Normalization and Plotting

Workshop on Sociophonetic Methodology, LSA Summer Institute, Boulder, USA
July 2011

Anne Fabricius, Roskilde University, Denmark
Tyler Kendall, University of Oregon, USA
Dominic Watt, University of York, UK
Topics for this session

Part 1: Plotting
  ▪ 1.1 A quick overview of its role in sociophonetics

Part 2: Normalization
  ▪ 2.1 Why and wherefore
  ▪ 2.2 Defining a normalization algorithm typologically
  ▪ 2.3 Evaluating an algorithm’s performance
    ▪ Fabricius, Watt & Johnson (2009)
    ▪ Flynn (2011)
    ▪ Flynn & Foulkes (2011)
    ▪ Fruehwald (ms)

Part 3: Metrics and the vowel space
  ▪ 3.1 Fabricius (2007)
  ▪ 3.2 Watt & Fabricius (2011)
  ▪ 3.3 Fridland & Kendall (under review)

Question & discussion time
Part 1: Plotting

- graphical formats (i.e. plots) provide “a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.” (Cleveland 1993: 1)

- “nothing beats a picture” (K. Johnson 2008).
(F1 \times F2) Vowel plots are ubiquitous

Thomas
Male
(used by

vowels.R (+ durplot plug-in)
- NORM (& Vowels R library)
  - http://ncslaap.lib.ncsu.edu/tools/norm/
- Plotnik
- Akustyk (Praat plug-in)
- Origin
- SigmaPlot
- Excel...
Part 2: Normalization

- **normalization:**
  - Chiefly *Math.* and *Physics*. To multiply (a series, function, variable, etc.) by a factor that makes the norm or some associated quantity (such as an integral) equal to a particular value, usually one. [OED online]

- **here:** **factoring out** of physical (anatomical > acoustic) differences between samples

- listeners unconsciously compensate for absolute formant frequency differences for any given vowel category
Part 2: Normalization

- cognitive processes underlying this faculty still not well understood (Johnson & Mullennix 1997; Wong et al. 2004; Ames & Grossberg 2008; Monahan & Idsardi 2010)

- not the aim of our own work to simulate this process directly

- rather, it is to enable qualitative (visual) and quantitative comparisons of speakers’ and groups’ vowel productions
2.1 Why and wherefore

- cross-gender and cross-age comparisons
- same-gender comparisons
- same speaker over time (age-dependent)
- sociophonetic community studies will encompass at least some of these
- choice *not* to normalize should be justified

RP data for 3 older RP-speaking men (Hawkins & Midgley 2005)
2.1 Two types of transforms/ algorithms...

- reflecting the stages of processing carried out on an incoming acoustic signal
  - the peripheral auditory system (transform)
  - the auditory processing centers of the brain (normalization proper)
2.2 Defining a Normalization algorithm typologically

- 3-way cross-cutting set of terms
- defining where the algorithm derives its information from
  - **speaker** extrinsic vs. intrinsic
  - **vowel** extrinsic vs. intrinsic
  - **formant** extrinsic vs. intrinsic
psychoperceptual scales (Bark, ERB, mel)

approximate non-linear frequency response of inner ear

much more sensitive to changes in frequency at the lower end of the spectrum

- 1 critical band = 100 Hz between 150 and 250 Hz

- but = 350 Hz between 2150 and 2500 Hz
Vowel intrinsic/ formant extrinsic

- Bark Difference Metric
- Syrdal and Gopal (1986)

- two slightly different versions of same idea:
  - Hertz values converted into Bark
  - $Z_3-Z_2$ or $Z_2-Z_1$ modelling advancement
  - $Z_1-Z_0$ modelling vowel height (NORM uses $Z_3-Z_1$)
Vowel extrinsic/ formant intrinsic

- most successful category for sociophonetic purposes (Adank 2003, Flynn 2011)
- ranges (Gerstman)
- mean / standard dev (Lobanov)
- individual log-means (Nearey CLIH_{i4})
- centroids (W&F, mW&F, Bigham)
Vowel extrinsic/ formant extrinsic

- Nordström & Lindblom’s (1975) vocal tract scaling transformation

- Nearey’s shared log-mean model (‘Constant Log Interval Hypothesis’ or CLIH\(_{s4/s2}\))

- ANAE/Telsur G value (also speaker extrinsic) (Labov, Ash & Boberg 2006: 39-40)

- conceptualized as a ‘sliding template’ approach using a scaling factor
Not a bene: Two versions of Nearey

- Nearey 1 (NORM): CLIH_{i_2} – formant intrinsic
- Nearey 2 (NORM): CLIH_{s_2} – formant extrinsic

- Adank (2003) rates CLIH_{i_4} (≈ Nearey 1) as more successful than CLIH_{s_4}
- Nearey 1 of same typology as Lobanov
- Nearey 2 is the basis for the ANAE method and implemented in Plotnik
### Summary Table

<table>
<thead>
<tr>
<th>Scale/Method</th>
<th>Speaker</th>
<th>Vowel</th>
<th>Formant</th>
<th>Reference; NORM suite name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>Traunmüller 1990, 1997</td>
</tr>
<tr>
<td>mel</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>Stevens and Volkman 1940</td>
</tr>
<tr>
<td>Koenig</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>Koenig 1949</td>
</tr>
<tr>
<td>ERB</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>Moore and Glasberg 1983</td>
</tr>
<tr>
<td>Syrdal and Gopal</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>Syrdal and Gopal 1986</td>
</tr>
<tr>
<td>Bark Difference Metric</td>
<td>intrinsic</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td><em>Bark Difference Metric</em></td>
</tr>
<tr>
<td>Gerstman</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>intrinsic</td>
<td>Gerstman 1968</td>
</tr>
<tr>
<td>Lobanov</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>intrinsic</td>
<td>Lobanov</td>
</tr>
<tr>
<td>Neary CLIH$_{14}$</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>intrinsic</td>
<td>Neary 1977/8; <em>Nearey 1</em></td>
</tr>
<tr>
<td>Watt and Fabricius</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>intrinsic</td>
<td>Watt and Fabricius 2002;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>Watt and Fabricius</em></td>
</tr>
<tr>
<td>Nordström and Lindblom</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>extrinsic</td>
<td>Nordström and Lindblom 1975</td>
</tr>
<tr>
<td>Neary CLIH$_{34}$</td>
<td>intrinsic</td>
<td>extrinsic</td>
<td>extrinsic</td>
<td>Neary 1977/8; <em>Nearey 2</em></td>
</tr>
<tr>
<td>Labov ANAE methods</td>
<td>extrinsic</td>
<td>extrinsic</td>
<td>extrinsic</td>
<td>ANAE</td>
</tr>
</tbody>
</table>

[http://normtable.notlong.com](http://normtable.notlong.com)
2.3 Evaluating an algorithm’s performance

- a recent renewed interest in this area
- earlier work with emphasis more on speech perception (Disner 1980, Deterding 1990, but see Hindle 1978)
- now, comparisons tending towards sociophonetically-relevant parameters, both visual and quantitative
- Adank (2003) used measures of sorting efficiency using linear discriminant analysis
- more evaluation possibilities emerging all the time... (see also Clopper 2009)
2.3 Evaluating an algorithm’s performance

- 1 = disparity in area agreement and poor overlap
- 2 = good area agreement but poor overlap
- 3 = good fit on both counts
Example 1: Fabricius, Watt and Johnson 2009

- tested \( W&F, mW&F \) against Lobanov and Nearey1 on following parameters:
  - reduction of variance in area ratios of vowel polygons
  - improvement of intersection of vowel polygons
  - conservation of angular relationships between selected points on the \( F1/F2 \) plane, after normalization
Methods and Data

- **area**
  - proportional reduction in variance
  - Pitman-Morgan’s test of homogeneity of variance between correlated samples (Cohen, 1990).

- **intersection**
  - intersection of two vowel polygons divided by the union of the same polygons → intersection values compared statistically

- **vowel juxtapositions**
  - planar locations compared across methods (DRESS-LOT; TRAP-STRUT and LOT-FOOT)

- tested on data from RP and Aberdeen English
Results 1 and 2

Test 1: Equalizing vowel space areas

\[ \text{Lobanov} \geq \text{W\&F}, \text{mW\&F} > \text{Nearey1} > \text{Hertz} \]

Test 2: Improving vowel space overlap

\[ \text{Lobanov} > \text{mW\&F} \geq \text{W\&F}, \text{Nearey1} > \text{Hertz} \]

Overall:

\[ \text{Lobanov} > \text{mW\&F} \geq \text{W\&F} \geq \text{Nearey1} \]
Testing angular relations

Schematized RP (SSBE) short vowel trajectories during the 20th century

KIT/BIT
DRESS/BET
TRAP/BAT
FOOT/PUT
LOT/BOT
STRUT/BUT
## Results 3

**TABLE 8. Mean differences in angles, older speakers compared with younger speakers**

<table>
<thead>
<tr>
<th>TRAP-STRUT angle relative to horizontal</th>
<th>Hz</th>
<th>Neareyl</th>
<th>W&amp;F</th>
<th>Lobanov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean differences between older and younger groups</td>
<td>38</td>
<td>61</td>
<td>61</td>
<td>64</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>12</td>
<td>25</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>LOT-FOOT angle relative to vertical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean differences between older and younger groups</td>
<td>49</td>
<td>51</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>24</td>
<td>13</td>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Fabricius, Watt & Johnson (2009: 429)
Example 2: Flynn 2011

- compares 20 normalization methods (6 vowel intrinsic, 14 vowel extrinsic)
  - includes some innovative normalization techniques
    - Bigham (2008)
    - and additional new possible methods

- 20 speakers of Nottingham English (age, gender); 180 vowel tokens per speaker
Example 2: Flynn 2011

- methods
  - for equalising vowel space areas: squared coefficients of variance (Fabricius et al. 2009)
  - Python v2.6.4 incorporating the Shapely v1.2.6 package used to determine intersection and union of all 20 speaker vowel space areas
Flynn (2011: 16)
Results of testing equalization of vowel space areas, 10 best methods only

<table>
<thead>
<tr>
<th>Method</th>
<th>SCV</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Hertz)</td>
<td>0.06212</td>
<td>N/A</td>
</tr>
<tr>
<td>Gerstman</td>
<td>0.01020</td>
<td>1</td>
</tr>
<tr>
<td>LCE</td>
<td>0.01487</td>
<td>2</td>
</tr>
<tr>
<td>Lobanov</td>
<td>0.02032</td>
<td>3</td>
</tr>
<tr>
<td>Bigham</td>
<td>0.02556</td>
<td>4</td>
</tr>
<tr>
<td>1mW&amp;F</td>
<td>0.02587</td>
<td>5</td>
</tr>
<tr>
<td>Letter</td>
<td>0.02637</td>
<td>6</td>
</tr>
<tr>
<td>origW&amp;F</td>
<td>0.02671</td>
<td>7</td>
</tr>
<tr>
<td>2mW&amp;F</td>
<td>0.02818</td>
<td>8</td>
</tr>
<tr>
<td>ERB</td>
<td>0.03233</td>
<td>9</td>
</tr>
<tr>
<td>Nearey1</td>
<td>0.03250</td>
<td>=10</td>
</tr>
<tr>
<td>NeareyGM</td>
<td>0.03250</td>
<td>=10</td>
</tr>
<tr>
<td>Log</td>
<td>0.03250</td>
<td>=10</td>
</tr>
<tr>
<td>Ln</td>
<td>0.03250</td>
<td>=10</td>
</tr>
<tr>
<td>method</td>
<td>% overlapping</td>
<td>rank</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>------</td>
</tr>
<tr>
<td>Bigham</td>
<td>45.8%</td>
<td>1</td>
</tr>
<tr>
<td>2mW&amp;F</td>
<td>43.8%</td>
<td>2</td>
</tr>
<tr>
<td>origW&amp;F</td>
<td>43.4%</td>
<td>3</td>
</tr>
<tr>
<td>1mW&amp;F</td>
<td>42.3%</td>
<td>4</td>
</tr>
<tr>
<td>Gerstman</td>
<td>30.0%</td>
<td>5</td>
</tr>
<tr>
<td>Lobanov</td>
<td>29.2%</td>
<td>6</td>
</tr>
<tr>
<td>Nordstrom</td>
<td>28.7%</td>
<td>7</td>
</tr>
<tr>
<td>exp{Nearey1}</td>
<td>27.6%</td>
<td>8</td>
</tr>
<tr>
<td>Nearey1</td>
<td>27.1%</td>
<td>9</td>
</tr>
<tr>
<td>exp{NGM}</td>
<td>26.9%</td>
<td>10</td>
</tr>
<tr>
<td>Bladon</td>
<td>25.9%</td>
<td>11</td>
</tr>
<tr>
<td>NeareyGM</td>
<td>25.7%</td>
<td>12</td>
</tr>
<tr>
<td>Letter</td>
<td>24.1%</td>
<td>13</td>
</tr>
<tr>
<td>LCE</td>
<td>23.1%</td>
<td>14</td>
</tr>
<tr>
<td>Bark-diff</td>
<td>13.5%</td>
<td>15</td>
</tr>
<tr>
<td>Bark</td>
<td>13.2%</td>
<td>16</td>
</tr>
<tr>
<td>Mel</td>
<td>13.1%</td>
<td>17</td>
</tr>
<tr>
<td>ERB</td>
<td>12.8%</td>
<td>18</td>
</tr>
<tr>
<td>Ln</td>
<td>12.2%</td>
<td>=19</td>
</tr>
<tr>
<td>Log</td>
<td>12.2%</td>
<td>=19</td>
</tr>
<tr>
<td>Hertz</td>
<td>12.6%</td>
<td></td>
</tr>
</tbody>
</table>

Flynn (2011: 17)

Results of testing overlap in vowel space areas
<table>
<thead>
<tr>
<th>method</th>
<th>SCV</th>
<th>area</th>
<th>total</th>
<th>overall rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigham</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Gerstman</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>1mW&amp;F</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>=3</td>
</tr>
<tr>
<td>Lobanov</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>=3</td>
</tr>
<tr>
<td>2mW&amp;F</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>=5</td>
</tr>
<tr>
<td>origW&amp;F</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>=5</td>
</tr>
<tr>
<td>LCE</td>
<td>2</td>
<td>14</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Letter</td>
<td>6</td>
<td>13</td>
<td>19</td>
<td>=8</td>
</tr>
<tr>
<td>Nearey1</td>
<td>=10</td>
<td>9</td>
<td>19</td>
<td>=8</td>
</tr>
<tr>
<td>NeareyGM</td>
<td>=10</td>
<td>12</td>
<td>22</td>
<td>10</td>
</tr>
</tbody>
</table>

Flynn (2011: 21). Overall rankings converted to points and then ranked. 10 best performing methods only
Example 3: Flynn and Foulkes 2011

- condensed version of Flynn (2011)
- illustrations of Hertz/Bigham/ERB

- ‘These results demonstrate the possibility of methods performing to **different levels of effectiveness** depending on the method of comparison used, and suggest evaluation of methods should ideally be based on a **range** of comparative tests.’
Example 4: Fruehwald (ms)

- to evaluate normalization algorithms’ efficacy in reducing male-female vowel space differences
  - compares **density functions** for F1, F2, F3
  - i.e. the likelihood of a formant appearing at a particular frequency
- uses some known methods and introduces adjustments to these by varying scaling factors
- data: 17 Philadelphia speakers (12 women, 5 men)
Figure 3. Density Distribution: All Formants

Fruehwald (ms, p.2): Distribution of formants on Hz scale
**Figure 4. ECDF: All formants**

Fruehwald (ms, p. 3): Empirical cumulative density function (Hz)
Fruehwald (ms, p. X): variant of W&F using a difference metric (two-factor scaling)
the vowel extrinsic techniques were more effective than the vowel intrinsic technique,
the formant intrinsic techniques were more effective than the formant extrinsic techniques,
two factor scaling techniques were more effective than single factor scaling techniques.

For more details, contact Josef at joseff@babel.ling.upenn.edu
what matters in the choice of a normalization algorithm:
  - explicitly testing a range of normalization procedures
  - using a range of test types
  - finding arguments to support a choice based on a range of factors
    - suitability of the typological choice (vowel extrinsic, formant intrinsic preferred)
    - purposes for which the data is being analyzed
    - optimal performance within those boundaries
    - time and budget parameters
  - aiming also to optimize comparability of results

the choice will thus be (to some extent) individual
Part 3: Metrics and the vowel space

- Phoneticians and sociolinguists’ practices in this area have tended to differ somewhat

- Measuring formant differences across one formant at a time
  OR

- Describing the vowel space two-dimensionally
Figure 4 Mean F1 and F2 frequencies (ERB) for each speaker in each age group for vowels showing the greatest differences between age groups, /ɛ ə ʊ/, as indicated in each panel. Each ellipse represents two standard deviations in F1 and F2 ERB-rate, calculated as described for figure 3. The scales are the same in each panel but encompass different parts of the frequency range. Solid symbols in each panel represent means for individual speakers within the age group identified as a break group for that vowel. See text for explanation.

Example: Hawkins and Midgley 2005:190
Can we bring in some of the benefits of Labovian-type holistic views?
Can we make two-dimensional comparisons of vowel plots more stringent and replicable by using mathematical methods?

Labov 1994: 167

“In Andersen’s speech, the membership of the New York City /æh/ class is quite regular, in that words like man, pass, half are lengthened and more peripheral than other words; but the raising to mid and high position characteristic of younger speakers has not actually begun, and (oh) is equally conservative. The conservative orientation of (ay) and (aw) is equally clear. They are both squarely located in central position, with no tendency toward fronting or backing.”

Figure: Labov 1984:168
Recent moves in this area

- Pillai-Bartlett statistic/Pillai scores (Hall-Lew 2009, Hay, Warren & Drager 2006)

- Quantifying angular relations between vowel points (Fabricius 2007) and using the centroid of the vowel space (Watt & Fabricius 2011)

- Mahalanobis distances (e.g. Esling 1986) or scaling to enable use of Euclidean distance (Fridland & Kendall, submitted)

- Procrustean analysis, relating acoustic and articulatory data (Geng & Mooshammer 2009)

TRAP and STRUT relative to the horizontal

(1) $\tan \Theta = \frac{(F_1 \text{ TRAP} - F_1 \text{ STRUT})}{(F_2 \text{ TRAP} - F_2 \text{ STRUT})}$

LOT and FOOT relative to the vertical

(2) $\tan \Theta = \frac{(F_2 \text{ FOOT} - F_2 \text{ LOT})}{(F_1 \text{ LOT} - F_1 \text{ FOOT})}$

Euclidean distance

(3) $\text{DISTANCE} \ (x,y) = \sqrt{(F_1 \ x - F_1 \ y)^2 + (F_2 \ x - F_2 \ y)^2}$

3.2 Watt and Fabricius 2011


- Derives its methodology from the S-centroid calculated by the normalization algorithm (1,1)
- ((Lobanov normalization also provides a centroid at (0,0) ))
Figure 5: Formant data for 20 RP speakers, $mW&F$-normalized; lines join speakers’ LOT and FOOT means.

Watt & Fabricius 1 normalized formant values
Figure 4: $S$-centroid anchor and angle values in 360° space
Figure 6: FOOT angles to centroid for 20 RP speakers

Figure 7: LOT angles to centroid for 20 RP speakers
3.3 Fridland and Kendall, under rev

- work examining the relationship between perception and production of vowel categories during regional vowel shifts
Southern Vowel Shift (SVS)  Northern Cities Shift (NCS)

Euclidean distance relies on scale equivalence: Lobanov handles this nicely (W & F actually performs slightly better, at least on datasets tested)
Not all speakers within a region participate in the regional shift, and speakers who do participate do so to varying degrees.
/ɛ/-/ɛ/ onset distances ordered by region

South  West  North

/el/ - /E/ onset distances (Lobanov norm'd)

Perception is related to an individual’s production, seen in the non-linear (fit by a 2\textsuperscript{nd} order polynomial) effect for Euclidean distance.
normalization (and plotting) decisions are crucial components of any sociophonetic vowel (and beyond?) study
- not normalizing is a valid decision, but crucially is still a decision

as sociophonetics matures, it is important that we strive to develop both shared best-practices and an innovative eye toward rigorous and appropriate quantitative techniques


Fridland, Valerie and Tyler Kendall. Under review. Exploring the relationship between production and perception in the mid front vowels of U.S. English


Bibliography III


See also [http://ncslaap.lib.ncsu.edu/tools/norm/biblio1.php](http://ncslaap.lib.ncsu.edu/tools/norm/biblio1.php)
Thanks

- Questions and discussion?

- Anne H. Fabricius, fabri@ruc.dk
- Tyler Kendall, tsk@uoregon.edu
- Dom Watt, dominic.watt@york.ac.uk
Normalization and Plotting

Workshop on Sociophonetic Methodology, LSA Summer Institute, Boulder, USA
July 2011

Anne Fabricius, Roskilde University, Denmark
Tyler Kendall, University of Oregon, USA
Dominic Watt, University of York, UK