Inference of VTL from formant frequencies

Rich Turner (turner@gatsby.ucl.ac.uk)

CNBH, 19/12/2006
How linear is $f(a)$?

How uniform is $f(a)$?

What about measurement noise & other sources of variability?
Formant Ratio Theory

Ratios of formants much less variable than the formants themselves.

i.e. \( \frac{\lambda_i}{\lambda_j} \cong c \) or

\[
\lambda = \begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\lambda_3
\end{bmatrix} = g \cdot a
\]

⇒ Vowels lie on lines that go through the origin
Measurement Noise

$P(f)$

- Formants can only be known to $\sim (\frac{1}{2} - \frac{1}{4}) f_0 = \sigma_{\beta_1}$

- Biases formant ratios & wavelengths as 

  the $\%$ error in $f_1 \gg \%$ error $f_2$

$\Rightarrow$ Need to model the data statistically.
Fitch's MRI Data

Is VTL growth linear?

\[ L_i = \frac{dL_i}{da} \alpha + \langle L_i \rangle \]

↑ cavities grow at different rates

Uniform \(\Rightarrow\) \[ \frac{dL_i}{da} \alpha \langle L_i \rangle \text{ i.e. } 2\times \text{longer} \]

↓ \[ 2\times \text{faster growth} \]
PREDICTION OF LINEAR GROWTH

\[ l_i = \frac{dl_i}{da} + \langle l_i \rangle \]

\[ L = \sum_i l_i = \sum_i \frac{dl_i}{da} a + \sum_i \langle l_i \rangle = \frac{dl_i}{da} a + \langle L \rangle \]

\[ \Rightarrow \frac{l_i - \langle l_i \rangle}{dl_i/da} = \frac{L - \langle L \rangle}{dL/da} \]

\[ \Rightarrow \frac{l_i}{L} = \frac{1}{L} \times \Delta + \beta \]

\[ \Delta = l_i - \langle L \rangle \frac{dl_i}{dL} \]
Reciprocal of the total VTL \( \frac{1}{m} \)

VTL section/VTL /no units

lip
blade
dorsum
velum
pharynx
oral length
women
men
linear
quadratic

Rich T.
First Model - Eqs.

\[
\begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\lambda_3
\end{bmatrix} =
\begin{bmatrix}
\frac{d\lambda_1}{da} \\
\frac{d\lambda_2}{da} \\
\frac{d\lambda_3}{da}
\end{bmatrix} a +
\begin{bmatrix}
\langle \lambda_1 \rangle \\
\langle \lambda_2 \rangle \\
\langle \lambda_3 \rangle
\end{bmatrix} +
\begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\varepsilon_3
\end{bmatrix}
\]

- assume \( \lambda_i = C \times \text{left} \)

\( \Rightarrow \) From Fitch, \( \lambda \propto a \) i.e linear

Uniform \( \Rightarrow \) \[ \begin{bmatrix}
\frac{d\lambda_1}{da} \\
\frac{d\lambda_2}{da} \\
\frac{d\lambda_3}{da}
\end{bmatrix} \] parallel to \[ \begin{bmatrix}
\langle \lambda_1 \rangle \\
\langle \lambda_2 \rangle \\
\langle \lambda_3 \rangle
\end{bmatrix} \]
First Model - Pictures

No Variability

\( \lambda_2 m_2 \)

\( \lambda_1 \)

everyone is identical
First Model - Pictures

No Variability  VTL Variability

\[ \lambda_0 \]
\[ \lambda_1 \]
\[ \lambda_2 \]

\( m_1 \)

everyone is identical

Vowels live on a line
**FIRST MODEL - PICTURES**

- **NO Variability**
- **VTL Variability**
- **Measurement Noise**

Diagram:
- Points labeled with indices: $\lambda_1$, $\lambda_2$, $\lambda_3$.
- Arrows indicating movement or direction:
  - $\lambda_1$:
    - "everyone is identical"
  - Vowels live on a line
  - "Fuzzy Pencil"
iy: distribution of VTLS is non-Gaussian
iy: there are outliers
iy: there is considerable natural variability
lots of noise in second formant or uniform scaling?

\[ \Rightarrow \text{Ambiguous} \]
Problems with 1st model

1. Distribution of UTLS is non-Gaussian

2. There are outliers

3. There is considerable natural variability

4. VTL variability & measurement noise can be confounded.
NEW MODEL

\[
\begin{bmatrix}
    f_1 \\
    f_2 \\
    f_3
\end{bmatrix}
= \mathbf{g} \mathbf{a} + \mathbf{M} + \mathbf{\xi} + \sum_{i=1}^{I} h_i b_i
\]

LEARN VIA VARIATIONAL EM

\[
p(\text{latent variables} \mid Y) = \prod_i q(S_i) q(a_i)
\]

mixture components \( S_i \)

UTLS \& \( a_i \)
Uniformity of the scaling

- PCA wavelength space: $\langle \theta \rangle = 8.4$
- FA wavelength space: $\langle \theta \rangle = 7.4$
- PCA frequency space: $\langle \theta \rangle = 6.0$
- FA-MOG wavelength space: $\langle \theta \rangle = 4.6$

Noise, $\sigma_y = 50.3 \, Hz = \text{prediction from Nyquist}$
CONCLUSIONS

1. Vowel formants depend on VTL almost LINEARLY

2. This linear dependence is almost UNIFORM

3. One needs a statistical model to see this in its full extent, which removes the effects of natural variability, measurement noise & outliers.
TO BE DONE ....

1. Automatic, soft outlier detection

2. Sharing of VTLs between clusters

3. Model Comparison

\[ f \text{ vs } f & \text{ } \text{ } \text{ free} \]

# of natural variability factors

4. See what the natural variability parameters are telling us.

5. Prediction: from formants of known speaker, can we predict others?

\[ \Rightarrow \text{ validation} \]