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
20000 NW Walker Road, Beaverton, OR 97006 USA
{bush, hosom, kain, akusumoto}@cslu.ogi.edu

Abstract

We present a real-coded genetic algorithm that efficiently estimates parameters of a formant trajectory model. The genetic algorithm uses roulette-wheel selection and elitism to minimize the root mean square error between the observed formant trajectory and the model trajectory. Parameters, including vowel and consonant target values and coarticulation parameters, are estimated for a corpus of English clear and conversational CVC words. Results show consistent consonant formant targets, even when those consonants do not themselves have formant structure. We also present findings of a relationship between a coarticulation parameter and the consonant identity.

Index Terms: coarticulation, formant targets, genetic algorithms.

1. Introduction

While automatic speech recognition (ASR) has made tremendous progress in the last two decades through the use of machine learning techniques and other stochastic approaches, state of the art systems are still not yet fluent conversational partners. Broadcast news transcription average word error rates of less than 10 percent can be achieved [1]. However, human performance on broadcast news transcription is less than 1% [2], indicating that there is still a large amount of room for improvement. In addition, error rates on conversational speech are much higher than 10% (e. g. [3]), highlighting the need for systems to recognize both carefully-articulated speech and more typical speaking styles. Technology for resource-constrained languages may also benefit from explicit modeling of speech dynamics, instead of having to learn such dynamics from a large corpus of data.

Current ASR systems have largely ignored explicit modeling of coarticulation, which is a primary source of acoustic variability in both consonants and vowels. Modeling the effects of coarticulation on vowels is believed to be a crucial step towards more sophisticated dynamic speech models.

In this study, we explore a new approach to estimating parameters of a formant trajectory model which describes the formant trajectories of CVC (Consonant-Vowel-Consonant) words [4]. We will discuss the formant trajectory model (Section 2), the methodology used to perform parameter estimation (Section 3), and results from model validation (Section 4.1). We then discuss results of the parameter estimation, in particular the clustering of consonant targets and relationships between model parameters for degree of coarticulation (Section 4).

This work was supported by NSF Grant IIS-0915754. Akiko Amano-Kusumoto is now with the House Ear Institute.

2. Formant Frequency Trajectory Model

The formant trajectory model that we employ is an extension of the model presented in [4]. The equation for any particular formant is:

$$\hat{f}(t) = d(t; s_1, p_1) \cdot (T_{C_1} - T_V) + T_V + d(t, -s_2, p_2) \cdot (T_{C_2} - T_V) \quad (1)$$

where $\hat{f}(t)$ 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:

$$d(t; s, p) = \frac{1}{1 + e^{s(t-p)}} \quad (2)$$

wherein parameter s represents the sigmoid slope and p the sigmoid midpoint position. In the new version of this model, we describe the rate of change of each formant frequency trajectory asynchronously. Therefore, s_1 and s_2 are specific to each formant. The time-point of maximum change, p, is currently synchronous for all formants. In this paper, we focus our analysis of coarticulation parameters on s_1 of the second formant.

This model allows the estimation of consonant formant targets even when the consonants do not have formant structure. In this case, the consonant targets are based on the patterns of formant movement in the neighboring vowel. Such "virtual" formants and formant targets are an underlying assumption of the locus theory, although estimation of these targets from data is a challenging task.

3. Parameter Estimation

Estimation of the coarticulation parameters s_1, s_2, p_1, p_2 are performed on a per token basis, because there are no known quantitative rules for how coarticulation should behave in a particular speaking style and phonetic context. The formant target parameters (T_{C_1}, T_{C_2}, T_V) are estimated globally for each phoneme, independent of speaking style, under the common assumption that targets are invariant of context or style. Estimation of coarticulation parameters and global targets is treated separately. In both cases, we employ a real-coded genetic algorithm to perform the parameter estimation. A genetic-algorithm approach was applied in order to not assume, for example, a unimodal error surface that could be solved by simpler techniques such as hill-climbing.

3.1. Real-Coded Genetic Algorithm

Genetic algorithms (GA) have been employed to solve a wide variety of engineering problems that are difficult to solve using

traditional optimization approaches. A GA maintains a population of solutions and implements a survival-of-the-fittest strategy in search for the best-fitting solutions. The fittest individuals in the population tend to reproduce and survive to the next generation, thus improving successive generations. Inferior individuals are also allowed to survive and reproduce, which provides diversity within the population. For any GA implementation, six components must be specified: chromosome representation, genetic operators, selection, initialization, termination and fitness function. We shall now discuss these six components with respect to our specific task of estimating parameters for a formant trajectory model.

3.1.1. Chromosome Representation

The chromosome representation determines how a problem is encoded in a GA and also determines how the genetic operators are to be employed. Each chromosome is composed of a sequence of genes. It is reported that, for some problems, real-valued techniques outperform conventional binary representations [5]. In our case, a chromosome can be visualized as a vector of floating-point values. Each gene maps directly to a model parameter, e. g. s_1 or a specific formant target value.

3.1.2. Genetic Operators

The basic search mechanism in a GA uses genetic operators. There are two types of operators: mutation and crossover. These operators are used to create new solutions based on existing solutions that exist within the population. Mutation randomly selects a variable and uniformly perturbs the variable by a specified factor. In our case, mutation of a formant target is a uniformly random shift in a random direction if within constraints.

Crossover takes two chromosomes at random from the population and produces two new chromosomes that are a mixture of the two former. In our implementation, the crossover operator exchanges the parameters for a d function (Equation 2) between two chromosomes, e. g. s_1, p_1.

3.1.3. Selection

Selection is the key mechanism that chooses which chromosomes will survive and move onto the next generation. A probabilistic selection is employed based a chromosome's fitness, computed as described in Section 3.1.5. Our GA uses roulette wheel selection, where the chance of a chromosome being chosen is directly proportional to the percentage of its fitness compared to the total fitness achieved by the generation. For example, if a chromosome scored a fitness of 10 and the generation total fitness was 100, then the chance of selecting the chromosome would be $1/10$. We also employ elitism, wherein we pass the top *n* chromosomes unaltered into the next generation.

3.1.4. Initialization and Termination

An initial population is required to start the GA process. The initial population can be generated randomly, or can use known good starting values. The parameters have the following initial values:

- s_1 and s_2 values are set to 50 Hz/msec for all formants
- p_1 is defined as the C_1V boundary
- p_2 is defined as the VC_2 boundary
- formant targets T_{C_1}, T_{C_2}, and T_V are initialized to values from [6]

Also, the following limits are set:

- min_{p_1} is the middle of C_1; max_{p_2} is the middle of C_2
- max_{p_1} and min_{p_2} are both the middle of V
- t_1, the initial time point for analysis, is the C_1V boundary, unless C_1 is an approximant, in which case t_1 is defined as the middle of C_1
- t_2, the final time point for analysis, is the VC_2 boundary, unless C_2 is an approximant, in which case t_2 is defined as the middle of C_2

The t_1 and t_2 limits are based on (a) where formants can be seen in the speech signal for non-approximants and (b) minimizing the coarticulatory effects of preceding or following phonemes if C_1 or C_2 are approximants.

During parameter estimation, a model is altered many times by genetic operators. Each change is accepted only if all of the following constraints are met:

- $d(t; s_1, p_1) + d(t; -s_2, p_2) \leq 1.0$
- $s \geq 0$ and $s \leq 500$ (in Equation 2)
- $\max_{t \in t_1 tot_2}(1 - d(t; s_1, p_1) - d(t; -s_2, p_2)) \geq 0.6$
- $\max_{t \in t_1 tot_2} d(t; s_1, p_1) \geq 0.9$
- $\max_{t \in t_1 tot_2} d(t; -s_2, p_2) \geq 0.9$
- $p_1 \in [min_{p_1}, max_{p_1}]$
- $p_2 \in [min_{p_2}, max_{p_2}]$

The first constraint simply says that the combined contribution of the consonants must always be less than or equal to 100%. The third constraint says that the contribution of the vowel must be at least 60% at some point during the trajectory. This constraint still allows for highly coarticulated speech. The fourth and fifth constraints say that the consonants must each have a contribution of at least 90% for at least one time point; even in conversational speech, consonants tend to require a high degree of articulatory precision, such as the tongue tip always touching the alveolar ridge for /t/, /d/, or /n/.

Formant target constraints for F1 and F2 are as follows:

- $F2 - F1 > 200$
- $200 < F1 < 900$ and $400 < F2 < 3000$

The GA iterates through generations until stopping criteria are met. Our stopping criterion is a fixed number of generations.

3.1.5. Fitness Function

A fitness function is a method by which we can evaluate the effectiveness of a particular chromosome. This allows the GA to order all members in the population according to their ability to solve the problem.

The trajectory error for one CVC word is:

$$Err_1 = \sum_{i=1}^{F} \sqrt{\frac{\sum_{t=t_1}^{t_2} |\hat{f}_i(t) - f_i(t)|^2}{t_2 - t_1}} \tag{3}$$

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 the error over the first two formants ($F = 2$). We can calculate a global trajectory error:

$$Err_2 = \sum_{k=1}^{K} Err_1^k \tag{4}$$

where the error is summed over all K words in the training set, independent of speaking style.

3.1.6. Usage

Optimal parameter values are found using two sequentially interleaved genetic algorithms: (1) for each token, estimate coarticulation parameters s,p independently, minimizing Equation 3 while keeping formant targets constant and (2) estimate global

formant targets while keeping coarticulation parameters constant, minimizing Equation 4. This process runs serially for a specified number of iterations, typically 100.

The coarticulation parameters are specific to a single CVC token and are estimated separately from the formant targets. Considering a specific instance, parameters are initialized as described in Section 3.1.4. A population of size 50 of these initial chromosomes is created each with some random variation from the initial values. Each chromosome in the population is evaluated to see how it performs at minimizing error in Equation 3. The population is transformed into a new population by using selection, mutation, crossover and elitism. Selection chooses individuals as a function of their fitness. Mutation and crossover change the chromosomes as they enter this new generation, while abiding by constraints in Section 3.1.4. Elitism retains the most successful chromosomes in this new generation. This transformation process from one generation to the next continues until we reach 20 generations, wherein we preserve the best coarticulation parameters found for this specific CVC token. This process continues for all tokens in the corpus.

Next we turn to estimating the global formant targets. A population of formant targets of size 400 is created with initial values from Allen [6] and some random variation. Much akin to the above genetic algorithm, we evaluate each set of targets with respect to the new improved coarticulation parameters in each CVC token; however we now measure how well each set of targets minimizes Equation 4. The population of targets is transformed into a new population via selection, mutation, crossover and elitism. The number of generations is also fixed at 20.

4. Experiments

The corpus we used is composed of 242 CVC words spoken twice in both clear and conversational style by a male speaker [4]. There are eight vowels (both front and back) represented in the corpus. Diphthongs are not represented in this corpus. The formants and phoneme boundaries were automatically extracted and then hand-corrected.

In the following experiments, we are concerned with the *consistency* of estimated parameters from different random seeds.

4.1. Model Validation

We checked the validity of results by initializing the model with random values (for both s and formant target parameters) instead of the values defined in Section 3.1.4, and comparing the resulting error values. To accommodate the suboptimal starting point, the stopping criterion for random initialization was set to 800 iterations of both GAs with 20 generations each. The random initialization yielded a total error of 546.9, which is within 4% of the total error based on standard initialization, 527.8. Visual inspection of the targets in the F1/F2 space indicated that the method of initialization had little impact on most of the formant target values, although greater variance was noted for some phonemes when using random initialization. We conclude that, because the GA yields similar results with different initializations, this method of parameter estimation yields generally valid results.

4.2. Result 1: Consonant Formant Target Estimation

We used 60 random seeds with the standard initialization to obtain 60 sets of target vectors in the F1/F2 space, $\mathbf{T}_{C_1}$, $\mathbf{T}_V$, $\mathbf{T}_{C_2}$. Figures 1 through 3 show the estimated targets for sixteen phonemes in C_1. Since there is no gold standard with which to evaluate these results, (a) we note the consistency with which the targets are estimated in most cases, and (b) we evaluate the target estimation performance in terms of how the targets fit our expectations from knowledge of acoustic-phonetics.

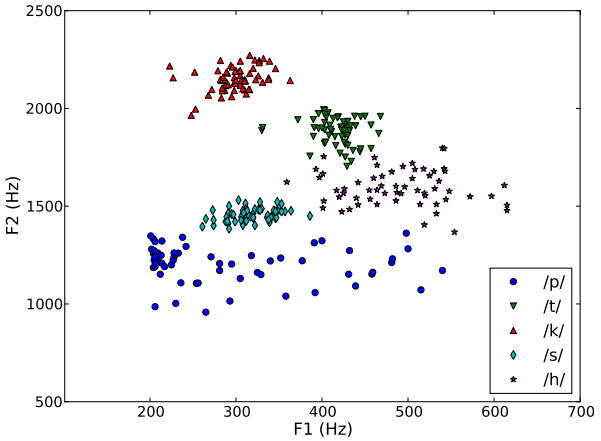

Figure 1: Estimated formant targets in F1/F2 space for C_1 phonemes /p/, /t/, /k/, /s/ and /h/.

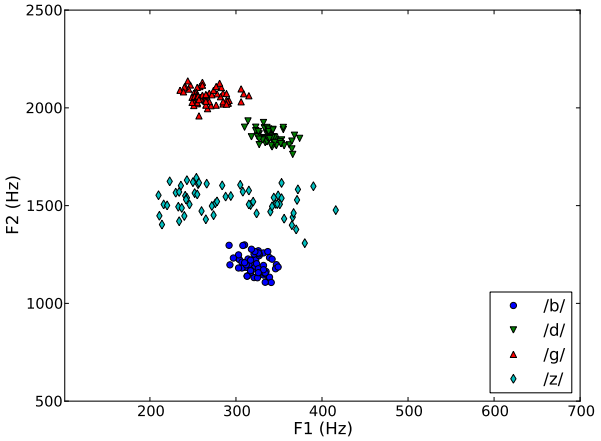

Figure 2: Estimated formant targets in F1/F2 space for C_1 phonemes /b/, /d/, /g/ and /z/.

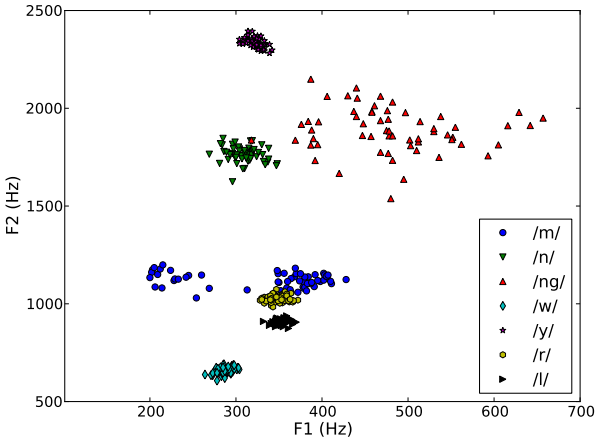

Figure 3: Estimated formant targets in F1/F2 space for C_1 phonemes /m/, /n/, /ng/, /w/, /y/, /r/ and /l/

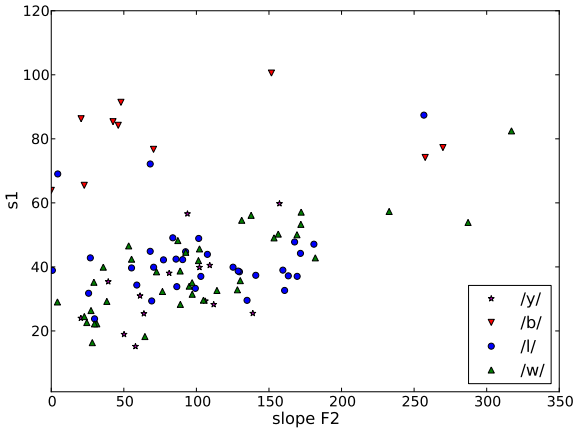

Figure 4: Linear relationship in some phonemes (/w/, /l/, /y/, and /b/) between observed F2 slope at vowel onset and the estimated model parameter s_1 for F2.

For the bilabials, there is consistency in the placement of F2 (at around 1200 Hz), although /p/ does not cluster as well as other phonemes and /m/ has clusters at two values of F1.

For the alveolars /t/, /d/, and /n/, F2 is consistently around 1800 Hz, as would be expected from acoustic-phonetic knowledge. The alveolar fricatives /s/ and /z/ are located in the same region of the F1/F2 space, as expected because they differ only in voicing. However, there is a wide range of F1 values for /z/, and F2 at 1500 Hz is lower than the expected value of 1800 Hz.

The velars should have two clusters of targets, a high target for front vowels and a low target for back vowels. Perhaps because the model has been constrained to one target for all vowels, a single target that is higher than the alveolars is seen. The phonemes /k/ and /g/ have a close F2 target (with 2200 Hz for /k/ and 2100 Hz for /g/), but the F2 target for /ng/ has a wide variation (1600 to 2200 Hz) that is still unexplained.

The unvoiced stops (in Figure 1) do not cluster as well as the voiced stops (in Figure 2). The reason for this is unclear. The approximants (in Figure 3) show targets tightly clustered and at regions near their expected values; low F1 and F2 for /r/ and /l/ ([7]), even lower F1 and F2 values for /w/, and a low F1 and very high F2 for /y/.The phoneme /h/ (in Figure 1) shows targets scattered around the middle of the vowel space, which may represent the best single-target compromise for a phoneme that takes on the targets of its neighboring vowel.

In summary, the targets obtained by the data-driven GA approach conform reasonably well to our expectations of where consonant targets should be based on acoustic-phonetic knowledge. There are some exceptions, e. g. F2 targets below 1800 Hz for /s/ and /z/; and while most phonemes cluster well around a single point, /p/, /ng/, and /z/ do not cluster as well; and /m/ shows two separate clusters. It is unclear whether this behavior is due to limitations of the model, limitations of the estimation method, or limitations of the general knowledge provided by acoustic-phonetics. Despite these exceptions, data-driven estimation of consonant targets appears to be generally successful.

4.3. Result 2: Patterns of Coarticulation Parameters

The parameter s_1 for the second formant shows a linear relationship with features of the speech signal for only /w/, /l/, /y/, and /b/ (Figure 4). Despite the lack of a clear relationship for all phonemes, the value of s_1 does seem to depend on the identity

phoneme	/p/	/t/	/k/	/b/	/d/
mean	49.3	53.6	65.1	76.5	54.7
std dev	20.0	29.5	30.1	14.5	20.1
phoneme	/g/	/l/	/r/	/y/	/w/
mean	62.1	40.8	33.6	34.9	40.4
std dev	21.9	10.7	10.1	14.7	11.3

Table 1: Mean and standard deviation of second-formant s_1 values for ten phonemes

of the consonant. Table 1 shows the mean and standard deviation of s_1 for ten phonemes. It can be seen that the average s_1 can be different by a factor of two (34 for /r/ and 76 for /b/), and that the different average values of s_1 are not easily attributed simply to the variance of this parameter. While some relationships appear based on the place of articulation (both /t/ and /d/ have similar average s_1 values, and both /k/ and /g/ have similar average s_1 values), it is difficult to generalize such relationships (e. g. /p/ and /b/ have quite different average s_1 values). Characterization of s_1 will be important when the model is used to estimate parameters on a per-token basis for automatic speech recognition; our initial work on such parametrization indicates that there is too much flexibility in the model unless s_1 and s_2 are further constrained.

5. Conclusions

In this study, we presented a methodology to estimate parameters for a formant trajectory model. This method is a data-driven approach to discover formant target and coarticulation parameters. The usefulness of the proposed method is indicated by the model validation, the clustering of targets, and the locations of estimated targets generally conforming with expectations based on acoustic-phonetic knowledge.

6. References

[1] L. Nguyen, S. Abdou, and M. e. a. Afify, "The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System," in *2004 Rich Transcriptions Workshop, Pallisades, NY*, 2004.

[2] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, July 1997.

[3] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRIs conversational speech recognition system," in *in Proc. Interspeech*, pp. 2141–2144, 2005.

[4] A. Amano-Kusumoto and J. Hosom, "Effect of speaking style and speaking rate on formant contours," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4202–4205, IEEE, 2010.

[5] A. Wright, "Genetic algorithms for real parameter optimization," *Foundations of genetic algorithms*, vol. 1, pp. 205–218, 1991.

[6] J. Allen, M. Hunnicutt, and D. Klatt, "From Text to Speech: The MITalk System," 1987.

[7] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech. A Dynamic Approach*. New York etc.: Springer, 1992.