Modeling Coarticulation in Continuous Speech
Brian O. Bush and Alexander Kain

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR

Abstract
Modeling coarticulation in speech has been largely limited to short sequences and/or limited phonetic context. We introduce a methodology for modeling both formant frequency and bandwidth in continuous speech, allowing examination of sentence-level coarticulation. The model represents continuous trajectories as a combination of overlapping local trajectories, which are represented by a weighted-addition of acoustic event targets by sigmoidal coarticulation functions characterized by slope and position. Estimation is achieved using a combination of hill-climbing and grid-search, with global target, joint slope for identical contexts, and local position parameters. We evaluate model performance for two speakers using an intelligibility test that compares vocoded model output to a purely vocoded and a natural condition.

Index Terms: coarticulation, formants, continuous speech

1. Introduction
We present a methodology that models formant trajectories of continuous speech as a sum of phoneme-specific formant frequency and bandwidth targets weighted by coarticulation functions. Historically, formant frequency targets of a few selected phonemes were estimated using isolated words, often with some manual intervention [1, 2, 3, 4, 5, 6, 7, 8]. Most notable is the work of Broad and Clermont [5] in which the authors produced several models of formant transitions in consonant-vowel (CV) and CV/d/ contexts, where the prevocalic consonant was limited to voiced stops. Their more elaborate model used a linear combination of coarticulation functions (modeled with exponential functions) and target values. Niu and van Santen [9] expanded Broad and Clermont’s model to a broader set of consonants, but limited the modeling to the vowel centers in their sonants, but limited the modeling to the vowel centers in their expanded Broad and Clermont’s model to a broader set of consonant-vowel (CV) and CV/d/ contexts, where the prevocalic consonant was limited to voiced stops. Their more elaborate model used a linear combination of coarticulation functions (modeled with exponential functions) and target values. Niu and van Santen [9] expanded Broad and Clermont’s model to a broader set of consonants, but limited the modeling to the vowel centers in their application to quantifying coarticulation in dysarthric speech. Amano-Kusumoto and Hosom [10] further improved upon that approach by modeling the entire formant trajectory of the vowel region and moving to a sigmoid function. A sigmoid was chosen because it restricts the coarticulatory effects to be smoothly and monotonically increasing or decreasing, specifically in cases where the consonant being modeled is an approximant (b, l, w, r).

Our previous work [11, 12] investigated fully-automatic and comprehensive estimation of formant frequency targets for a single male speaker, but was limited to modeling CVC triphones. In this paper, we (1) present a novel approach to modeling coarticulation in continuous speech, (2) extend the model to include formant bandwidths, and (3) improve model parameter accuracy through joint estimation.

2. Coarticulation Model
While in principle the proposed model could be applied to any interpolatable feature representing the speech signal, we use traditional formant frequency (F1–F4) and bandwidth (B1–B4) trajectories as speech features. In addition, we assume a phoneme segmentation is provided. Because some phonemes contain more than one acoustic event, we map the phoneme segmentation to an acoustic event segmentation \( e_1, e_2, \ldots, e_N \) with event boundary times \( b_1, b_2, \ldots, b_{N+1} \). Most phonemes map to one acoustic event, but diphthongs, affricates, and released stops map to more events (e.g., closure and release for released stops). For convenience, we also define event center times \( c_i = (b_i + b_{i+1})/2, i = 1, 2, \ldots, N \).

2.1. Local Coarticulation
During analysis, an individual continuous feature trajectory \( X(t) \) is divided into consecutively overlapping local regions comprising three acoustic events such that the local feature trajectories \( \hat{x}_i(t) \) are defined for the interval \( t \in [e_{i-1}, e_{i+1}], i = 2, \ldots, N-1 \). This local trajectory is modeled as

\[
\hat{x}_i(t) = f_{i,L}(t; s_i, L, p_i, L)T_L + f_{i,C}(t; C)T_C + f_{i,R}(t; s_i, R, p_i, R)T_R
\]

with \( L = e_{i-1}, C = e_i, R = e_{i+1} \). Eq. 1 is a convex linear combination of \( T_L, T_C, \) and \( T_R \), representing global feature target values for the left, center, and right acoustic event type, respectively, weighted by \( f_{i,L}(t), f_{i,C}(t), \) and \( f_{i,R}(t) \), representing local coarticulation functions \( i \) for the left, center, and right acoustic events, respectively, in this local event context. For all contexts, coarticulation functions are based on the sigmoid

\[
\sigma(t; s, p) = (1 + e^{(t-p)})^{-1}
\]

with

\[
f_{i,L}(t; s_L, p_L) = \sigma(t; s_L, p_L), \quad s_L > 0
\]
\[
f_{i,R}(t; s_R, p_R) = \sigma(t; -s_R, p_R), \quad s_R > 0
\]
\[
f_{i,C}(t) = 1 - f_{i,L}(t; s_L, p_L) - f_{i,R}(t; s_R, p_R)
\]

where \( s_L, s_R \) represent sigmoid slope (slow versus fast transition, from the left to center acoustic event and from the center to right acoustic event, respectively), and \( p_L, p_R \) represent sigmoid midpoint position (point of maximum slope), measured relative to their respective phoneme boundaries \( b_i \) and \( b_{i+1} \) for the \( i \)th context. Figure 1a illustrates the concept.

We assume the existence of a single target value \( T \) for each acoustic event type; thus they are defined globally. Furthermore, we assume that slope parameters \( s_L \) and \( s_R \) are shared among contexts of the same type (i.e., different tokens with identical acoustic event labels). Finally, position parameters \( p_L \) and \( p_R \) are modeled separately for each token.

2.2. Continuous Coarticulation
In a second step, local coarticulation functions modeling identical acoustic events are cross-faded to form global, continuous...
global coarticulation functions

where \( \beta \) (magenta bars). Lower panel shows coarticulation functions with sigmoid centers at the right side of the interval. This operation can be thought of as a smooth transfer of knowledge about a particular acoustic event from one local model to the next. For the second case in Eq. 3, the \( i-2^{nd} \) local model contains no information about \( e_i \), and thus we are cross-fading with zero, equivalent to a fade-in; analogously for the fifth case and the \( i+2^{nd} \) local model, where we are performing a fade-out. For \( i = 1,2,N-1,N \) the formula is modified to handle these edge cases correctly. Note that cross-fading preserves convexity, i.e. \( \sum_{t=1}^{T} g_{e_i}(t) = 1, t \in [c_1,c_N] \). For an example illustration refer to Figure 2.

The final model representation of the continuous feature trajectory is calculated via \( \hat{X}(t) = \sum_{i=1}^{N} g_{e_i}(t) \cdot T_{e_i} \), implementable by matrix multiplication for discrete \( t \) (such as consecutive synthesis frame times). An example is shown in Figure 1b.

coarticulation functions

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

where \( \beta(t) = 1-\alpha(t) \), for \( i = 3,4,...,N-2 \), 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. This operation can be thought of as a smooth transfer of knowledge about a particular acoustic event from one local model to the next. For the second case in Eq. 3, the \( i-2^{nd} \) local model contains no information about \( e_i \), and thus we are cross-fading with zero, equivalent to a fade-in; analogously for the fifth case and the \( i+2^{nd} \) local model, where we are performing a fade-out. For \( i = 1,2,N-1,N \) the formula is modified to handle these edge cases correctly. Note that cross-fading preserves convexity, i.e. \( \sum_{t=1}^{T} g_{e_i}(t) = 1, t \in [c_1,c_N] \). For an example illustration refer to Figure 2.
3. Experiment

3.1. Parallel Style Corpus

One male and one female speaker produced identical speech material in two different speaking styles. For conversational speech (CNV), the speaker was asked to speak as if “talking with a colleague at a natural pace”. For clear speech (CLR), the speaker was asked to “enunciate consonants more carefully and with greater effort than for CNV speech and avoid slurring words while decreasing the target step size by half with each iteration”. The corpus was composed of 242 single-syllable American English words, spoken in a carrier sentence [10, 15]. Each token was rendered twice in both styles, for a total of 242 words × 2 styles × 2 renditions = 968 tokens. Formants were automatically estimated using a standard formant tracker [16, 17]. A trained labeler added phonetic transcriptions.

Any two words spoken separately in the same style will have a specific amount of variation that occurs naturally. To quantify this variation, we define the rendition error as the root mean square error (RMSE) between two formant trajectories of two renditions of the same word: \( E(X_1,X_2) = \sqrt{1/(c_{i+1}-c_{i-1}+1) \sum_{t=c_{i-1}}^{c_{i+1}} (X_1(t) - X_2(t))^2} \) where \( X_1 \) and \( X_2 \) are two time-aligned formant trajectories over frames \( c_{i-1} \) to \( c_{i+1} \). We use RMSE as our error measure since it is measured in the same units as the data (Hz), rather than in squared units, and is representative of the size of a typical error. Average rendition errors are shown in Table 1.

We partitioned the 242 words of the corpus using an approximate 80/20 training/test ratio, resulting in 194 training and 48 test words. The partitioning was created by clustering the corpus into 48 equal-sized clusters, with the word corresponding to the cluster medoid used for testing. The clustering distance measure calculated word distances by averaging individual phoneme distances defined as the Euclidean distance between four-dimensional, manually-derived phonetic features describing the sonority, manner, place and height of each phoneme.

3.2. Parameter Estimation

For a local context, we define the error as \( E(x(t),\tilde{x}(t,L)) \) identical to the rendition error, where \( x(t) \) and \( \tilde{x}(t) \) are the observed and modeled individual formant trajectories, respectively. For a specific feature trajectory, we can now calculate a global error by summing the local error over all \( K \) tokens in our training set. Model parameters are estimated with the objective to minimize this total error. Initially, the target parameters \( T \) are set to the mean observed values at the center of each phone using the CLR-style words. Our parameter estimation approach is an iterative process that first estimates global targets \( T \) using hill climbing, then estimates local parameters, \( s_L, s_R, p_L \) and \( p_R \) for each token using grid search, while decreasing the target step size by half with each iteration. The hill climbing process to estimate targets uses the discretized parameter space defined as follows: for all \( T \) in Hz, \( F_1 = [200,1000], F_2 = [\max(F_1+200,400),2800], F_3 = [\max(F_2+200,900),4000], \) and \( F_4 = [\max(F_3+200,3000),6000] \), with step sizes initially at 50 Hz. For any iteration, the hill climbing process is halted after an error change of less than one Hz. Coarticulation parameters are estimated by jointly estimating \( s_L \) and \( s_R \) of identical triphones by grid searching over the interval \( s = 10,20,30,...,120 \). For each \( s_L \) and \( s_R \) parameter set considered, both \( p_L \) and \( p_R \) are grid-searched to find the parameters that have the lowest total error. We define valid intervals of \( p_L = [c_{i-1},c_i] \) and \( p_R = [c_i,c_{i+1}] \), calculated by subtracting the corresponding phoneme boundary time from the corresponding \( p \) in Equation 2. We performed this process for F1, F2, F3, and F4 consecutively, as well as for B1, B2, B3, and B4.

The entire parameter estimation was completed ten times for each speaker, with final targets set to the outcome of the best run. The only source of variation in each run was the initial target values \( T \), which were randomly varied uniformly by ±[20,100] Hz. The average training error and the average standard deviation of all formant frequency targets are shown in Table 1; we observe that despite different initial values for \( T \), their final values are relatively consistent.

3.3. Analysis of Model Parameters

Final formant frequency targets are shown in Figure 3. The formant targets for vowels and approximants are largely in their expected locations [18, 19]. We observe that some consonants are at extremes, possibly due to the effects of joint optimization, convexity constraints, and/or systematic formant tracking errors.

When analyzing coarticulation parameters, we only considered local models with target differences of at least 300 Hz to obtain meaningful values. First, as a measure of asynchronicity, we calculated the averages of the per-token standard deviations of \( p_L \) and \( p_R \) combined for F1–F4 to be 73 ms and 68 ms for the male and female speaker, respectively. This variation strengthens the case for asynchronous modeling. Second, examining the maximum value of the center coarticulation function, \( \max_2(f_c(t)) \), of F2 by style, we found the mean values \( \text{CLR} = 0.88 \) and \( \text{CNV} = 0.71 \) for the male speaker, where the female has mean values \( \text{CLR} = 0.86 \) and \( \text{CNV} = 0.75 \). As these two distributions failed normality tests, we used a non-parametric test to measure significance. For the male speaker, we found Mann-Whitney \( U = 636,924, p < 0.05 \); Cohen’s \( d = 0.79 \) [20], a large effect size [21], for the female speaker \( U = 759,380, p < 0.05 \) and Cohen’s \( d = 0.47 \), a medium effect size; i.e. \( \max_2(f_c(t)) \) of F2 was a statistically significant discriminator between the two styles.

3.4. Evaluation

To objectively evaluate model performance, we estimated the best \( s_L, p_L, s_R \) and \( p_R \) by grid search for each local context contained in our test set, using the targets \( T \) derived from the

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>rendition error</td>
<td>184 (114)</td>
<td>154 (106)</td>
</tr>
<tr>
<td>training error</td>
<td>219 (1)</td>
<td>187 (1)</td>
</tr>
<tr>
<td>test error</td>
<td>202 (119)</td>
<td>186 (101)</td>
</tr>
<tr>
<td>mean target std</td>
<td>33</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 1: Top rows contain rendition, training, and test error (standard deviations in parentheses); bottom row contains average standard deviations of estimated formant frequency targets based on ten restarts (all values are in Hz).
training set (see Table 1). The test error was reasonably close to the rendition error, which serves as an estimate of the lower bound of best-possible-performance.

We also conducted a perceptual evaluation using a speech intelligibility test to compare vocoded speech from model trajectories with vocoded speech from observed trajectories, as well as with natural speech. We prepared stimuli for these 3 conditions (model-vocoded, observed vocoded, and natural), resulting in a total of 48 test set words × 2 speakers × 2 styles × 3 conditions = 576 stimuli, which were divided into 12 partitions consisting of 48 stimuli in a Latin squares design. We ran three complete experiments with the model-vocoded condition replacing (1) formant frequencies and bandwidths, (2) bandwidths only, and (3) with frequencies only.

All stimuli were loudness normalized using an A-weighted RMS measure [22], and 12-talker babble noise was added to prevent saturation effects. The energy of the noise was set to a signal-to-noise ratio of +3 dB. Vocoder resynthesis was accomplished by using a hybrid linear predictive coding/formant analysis-synthesis vocoder with energy and pitch trajectories preserved. In the model-vocoded condition, automatically estimated formant frequency trajectories and bandwidths were both replaced with either formant frequency, bandwidth, or both with trajectories generated from the trained model(s), dependent on experiment.

Amazon’s Mechanical Turk crowd-sourcing service was used to provide human subjects. Workers were paid $0.40 USD per Human Intelligence Task (HIT) completed. Each HIT was completed by unique individuals located in the United States who had a previous approval rating of 95% or higher and had completed at least 500 HITs. Each HIT contained 48 stimuli with four words per speaker, style, and condition combination, presented in random order. After hearing a stimulus, a listener was asked to choose one of five possible answers to the question “What did you hear?”, with four decoy terms among the correct term (e.g. “fan”, “van”, “pan”, “than”, and “ban”). Decoy terms were selected based on the closest phonetic similarity to the target term, using a list of common single-syllable words. (The similarity measure is described in Sec 3.1.) HITs that had high error rates (>50% incorrect) on the eight stimuli in the CLR natural condition were discarded [12]; discarded HITs were resubmitted until completed.

A total of 120 adults (10 blocks) per test participated, with self-reported normal hearing and unfamiliar with the goals of the study. The average proportion of words heard correctly, combining speakers and styles, were as follows: 93% (σ=7), 93% (σ=8), and 91% (σ=10) for natural, 81% (σ=12), 81% (σ=13), 80% (σ=13) for observed-vocoded, and 66% (σ=12), 79% (σ=14), 70% (σ=15) for model-vocoded, when replacing formant frequencies and bandwidths, formant bandwidths only, and formant frequencies only, respectively. This means that our model is capable of modeling formant bandwidths relatively effectively, but that improvements are needed in the modeling of formant frequencies.

4. Conclusions and Future Work

We formulated and applied a continuous coarticulation model to formant frequency and bandwidth trajectories of two speakers. In future work, we will apply our model to a parallel CLR/CNV sentences and expand the number of speakers. We plan on exploring the joint estimation of \( p_R \) of the current local context and \( p_L \) of the following local context. Additionally, we will calculate a per-frame weighting to the model error to reflect the time-varying confidence in the observed formant frequency and bandwidth trajectories. We speculate that this will improve formant frequency modeling.
5. References


