ESTIMATING PHONEME FORMANT TARGETS AND COARTICULATION PARAMETERS OF CONVERSATIONAL AND CLEAR SPEECH

Brian O. Bush and Alexander Kain
Oregon Health & Science University

INTRODUCTION

- Goal: Estimate (1) global speaker-specific phonetic formant targets and (2) degree of coarticulation for both CLR and CNV speech using an explicit coarticulation model.
- Hypothesis: Phoneme formant targets are speaker dependent, but consistent between speaking styles.

TRAJECTORY MODEL

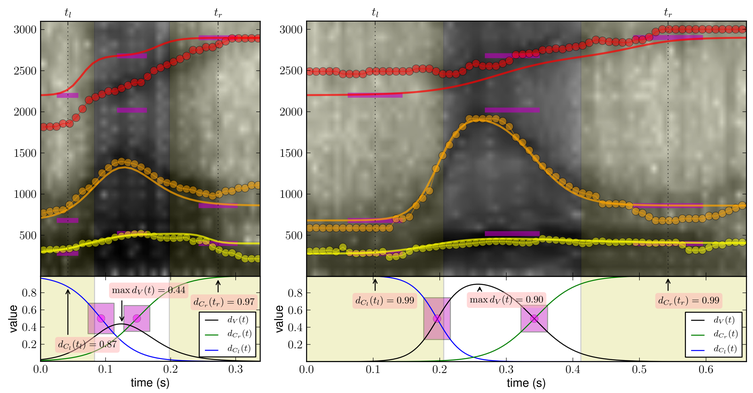
- An individual formant trajectory $X(t)$ of a C_lVC_r word is modeled as a convex linear combination of target formant values

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

- $d_{C_l}(t)$, $d_V(t)$ and $d_{C_r}(t)$ are *coarticulation functions*, based on the sigmoid $d(t;s,p) = (1+e^{s\cdot(t-p)})^{-1}$

$$d_{C_l}(t;s_l,p_l) = d(t;s_l,p_l)$$
$$d_{C_r}(t;s_r,p_r) = d(t;-s_r,p_r)$$
$$d_V(t) = 1 - d_{C_l}(t) - d_{C_r}(t)$$

- 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 → *asynchronous* model.



Example of the word "will" in CNV (left) and CLR (right) speech.

PARALLEL STYLE CORPUS

- 1 male, native speaker of American English
- Keywords are common English CVC words with 21 initial and final consonants and 8 monophthongs (no diphthongs or affricates)
- All sentences spoken *twice* in both clear and conversational styles

ESTIMATING MODEL PARAMETERS

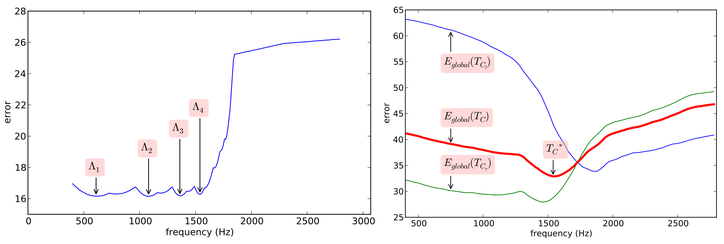
- We define the per-token root-mean-squared (RMS) model error as

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

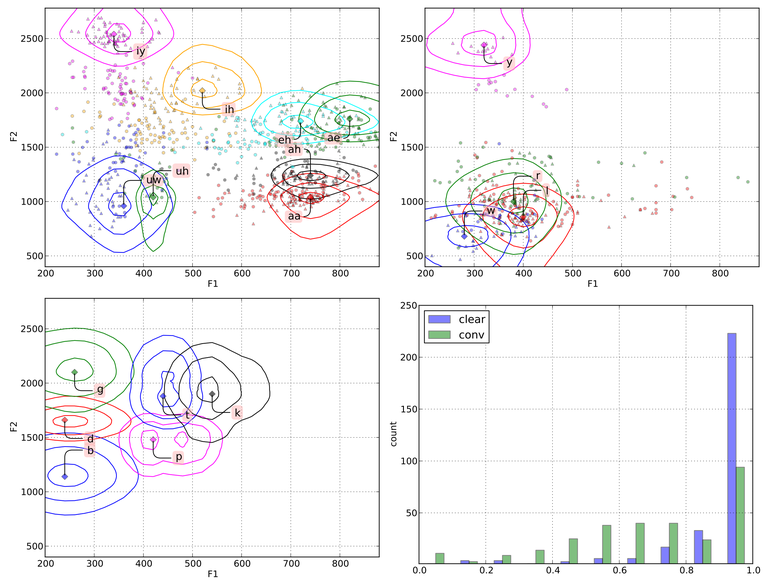
  - t_l is the middle of C_l and t_r is the middle of C_r, w is the confidence of formant measurements at time t, calculated from formant bandwidth
- Given a single token, we sweep parameter $\lambda \in \Lambda$ along a prescribed interval, while grid-searching for the lowest model error for each λ

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

and E_{global} is the average over all sweeps per phoneme



$E_{sweep}(T_{C_l})$ for the CLR token /n/-/eh/-/k/ for F2. $E_{global}(T_{C_l}), E_{global}(T_{C_r})$ and their pre- and postvocalic combination $E_{global}(T_C)$ for C=/n/ for F2 (there are more /n/ in postvocalic context).



Iso-contours based on the global minimum of E_{sweep}. Histogram of $\max d_v(t)$.

PERCEPTUAL VALIDATION

- Stimuli were loudness normalized using an A-weighted RMS measure and 12-talker babble noise was added at a SNR of +3 dB
- 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 stimulus, 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. "chief", "safe", "chip", "chef" and "ship")
  - Similarity measure was the average phonetic distance from target term, where distance between two phonemes was Euclidean distance of a four-dimensional description (sonority, manner, place and height)
- Average intelligibility rates for each condition across all listeners:

Style \ Cond	Natural	Vocoded	Model
CLR	94.1% (3.7)	88.9% (5.4)	84.8% (4.8)
CNV	85.5% (7.6)	68.8% (7.3)	66.7% (9.2)
CLR & CNV	89.7% (4.1)	78.8% (5.6)	75.7% (5.7)

CONCLUSIONS

- 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 appear *consistent* between styles
- Formant targets for vowels and consonants are mostly in agreement with acoustic-phonetic expectations
- CLR style models had higher values of $\max d_v(t)$ than the CNV style
- C_lV coarticulation slopes (s_l) were significantly steeper for CLR tokens than their CNV counterparts

FUTURE WORK

- Investigate optimized techniques for estimating targets
- Generalization of the model to continuous speech
- 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

This work was supported by NSF Grant IIS-0915754

email: bush@ohsu.edu, kaina@ohsu.edu
www: ohsu.edu/cslu