Five-dimensional articulography

Phil Hoole, Andreas Zierdt

Phonetics Institute, Munich University

hoole@phonetik.uni-muenchen.de
EMMA → EMA

Five dimensions: Both benefit and bane
EMMA → EMA

Five dimensions: Both benefit and bane

Benefit:

High information density per sensor
  3 Cartesian coordinates (x, y, z)
  2 angular coordinates (azimuth, elevation)

Freedom of head movement → potential for new kinds of experiments
EMMA \rightarrow EMA

Five dimensions: Both benefit and bane

Benefit:
High information density per sensor
  3 Cartesian coordinates (x, y, z)
  2 angular coordinates (azimuth, elevation)

Freedom of head movement \rightarrow potential for new kinds of experiments

Bane:
“A complicated non-linear problem must be solved”
(Kaburagi et al., 2005)
The Munich EMA lab
PC1
AG500
Control

PC2
AG500
IDA

Receiv-
Trans-
ers

Sybox

PC3 (dual screen)
video timer control
AG500 status monitor
Stimulus display to subject
Session log

PC4
Realtime
Oscilloscope

"say cheese
again"

Mikes, Analog signals

Multichannel
DAT

Trans-
mitters

Receiv-
ers

PC2
AG500
IDA

Sybox

Ethernet

PC1
AG500
Control

Synch
PC1
AG500
Control

PC2
AG500
IDA

Receiv-
ers

Trans-
mitters

Sybox

PC3 (dual screen)
video timer control
AG500 status monitor
Stimulus display to subject
Session log

PC4
Realtime
Oscilloscope

Multichannel
DAT

Synch

PC2
AG500
IDA

Ethernet

PC1
AG500
Control

PC5
Position
Display

USB
Amp data

PC6 ... PCn
Position
Calculation

Position data

Amp data

"say cheese again"
High Information Density per Sensor

Example: Coronal consonants in Moroccan Arabic

(collaboration with Adamantios Gafos, NYU, and Chakir Zeroual, Paris/Taza)

Order of presentation in movie:

D, S, T, c, d, j, l, n, r, s, t, z

Upper case = emphatic

c = /ʃ /
j = /ʒ /
Traditional sagittal view of tongue-tip
Representation using angular parameters (tip and back)

- Even clearer separation of categories
- Very succinct characterization of tongue *shape*
Sagittal view without head correction: Chaos!
Angular parameters without head correction: Much more robust
Monitoring Head Movement

Further example of high information density per sensor

How many sensors needed to capture the 6 DOF of head movement (3 translations, 3 rotations)?
Monitoring Head Movement

Further example of high information density per sensor

How many sensors needed to capture the 6 DOF of head movement (3 translations, 3 rotations)?

Ideally, two.
(cf. 2D EMMA, Optotrak)

- Freedom of head movement for subject
  ➔ More natural speaking situation. Longer experiments!
- Communicative relevance of head movement
Reference sensors: Ideally two sufficient, but in practice ...
Reference sensors: Ideally two sufficient, but in practice ... Euclidean distances should be constant over the experiment. But some pairs may be more equal than others → Record more than two, and select the most stable ones
Relations between head movement and linguistic behaviour
a topic in its own right

Selected results from:

Tone-Vowel Interaction in Standard Chinese
Phil Hoole & Fang Hu
Conference on Tonal Aspects of Language, Beijing, 2004
Overall Question

To what extent is tone production reflected in supraglottal articulation?

Approach 1 (skipped here):

Does position of tongue (and jaw and lips) vary systematically with tone?
**Approach 2:**

To what extent is tone production associated with visible movements?

Burnham et al. (2001) found evidence that limited tonal perception is possible based on visual information alone (see also Mixdorff & Charnvivit, 2004).

But what visual information is actually involved here? One candidate is **head movement**.

Yehia et al., 2002, found systematic relations between head-position and F0.
Results Approach 2
Head Movement

Preprocessing:

Raw position data converted to deviation from average position in each block of repetitions
Vertical Head Position

Tone 1  Tone 2  Tone 3  Tone 4

a  i  u  w  y
The clearest differences are between Tone 3 and Tone 2:

Head position lower and more retracted for Tone 3 than Tone 2 for all vowels except /a/.

Magnitude of the differences is small, but clearly significant (p<0.01)

Does analysis at vowel mid-point miss anything?
Head movement for each tone. Vowel i

Vowel Onset

Vertical Position (mm)

Time (s)
Head movement for each tone. Vowel i

Anterior–Posterior Position (mm)

Time (s)

Vowel Onset

Anterior →

Posterior ↑

Anterior ↓

1 2 3 4
Overall time-course of head movement is rather different for Tone 3

This is also reflected in velocity values:

Velocity of upward and forward movement higher for Tone 3

Velocity patterns are attractive because they may be communicatively more robust than subtle differences in position (cf. Keating et al., 2003).

Head-movement certainly does not simply mimic the F0-contour.
Freedom of head movement: Longer recordings

Advantageous for paradigms with low ratio of "interesting" material to total recorded material

   e.g spontaneous speech, speech errors.
Another kind of example:

Articulatory data for unit-selection speech synthesis (Korin Richmond, CSTR, Edinburgh).

Hypothesis: Defining joint costs in the articulatory domain may be beneficial for unit selection.

Example of dataset: subset of phonetically rich corpus recorded with velum sensor and 2 tongue sensors (plus lip/jaw). Approx. 50 min. net speaking time.

(As an aside: 5D system may make handling velum sensors less tricky)
To conclude this section: Further examples of diversification?
Hixon, Goldman, Mead
Kinematics of the chest wall during speech production
JSHR 16, 78-115 (1973)

**Figure 3-3.** Schematic illustration of tilt table configuration, subject positioning, and magnetometer coil placements.
Part 2

5D - Why processing can be tough

Introduction to non-linear optimization

Two typical problems and how to handle them
Non-linear optimization

A toy example in one dimension

In practice: Determine the coordinates of the sensor in five-dimensional space from 6 transmitter signals

Conventions for the following examples:

\( x \) \hspace{1cm} \text{Position}
\( f(x) \) \hspace{1cm} \text{A non-linear function of } x \ (\text{e.g magnetic field model})
\( y \) \hspace{1cm} \text{Measured sensor signal}

To determine \( x \) solve \( f(x) - y = 0 \)
Finding the root of a function, using Newton’s method

\[ f(x) \]

<table>
<thead>
<tr>
<th>Start-Value</th>
<th>True Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_0 = 0.00</td>
<td></td>
</tr>
</tbody>
</table>
Finding the root of a function, using Newton’s method

Start-Value   True Solution
x₀ = 0.00      x₂ = 1.59
x₁ = 0.85      x₃ = 2.57
x₂ = 1.59
Finding the root of a function, using Newton’s method
Finding the root of a function, using Newton’s method

\[ Err_x(x_3) = 0.12 \text{ (Accuracy)} \]
Finding the root of a function, using Newton’s method

\[ \text{Err}_x(x_3) = 0.12 \text{ (Accuracy)} \]

\[ \text{Err}_y(x_3) = 0.25 \text{ (Residual)} \]
Finding the root of a function, using Newton’s method

\[ \text{Err}_x(x_3) = 0.12 \text{ (Accuracy)} \]

\[ \text{Err}_y(x_3) = 0.25 \text{ (Residual)} \]

Accuracy \sim \text{Residual / Gradient}
Example 2: Getting lost in the woods
Finding the root of a function, using Newton’s method

Start-Value $x_0 = 0.00$

True Solution
Finding the root of a function, using Newton’s method

\[ f_1(x) \]

\[
\begin{align*}
    x_0 &= 0.00 \\
    x_1 &= 0.86 \\
    x_2 &= 1.51
\end{align*}
\]
Finding the root of a function, using Newton’s method

\[ f_1(x) \]

\begin{align*}
  x_0 &= 0.00 \\
  x_1 &= 0.86 \\
  x_2 &= 1.51 \\
  x_3 &= 2.48 \\
\end{align*}
Finding the root of a function, using Newton's method

$$f_1(x) = 1.44$$
Finding the root of a function, using Newton’s method

Err \( x_4 \) = 3.74 (Accuracy)
Err \( y_4 \) = 7.39 (Residual)

Residual good but Accuracy bad
Finding the root of a function, using Newton's method

Err_x(x_4) = 3.74 (Accuracy)
Err_y(x_4) = 7.39 (Residual)

Residual good but Accuracy bad

Convergence radius

x_4 = 1.44
Example 3: Noise may have unpredictable effects
Finding the root of a function, using Newton’s method

$x_0 = 1.08$

True Solution
Finding the root of a function, using Newton’s method

$x_0 = 1.08$

Start-Value
True Solution
Typical problem 1

Calculated positions (here for tongue-mid sensor) appear unstable, compared to neighbouring sensors.
Lateral Position

mm

0 0.5 1 1.5 2

−15 −10 −5 0
A closer look at the residuals. Do they show a pattern?
Example: Residual for Transmitter 3 signal plotted vs. all 6 transmitter signals
Adjust the originally measured amplitudes by the predictable component of the residual

Re-calculate positions.
Typical problem 2

Some instabilities remain even after amplitude correction
Raw amplitudes clearly continuous, unlike the calculated positions

Repair procedure:

- Eliminate unreliable data
- Use linear regression to compute estimate of sensor velocities directly from first derivative of raw amplitudes
- Repair unreliable position data using predicted velocities (after integration)
Position X

Velocity X

Raw Amplitudes

Time (s)
Summary of repair procedures

Depend on a fair proportion of data being basically accurate

Depend on being able to specify plausible ranges for velocity and inter-sensor distances in speech movements.

Instead of patching after the event it would be more elegant to build constraints into the position calculation itself.
<table>
<thead>
<tr>
<th>Language</th>
<th>Collaboration</th>
<th>n</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin Chinese</td>
<td>Hu</td>
<td>1</td>
<td>Vowels and tone. 200 trials, 10s each</td>
</tr>
<tr>
<td>English</td>
<td>Richmond</td>
<td>1</td>
<td>1200 phonetically rich sentences (ave. 4s each) + 800 with velum</td>
</tr>
<tr>
<td>Moroccan Arabic</td>
<td>Gafos, Zeroual</td>
<td>5</td>
<td>1000 consonant cluster target words in carrier phrase</td>
</tr>
<tr>
<td>French</td>
<td>(1) Kühnert</td>
<td>2</td>
<td>(1) Consonant cluster target words in carrier phrase; 300 phrases, 3 targets per phrase</td>
</tr>
<tr>
<td></td>
<td>(2) Fougeron</td>
<td></td>
<td>(2) Articulatory prominence of prosodic boundaries: 250 phrases (3.5s each)</td>
</tr>
<tr>
<td></td>
<td>(3) Nguyen</td>
<td></td>
<td>(3) 300 pseudowords</td>
</tr>
<tr>
<td>German</td>
<td>Mooshammer</td>
<td>1</td>
<td>Focus/accent variation. 300 phrases, 2 targets per phrase. With RIP</td>
</tr>
<tr>
<td>German</td>
<td></td>
<td>2</td>
<td>CVC pseudoword corpus in carrier phrase. 500 phrases. Speakers prompted with their own utterances from previous lingual EMG exp. with same corpus</td>
</tr>
<tr>
<td>English</td>
<td>Pouplier</td>
<td>2</td>
<td>Speech error elicitation. 80 trials, approx. 30s each</td>
</tr>
<tr>
<td>German</td>
<td>Brunner</td>
<td>3</td>
<td>3 EMA sessions per speaker. Ave. 500 utterances per session. CVC pseudoword targets in carrier phrase. Speakers wore palatal prosthesis over 2-week period. (4 more speakers with 2D EMMA in Berlin)</td>
</tr>
</tbody>
</table>