Towards the Recovery of Targets from Coarticulated Speech for Automatic Speech Recognition

162nd Meeting of ASA / CSLU

Alexander Kain, Brian Bush, and John-Paul Hosom

Center for Spoken Language Understanding
Department of Biomedical Engineering
School of Medicine
Oregon Health & Science University

November 8, 2011

1 now with Sensory, Inc.
Outline

1. Introduction

2. Background
   - Previous approaches to model coarticulation in ASR
   - Clear speech
   - Formant targets and locus theory

3. Our Study
   - Experiment
   - Model fitting
   - Vowel recognition

4. Conclusion
Phoneme ASR on the TIMIT database

- Trained HMM/ANN system on TIMIT, decoded with bigram language model [Hosom et al. 2010]
- Accuracy of 74% is high for HMM-based systems
- Vowel substitutions account for 35% of errors, covering all distinctive-phonetic feature dimensions:
  - 62% front/back error
  - 52% tense/lax error
  - 68% height error
  - 31% vowel/consonant error
- Errors are not confined to specific type
The cause of ASR errors

Hypothesis

- The main cause of ASR errors is not the feature space or the classification technique (which might result in more distinct error patterns), but in *noisy* probability estimates.
- This noise is caused by *variability in the features*, which can be reduced by estimating phoneme *targets* instead of the observed values.

- For now, we work in formant space, although targets need not be limited to this space.
- To control for speaking style, we are interested in both “clear” and “conversational” speech.
Coarticulation in ASR

Most of the coarticulation models currently employed in ASR follow Wickelgren’s theory [Wickelgren 1969]

- speech units are code as context-dependent units
- e.g. phoneme /ae/ in the context of /t/-/ae/-/b/ is represented using a unit that is entirely different from a phoneme /ae/ in the context /b/-/ae/-/t/.
- In ASR the states in an Hidden Markov Model (HMM) are associated with these context-dependent units.
- According to Wickelgren "By assuming (context-sensitive) allophones to be the basic unit of articulation, ... it is trivial to account for how the 'same phoneme' in different phonemic environments can be ... different in some respects at all levels of the speech process. [Wickelgren 1969]
Disadvantages to Wickelgren’s Approach

- Wickelgren’s model, although conceptually simple, has a number of disadvantages.
  
  - the number of units required for a full representation of all phonemes in all contexts (up to 100,000 units for a set of 45 phonemes) requires massive amounts of training data.
  - Even if such training data are available and provide sufficient coverage of the major phonemic contexts in a language, the Wickelgren model assumes that coarticulation is limited to adjacent phonemes.
    
    - It has been shown that, to the contrary, coarticulation may be present at a distance of up to six phonemes from the primary phoneme [Kent and Minifie 1977].
  - Expanding the Wickelgren model to account for contexts over longer phonemic durations requires exponentially more units and data.
Previous approaches to model coarticulation in ASR

**Approaches in Explicit Modeling: Trended HMM**

In the "trended HMM" by Deng et al [1994], speech parameters are captured through the use of polynomial equations that prescribe how the spectral features within a phoneme may change over time.

- Basically an extension of a semi-Markov Model [Rabiner 1993] where duration of each phoneme is explicitly modeled.
- Improved ability to model both duration and speech dynamics but did not yield significantly better performance over standard context-dependent phonemic units.

![Figure](image.png)

**Figure:** Example of fitting 3 states to the first cepstral coefficient of the vowel /iy/ with polynomial order 0, 1, 2, and 3. [Deng et al.]
Previous approaches to model coarticulation in ASR

Trended HMM: Disadvantages

- Polynomials used to describe the speech-feature trajectories embody no underlying understanding of the speech signal, other than constraints on the dynamics. The polynomials may be modeling speaker-dependent attributes, influences of speech rate, or background or channel noise, in addition to coarticulatory effects.
  - Several polynomial equations are used in succession to describe feature dynamics within a phoneme. No boundary constraints for these polynomial equations, and the model allows for large and possibly unrealistic discontinuities in the features within a phoneme.
  - No temporal normalization, making training on variable-rate speech difficult.
  - Foremost in these issues may be the lack of an explicit model of coarticulation.
Richards and Bridle [Richards et al. 1999] presented the concept of target values that are modified by coarticulation is explicitly modeled.

- Filter used to smooth the target values over time is not constrained to follow known properties of speech.
- Coarticulation is assumed to occur *synchronously* for all speech parameters within a frame of speech. NOTE: properties of manner, place, and height of articulation are all quasi-independent and vary asynchronously.
Previous work in modeling coarticulation for ASR has not resulted in models that explain how the speech signal is influenced by phonetic context. This lack of explanatory power then results in:

- need for massive databases in training and the expectation that data-driven learning algorithms can extract relevant properties from the data
- limits on how well the model fits the observed data. Because these models lack predictive power, they are unnecessarily large and/or complex, or they focus on a small part of the coarticulation problem at the expense of a more general, effective solution.
Clear speech has expanded vowel space and longer phoneme durations, as compared to conversational speech.

Clear speech is significantly more intelligible than conversational speech when hearing is impaired.

**Figure:** Observed F1/F2 values at vowel centers
Previously, we searched for acoustic features that cause the increased speech intelligibility of clear speech, using

1. a “hybridization” algorithm, that combined features of clear and conversational speech
2. perceptual testing of sentence intelligibility

Findings
The most relevant features for intelligibility are the combination of short-term spectrum and duration [Kain et al. 2008, Amano-Kusumoto et al. 2009]

This has led us to study a coarticulation model to quantitatively account for the changes of formants over time
Formant trajectories in long and short vowels

- Formants typically don’t reach their target in short vowels

Figure: Vowels of different durations [Klatt 1987]
Many consonants (except approximants) don’t have visible formants, but they have “virtual” formants identified by coarticulation in the vowel.

**Figure:** Virtual formants [Delattre et al. 1955 / Johnson 1997]
Locus Theory

1. Vowels and consonants have formant targets; most consonants have “virtual” formants
2. Coarticulation yields smooth change between targets when formants are visible
3. If duration is too short, formants do not reach their targets, yielding undershoot
4. Both the targets and the rate of change are important for intelligibility
Objectives

Key idea

Reduce the variability of observed features by applying an explicit model of feature trajectories, involving targets and coarticulation functions
Corpus

- 1 male, native speaker of American English
- Sentences contain a neutral carrier phrase (5 total) followed by a keyword (242 total)
- Keywords are common English CVC words with 23 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: $242 \times 2 \times 2 = 968$
- Formant trajectories and phoneme boundaries automatically estimated, manually corrected with verification
Coarticulation model

**Definition**

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

$$\hat{X}(t; \Lambda) = d_{C_1}(t) \cdot T_{C_1} + d_V(t) \cdot T_V + d_{C_2}(t) \cdot T_{C_2} \quad (1)$$

- $T_{C_1}$, $T_V$, and $T_{C_2}$ are the per-formant target values for
  - prevocalic consonant $C_1$
  - vowel $V$
  - postvocalic consonant $C_2$
- $d_{C_1}(t)$, $d_V(t)$, and $d_{C_2}(t)$ are coarticulation functions
Coarticulation functions

- The coarticulation functions are based on sigmoids

\[
\begin{align*}
d_{C_1}(t; s_1, p_1) &= \frac{1}{1 + e^{s_1 \cdot (t - p_1)}} \\
d_{C_2}(t; s_2, p_2) &= \frac{1}{1 + e^{-s_2 \cdot (t - p_2)}} \\
d_V(t) &= 1 - d_{C_1}(t) - d_{C_2}(t)
\end{align*}
\]

- \( s \) represents sigmoid slope and \( p \) sigmoid midpoint position
- An exponential function was previously used to model formant trajectories of vowels in /b,d,g/ contexts [Broad 1987]
- Parameters \( \Lambda = \{ T_{C_1}, T_V, T_{C_2}, s_1, p_1, s_2, p_2 \} \) are specific to a single formant trajectory, and thus the model approximates concurrent formant trajectories asynchronously
Token “will” (clear): formant trajectories

Figure: Formant trajectories, targets, and sigmoid midpoints
Token “will” (clear): coarticulation functions

Figure: Coarticulation functions and sigmoid midpoints
We define the per-token root-mean-squared model error as

\[ E(\Lambda) = \sqrt{\frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \left( X(t) - \hat{X}(t; \Lambda) \right)^2} \]

- \( t_1 \) is the \( C_1 \) boundary, unless \( C_1 \) is an approximant, in which case \( t_1 \) is the middle of \( C_1 \)
- \( t_2 \) is the \( VC_2 \) boundary, unless \( C_2 \) is an approximant, in which case \( t_2 \) is the middle of \( C_2 \)
Varying any two parameters $\lambda_1, \lambda_2 \in \Lambda$, we can display an image of $E(\lambda_1, \lambda_2; \Lambda \setminus \lambda_1, \lambda_2)$

(a) “will”: single optimum  
(b) “neck”: multiple optima

Figure: 2-dimensional error surfaces showing effects of varying $s_1$ and $p_1$ while other parameters are held constant
Minimum error sweep

- Given a single token, we sweep a chosen parameter $\lambda \in \Lambda$ along a prescribed interval, while grid-searching for the lowest model error at each point

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

- We perform this for both F1 and F2 separately
  - with intervals
    - $T_1 = 200, 220, \ldots, 900$ Hz
    - $T_2 = 400, 420, \ldots, 2800$ Hz
    - $s = 10, 30, \ldots, 110$ Hz/s
    - $p = -40, -30, \ldots, 40$ ms, relative to the phoneme boundary
  - and condition
    - $T_2 - T_1 > 200$ Hz
Token “neck” (clear): minimum error sweep

Figure: Minimum error sweep for target value $T_{C_1}$ of F2
Token “neck” (clear): multiple optima

- Some tokens have multiple optima
  - consequently, one or more parameters are “free”
  - need to introduce additional constraints

<table>
<thead>
<tr>
<th></th>
<th>$T_{C_1}$</th>
<th>$T_V$</th>
<th>$T_{C_2}$</th>
<th>$s_1$</th>
<th>$p_1$</th>
<th>$s_2$</th>
<th>$p_2$</th>
<th>$E(\Lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda_1$</td>
<td>610</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>4.46</td>
<td>30</td>
<td>4.64</td>
<td>16.1558</td>
</tr>
<tr>
<td>$\Lambda_2$</td>
<td>1080</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>4.47</td>
<td>30</td>
<td>4.64</td>
<td>16.1548</td>
</tr>
<tr>
<td>$\Lambda_3$</td>
<td>1360</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>4.48</td>
<td>30</td>
<td>4.64</td>
<td>16.1837</td>
</tr>
<tr>
<td>$\Lambda_4$</td>
<td>1540</td>
<td>1880</td>
<td>1680</td>
<td>50</td>
<td>4.49</td>
<td>30</td>
<td>4.64</td>
<td>16.2816</td>
</tr>
</tbody>
</table>

Table: Parameter values for four second formant $T_{C_1}$ frequency values that have near equal minimum error values
Estimating model parameters

1. 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 that particular target
   \[
   E_{global}(T) = \frac{1}{N} \sum_{T \in \text{token}} E_{\text{sweep}}(T)
   \]
   and finally let
   \[
   T^* = \arg \min_T E_{global}(T)
   \]

2. Coarticulation parameters $\{s_1, p_1, s_2, p_2\}$ remain per-token
   - found via grid-search using global target values

Total model error, averaged over all tokens
- 32 Hz for F1, 46 Hz for F2
- 18 Hz for F1, 20 Hz for F2 without global target assumption
Global minimum error sweep: example 1

Figure: Global minimum error sweep for second formant $T_V$ for /A/
Global minimum error sweep: example 2

Figure: Global minimum error sweep for second formant $T_V$ for /d/
Global minimum error sweep: example 3

Figure: Global minimum error sweep for second formant \( T_V \) for /k/
Model fitting

F1/F2 global minimum error sweeps

(a) /aa/

(b) /k/

Figure: F1/F2 global minimum error sweeps
Resulting targets: observable formants

(a) vowels

(b) approximants

**Figure**: F1/F2 target values and observed formant values at phoneme centers
Model fitting

Resulting targets: unobservable formants

(a) stops and affricates

(b) nasals and fricatives

Figure: F1/F2 target values
Vowel targets comparison F1/F2

**Figure:** Resulting F1/F2 vowel targets and comparison with steady-state vowel formant frequencies from Allen [1987] (○) and Ladefoged [1993] (⋆).
Vowel targets comparison side-by-side

(a) From Allen 1987

(b) our results

Figure: vowel formant frequency targets
Maximum vowel contribution

Figure: Histograms of clear and conversational max $d_V$ values
Vowel recognition

Baseline system

- GMM framework, using 10-fold validation
- Single Gaussian components trained on observed formant values at centers of each vowel

Accuracy

- mixed styles: 81.8%
- clear only: 92.6%
- conversational only: 74.4%
- mismatched train=clear, test=conversational: 61.9%
- mismatched train=conversational, test=clear: 77.0%
Vowel recognition

Vowel GMM

Figure: Training data and resulting Gaussian mixture
Model error based approach 1

1. Assume we know consonants on either side $\rightarrow T_{C_1}, T_{C_2}$

2. Hypothesize a possible vowel $\rightarrow T_v$
   - search for coarticulation function parameters $\{s_1, p_1, s_2, p_2\}$ that minimize the model error $E_{\text{token}}$
   - associate this error value with the hypothesized vowel

3. The vowel with the lowest $E_{\text{token}}$ is regarded as the solution

Accuracy: 70%
**Model error based approach 2**

1. Assume we know consonants on either side → $T_{C_1}, T_{C_2}$
2. Hypothesize a possible vowel → $T_v$
   1. take coarticulation function parameter sets {$s_1, p_1, s_2, p_2$} as seen during training for that CVC, and synthesize a new trajectory
      - differing from original trajectory by its duration
      - could be clear or conversational
   2. associate this error value with the hypothesized vowel
3. The vowel with the lowest $E_{token}$ is regarded as the solution
   - Accuracy: 76%
Maximum likelihood mixture model based approach

1. Convert global minimum error sweeps into PDFs \( p(x|c) \)
   - Inverting
   - Normalizing area to 1

2. Given an input token
   - Assume we know consonants on either side \( \rightarrow T_{C_1}, T_{C_2} \)
   - Search for optimal \( T_v, s_1, p_1, s_2, p_2 \) per formant

3. Calculate posterior probabilities of \( p(c|x) \) and take max
   - Accuracy: 68%
   - Accuracy with just Euclidian distance to target: 65%
Conclusion

- Introduced a formant-based model of coarticulation of CVC words
- *We automatically* estimated formant targets of *every phoneme*
- Targets mostly fit expectations about acoustic-phonetics, e.g.
  - F2 of bilabials around 1200 Hz
  - F2 of alveolars around 1800 Hz
- A simple model error based approach is not able to beat the baseline GMM recognizer
- Initial probabilistic framework is not successful
Future work

- Create a successful probabilistic framework
- For continuous speech recognition, slide the 3-phoneme window along the phoneme sequence and perform joint optimization / recognition
- Move from formants to another feature domain
Acknowledgements

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.