Using a Genetic Algorithm to Estimate Parameters of a Coarticulation Model

Brian O. Bush, John-Paul Hosom, Alexander Kain, Akiko Amano-Kusumoto

Department of Biomedical Engineering, Oregon Health & Science University

Introduction

- Current ASR systems have largely ignored explicit modeling of coarticulation
- Estimation of phoneme targets could be used to improve classification performance
- We present a coarticulation model and a genetic algorithm approach to parameter estimation, including targets

Formant Trajectory Model

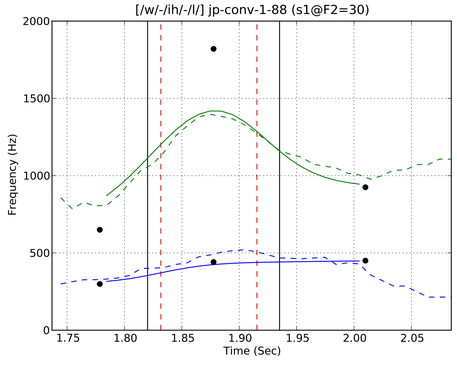
The formant trajectory model:

$$\hat{f}(t) = d_1(t; s_1, p_1) \cdot (T_{C_1} - T_V) + T_V + d_2(t; -s_2, p_2) \cdot (T_{C_2} - T_V)$$

- $\hat{f}(t)$ is the estimated formant trajectory for a specific formant of a CVC word over time t
- T_{C_1}, T_V, and T_{C_2} are the target formant values for consonant C_1, vowel V and consonant C_2
- The coarticulation function, $d(t; s, p)$, for consonants C_1 and C_2 is defined as:

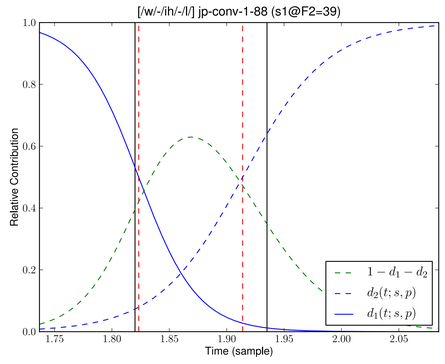
$$d(t; s, p) = \frac{1}{1 + e^{s(t-p)}}$$

- s represents the sigmoid slope and p the sigmoid midpoint position
- The rate of change of each formant frequency trajectory is asynchronous, i.e., s_1 and s_2 are specific to each formant
- The time-point of maximum change, p, is currently synchronous for all formants
- Example observed and estimated formant trajectory plot:



Formant Trajectory Model (cont)

- Example degree of coarticulation using d_1 and d_2 functions:



Corpus

- 242 CVC words with 23 initial and final consonants, eight monophthong vowels (e. g., "cat", "well", etc.)
- Each CVC word recorded twice by a male speaker in two speaking styles: "clear" and "conversational"
- Formants automatically estimated, manually corrected
- Total of 968 ($242 \times 2 \times 2$) CVC tokens
- Training and testing employed k-fold validation, with $k = 20$

Error Surface

- Error surface between estimated model and observed data as a function of two parameters: position p and slope s
- Note the variability of valid optimal s values:

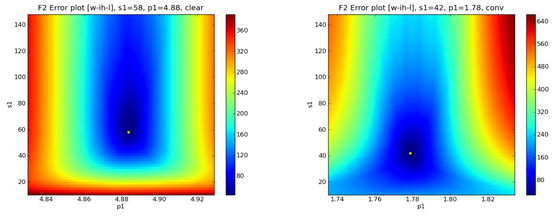


Figure: Model fitting error as a function of coarticulation position (horizontal axis) and slope (vertical axis) for clear and conversational elicitations of the syllable /w ih l/.

Parameter Estimation

- Two sets of parameters to estimate:
  - Coarticulation parameters s_1, s_2, p_1, p_2 are estimated on a per-token basis
  - Formant target parameters (T_{C_1}, T_{C_2}, T_V) are estimated globally for each phoneme, independent of speaking style
- A real-coded GA was used to perform the parameter estimation for both coarticulation and target parameters
- The trajectory error for one CVC word is:

$$Err = \sum_{i=1}^{F} \sqrt{\frac{\sum_{t=t_1}^{t_2} |\hat{f}_i(t) - f_i(t)|^2}{t_2 - t_1}}$$

where $f_i(t)$ and $\hat{f}_i(t)$ are the observed and estimated i^{th} formant trajectories over frames t_1 to t_2; We sum error over first two formants, $F = 2$

- Fitness function is error summed over all tokens in training set
- Mutation selects a variable and uniformly perturbs the variable by a specific factor
- Crossover exchanges parameters for d function between two individuals, e. g., s_1, p_1
- Elitism is employed to keep best solutions found during estimation

Results

- Performed 60 randomly-initialized starts in parameter estimation to arrive at 60 points per phoneme
- Note consistency with which targets are estimated in most cases, which fit our expectations from knowledge of acoustic-phonetics
- Bilabials are consistently clustered around 1200 Hz for F2
- F2 for alveolars (/t/, /d/, and /n/) consistently around 1800 Hz
- Alveolar fricatives /s/ and /z/ are located in the same region of F1/F2 space, as expected as they differ only in voicing
- Approximants tightly clustered at expected locations; /h/ (which does not have a specific target) poorly clustered in middle of F1/F2 space
- However, /p/ did not cluster well, and /s/ and /z/ lower in F2 than expected

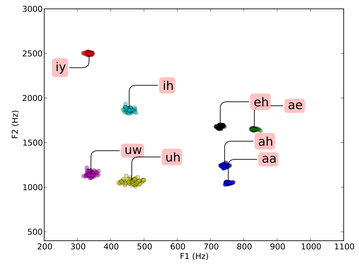


Figure: Estimated targets for monophthong vowels.

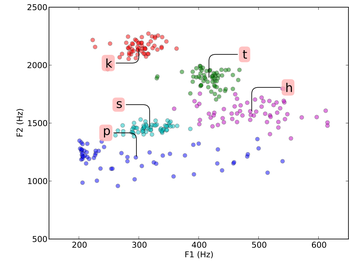


Figure: Estimated targets for /p/, /t/, /k/, /s/, and /h/.

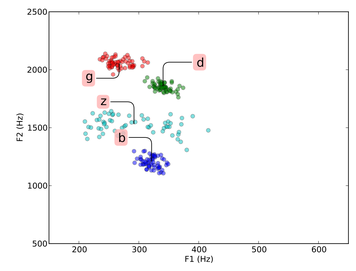


Figure: Estimated targets for /b/, /d/, /g/, and /z/.

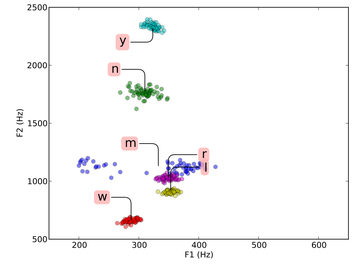


Figure: Estimated targets for /m/, /n/, /w/, /y/, /r/, and /l/

Conclusions & Acknowledgment

- Developed new methodology to estimate parameters of a formant trajectory model
- Clustering of targets and locations of estimated targets indicate the usefulness of this method

- This work supported by NSF grant IIS-091575