3aSC10.
Information conveyed by f0 for vowel identification.
Session: Wednesday Morning, Jun 07

Terrance M. Nearey
Univ. of Alberta
Peter F. Assmann
The Univ. of Texas at Dallas
[Work supported by SSHRC, NSF]

Background

• Speaker variation and perception
  ◦ Evidence f0 may help but role may vary [8]
  ◦ Vocoding: Changing f0 - formant relations degrades ID performance [1-5]
• We present possible accounts
  ◦ Speaker normalization perspective developed in [12]

Training Datasets

• /i, ɪ, ɛ, æ, ʌ, ɔ, ʊ, o, ɔ/ in h-V-d syllables
• H95 Hillenbrand et al. 1995 [7]
  ◦ 45 men, 48 women, and 46 children from western Michigan. (1 ‘take’ per v per speaker) (http://homepages.wmich.edu/~hillenbr/voweldata.html)
• A00 Assmann and Katz 2000 [2]
  ◦ 10 men, 10 women, and 30 children (ages 3, 5, and 7 years) from north Texas (multiple takes)
• P52 Peterson and Barney (1952) [13]
  ◦ 33 men, 28 women, 15 children (no / ɛ/, ɔ/) 2 takes each
Measurements

- \( f_0 = \) average \( f_0 \) of vowel
- \( F_1 \, F_2 \, F_3 \) - first 3 formants
  - For A00 and H95 two sections
    - \( F_1a \, F_2a \, F_3a \) at 20% of duration
    - \( F_1b \, F_2b \, F_3b \) at 80%
  - For P52 only one section (steady-state)

**NOTE:** Log values
- \([g_0, G_1, G_2, G_3] = \ln([f_0, F_1, F_2, F_3])\)

Three normalization methods

- Method A: log-mean normalization
  - (Some extrinsic direct info from speaker required)
- Method C: no explicit normalization
  - Intrinsic info only, apparently direct only
    - Arguably, indirect normalization implicit in covariance among formants frequencies and \( f_0 \)
- Method B: middle way
  - Intrinsic info only
  - Well-defined imputation of info related to A
- (Additional model studied in [12])

Notation: kinds of info

*(after Johnson 1990 [8])*

- Identifying for current vowel \( \nu \)
  - Intrinsic - info from \( \nu \) only
  - Extrinsic - some info from beyond \( \nu \)
  - Direct - info from signal only
  - Indirect - some info ‘imputed’ from beyond signal (‘knowledge’)

Method A:

**log-mean normalization**

Let

\[
\hat{\psi}_{s} = \bar{G}_{s} = \frac{1}{(3 \cdot V \cdot T)} \sum_{v} \sum_{t} G_{kvst}
\]

where \( G_{kvst} \) is the measured frequency of formant \( k \)

of “take” \( t \) of vowel \( \nu \) by speaker \( s \).

Define:

\[
\hat{N}_{kvs} = G_{kvs} - \hat{\psi}_{s}
\]

where \( \hat{N}_{kvs} \) is the normalized value of formant \( k \) for vowel \( \nu \) of

subject \( s \).
Constant log-interval hypothesis (CLH = log version of CRH [11])

\[ G_{k,v} = G^*_k + \psi_s \]

where

- \( G^*_k \) is the \( k \)-th log formant value for the reference pattern of vowel \( v \);
- \( \psi_s \) is a single-speaker-dependent displacement constant.

(Note: Each of the terms is the log of the corresponding elements in CRH = constant ratio hypothesis, \( F_{k,v} = F^*_k \cdot \rho_s \), where \( \rho_s \) is a single-speaker-dependent scale factor.)

Application of Method A

- Normalize each speaker’s data using log-mean normalization
- Measurement vector for A00, H95 is then
  - [N1a, N2a, N3a, N2a, N2b, N3b] for each vowel token
  - Note: No f0 (g0) information used!
- Apply LDFA
  - (linear discriminant function analysis)

Graphical interpretation of CLH (sliding template)

- Movement along diagonal for different speakers
- Position corresponds to \( \psi_s \) in CLH
- Fixed pattern of ‘holes’ in template correspond to reference pattern \( G^* \)

Method C: No explicit normalization

- Use raw log format and g0 measures
- Measurement vector per vowel token is:
  - [g0, G1a, G2a, G3b, G1b, G2b, G3b]
  - All info from single syllable
- Apply LDFA
Graphs related to LDFA info

- Illustration of in Method A and Method C in reduced space
- Next 2 panels
  - Plot of 1-s.d. ellipses in N1a x N2a space of mean and covariance pdfs of Method A
  - Plot of ellipsoids in g0 x G1a x G2a space of Method A
  - **NB**: these are empirical estimates from A00 data with no editing

(i) Method A [N1a x N2a]

Note: little correlation between N1a N2a

Full Method A is 6-dimensional

Method B: Preliminaries

- Method A reduces correlation among variates due to speaker variation (expected by CRH)
- Problem: Don’t have good estimate of log mean $\bar{G} (\approx \psi_s)$ until you have lots of data
- But we can estimate $\psi_s$ from stable statistical patterns shown below
  - And (!) from good match of template with incoming vowel at estimated $\psi_s$
(ii) Relation between $g_0$ and $\bar{G}$

Data from all speakers in H95, A00 and P52

Regression lines
- $G$ on $g_0$
- $g_0$ on $G$

Method B strategy

- Apply *structure* of Meth. A (normalized training data)
  - ...even if speaker averages are not available (e.g., talkers randomized)
  - Requires estimation of a latent $\psi_T$ ($\approx \bar{G}$)
- Choose $\psi_S$ (position of template) to provide:
  - (i) Good match to a vowel category
  - (ii) Good match with average relation of $g_0$ and $\bar{G}$
  - (iii) Good match usual ranges of $\psi_S$
  - (Compare: Roman numbering on prior panels)

(Technical sketch at end of poster)

(iii) Marginal distribution of $\psi_S$:
limits on human range?

Data from all speakers in H95, A00 and P52

Cross-validated classification

- LDFA used with each model
  - Hedge: Method B requires optimization of a latent $\psi_T$ for each vowel category for each signal
    - But uses method similar to LDFA in Method A
  - Compare cross validation classification rates of the three methods
    - Leave-out-one-talker, train on rest
    - Classify that token
    - Error bars show estimated standard error of classification rate
**LDFA Classification of H95**

![Bar chart showing classification accuracy for H95](image)

**LDFA Classification of P52**

![Bar chart showing classification accuracy for P52](image)

**LDFA Classification of A00**

![Bar chart showing classification accuracy for A00](image)

**Summary Classification**

- Method B does about same as Method C, using f0 info only indirectly
- Involves explicit estimation of speaker scaling parameter $\psi_s$
  - Makes predictions about apparent speaker ‘size’
- With more speech from same speaker: $\psi_s (\approx \overline{O})$ can be updated
- Eventually, Method B should match Method A
Preliminary Perceptual Results

- 11 vowels in hVd words, 2 male voices
- STRAIGHT vocoder (Kawahara, 1997)
  - formant frequency scale factors (0.6–2.0)
  - f0 scale factors (0.25–8.0)
- Listeners’ ID rate declines:
  - When f0 and formants are scaled beyond natural human range
  - Even more when f0/formant scaling violates natural covariation between f0 and formants in speakers of different age/sex classes.

Method C predictions * v. listeners’ resp. O

Discussion of perceptual results

- Compared to listeners
  - Method C trained on production data shows too great a sensitivity to f0 x formant covariation
  - Method B trained on production data is a little less sensitive (not illustrated)
  - Method B can be ‘tuned’ (as illustrated) by inflating error variance of regression of g0 on $\bar{G}$
Caveats on perceptual modeling

- Method B does better job than Method C in predicting listeners’ responses
- But there are still clear mismatches between tuned Method B behavior and listeners
  - Listeners do worse than predicted in upper left and and lower right corners
    - Especially with high f0 where F1 may be poorly defined
  - Predicted confusion matrices show definite discrepancies with listeners
  - Research in progress to solve remaining problems

References


Optimization formula for Method B

\[ \hat{\psi}_{v,d} = \arg \max_{\psi} [P(G_1 | v, \psi^d) \cdot P(g_0 | \psi^d) \cdot P(\psi)] \]

- \( G_1 \) represents the log formants of the current token \([G1a, G2a, G3a, G1b, G2b, G3b]\);
- \( g_0 \) is the log f0 of the current token;
- \( \arg \max \) returns the value of \( \psi \) that maximizes the quantity in \([ \cdot ]\).

There are three components related to three plots above:

- (i) \( P(G_1 | v, \psi^d) \) is the multivariate normal probability density for vowel \( v \) when \( G_1 \) is normalized by subtracting current \( \psi^d \) (generalization of 2-D ellipse plot)
- (ii) \( P(g_0 | \psi^d) \) is a regression estimate of likelihood of the current g0, assining \( \psi^d \) (see regression plot)
- (iii) \( P(\psi^d) \) is the marginal probability of \( \psi^d \) (see histogram plot)