Modeling Coarticulation in Continuous Speech

Brian O. Bush

Oregon Health & Science University
Center for Spoken Language Understanding

December 16, 2013

Outline

Introduction

- Coarticulation is the influence of one phoneme on another

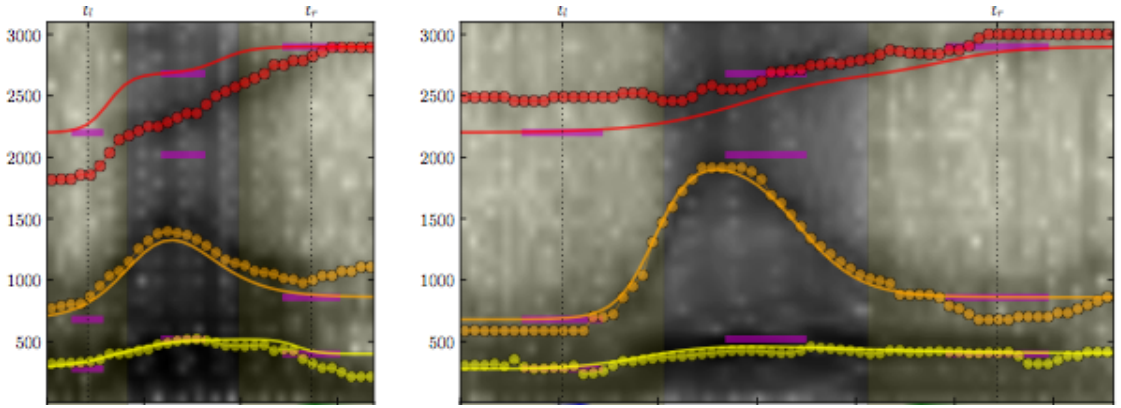


Figure: Example of coarticulation in CNV (left) and CLR (right)

- Degree of coarticulation depends on phonetic context
  - For example, /w/ has a stronger effect on the following vowel than /z/
- **Motivation:** To-date there is no comprehensive, data-driven model that explains the timing and degree of coarticulation effect of one phoneme on its neighbors

Speaking Styles: Defined

Clear speech and conversational speech are defined as

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

Applications

There are several application areas for coarticulation modeling:

- Convert conversational style speech to clear – increase intelligibility
- Dysarthria diagnosis – assessing the presence or severity
- Formant tracking – automatically correct errors in tracking
- Text-to-speech – vary the degree of coarticulation from conversational to clear
- Index of intelligibility – infer the intelligibility of speech based on coarticulation

Broad and Clermont (1987)

- 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)

- Niu and van Santen (2003) applied Broad and Clermont's CV/d/ model to Dysarthria to measure coarticulation
- Expanded model application to generic CVC tokens
- Modeling was limited to vowel centers
- Results: Coarticulation effects of a dysarthric speaker were higher than normal speaker

Amano and Hosom (2010)

- 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

Bush and Kain (2013)

- Expanded model trajectories over CVC region
- Relaxed synchronicity; previous work modeled all formants (F1-F4) synchronously
- Validated model using an intelligibility test to show that model is capturing important features from spectral domain
- Resynthesis results: 74.6% for observed formants, 70.8% for modeled formants

Proposed Model

- Uses triphone models as local models of coarticulation during analysis
- Creates continuous speech by cross-fading triphone models during synthesis
- Includes formant bandwidth
- Handles two primary components of plosives (/b,d,g,p,t,k/) separately
- Optimization uses joint-optimization over identical types

Triphone Trajectory Model

Definition

An individual formant trajectory $X(t)$ of a triphone is modeled as

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

which is a convex linear combination of T_L, T_C, and T_R representing global formant target values for three consecutive acoustic events. A phone includes one or more distinct 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 $f(t; s, p) = (1 + e^{s \cdot (t-p)})^{-1}$

$$\begin{aligned} f_L(t; s_L, p_L) &= f(t; s_L, p_L) \\ f_R(t; s_R, p_R) &= f(t; -s_R, p_R) \\ f_C(t) &= 1 - f_L(t) - f_R(t) \end{aligned}$$

  - 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 Model – CLR Example

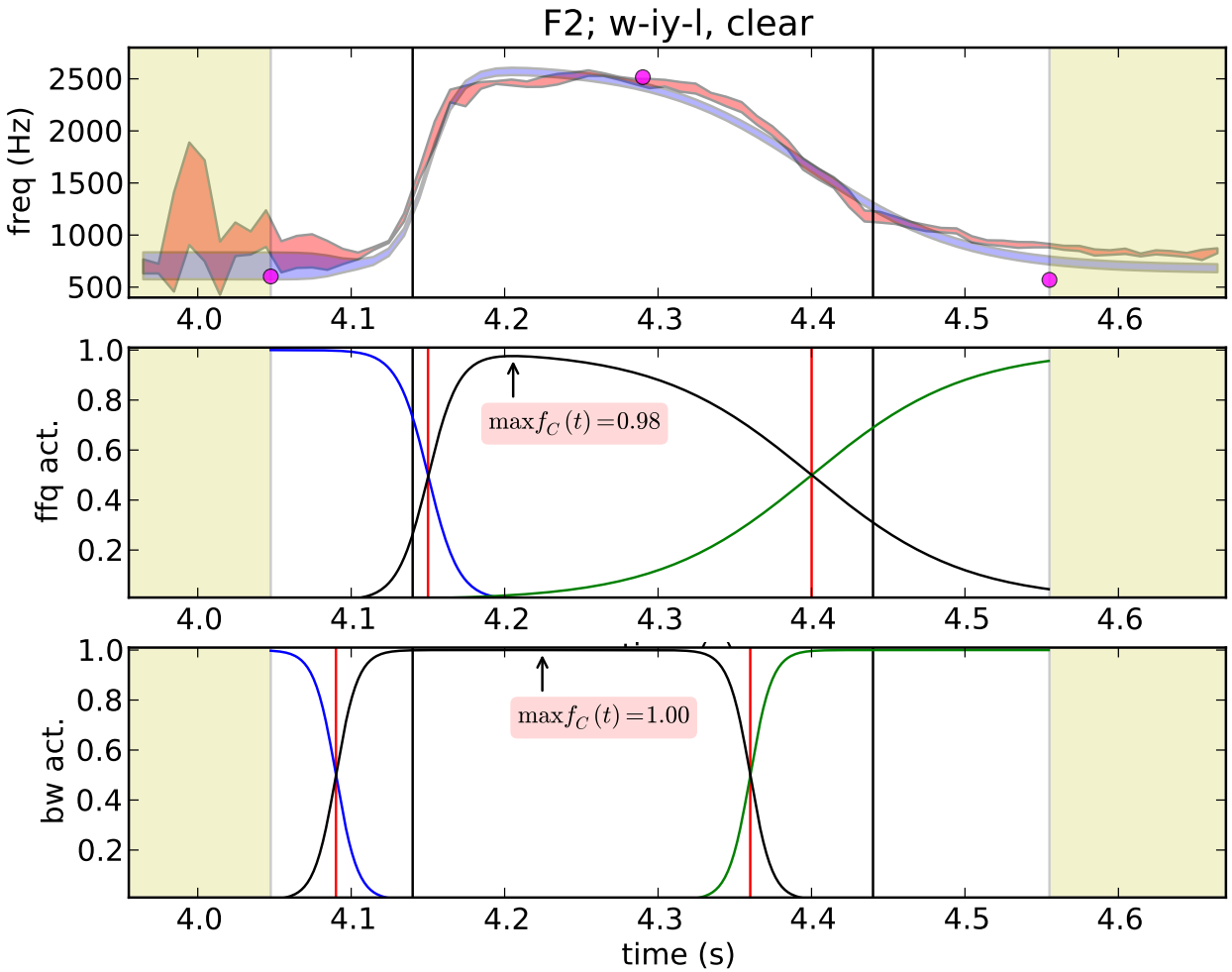


Figure: F2 model on the triphone "w-iy-l" in CLR speech

Triphone Model – CNV Example

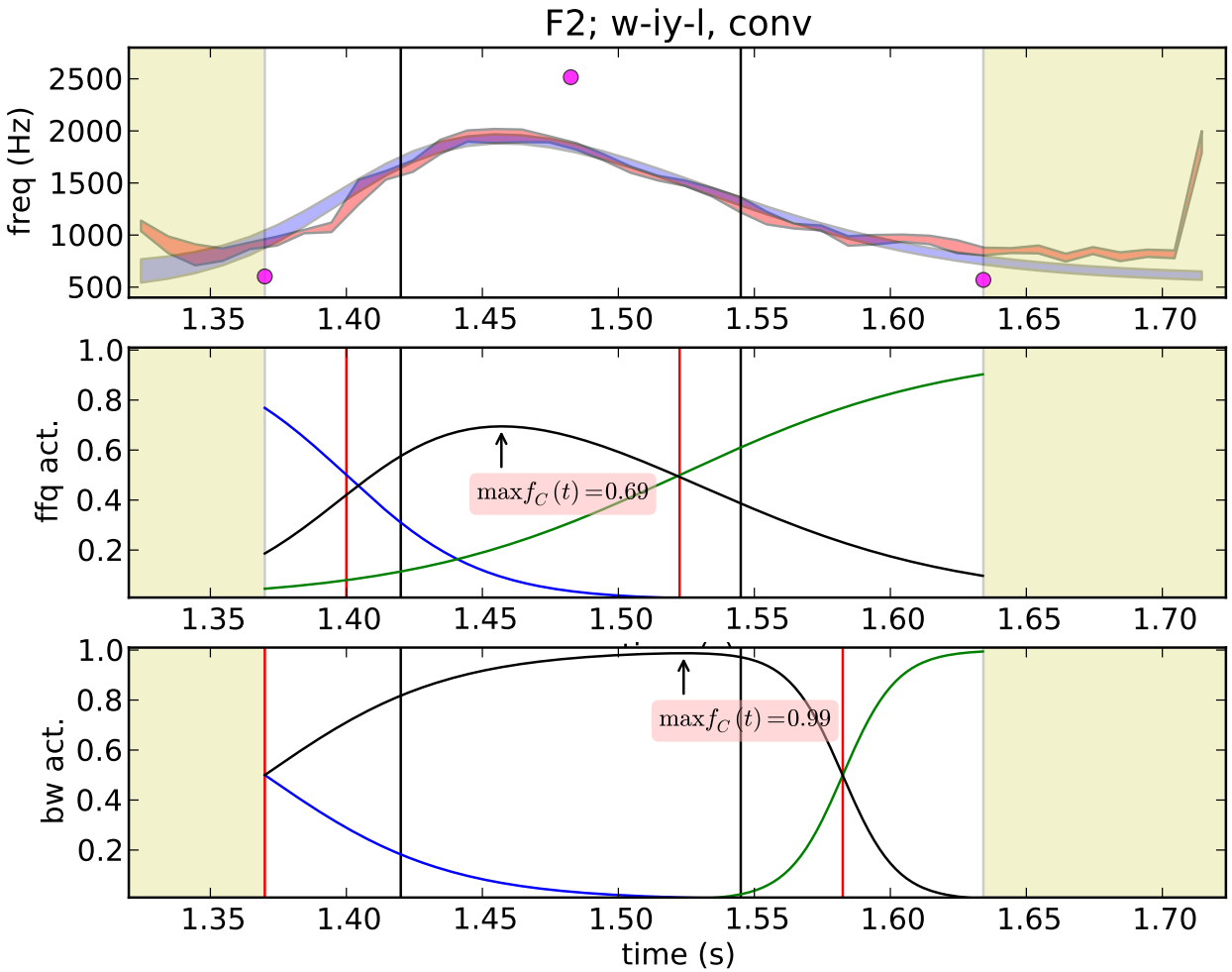


Figure: F2 model on the triphone "w-iy-l" in CNV speech

Triphone Model Error

Definition

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.

Estimating Model Parameters

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

- Targets (global)
- Coarticulation functions (local to each triphone token)

Estimating Targets

- Initialize targets to median formant frequency and bandwidth 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:
  - $F1 = 200, 250, \ldots, 1000$ Hz
  - $F2 = 400, 450, \ldots, 3000$ Hz
  - $F3 = 900, 950, \ldots, 4000$ Hz
  - $F4 = 3000, 3050, \ldots, 6000$ Hz
  - $B1, B2, B3, B4 = 10, 60, \ldots, 460$ 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

Aligned Local Triphone Models

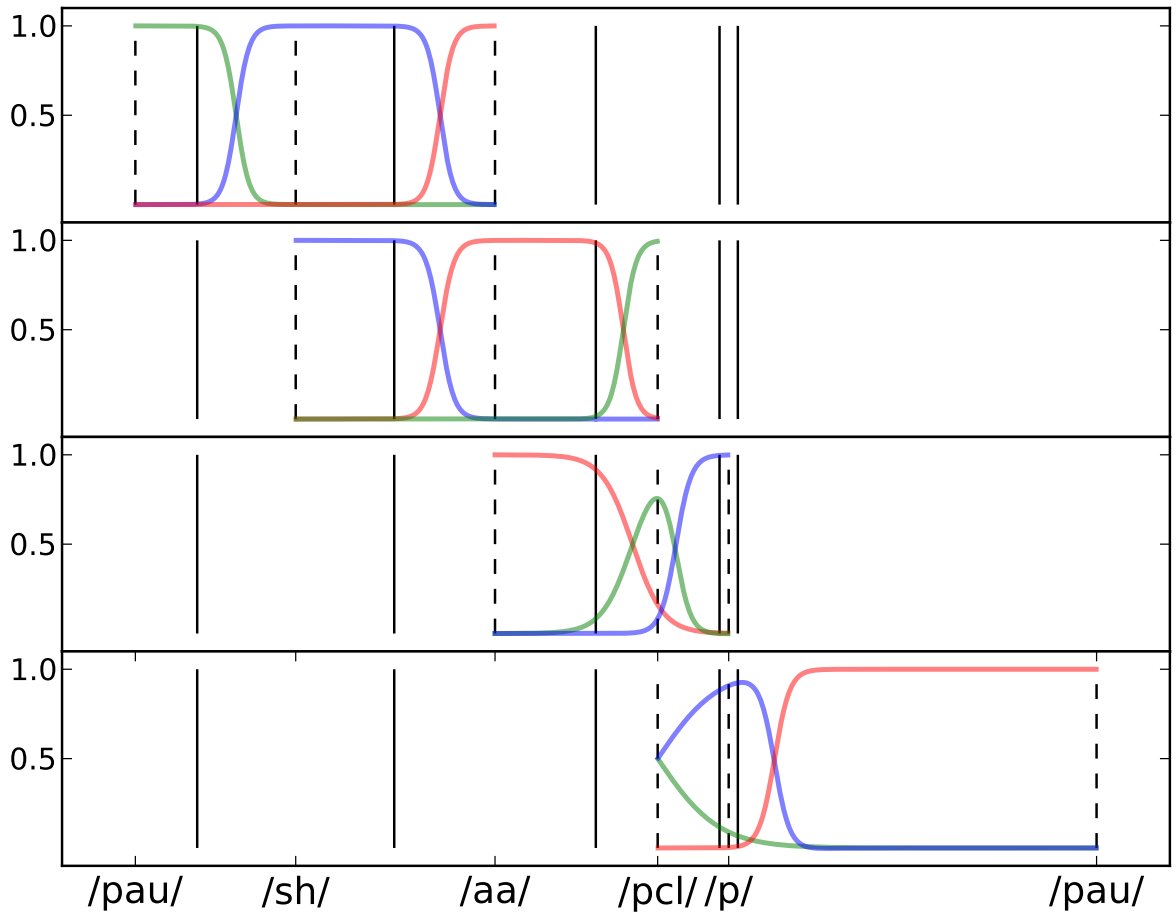


Figure: F2 coarticulation functions for sequence pau-sh-aa-pcl-p-pau

Overlaid Coarticulation Functions

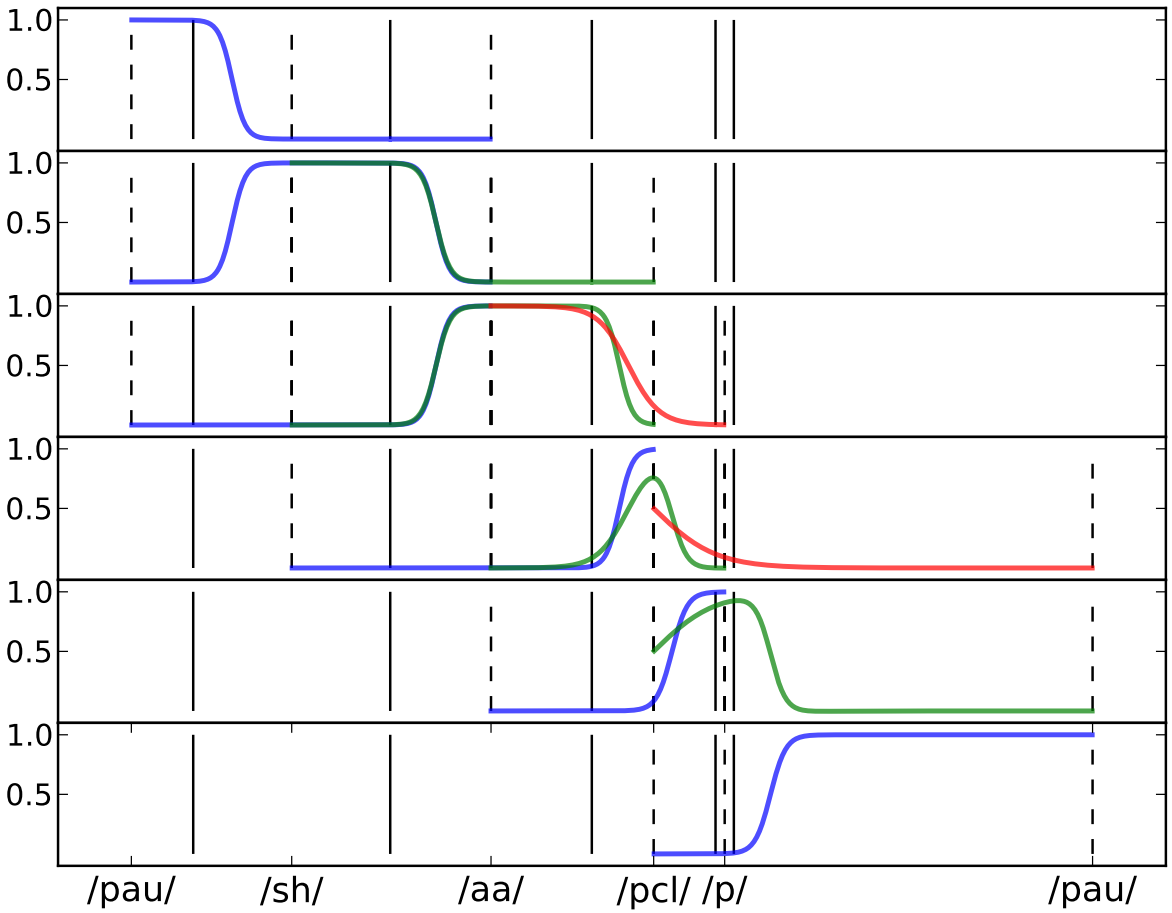


Figure: F2 coarticulation functions for sequence pau-sh-aa-pcl-p-pau

Continuous Coarticulation Functions $\mathbf{F}$

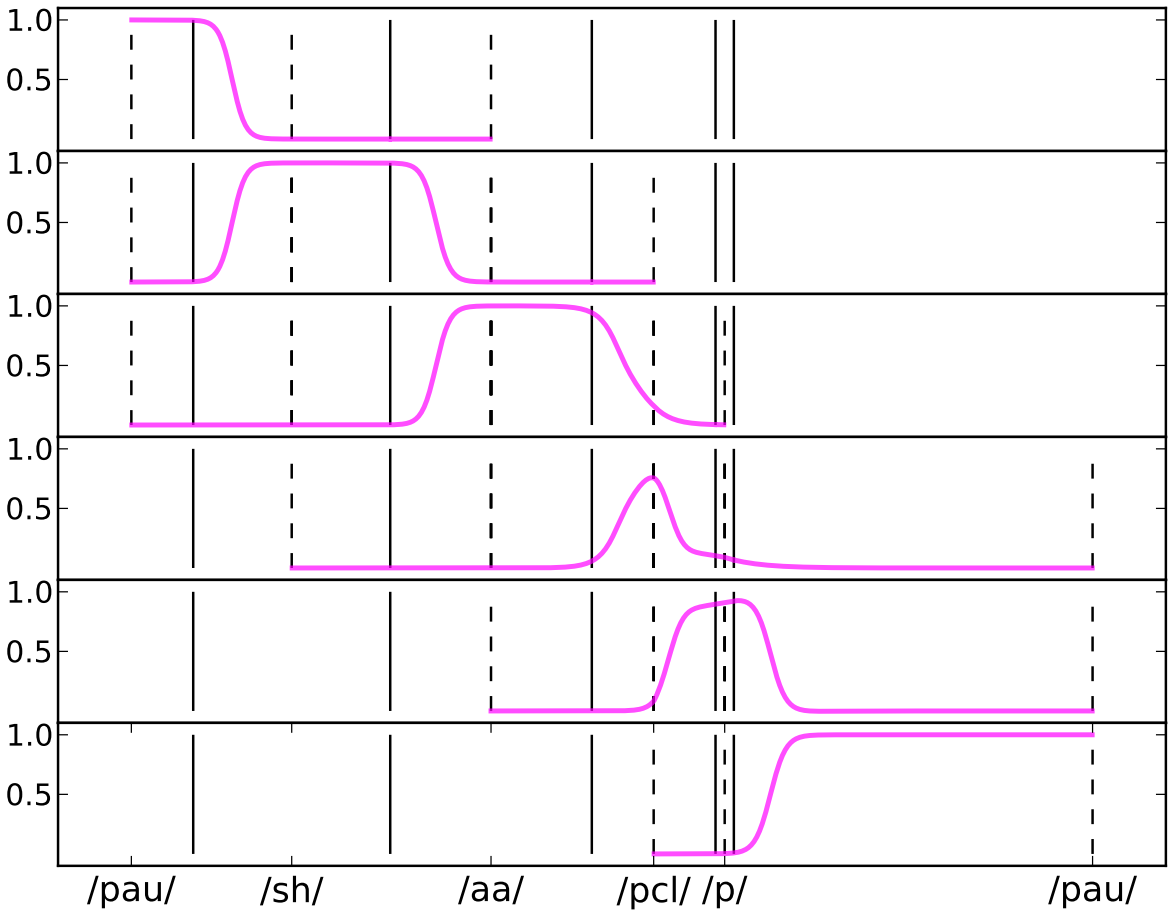


Figure: Cross-faded coarticulation functions

Original and Final Spectrum

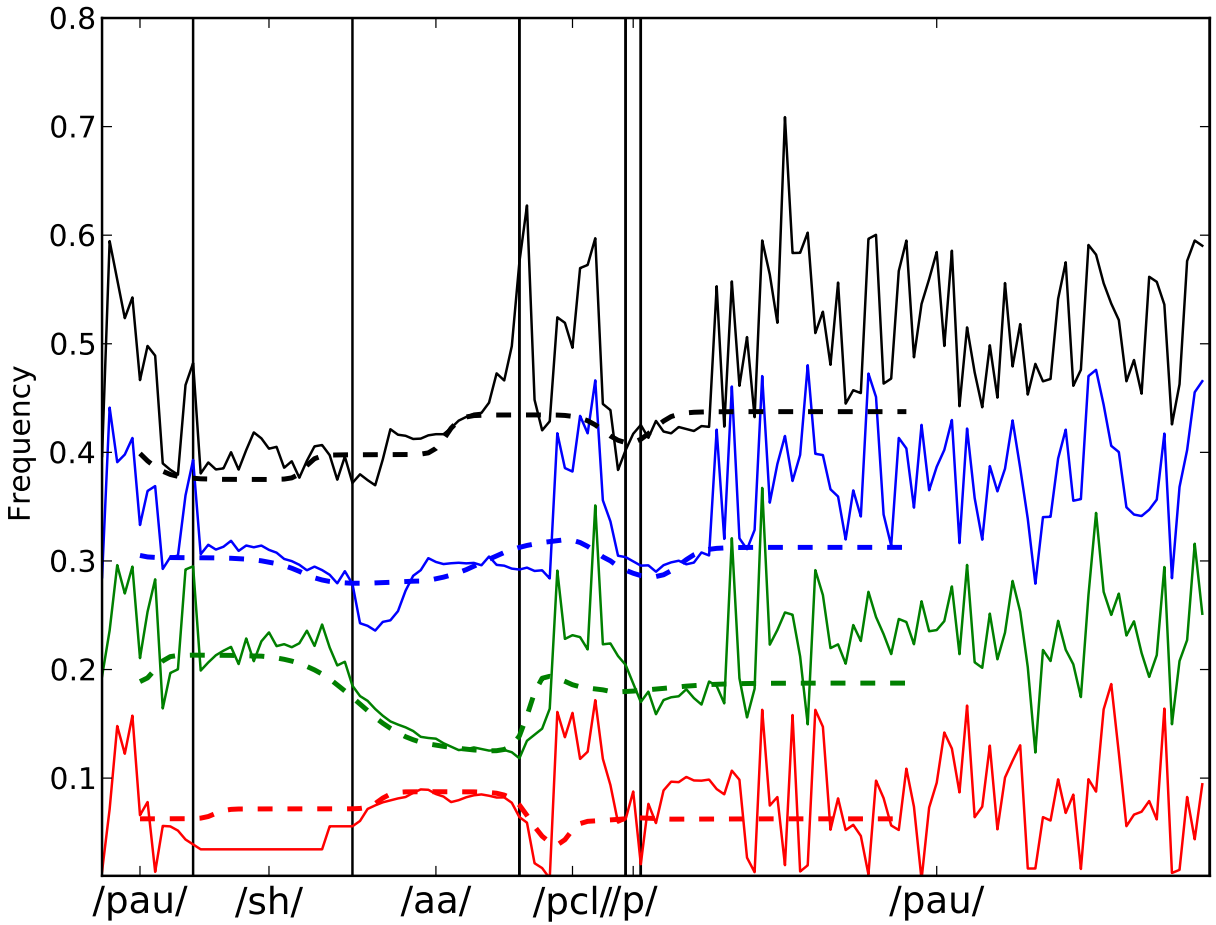
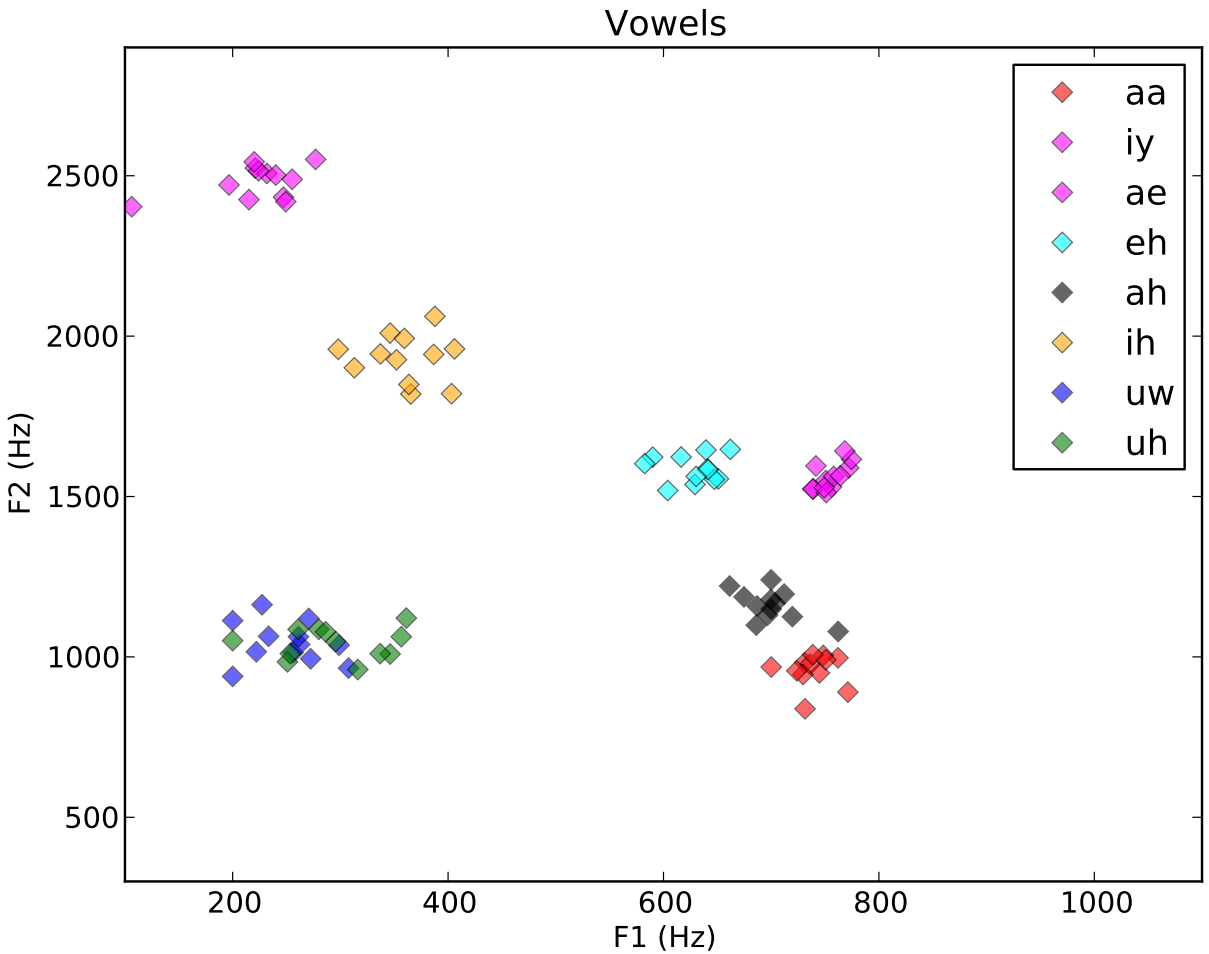


Figure: $\mathbf{X}_{N\times 4} = \mathbf{F}_{N\times P}\mathbf{T}_{P\times 4}$

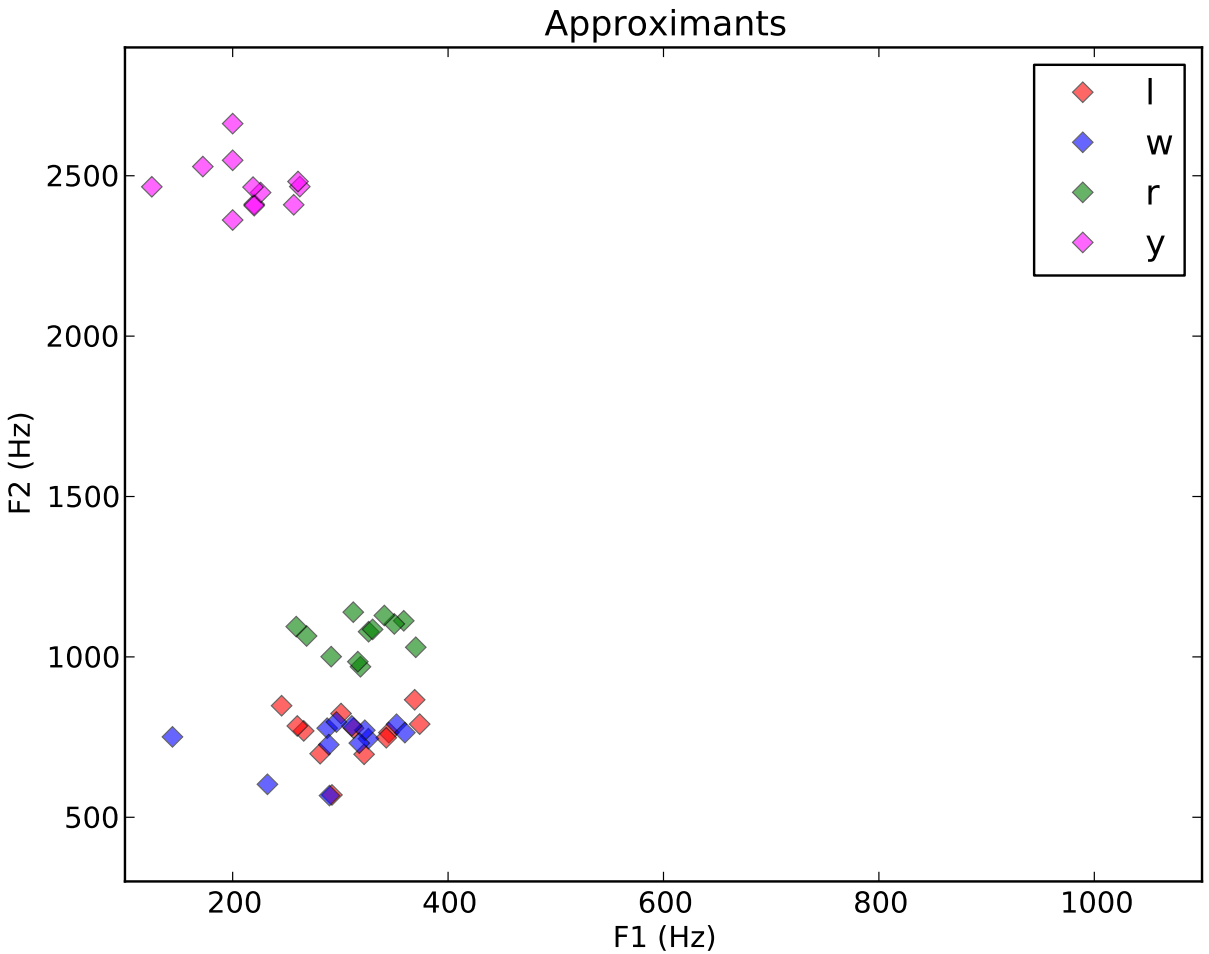
Parallel Style Corpus

- 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 monophthongs
  - All sentences spoken in both clear and conversational styles
  - Two recordings per style of each sentence
  - Total number of keyword tokens: $242 \times 2 \times 2 = 968$
  - Typical token generates four triphones
- Diphthongs not represented

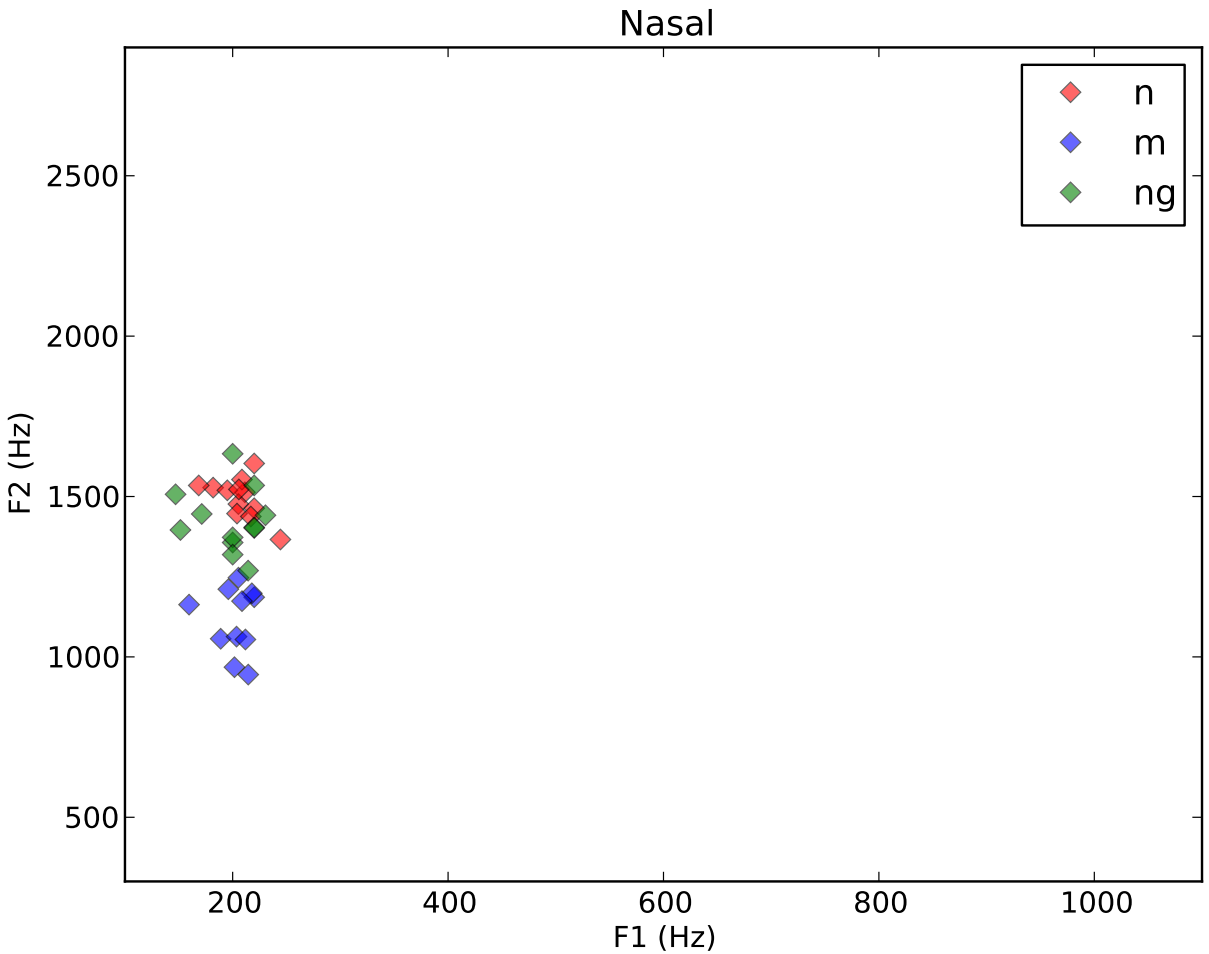
Global Targets – Vowels



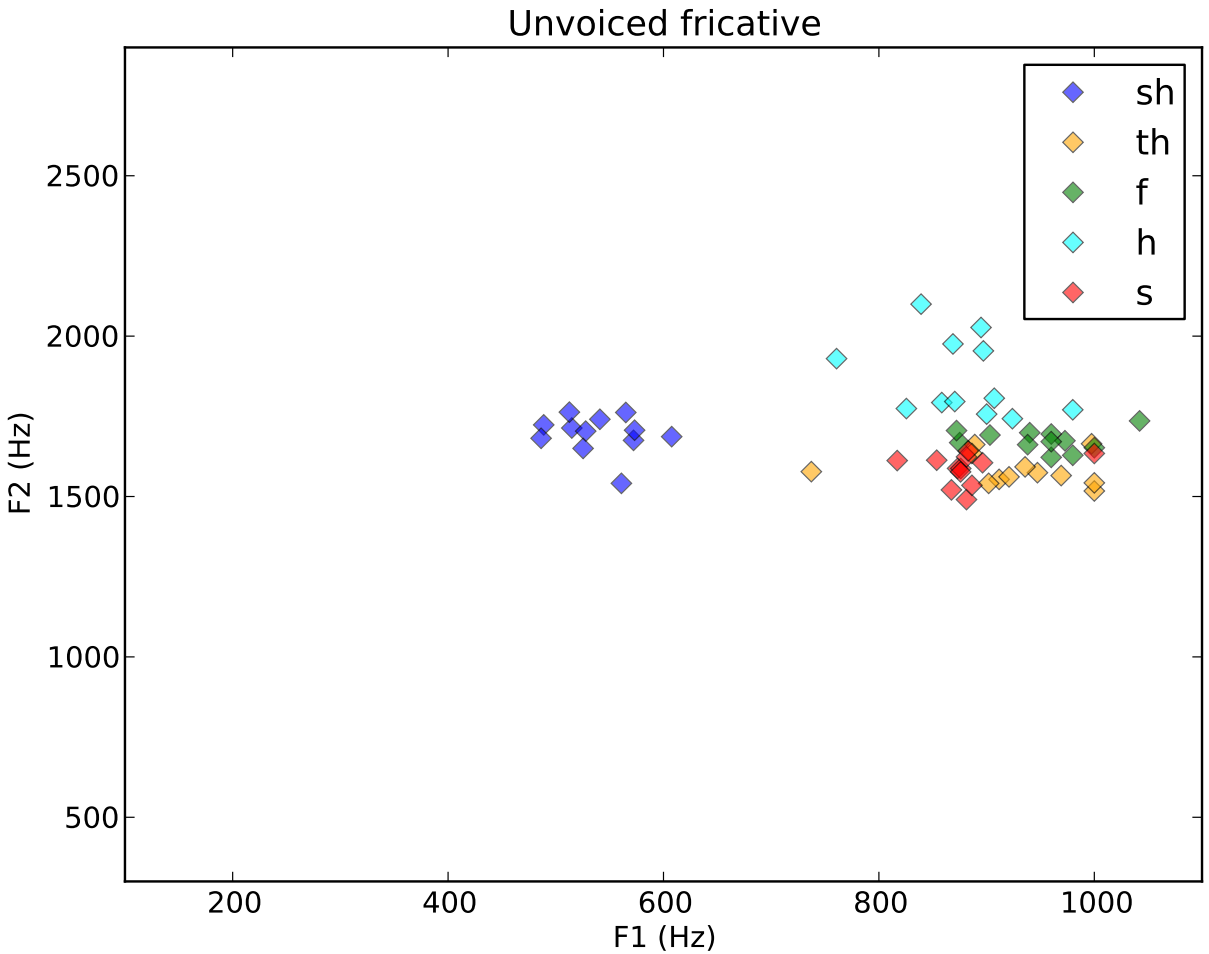
Global Targets – Approximants



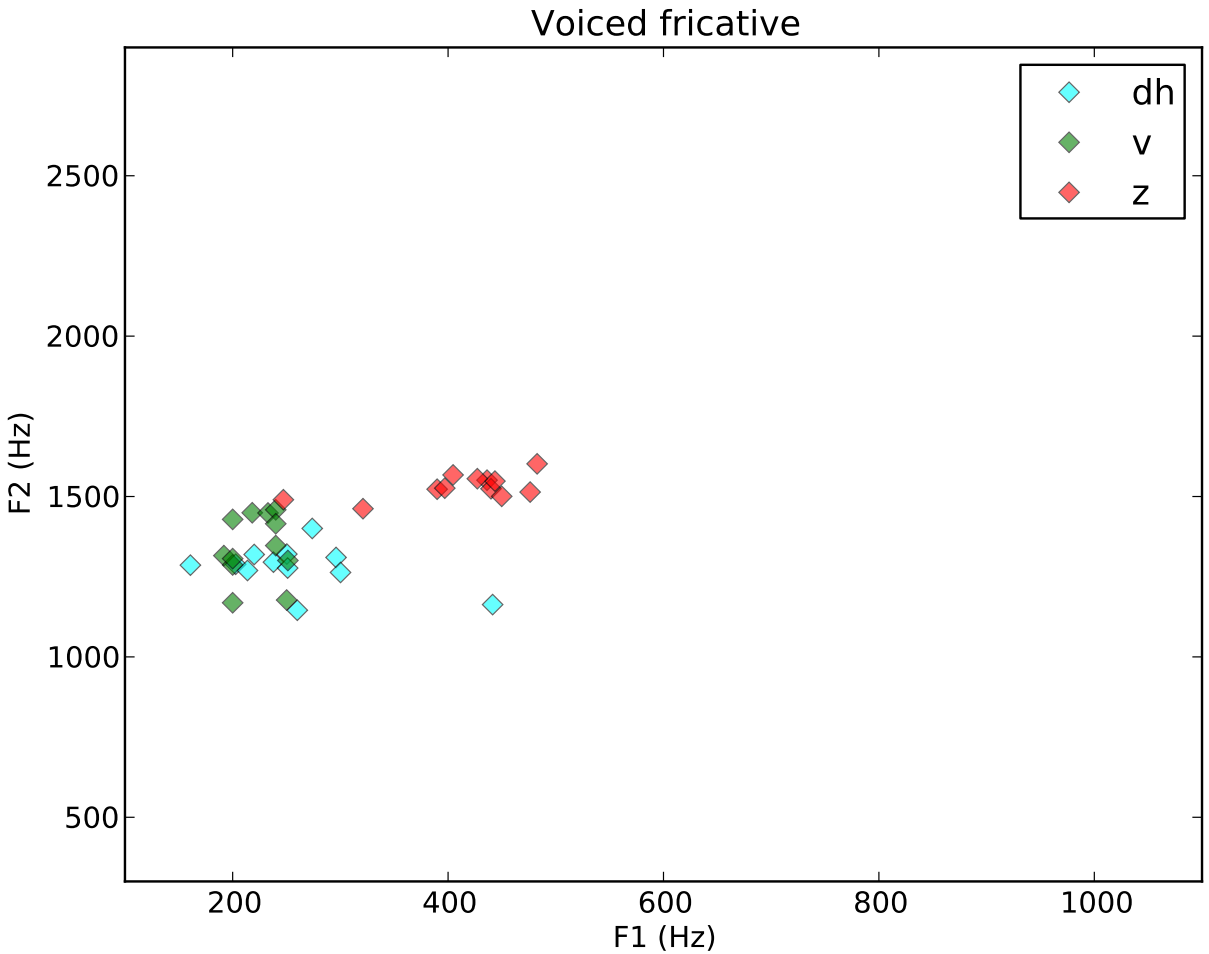
Global Targets – Nasals



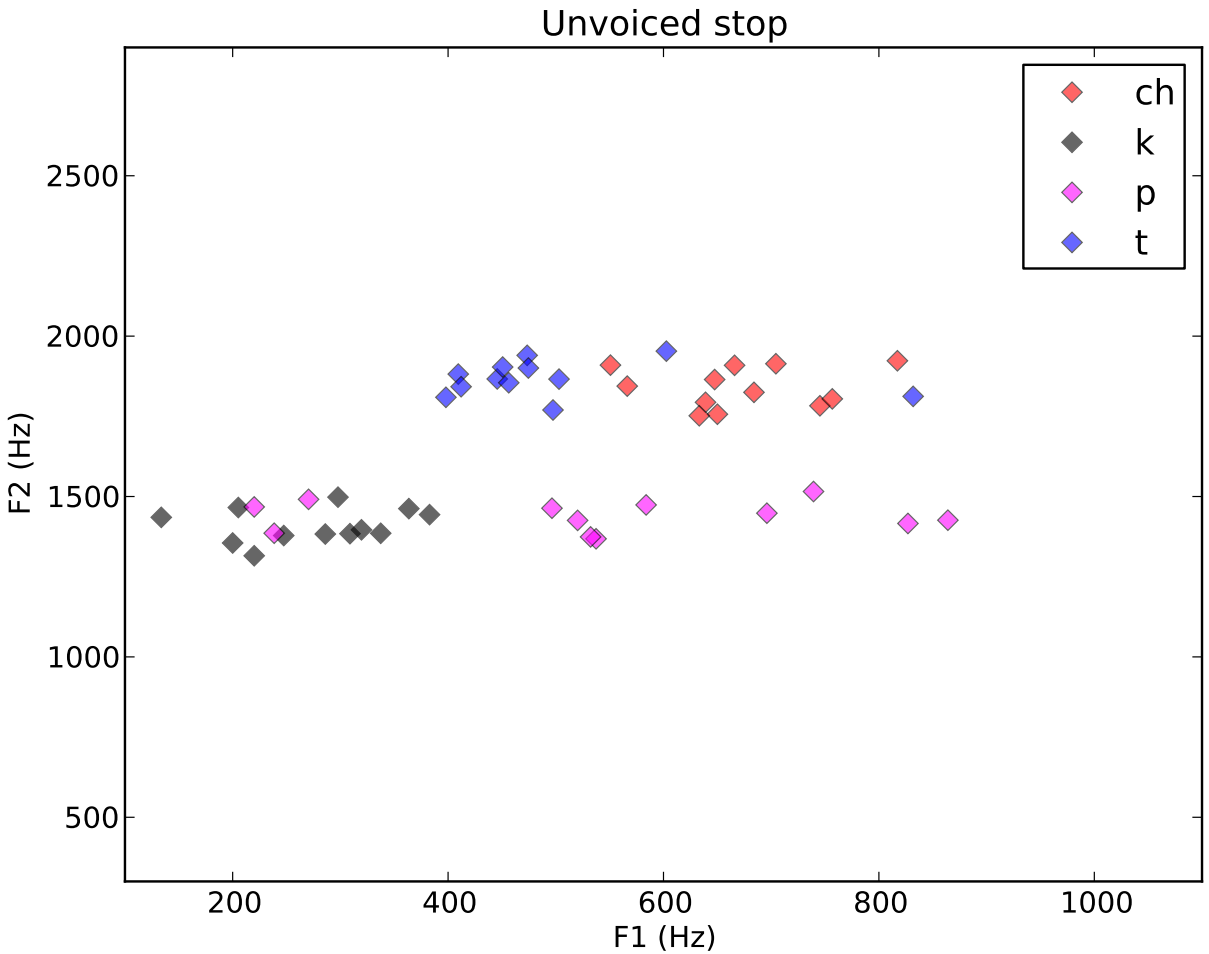
Global Targets – Unvoiced Fricatives



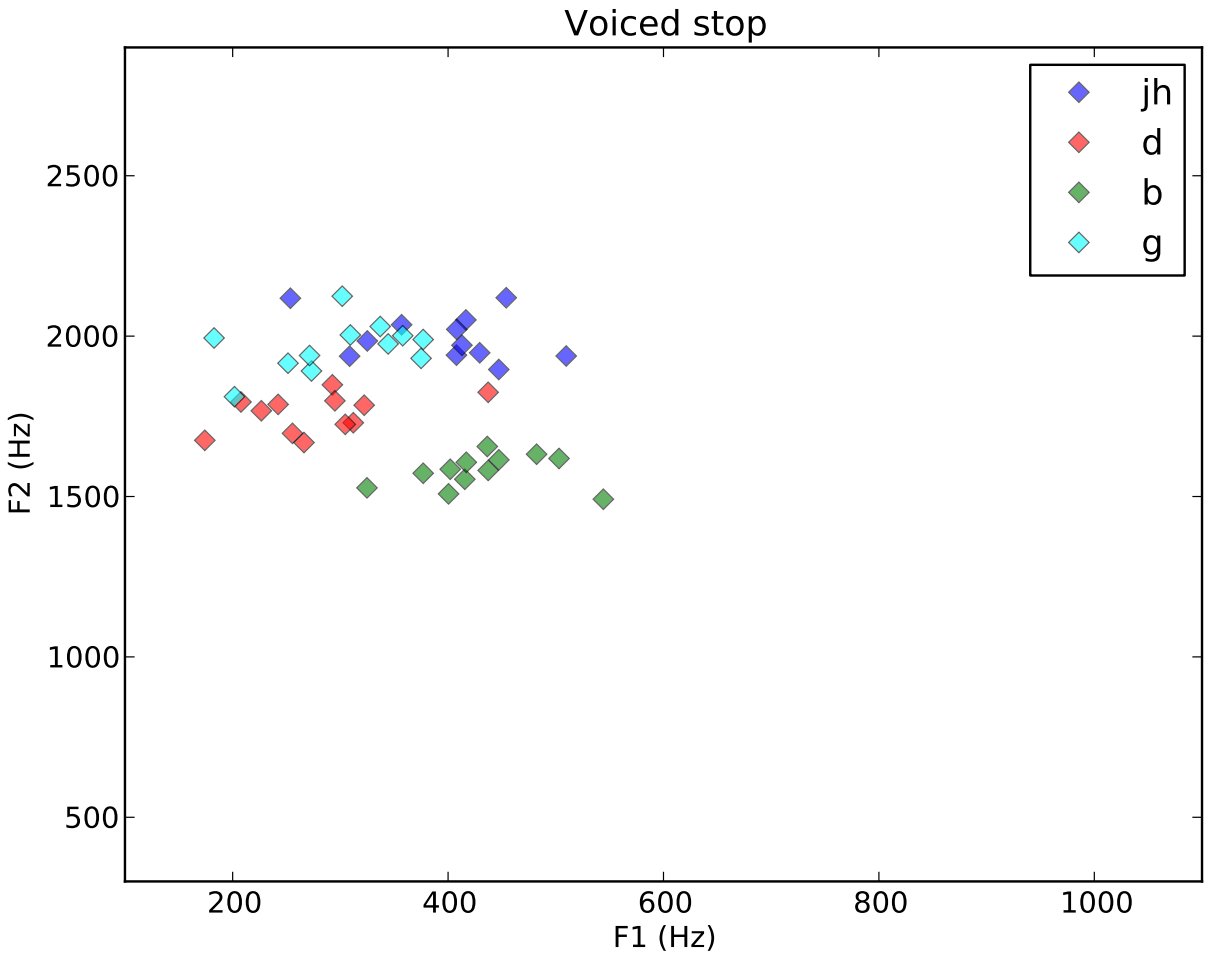
Global Targets – Voiced Fricatives



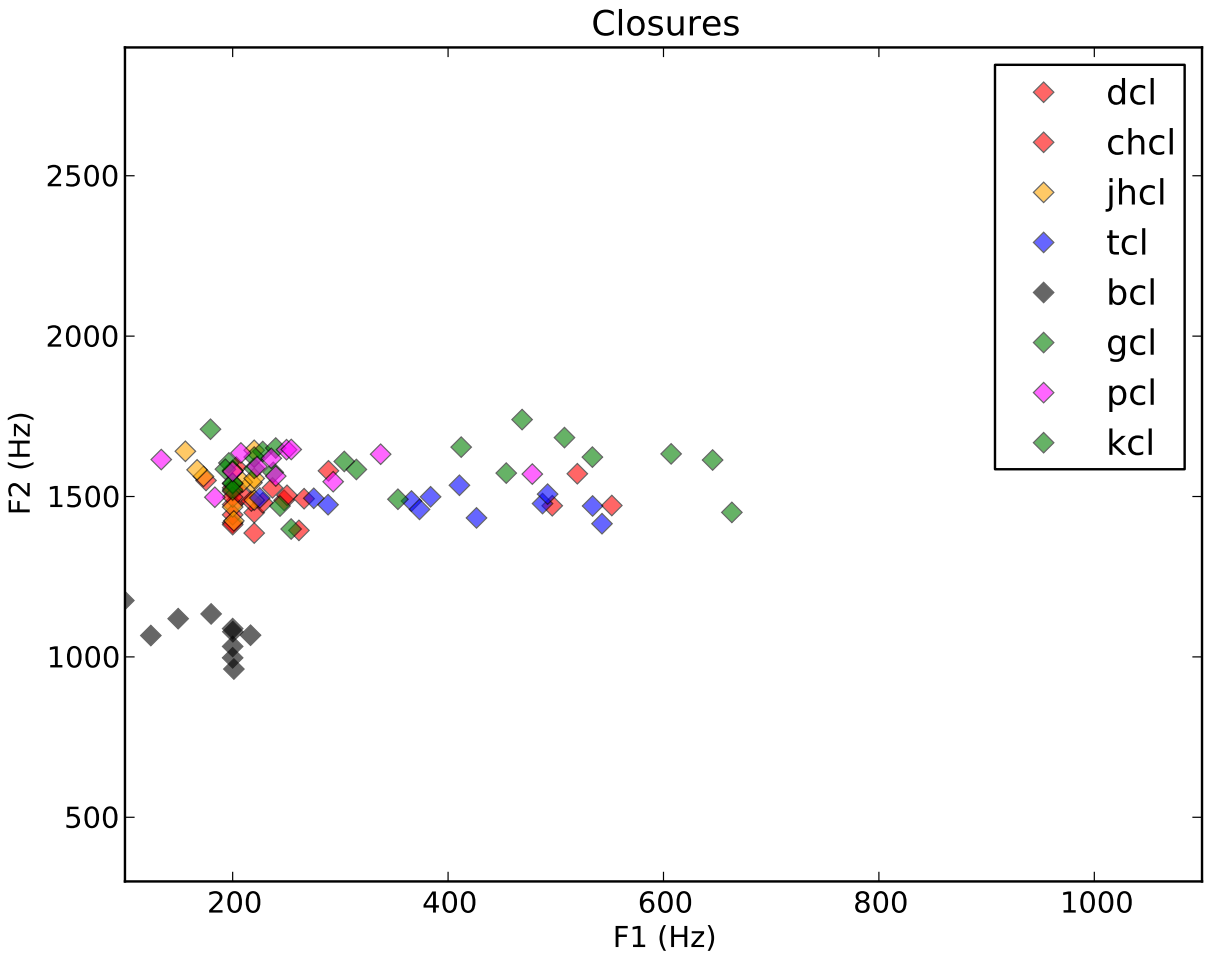
Global Targets – Unvoiced Stops



Global Targets – Voiced Stops



Global Targets – Closures



Global Targets – Vowel Comparison

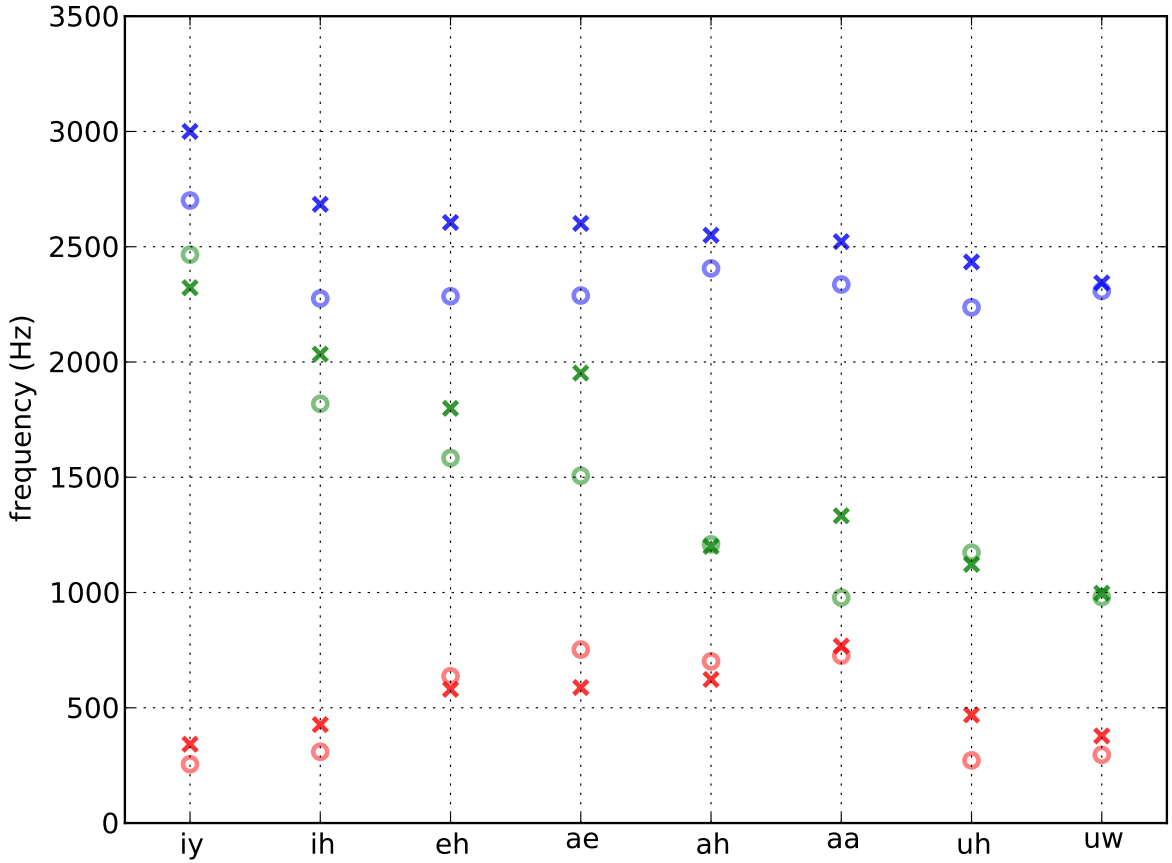


Figure: Vowel targets (circles) compared with Hillenbrand *et al* (x). F1 (red), F2 (green) and F3 (blue).

Global Targets – Consonant Comparison

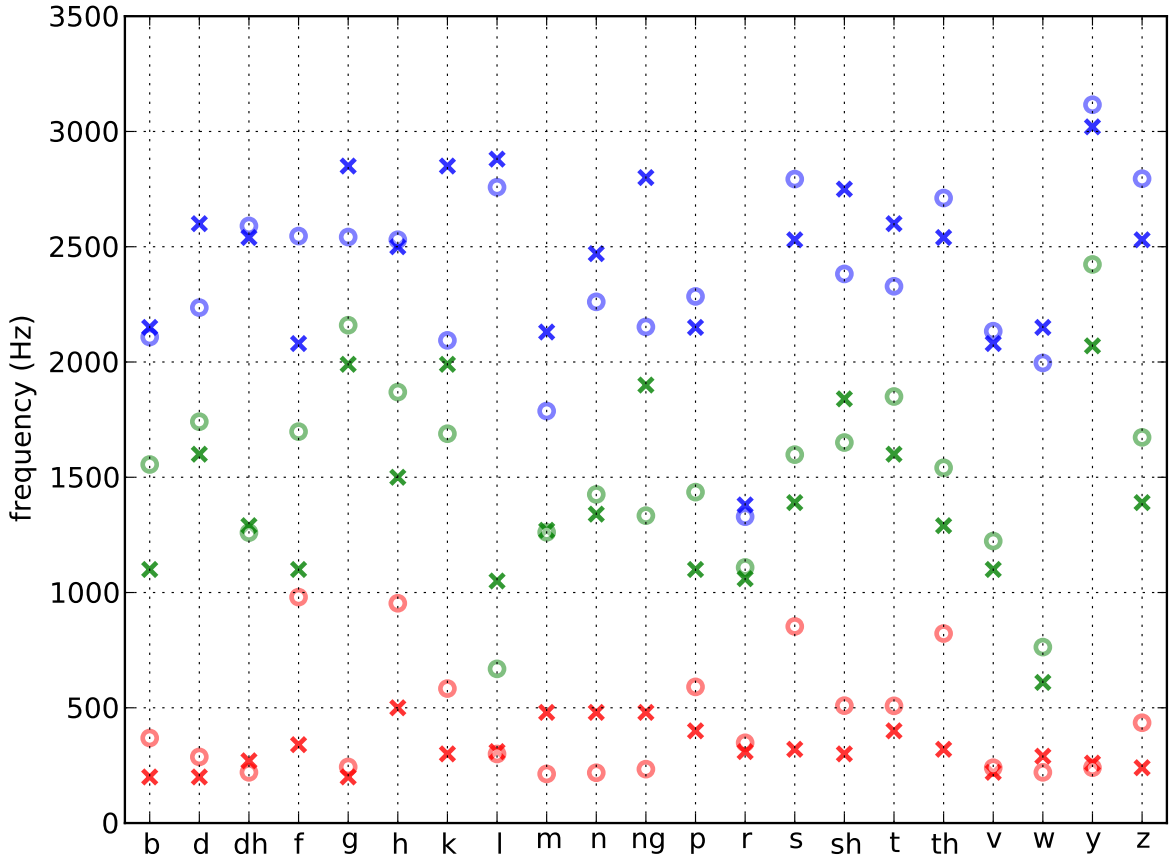


Figure: Consonant targets (circles) and Allen *et al* (x). F1 (red), F2 (green) and F3 (blue).

Asynchronous Formant Movement – CLR

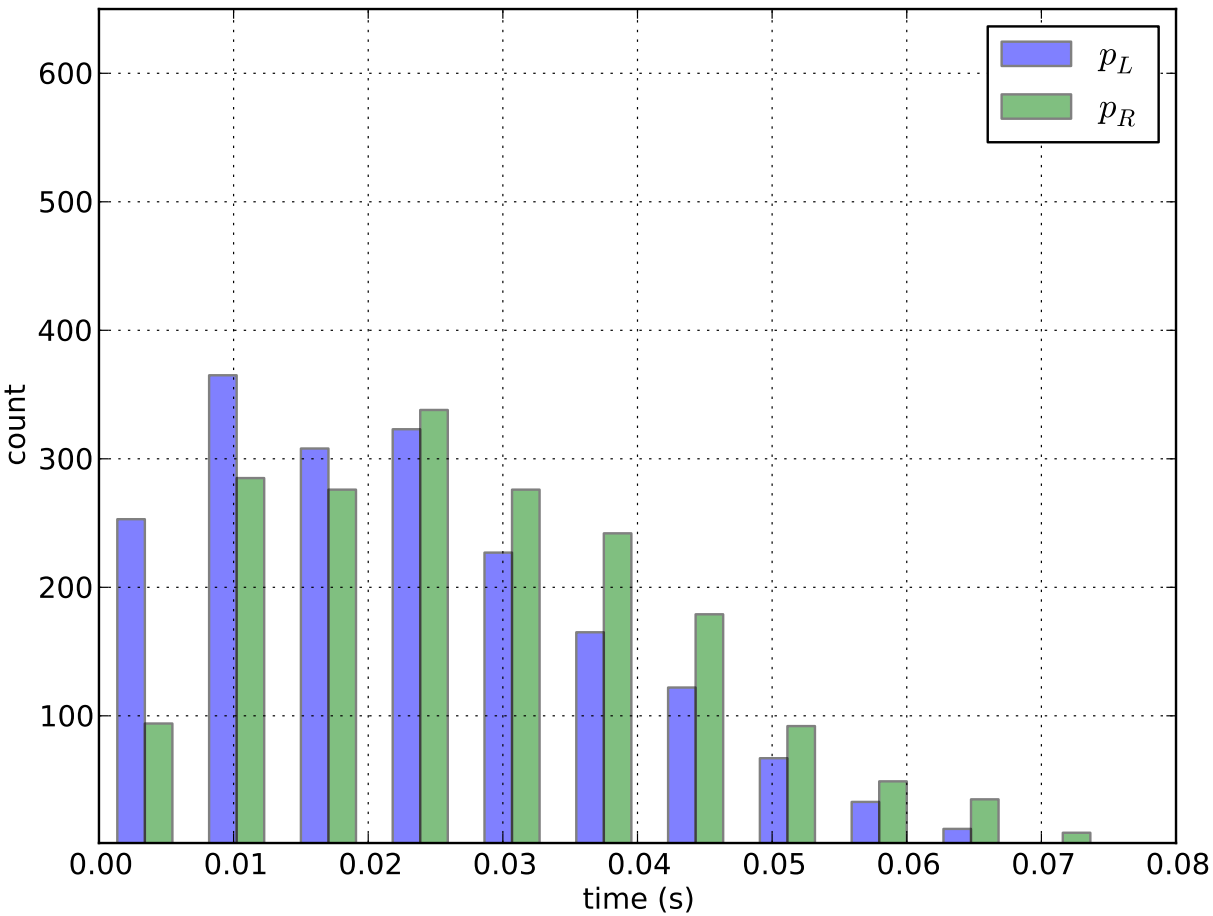


Figure: std dev of p_L and p_R values over F1-F4

Asynchronous Formant Movement – CNV

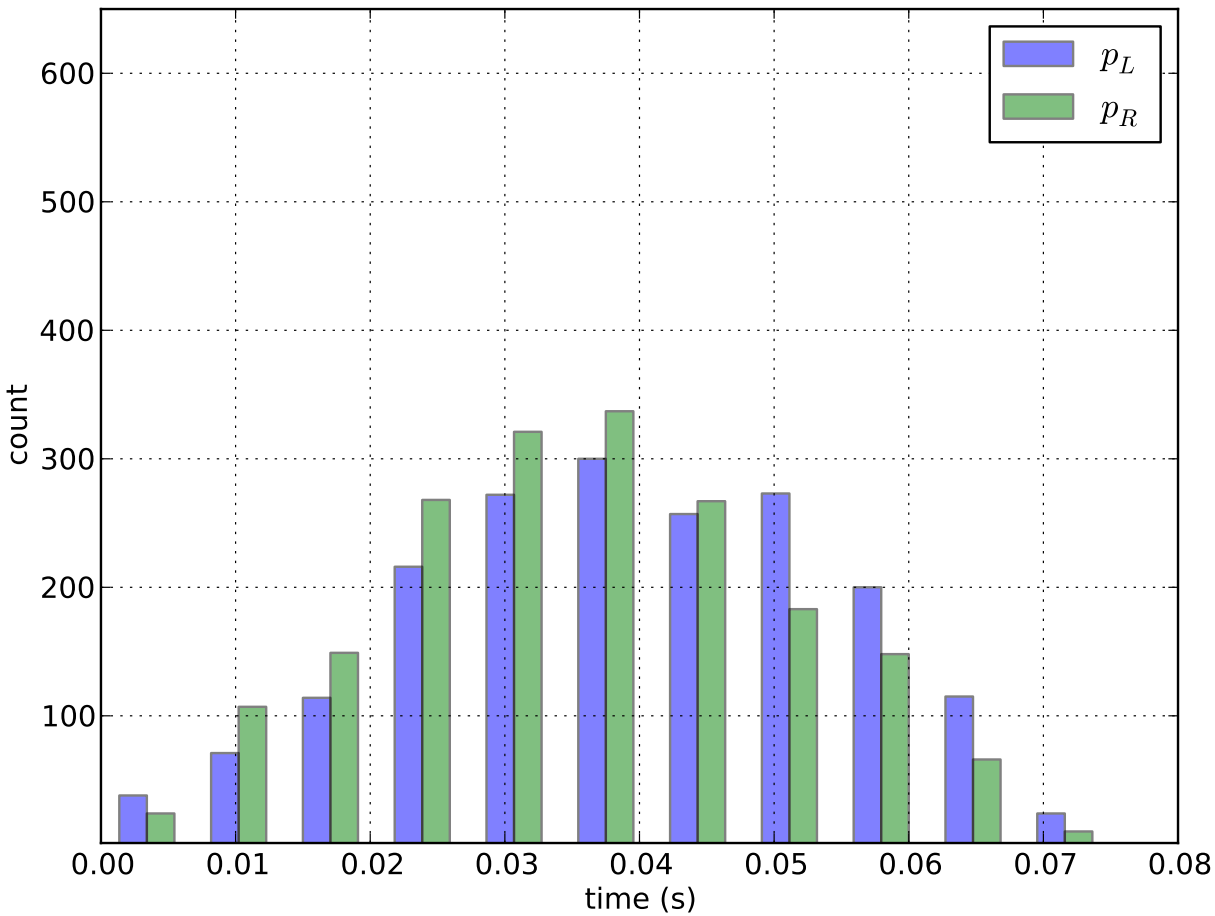
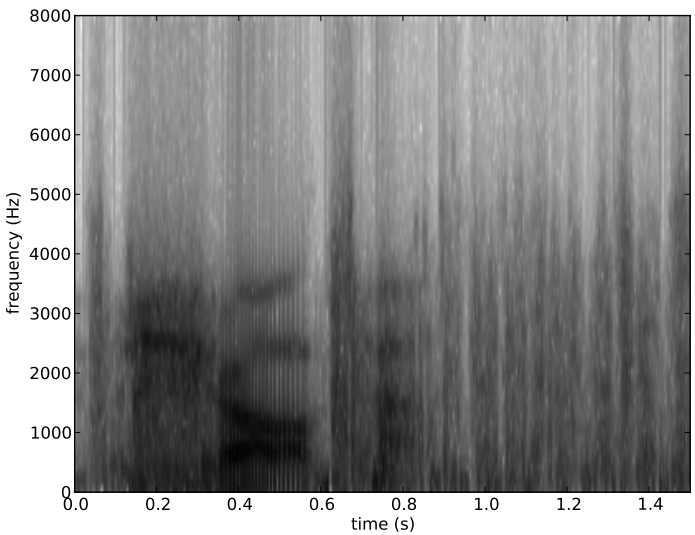
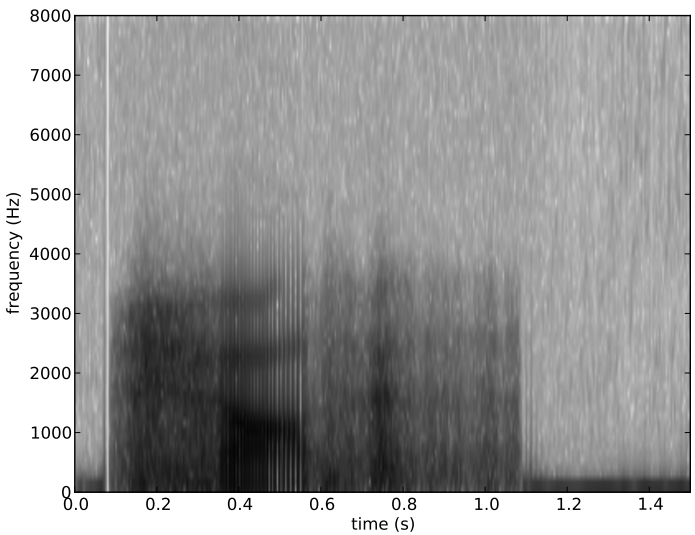


Figure: std dev of p_L and p_R values over F1-F4

Results – Audio Demo

- Resynthesis uses linear predictive coding and formant analysis
- Energy and pitch trajectories preserved
- Samples of both vocoded (left) and vocoded with model trajectories replacing observed trajectories (right)





Planned Experiment – Validation

- **Goal:** Test if resynthesis using model parameters and global targets produces intelligible speech
- Two vocoded stimulus conditions: observed and model formant trajectories
  - Two styles: CLR and CNV
  - Two speakers: male and female
- Use 20% testing material from parallel corpus; 80% training used in determining targets
- Stimuli to be loudness normalized and 12-talker babble noise added at +3 dB SNR
- AMT listening to speech samples; must choose the term heard from a list of five terms – four being decoy terms
- Decoy terms are selected based on closest phonetic similarity to the target term, using common CVC words (e. g. "fan", "van", "than", "pan" and "ban")

Planned Experiment – Clear/Conv Targets

- **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

Matched		Mismatched		All
CNV/CNV	$\leftarrow$	CLR/CNV	$=$	CNV+CLR/CNV
$\downarrow$		$\uparrow$		$\downarrow$
CLR/CLR	$\leftarrow$	CNV/CLR	$\rightarrow$	CNV+CLR/CLR

Conclusions

- Developed a new data-driven methodology to estimate style and context-independent vowel and consonant formant targets
- Developed a joint optimization technique that robustly estimates coarticulation parameters
- Demonstrated analysis and synthesis of speech using new continuous coarticulation model
- Outlined experiments to be conducted to validate model and study style-mismatched targets

Questions

Thank you!