Estimating Phoneme Formant Targets and Coarticulation Parameters of Clear and Conversational Speech

Brian O. Bush

Oregon Health & Science University
Center for Spoken Language Understanding

December 7, 2012
Outline

1 Introduction

2 Background
   - Speaking Styles
   - Coarticulation
   - Locus Theory

3 Our Study
   - Formant Trajectory Model
   - Estimating Model Parameters
   - Analytic Results
   - Perceptual Study

4 Conclusion
Introduction: Motivation

- When people speak clearly, speech becomes more intelligible.
- However, attempts to modify speech by computer to improve intelligibility are not yet very successful.
- We need to understand which parts of the speech signal should be modified and how to modify them.
- Overall, a better understanding of how acoustic features contribute to speech intelligibility could guide research on improving ASR and TTS.
Research Objective

To identify and quantitatively model the relative contribution of acoustic features to intelligibility by examining clear and conversational speech.
Outline

1 Introduction

2 Background
   - Speaking Styles
   - Coarticulation
   - Locus Theory

3 Our Study
   - Formant Trajectory Model
   - Estimating Model Parameters
   - Analytic Results
   - Perceptual Study

4 Conclusion
Speaking Styles: Defined

Clear speech and conversational speech are defined as

- **Clear** \((\text{CLR})\): speech spoken clearly when talking to a hearing-impaired listener
- **Conversational** \((\text{CNV})\): speech spoken when speaking with a colleague

There are several acoustic differences between “Clear” and “Conversational” speech. Clear speech usually differs from conversational in the following aspects:

- Typically has longer duration
- Expanded vowel space (F1/F2)
- Expanded fundamental frequency (F0) mean and variance
Figure: Comparing CNV and CLR styles. Note the expanded vowel space in CLR vs CNV; shift towards neutral vowel in CNV.
Figure: F1/F2 formant frequency relationship between CNV and CLR at vowel centers, arrows pointing from the former to the latter.
Coarticulation: Example

(a) Separately pronounced phonemes, /yu/-/uw/-/aa/-/r/.

(b) The phrase “you are” spoken at natural speed.
Locus Theory and Formant Targets

**Locus Theory** (Delattre *et al*, 1955)

- Each consonant has “target frequencies” independent of the neighboring vowels
- Formants transition from these target frequencies to the vowel target frequencies
- Consonants and vowels both have targets of articulator positions and corresponding formant frequency locations
- Given sufficient duration of a syllable, all phonemes reach their targets
Outline

1. Introduction

2. Background
   - Speaking Styles
   - Coarticulation
   - Locus Theory

3. Our Study
   - Formant Trajectory Model
   - Estimating Model Parameters
   - Analytic Results
   - Perceptual Study

4. Conclusion
Objective

Key idea

Estimate global speaker-specific phonetic formant targets for both CLR and CNV speech using an explicit coarticulation model.

Hypothesis

Phoneme formant targets are speaker dependent, but consistent between speaking styles.
Parallel Style Corpus

- 1 male, native speaker of American English
- Sentences contain neutral carrier phrase (5 total) followed by a keyword (212 total) in sentence final context
  - e.g. I know the meaning of the word will
- Affricates and diphthongs not represented
- Keywords are common English CVC words with 21 initial and final consonants and 8 monophthongs
  - All sentences spoken in both clear and conversational styles
  - Two recordings per style of each sentence
  - Total number of keyword tokens: $212 \times 2 \times 2 = 848$
We adopt a model, similar to which was used in Broad and Clermont (1987), Niu and Santen (2003), Amano et al (2010).

Broad and Clermont originally employed an exponential function to model coarticulation functions in the context of voiced stops (/b, d, g/), Amano et al changed the coarticulation function to be a sigmoid.

This study extends model by (1) modeling trajectories asynchronously and (2) expands the range of evaluation from the vowel region to include half of the neighboring consonants.
CVC Model – Defined

Definition

An individual formant trajectory $X(t)$ of a CVC word is modeled as a convex linear combination of target formant values

$$\hat{X}(t; \Lambda) = d_{C_l}(t) \cdot T_{C_l} + d_V(t) \cdot T_V + d_{C_r}(t) \cdot T_{C_r}$$

- $T_{C_l}$, $T_V$, and $T_{C_r}$ are the per-formant target values for
  - prevocalic consonant $C_l$
  - vowel $V$
  - postvocalic consonant $C_r$
- $d_{C_l}(t)$, $d_V(t)$ and $d_{C_r}(t)$ are coarticulation functions
CVC Model – Coarticulation Functions

- The coarticulation functions are based on the sigmoid
  \[ d(t; s, p) = \left(1 + e^{s \cdot (t - p)}\right)^{-1} \]

  \[
  d_{C_{l}}(t; s_{l}, p_{l}) = d(t; s_{l}, p_{l}) \\
  d_{C_{r}}(t; s_{r}, p_{r}) = d(t; -s_{r}, p_{r}) \\
  d_{V}(t) = 1 - d_{C_{l}}(t) - d_{C_{r}}(t)
  \]

- \( s \) represents sigmoid slope and \( p \) sigmoid midpoint position

- Parameters \( \Lambda = \{ T_{C_{l}}, T_{V}, T_{C_{r}}, s_{l}, p_{l}, s_{r}, p_{r} \} \) are specific to a single formant trajectory, and thus the model approximates concurrent formant trajectories asynchronously.
We define the per-token root-mean-squared (RMS) model error as

$$E(X, \Lambda) = \sqrt{\frac{1}{t_r - t_l} \sum_{t=t_l}^{t_r} w(t) \cdot (X(t) - \hat{X}(t; \Lambda))^2}$$

- $t_l$ is the middle of $C_l$ and $t_r$ is the middle of $C_r$
- $w$ is the confidence of formant measurements at time $t$, calculated from formant bandwidth
Token “will” Example

Figure: Example of the model on the word “will” in CNV (left) and CLR (right) speech.
Given a single token, we sweep parameter $\lambda \in \Lambda$ along a prescribed interval, while grid-searching for the lowest model error at each given value of $\lambda$

$$E_{\text{sweep}}(\lambda) = \min_{\Lambda \setminus \lambda} E(X, \Lambda)$$

We perform this for both $F1$, $F2$, and $F3$ separately

with intervals

- $T_{F1} = 200, 220, \ldots, 900$ Hz
- $T_{F2} = 400, 420, \ldots, 2800$ Hz
- $T_{F3} = 900, 920, \ldots, 3700$ Hz
- $s = 10, 30, \ldots, 110$ Hz/s
- $p = -40, -30, \ldots, 40$ ms, relative to the phoneme boundary

and condition

- $d_{C_l}(t) + d_{C_r}(t) \leq 1 \forall t$
Figure: Example token $E_{\text{sweep}}(T_{C_i})$ for /n/ in the token “neck”
Global Sweep

- We assume the existence of global targets for each phoneme.
- Define the global minimum error sweep as the average minimum error sweeps for target $T$, over all $N$ tokens that involve a particular phoneme target.

$$E_{global}(T) = \frac{1}{N} \sum_{t \in token} E_{sweep}(T)$$
**Global Minimum Error Sweep Example**

![Graph showing error sweep](image)

**Figure:** $E_{global}(T_{C_l})$, $E_{global}(T_{C_r})$, and their combination $E_{global}(T_C)$ for $C = /n/$ for F2 (there are more /n/ in postvocalic context)
Using global phoneme targets, we re-estimate the optimal set of coarticulation parameters $s_l, p_l, s_r, p_r$ on a per-token basis.

To better approximate the formant dynamics we substitute $\Delta X(t)$ and $\Delta \hat{X}(t)$ for $X(t)$ and $\hat{X}(t)$ in the per-token RMS model error formula:

$$
\Delta E(X, \Lambda) = \sqrt{\frac{1}{t_r - t_l} \sum_{t=t_l}^{t_r} w(t) \cdot (\Delta X(t) - \Delta \hat{X}(t; \Lambda))^2}
$$

The re-estimation uses fine-grained intervals for search, with $s = 10, 11, 12, \ldots, 110$ and $p = -80, -70, -60, \ldots, 80$ ms.
Vowel Targets

**Figure:** Vowel targets (circles) compared with Hillenbrand et al (x). F1 (red), F2 (green) and F3 (blue).
Consonant Targets

Figure: Consonant targets (circles) and Allen et al (x). F1 (red), F2 (green) and F3 (blue).
Figure: Vowel formant targets with observed data and iso-contours based on the global minimum of $E_{sweep}$
Approximant Contours

Figure: Approximant formant targets with observed data and iso-contours based on the global minimum of $E_{\text{sweep}}$
Nasal Contours

Figure: Nasal formant targets iso-contours based on the global minimum of $E_{\text{sweep}}$
Voiced and Unvoiced Stop Contours

Figure: Voiced and unvoiced stop formant targets iso-contours based on the global minimum of $E_{sweep}$
Perceptual Study – Goals

- **Goal**: Test if resynthesis using model parameters and global targets produces intelligible speech
- 212 CVC words used in perceptual listening test
- **Six stimulus conditions**: CLR, CNV, CLR-vocoded, CNV-vocoded, CLR-model, CNV-model
- **Total**: 212 words × 6 conditions = 1272 stimuli
- Stimuli were loudness normalized and 12-talker babble noise added (+3 dB SNR)
- Resynthesis used linear predictive coding with energy and pitch trajectories preserved
Perceptual Study – Administration

- Subjects: 18 adults aged 23-55, all native speakers of American English
- Listened to CVC stimuli through headphones in a quiet room
- Each listener presented 212 stimuli in random order
- For each stimuli, listener was presented five possible answers to the question “What did you hear?”
- Four of the terms were decoy terms, selected based on closest phonetic similarity to the target term, using a list of common CVC words (e.g. “fan”, “van”, “than”, “pan” and “ban”)
Perceptual Study – Raw Results

<table>
<thead>
<tr>
<th>Style</th>
<th>Natural</th>
<th>Vocodered</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR</td>
<td>94.05% (3.68)</td>
<td>88.93% (5.38)</td>
<td>84.83% (4.77)</td>
</tr>
<tr>
<td>CNV</td>
<td>85.53% (7.55)</td>
<td>68.75% (7.25)</td>
<td>66.77% (9.26)</td>
</tr>
<tr>
<td>CLR &amp; CNV</td>
<td>89.69% (4.07)</td>
<td>78.82% (5.57)</td>
<td>75.70% (5.74)</td>
</tr>
</tbody>
</table>

Table: Average intelligibility rate for each condition averaged across all listeners. Standard deviation in parenthesis.
Perceptual Study – Significance

- A one-sample two-tailed $t$-test was used to compare the vocoded and model conditions (value of 2.2, with 17 degrees of freedom, significant at 0.05; effect size of 0.6 considered “moderate”)
- Showed that vocoded and model condition means are significantly different
Conclusions

- Developed a new data-driven methodology to estimate style and context-independent vowel and consonant formant targets for one speaker.
- Intelligibility test validated CVC words using modeled formant trajectories were nearly as intelligible as observed formant trajectories.
- Demonstrated that targets are consistent between styles.
- Compression ratio of 1:12 for our corpus achieved using model parameters vs raw formant trajectories.
Future Work

- Can we increase intelligibility of \textit{CNV} CVC words by converting to \textit{CLR} using model?
- Investigate phonemes that seemingly have target ranges vs points (e.g. /k/)
- Investigate optimized techniques for estimating global phoneme targets
- Generalization of the model to continuous speech
  - Application of model to an expanded set of speakers (including female)
  - Expanded phoneme support for affricates (/j/ and /ch/) and diphthongs (e.g. /ow/)
- Application of model to various domains such as complete TTS, ASR, and dysarthria diagnosis
Questions

Thank you!
This material is based upon work supported by the National Science Foundation under Grant IIS-091575. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
This slide intentionally left blank.
Histograms of Coarticulation Function Values (cont)

Figure: Histogram of $d_{cl}(t_i)$
Figure: Histogram of $d_{cr}(t_r)$
### Minimum Error Sweep Example (cont)

<table>
<thead>
<tr>
<th>$T_{C_l}$</th>
<th>$T_V$</th>
<th>$T_{C_r}$</th>
<th>$s_1$</th>
<th>$p_1$</th>
<th>$s_r$</th>
<th>$p_r$</th>
<th>$E_{sweep}(T_{C_l})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda_1$</td>
<td>620</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>-0.04</td>
<td>30</td>
<td>-0.08</td>
</tr>
<tr>
<td>$\Lambda_2$</td>
<td>1080</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>-0.03</td>
<td>30</td>
<td>-0.08</td>
</tr>
<tr>
<td>$\Lambda_3$</td>
<td>1360</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>-0.02</td>
<td>30</td>
<td>-0.08</td>
</tr>
<tr>
<td>$\Lambda_4$</td>
<td>1540</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>-0.01</td>
<td>30</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

**Table:** $\Lambda$ values for four frequencies in $C_l$ that have near equal value in error. Note the interaction between $T_{C_l} \backslash p_1$
Analysis of $s_l$ parameters

Figure: Linear relationship in CLR approximants between observed F2 slope at vowel onset and the estimated model parameter $s_l$ for F2. CNV approximants do not exhibit such a defined relationship.
Figure: Unvoiced fricative formant targets iso-contours based on the global minimum of $E_{\text{sweep}}$
Figure: Voiced fricative formant targets iso-contours based on the global minimum of $E_{\text{sweep}}$
Histograms of Coarticulation Function Values

Figure: Histogram of max $d_v(t)$. Note the CLR vs CNV differences.
Perceptual Study – Clear Speech Effect

- Examining speech style separately, CLR style is consistently better than the corresponding condition in CNV style.
- Demonstrates the CLR clear speech effect, wherein a larger acoustic vowel space leads to increased intelligibility.
- Accuracy for natural, 94% CLR vs 85.5% CNV.