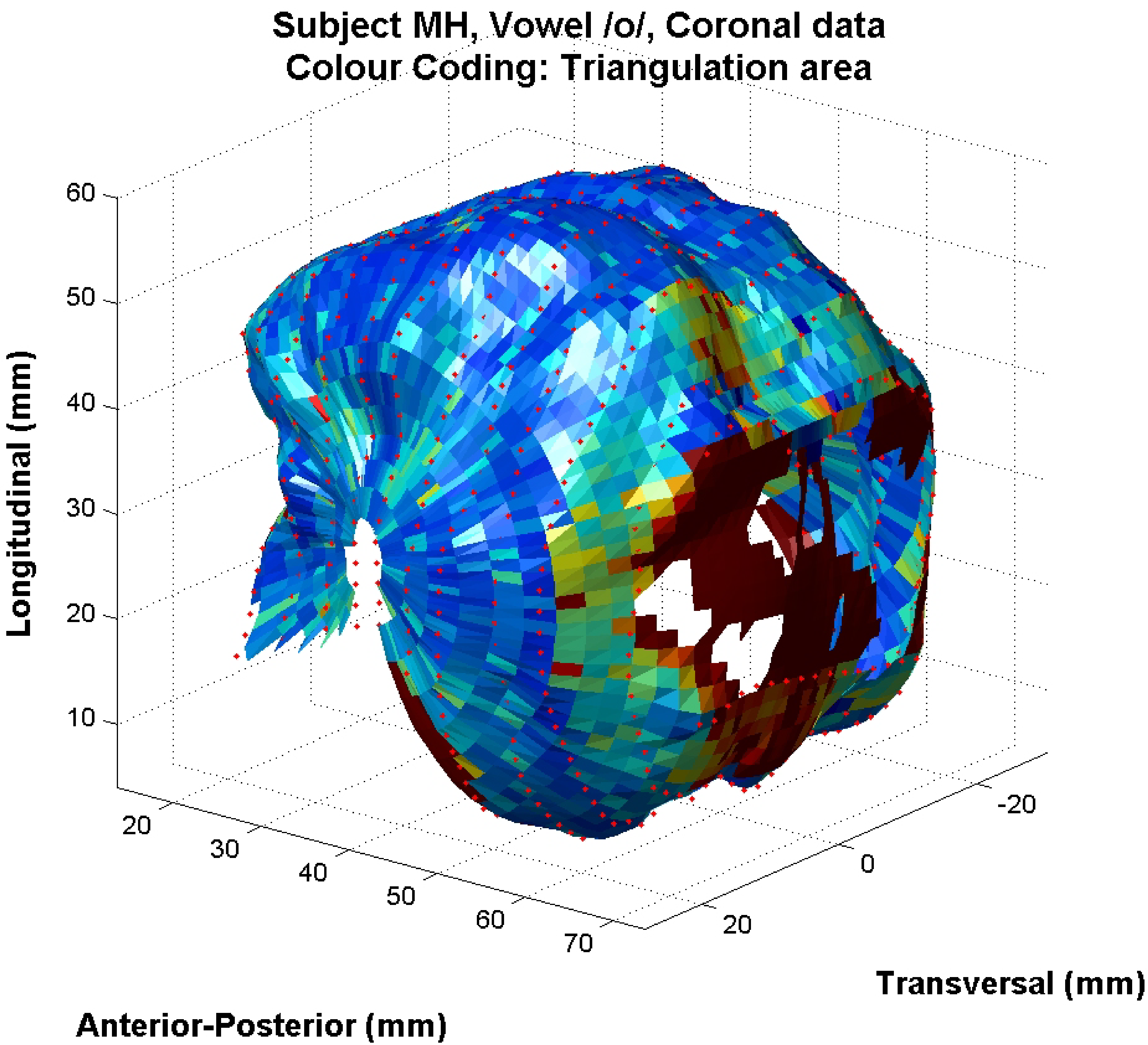

Coronal view of spherical grid
50
40
30
20
10
0
−10
−20
−30
−40
−50
longitudinal (mm)
−50
−40
−30
−20
−10
0
10
20
30
40
50
transversal (mm)


Sagittal view of spherical grid
longitudinal (mm)
60
40
20
0
−20
−40
−60
−80
−120
−100
−80
−60
−40
−20
0
anterior−posterior (mm)
Jaw
Reference

Advantages of using a spherical grid:

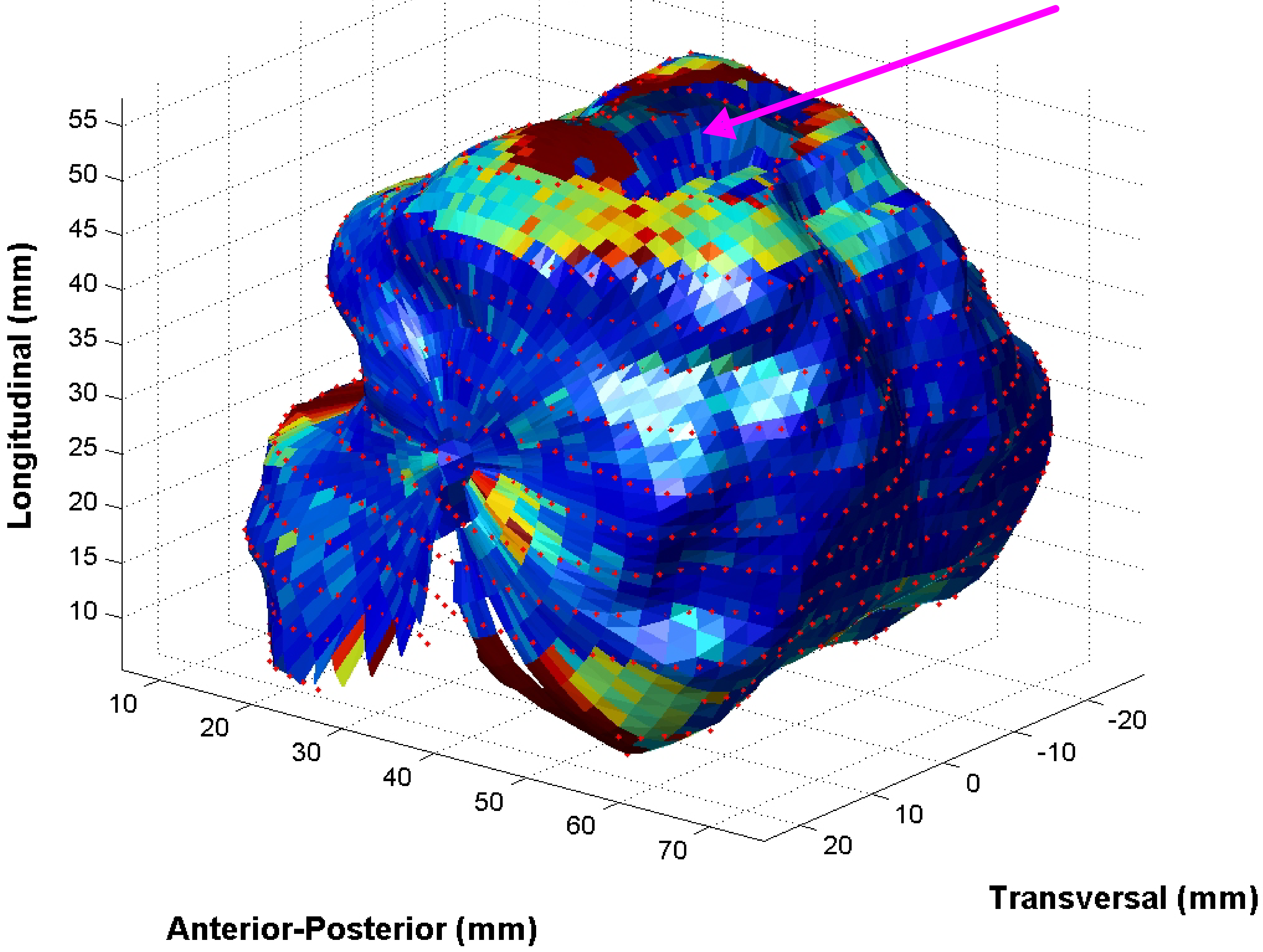
(1) Standard surface-fitting techniques can be used even though tongue surface is not an unambiguous function of height above the axial plane

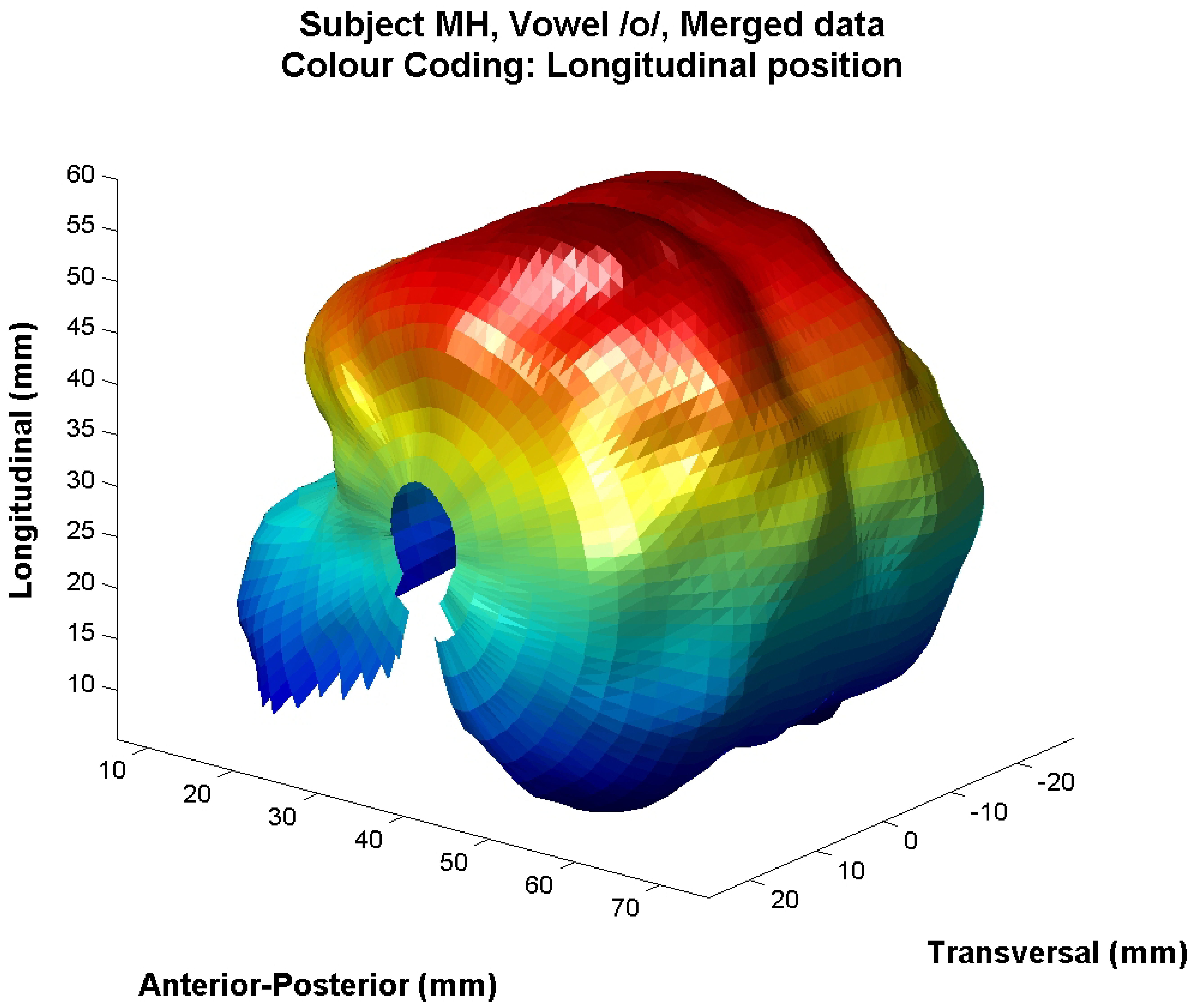
(2) Surface fitting involves triangulation of the input data.
The size of the triangle used to predict the location of each point on the grid can be used as a measure of the reliability of the grid points.
(Large triangle ➔ Grid point not well supported by the input data ➔ Low reliability)
Simple but flexible weighting scheme for merging axial- and coronal-based surfaces:
Axial data typically low reliability for oral surface of tongue
Coronal data typically low reliability for pharyngeal surface of tongue



Subject MH, Vowel /o/, Coronal data
Colour Coding: Triangulation area

Subject MH, Vowel /o/, Axial data
Colour Coding: Triangulation area




Subject MH, Vowel /o/, Merged data
Colour Coding: Longitudinal position
Longitudinal (mm)
60
55
50
45
40
35
30
25
20
15
10
10
20
30
40
50
60
70
Anterior-Posterior (mm)
-20
-10
0
10
20
Transversal (mm)

Alternatives to PARAFAC?

PARAFAC can be seen as a special case of the three-mode ***Tucker*** model:

$$\hat{x}_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} \qquad (1)$$

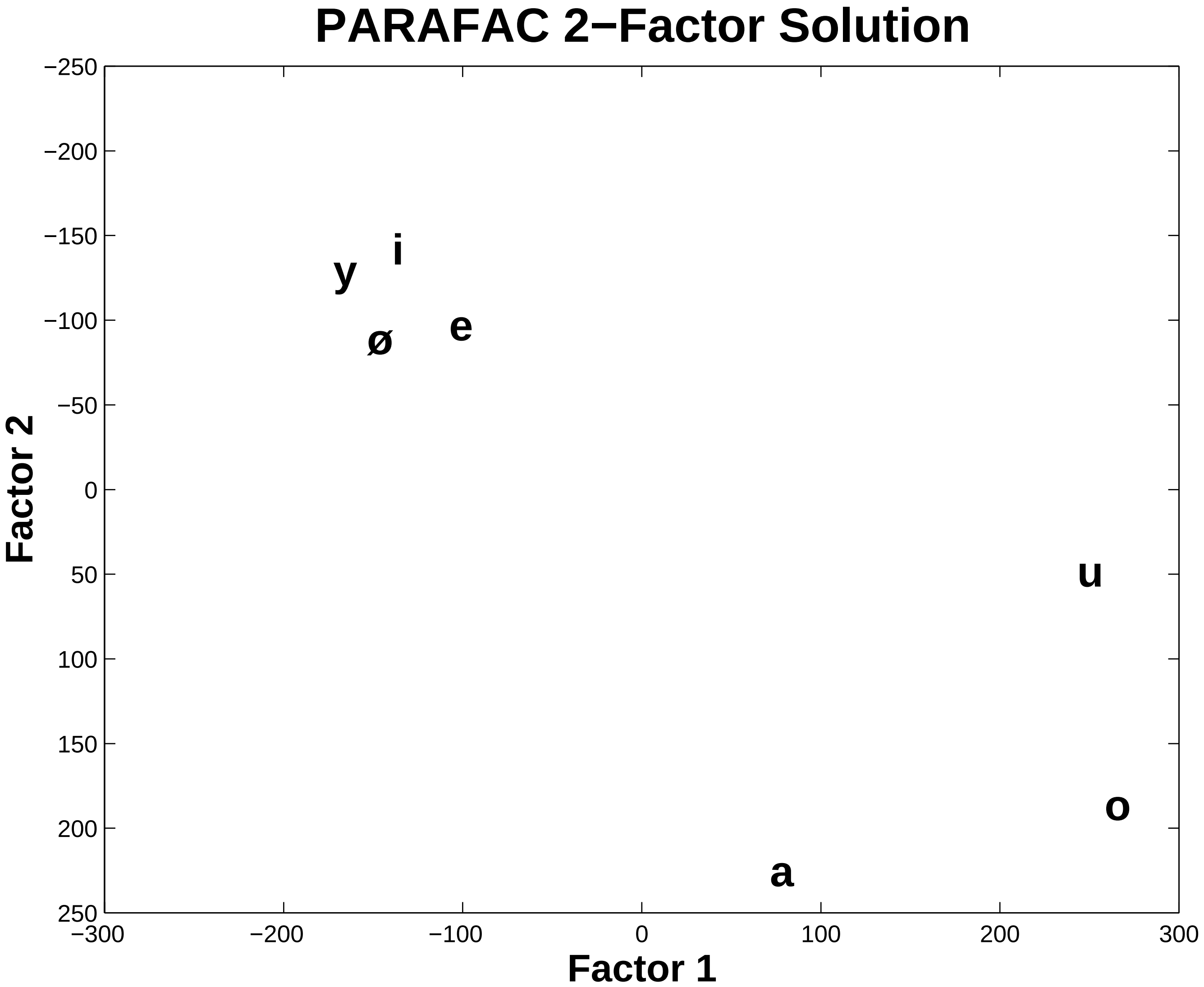
For PARAFAC, P=Q=R. The core matrix **G** only has elements on the main diagonal.

Compared to PARAFAC, the general Tucker3 model is more flexible (less constrained), but is less parsimonious, and may be less interpretable.

Recently, Zheng et al. (JASA, 2003) successfully extracted a two-factor PARAFAC solution for American English vowels from coronal MRI data.

This encouraged us to apply PARAFAC to our combined coronal/axial data.

Result: A stable two-factor solution explained about 80% of the variance.


PARAFAC 2−Factor Solution
−250
−200
−150
−100
−50
0
50
100
150
200
250
Factor 2
−300
−200
−100
0
100
200
300
Factor 1
y
i
ø
e
u
o
a

Animation of the factor shapes relative to the vowel space

The solution is very similar to the solution found in Hoole et al. (2001) based purely on **midsagittal** contours from this dataset.

(and also quite similar to the solution of Zheng et al.)

Tentative conclusion:

Confirms that tongue shapes for vowels appear amenable to the strong constraints of the PARAFAC method, even when using morphologically complex data.

Outlook
Linking 5D EMA and MRI?

EMA and MRI have complementary strengths and weaknesses.

On the practical level: Morphologically rich MRI can provide useful background for knowing what to look for in 5D EMA.

On the modelling level: Explore regression models for predicting MRI-derived tongue shapes from 5D EMA.
Example: Partial Least Squares Regression has recently been extended to 3-way data, i.e could be applied to multispeaker data (Bro, 1996)