Thesis Proposal: Modeling Coarticulation in Continuous Speech

Brian O. Bush

Oregon Health & Science University
Center for Spoken Language Understanding

June 2, 2014
Outline

1. Introduction
2. Background
3. Thesis Proposal
4. Accomplished Work
5. Proposed Work
6. Conclusion
# Table of Contents

1. **Introduction**

2. **Background**

3. **Thesis Proposal**

4. **Accomplished Work**

5. **Proposed Work**

6. **Conclusion**
Introduction

- At an abstract linguistic level, phonetic segments ([w], [ih], [l], [sh], etc.) are discrete, independent, interchangeable parts.
- The word **will** (/w/-/ih/-/l/) is simply built from three distinct parts.
- Swapping the /l/ for a /sh/, we can easily get **wish** (/w/-/ih/-/sh/).
- However, at the level of speech production this clean abstract linguistic level breaks down.
Coarticulation: Example (1)

Figure: Formants for “w-ih-l” (will) spoken clearly
Figure: Formants for “w-ih-sh” (wish) spoken clearly
Coarticulation Defined

- What is happening in will versus wish?
- Coarticulation is the “blending” of adjacent speech sounds, primarily due to gradual movement of articulators.
- Boundaries that we label speech with are simply an approximate point of transition between phonemes – there are no clean boundaries.
- This blending is mostly due to inertia – the tendency of a stationary body to resist movement or for a moving body to resist change in rate/direction.
- The more massive an articulator, the greater its inertia or resistance to movement. e.g. Tongue body is heavier than tongue tip.
Components of Coarticulation

Locus Theory (Delattre et al. 1955)

- Consonants and vowels both have “targets” of articulator positions and corresponding formant frequency locations
- Thus, each consonant has a “target frequencies” that is independent of the neighboring vowel(s)
- Formants transition from these target frequencies to the vowel target frequencies
- Given sufficient duration of a syllable, all phonemes reach their targets
Modeling Coarticulation in Continuous Speech

Brian O. Bush

Introduction

Background

Thesis Proposal

Accomplished Work

Interspeech 2011

ICASSP 2013

Interspeech 2014

Proposed Work

Conclusion

Table of Contents

1 Introduction

2 Background

3 Thesis Proposal

4 Accomplished Work

5 Proposed Work

6 Conclusion
Modeling Coarticulation

- There have been several attempts modeling the phenomenon
- Öhman (1966) attempted to characterize coarticulation by examining VCV words
- Modern ASR systems use Wickelgren’s model (1969) wherein speech units are coded as context-dependent units
- However, our path will focus on explicit models of coarticulation
Broad and Clermont (1987) produced several models of formant transition in vowels in CV and CV/d/ contexts. The most detailed model used a linear combination of coarticulation functions and target values. Coarticulation functions modeled with exponential functions. Consonants limited to voiced stops /b,d,g/.
Niu and van Santen (2003) applied Broad and Clermont’s CV/d/ model to Dysarthria to measure coarticulation.

- Expanded model application to a broader set of consonants
- However, modeling was limited to vowel centers
- **Results:** Coarticulation effects of a dysarthric speaker were higher than normal speaker.
Amano and Hosom (2010) expanded upon Niu and van Santen by modeling the entire vowel region of a CVC. Region of evaluation extended to consonant center if consonant was an Approximant (/w,y,l,r/). Changed exponential to sigmoid as coarticulation function.

Results: Applied model to formant tracking error detection and correction.
The primary objective of this thesis: is to show that targets of acoustic events, including classes of sounds where formants are not directly observable, can be automatically derived from continuous speech.

Secondary objectives are:

- Validation of modeling approach using a synthetic corpus
- Automatic error correction in formant tracking
- Investigating clear and conversational speech targets
- Dysarthria diagnosis wherein we might be able to infer reduction vowel space area
1. Introduction
2. Background
3. Thesis Proposal
4. Accomplished Work
5. Proposed Work
6. Conclusion
Bush et al. (Interspeech 2011)

- **Primary objective:** Estimate speaker F1/F2 targets from data
- Tied phoneme targets: hypothesized to be global, whereas coarticulatory functions are token specific
- Employed real-coded genetic algorithm versus hill-climbing, which showed signs of premature convergence
Parallel Style Corpus

- One male, native speaker of American English
- Sentences contain neutral carrier phrase (5 total) followed by a keyword (242 total) in sentence final context
  - e.g. I know the meaning of the word will
- Keywords are common English CVC words with 23 initial and final consonants and 8 monophthong vowels
  - All sentences spoken in both clear and conversational styles
  - Two recordings per style of each sentence
  - Total number of keyword tokens: $242 \times 2 \times 2 = 968$
- Diphthongs not represented
Triphone Trajectory Model – Defined

Definition

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

$$\hat{X}(t; \Lambda) = f_L(t) \cdot T_L + f_C(t) \cdot T_C + f_R(t) \cdot T_R$$

- $T_L$, $T_C$, and $T_R$ represent global formant target values for three consecutive acoustic events
- $f_L(t)$, $f_C(t)$ and $f_R(t)$ are coarticulation functions
Triphone Trajectory Model – Coarticulation Functions

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

  \[ f_L(t; s_L, p_L) = f(t; s_L, p_L) \]
  \[ f_R(t; s_R, p_R) = f(t; -s_R, p_R) \]
  \[ f_C(t) = 1 - f_L(t) - f_R(t) \]

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

- parameters \( \Lambda = \{T_L, T_C, T_R, s_L, p_L, s_R, p_R\} \) are specific to a single formant trajectory \( \rightarrow \) asynchronous model
Triphone Trajectory Model – Example

Figure: F2 model on the triphone “w-ih-l” (will) in CLR speech
We define the per-token model error as

$$E(X, \Lambda) = \sqrt{\frac{\sum_{t=t_L}^{t_R} \left( X(t) - \hat{X}(t, \Lambda) \right)^2}{t_R - t_L}}$$

where $X(t)$ and $\hat{X}(t, \Lambda)$ are observed and estimated individual formant trajectories. The error is evaluated over $t_L$ to $t_R$ where $t_L$ is the center of the first phone, and $t_R$ is the center of the final phone.
Results

Targets clustered well, which fit our expectations from knowledge of acoustic-phonetics

- Bilabials are consistently clustered around 1200 Hz for F2
- Alveolar fricatives /s/ and /z/ are located in the same region of F1/F2 space, as expected as they differ only in voicing. However, lower in F2 than expected
- Approximants tightly clustered at expected locations
- **Deficiencies:** Coarticulation parameters (primarily s) exhibited little consistency
Figure: Estimated formant targets in F1/F2 space for $C_1$ phonemes /p/, /t/, /k/, /s/ and /h/
Consonant Target Results (2)

Figure: Estimated formant targets in F1/F2 space for $C_1$ phonemes /b/, /d/, /g/ and /z/
Figure: Estimated formant targets in F1/F2 space for $C_1$ phonemes /m/, /n/, /ng/, /w/, /y/, /r/ and /l/
Bush and Kain (ICASSP 2013)

- **Primary objective:** Characterize parameter space of coarticulation model
- Used a brute force search of parameter space
- Weighting of trajectory using formant bandwidth as confidence measure
Triphone Trajectory Model – Weighted Model Error

Definition

We define the per-token *model error* as

\[
E(X, \Lambda) = \sqrt{\frac{1}{\sum_{t=t_L}^{t_R} w(t)} \sum_{t=t_L}^{t_R} w(t) \cdot \left( X(t) - \hat{X}(t; \Lambda) \right)^2}
\]

where \( X(t) \) and \( \hat{X}(t) \) are the observed and modeled individual formant trajectory. The weighting factor \( 0 \leq w(t) \leq 1 \) indicates our confidence in formant measurement.
Estimating Model Parameters

- 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 performed this for both F1, F2, and F3 separately
Figure: Example token $E_{sweep}(T_L)$ for /n/ in the token “neck” in F2
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

$$E_{global}(T) = \frac{1}{N} \sum_{t\in token} E_{sweep}(T)$$
Figure: $E_{global}(T_L)$, $E_{global}(T_R)$, and their combination $E_{global}(T)$ for /n/ for F2 (there are more /n/ in postvocalic context)
Figure: Vowel formant targets with observed data and iso-contours based on the global minimum of $E_{sweep}$
Figure: Voiced and unvoiced stop formant target iso-contours based on the global minimum of $E_{sweep}$
Perceptual Study – Stimuli

- **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-observed, CNV-observed, 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
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”)

Results: natural 90% (4), vocoded observed 79% (5.5) vocoded model 76% (5.7), Average intelligibility rate. Standard deviation in parenthesis
Primary objective: Coarticulation modeling of continuous speech

- Developed a joint optimization technique that robustly estimates coarticulation parameters
- Uses triphone models as local models of coarticulation
- Creates continuous speech by cross-fading overlapping triphone models
- Handles two primary components of plosives (closure, release) separately, this approach could be used for diphthongs
Example of local coarticulation functions for F2:

Figure: Example of the word “use” (/y-uw-z/)
Continuous Coarticulation (2)

Now take local coarticulation functions modeling identical acoustic events are cross-faded to form a global, continuous coarticulation function:

\[
ge_{i}(t) = \begin{cases} 
    0 & t < c_{i-2} \\
    \alpha(t) \cdot f_{i-1,e_i}(t) & c_{i-2} \leq t < c_{i-1} \\
    \alpha(t) \cdot f_{i,e_i}(t) + \beta(t) \cdot f_{i-1,e_i}(t) & c_{i-1} \leq t < c_i \\
    \alpha(t) \cdot f_{i+1,e_i}(t) + \beta(t) \cdot f_{i,e_i}(t) & c_i \leq t < c_{i+1} \\
    \beta(t) \cdot f_{i+1,e_i}(t) & c_{i+1} \leq t < c_{i+2} \\
    0 & t \geq c_{i+2} 
\end{cases}
\]

where \(\beta(t) = 1 - \alpha(t)\), for \(i = 3, 4, \ldots, N - 2\), where for each case, \(\alpha(t)\) is a distinct cross-fading function defined to be zero at the left side of the interval, and monotonically increasing to one at the right side of the interval.
Continuous Coarticulation (3)

Example of global coarticulation functions for F2:

**Figure:** Example of the word “use” (/y-uw-z/)
Parameter Estimation

The parameter estimation strategy is nested hill-climbing with restart. Parameter set consists of:

- **Targets (global)**
- **Coarticulation functions (local to each triphone model)**
Estimating Targets

- Initialize targets to median formant frequency at phoneme centers for all phonemes
- Use hill-climbing to optimize targets
- At each iteration, we find optimal coarticulation parameters: $s$ and $p$ parameters
- Intervals:
  - $F_1 = 200, 250, \ldots, 1000$ Hz
  - $F_2 = 400, 450, \ldots, 3000$ Hz
  - $F_3 = 900, 950, \ldots, 4000$ Hz
  - $F_4 = 3000, 3050, \ldots, 6000$ Hz
Estimating Coarticulation Parameters

- For each existing triphone type, we consider its target parameters \((T_L, T_C, T_R)\) and identify all trajectories belonging to that triphone.
- We then jointly estimate \(s_L\) and \(s_R\) values by a secondary hill-climbing method.
- Finally, for each \(s\) value we use a tertiary hill-climbing method to estimate optimal \(p_L\) and \(p_R\) parameters.
- Intervals:
  - \(s = 10, 20, \ldots, 150\)
  - \(p = -80, -70, \ldots, 80\) ms, relative to the phoneme boundary.
Perceptual Study

- Using Amazon’s Mechanical Turk crowd-sourcing service, we employed 120 adults in a perceptual test.
- Each HIT was $0.40 USD and contained 48 stimuli.
- 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”)
- **Results:** natural 91% (10), vocoded observed 80% (13) vocoded model 70% (15), average intelligibility rate. Standard deviation in parenthesis.
- **Caveats:** No control over listener’s environment.
Table of Contents

1 Introduction
2 Background
3 Thesis Proposal
4 Accomplished Work
5 Proposed Work
6 Conclusion
Target Estimation

We propose a procedure that combines global target and coarticulation function parameter estimation.

1. Initialize targets: (1) observed, or Allen (1987)
2. Measure initial global error $G_0$
3. Sequentially optimize one target in $T$:
   1. For each utterance
      1. Measure utterance error $g$
      2. Optimize $s_{1,L}$ and $p_{1,L}$
      3. Optimize $s_{2,R}$ and $p_{2,R}$, and $s_{3,L}$
      4. Continue moving right
      5. At end, optimize $s_{N-1,R}$ and $p_{N-1,R}$
   2. Global error $G = \sum g$
4. If $G_i - G_{i-1} < \epsilon$, the error is not changing significantly, exit, otherwise continue to step 3
Validation of modeling approach using a synthetic corpus

- Using known targets $T$, generate trajectories using a speech corpus
- Find targets using these synthetic trajectories
- Resultant targets should approximate original targets $T$
- This would demonstrate that we can recover unknown targets from speech
Application – Formant tracking

- **Goal:** Use coarticulation model to model formant trajectories and detect and correct when errors are present.
- Studies concerned with the fundamentals of speech production often use formant features.
- Traditional formant tracking is known to make many errors.
- **Caveat:** phoneme segmentation is provided.
Goal: Are CLR targets sufficient to model CNV style speech? Does the opposite hold?

Four stimulus conditions (train/test): CLR/CLR, CLR/CNV, CNV/CLR and CNV/CNV

<table>
<thead>
<tr>
<th>Matched</th>
<th>Mismatched</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV/CNV</td>
<td>CLR/CNV</td>
<td>CNV + CLR/CNV</td>
</tr>
<tr>
<td>CLR/CLR</td>
<td>CNV/CLR</td>
<td>CNV + CLR/CLR</td>
</tr>
</tbody>
</table>
Application – Dysarthria

- **Goal:** Measure reduction in vowel space area using coarticulation model
- Dysarthria is a diverse group of motor speech disorders that typically impair intelligibility
- Several factors of dysarthric speech:
  - *target shift* in dysarthric speakers might differ from non-disorder speakers
  - *increased coarticulation* as the contextual influence on articulation might be greater
Table of Contents

1 Introduction
2 Background
3 Thesis Proposal
4 Accomplished Work
5 Proposed Work
6 Conclusion
Conclusions

- Data-driven model to estimate context-independent vowel and consonant formant targets for continuous speech
- Primary focus of proposed work will be target estimation, validation and applying model to a diverse set of applications
Thank you!