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

Brian O. Bush and Alexander Kain

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

ABSTRACT

We present a data-driven formant model and methodology for discovering its parameters, namely phoneme targets and coarticulation functions for consonant-vowel-consonant (CVC) words from fully-automatic formant data. The model uses formant targets that are speaker dependent, but independent of speaking style and phonemic context. We used a global error measure to search for optimal formant targets for all phonemes, including classes of sounds where formants are not directly observable. Analysis of coarticulation parameters found significant differences in parameters between clear and conversational speech. Estimated formant targets were largely in agreement with acoustic-phonetic expectations. An intelligibility test validated that resynthesized CVC words using modeled formant trajectories were nearly as intelligible as resynthesized CVC words using observed formant trajectories.

Index Terms: coarticulation, formants, clear speech.

1. INTRODUCTION

We present a methodology that models formant trajectories as a sum of phoneme targets weighted by coarticulation functions. In contrast to other work on formant target estimation and coarticulation that has focused on context-specific clear speech [1, 2, 3, 4, 5, 6, 7, 8], we model speech independently of context and speaking style. Our previous work [9] used a genetic algorithm to estimate formant targets; however, we found multiple optimal solutions in formant targets. Therefore, in this study, we performed an exhaustive search of the formant target space using fully automatic formant estimation with the primary goal of finding context- and style-independent phoneme targets. Our model has also been expanded to consider data in unvoiced regions, whereas previously formant trajectories were modeled from voicing onset to offset of the vowel only [2, 10, 11, 12].

In this paper, we first introduce a parallel style corpus (Sec. 2), our proposed formant trajectory model (Sec. 3), and an approach to estimating the model's phoneme target and coarticulation parameters (Sec. 4). We then report on the resulting model parameters (Sec. 5) and the outcome of a speech intelligibility test (Sec. 6) before concluding (Sec. 7).

2. PARALLEL STYLE CORPUS

One male speaker produced the same speech material in two different speaking styles. For conversational speech (CNV), the speaker was asked to speak as if one were talking with a colleague at a natural pace [13]. 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 together" [14].

This work was supported by NSF Grant IIS-0915754.

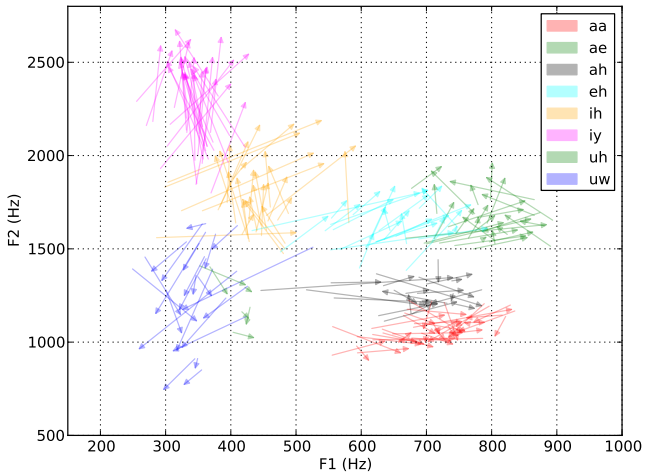


Fig. 1: F1/F2 formant frequency relationship between CNV and CLR at vowel centers, arrows pointing from the former to the latter.

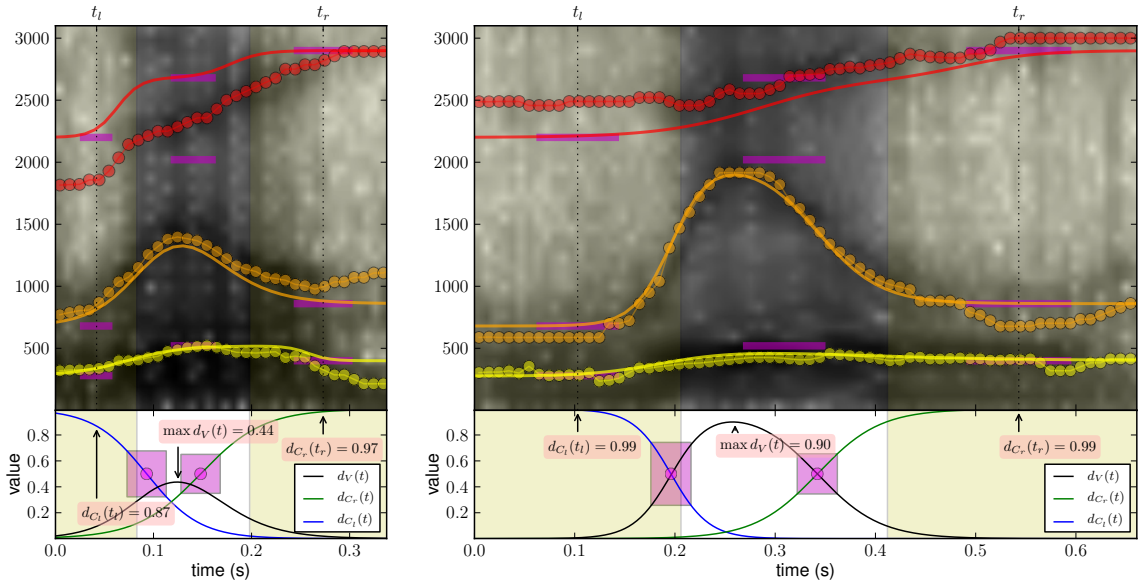
The corpus was composed of 212 consonant-vowel-consonant (CVC) words (e. g. "cat", "well") constructed from a combination of 21 initial and final consonants, and eight monophthong vowels, spoken in a carrier sentence [11, 12]. The carrier sentences provided neutral meaning and had a consistent phoneme /d/ before the target word in a sentence final context (e. g. "I know the meaning of the word *will*"). Since /ao/ is often pronounced as /aa/ in West-Coast American English we merged these two phonemes into /aa/. Affricates were not represented in this corpus since they are composed of two allophones and thus have two different formant targets. Each token was rendered *twice* in *both* styles, for a total of 212 words $\times$ 2 styles $\times$ 2 renditions = 848 CVC tokens. Formants were automatically estimated using a standard formant tracker [15, 16]. Fig. 1 shows the F1/F2 formant frequency relationship between the two styles at vowel centers. Note the expanded vowel space of CLR speech, as compared to CNV speech [17].

3. TRAJECTORY MODEL

The formant trajectory model presented here is an extension of previous work [11, 2, 10]. An individual formant trajectory $X(t)$ of a CVC word is modeled as

$$\hat{X}(t;\Lambda) = d_{C_l}(t) \cdot T_{C_l} + d_V(t) \cdot T_V + d_{C_r}(t) \cdot T_{C_r} \qquad (1)$$

which is a convex linear combination of T_{C_l}, T_V, and T_{C_r} representing formant target values for the prevocalic consonant C_l, the medial vowel V, and the postvocalic consonant C_r. The scalars $d_{C_l}(t)$, $d_V(t)$, and $d_{C_r}(t)$ are *coarticulation functions* based on the sigmoid



(a) CNV speech

(b) CLR speech

Fig. 2: Example of the model for the word "will" in CNV and CLR styles. Upper panels show spectrograms (with phoneme-specific shading), observed formant trajectories $X(t)$ (thin yellow (F1), orange (F2), and red lines (F3) with circles), model formant trajectories $\hat{X}(t)$ (thick line), and model targets T (magenta lines). Lower panels show coarticulation functions for F2, where magenta circles are locations of sigmoid centers p, and sigmoid slopes s indicated as diagonals of magenta boxes.

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

$$\begin{aligned} d_{C_l}(t;s_l,p_l) &= d(t;s_l,p_l) \\ d_{C_r}(t;s_r,p_r) &= d(t;-s_r,p_r) \\ d_V(t) &= 1-d_{C_l}(t)-d_{C_r}(t) \end{aligned} \tag{2}$$

where $\{s_l,s_r\}$ represent positive sigmoid *slope* (slow versus fast transition), and $\{p_l,p_r\}$ sigmoid midpoint *position* (and point of maximum slope), measured relative to their respective phoneme boundaries. Previously, an exponential function was used to model formant trajectories of vowels in /b,d,g/ contexts [2], however a sigmoid was selected to better fit cases where C_l or C_r is an approximant. Note that the rate of change represented by d functions is *not* normalized with respect to the length of the underlying transition. The complete set of parameters $\Lambda = \{T_{C_l},T_V,T_{C_r},s_l,p_l,s_r,p_r\}$ are specific to an individual formant trajectory, and thus the model approximates concurrent formant trajectories *asynchronously*. Model parameters are constrained by the following conditions: (1) the allowed range of p_l is from the center of C_l to the center of V, (2) similarly the allowed range for p_r is from the center of V to the center of C_r, and (3) $d_{C_l}(t)+d_{C_r}(t) \leq 1\,\forall t$ to ensure convexity. An example application of the model to the word "will" is shown in Fig. 2.

4. ESTIMATING MODEL PARAMETERS

In order to discover optimal formant targets, we define the per-token model weighted root-mean-square error (RMSE)

$$E(X,\Lambda) = \sqrt{\frac{1}{\sum_{t=t_l}^{t_r} w(t)} \sum_{t=t_l}^{t_r} w(t)\cdot\left(X(t)-\hat{X}(t;\Lambda)\right)^2} \tag{3}$$

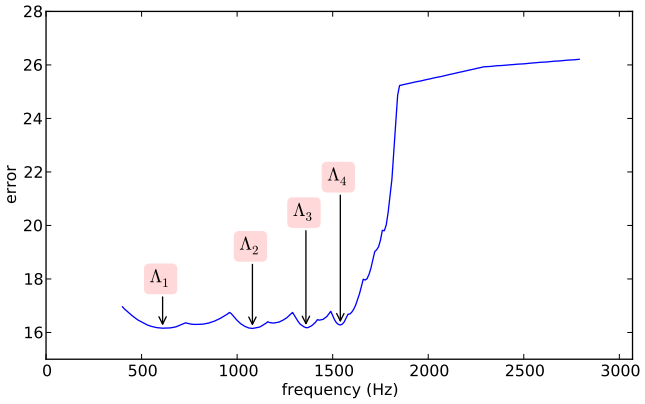
where $X(t)$ and $\hat{X}(t)$ are the observed and estimated individual formant trajectories, t_l is the center of C_l, and t_r is the center of C_r. The weighting factor $0 \leq w(t) \leq 1$ indicates our confidence in the formant measurement, calculated from formant bandwidth. We iterate over all combinations of discretized parameter values while calculating the model error, thus finding the parameter set Λ^* that provides the lowest error. We perform this for F1, F2, and F3 separately. Specifically, we iterate over all combinations of target parameters $\{T_{C_l}, T_V, T_{C_r}\}$ using F1=200,220,...,900 Hz, F2=400,420,...,2800 Hz, and F3=900,920,...,3700 Hz, and coarticulation parameters $\{s_l,p_l,s_r,p_r\}$ using $s = 10,30,\ldots,110$ and the *relative* $p = -40,-30,\ldots,40$ ms, calculated by subtracting the corresponding phoneme boundary time from the p in Eq. 2.

Given a specific token, we can sweep any individual parameter $\lambda \in \Lambda$ along its prescribed interval, while searching the remaining parameters for the lowest model error

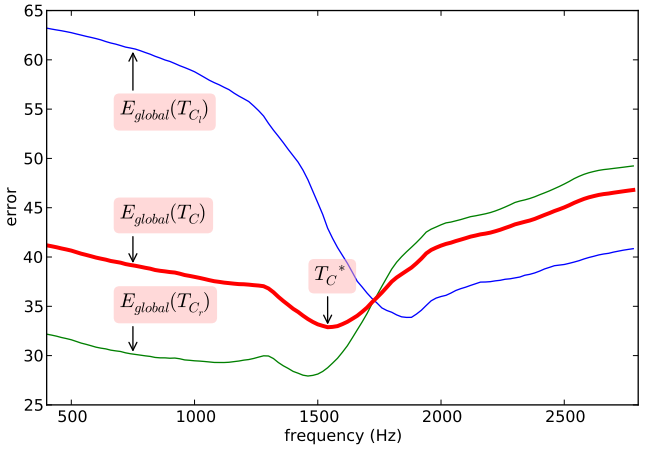
$$E_{sweep}(\lambda) = \min_{\Lambda\setminus\lambda} E(X,\Lambda) \tag{4}$$

where $\setminus$ represents set subtraction. The minimum error sweep $E_{sweep}(\lambda)$ allows consideration of the effect of an individual parameter on the lowest possible per-token RMSE (see Eq. 3). Fig. 3a shows an example plot of $E_{sweep}(T_{C_l})$; note that there are several solutions where different T_{C_l} values lead to similarly low errors. All solutions share $\{T_V = 1880, T_{c_r} = 1680, s_l = 50, s_r = 30, p_r = -0.08\}$, but there is a T_{C_l} / p_l interaction: to achieve a similarly low error the value of p_l increases as T_{C_l} increases, while all other parameters stay the same, e. g. $\Lambda_1 = \{T_{C_l} = 620, p_l = -0.04,\ldots\}$ and $\Lambda_4 = \{T_{C_l} = 1540, p_l = -0.01,\ldots\}$. These multiple solutions for T_{C_l} show that for some tokens it would prove difficult for a hill-climbing approach to discover the true optima, and token-specific target estimation could possibly yield several different targets, thus validating our choice of using a method that avoids local minima.

We collected individual token-based error sweeps for T_{C_l}, T_V and T_{C_r} for all tokens, and constructed a global error sweep E_{global} for



(a) $E_{sweep}(T_{C_l})$ for the CLR token /n/-/eh/-/k/ for F2



(b) $E_{global}(T_{C_l})$, $E_{global}(T_{C_r})$, and their pre- and postvocalic combination $E_{global}(T_C)$ for C=/n/ for F2 (there are more /n/ in postvocalic context)

Fig. 3: Minimum Error Sweeps

each phoneme by averaging all error sweeps for that phoneme independent of context and style. For consonants, we combined their pre- and postvocalic instances (see Fig. 3b). For each of the 29 phonemes in our corpus, a global phoneme target value was found at the minimum of its associated $E_{global}(T)$ function, for each formant F1, F2, and F3 independently. Phoneme targets /r/ and /g/ required slight manual adjustment by shifting F3 formant targets upwards to satisfy F3−F2>200 Hz. Finally, using the estimated global phoneme targets, we re-estimated each token's optimal set of coarticulation parameters $\{s_l, p_l, s_r, p_r\}$. However, to best match formant slopes, we substituted $\Delta X(t)$ and $\Delta\hat{X}(t)$ for $X(t)$ and $\hat{X}(t)$ in Eq. 3.

5. ANALYTIC RESULTS

Final formant target locations are shown in Fig. 4. The contours are generated by the normalized quantity $\max_T E_{global}(T) - E_{global}(T)$ for each point in F1/F2 space for the global phoneme-specific error sweeps, resulting in a hill-like contour with bands at 1, 5 and 10%. The formant targets for vowels and consonants are largely in their expected locations [18, 19]. We note that some consonants appear to have target *ranges* instead of points, e. g. /p/ and /k/.

For a meaningful analysis of coarticulation parameters $\{s_l, p_l, s_r, p_r\}$, we only included tokens with neighboring targets that differ by at least 300 Hz. Regarding the sigmoid slope parameters s_l and s_r, their means were as follows: for F1, $\overline{s_l}$=72 (σ=37) and $\overline{s_r}$=76 (σ=38), for F2, $\overline{s_l}$=46 (σ=31) and $\overline{s_r}$=55 (σ=37), and for F3, $\overline{s_l}$=62 (σ=41) and $\overline{s_r}$=68 (σ=38). We defined Δs_l as CLR s_l minus CNV s_l from parallel tokens, and analogously for Δs_r. For the former, we computed means $\overline{\Delta s_l}$=7 (σ=43) for F1, $\overline{\Delta s_l}$=10 (σ=41) for F2, and $\overline{\Delta s_l}$=19 (σ=38) for F3, indicating that CLR F1, F2, and F3 coarticulation functions have *faster* transitions than their CNV counterparts, for this speaker. This result was validated using a one-sample t-test. Significance tests for Δs_r exhibited no such relationship for F1, F2, or F3 (possibly due to the sentence-final position of the CVC word).

Fig. 5 shows histograms for coarticulation function values at consonant centers $d_{C_l}(t_l)$ and $d_{C_r}(t_r)$, and the maximum value during the vowel $\max d_v(t)$. We observe that the CLR style histograms have more occurrences of higher values than the CNV style. For vowels, this was expected since their trajectories are more likely to reach their target in CLR style ($\max d_V(t) \approx 1$), while undershoot is more prevalent in CNV style ($\max d_V(t) \ll 1$).

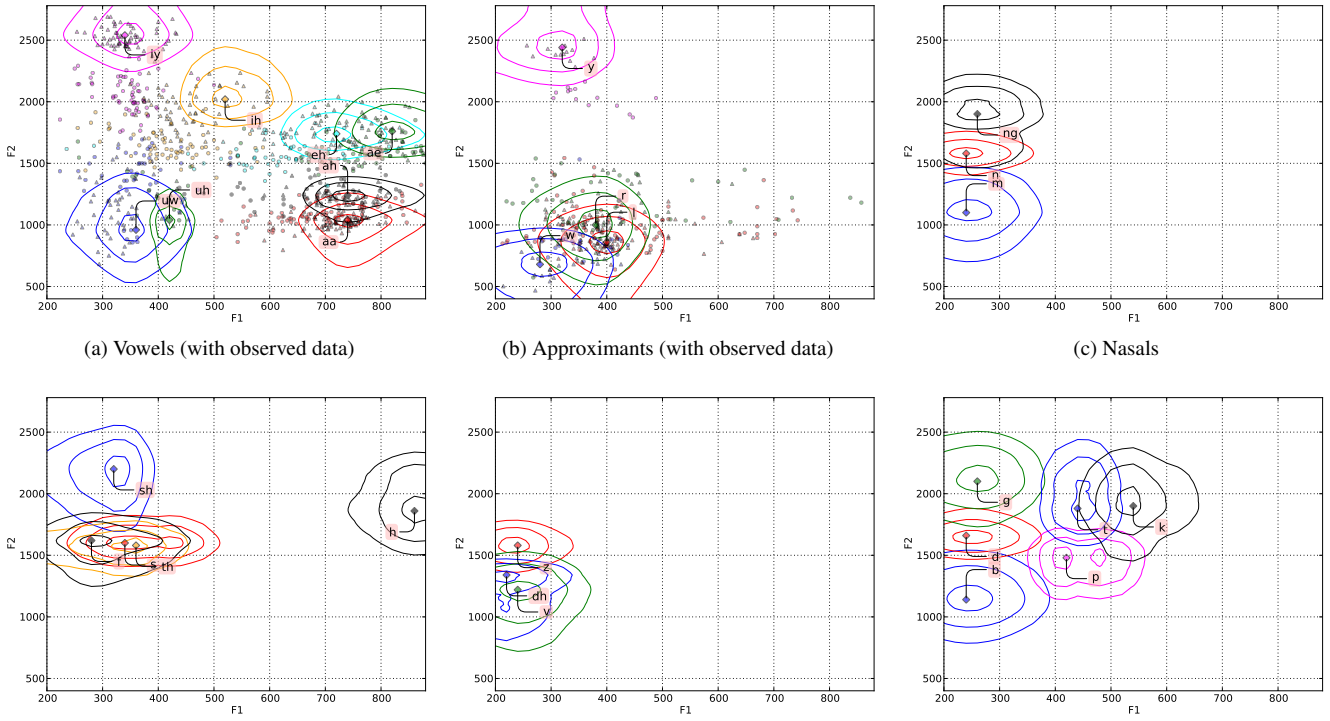
6. PERCEPTUAL VALIDATION

A perceptual evaluation was conducted by means of a speech intelligibility test to examine whether resynthesis from model parameters produces speech that is as intelligible as vocoded speech, thus validating the model and its estimation procedure.

Three stimulus conditions were created for 212 CVC words in CNV and CLR styles: (1) natural, (2) vocoded, and (3) model, for a total of 212 words × 2 styles × 3 conditions = 1272 stimuli in a Latin squares design. All stimuli were loudness normalized using an A-weighted [20] RMS measure 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. Resynthesis was accomplished by using a hybrid linear predictive coding / formant analysis-synthesis vocoder with energy and pitch trajectories preserved. For the model condition, automatically estimated formant frequency trajectories were replaced with those of the model.

During testing, an individual listened to stimulus waveforms through circumaural headphones, binaurally in a quiet room. Each listener was presented with 212 stimuli in randomized order. With each stimulus, the listener was provided a closed set 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 CVC words. The similarity measurement measured the average phonetic distance from the target term, where distance between any two phonemes was defined as the Euclidean distance of a four-dimensional manually derived description (sonority, manner, place and height) of each phoneme. 18 adults aged 23–55 with self-reported normal hearing, all native speakers of American English and unfamiliar with the goals of the study, participated in the experiment.

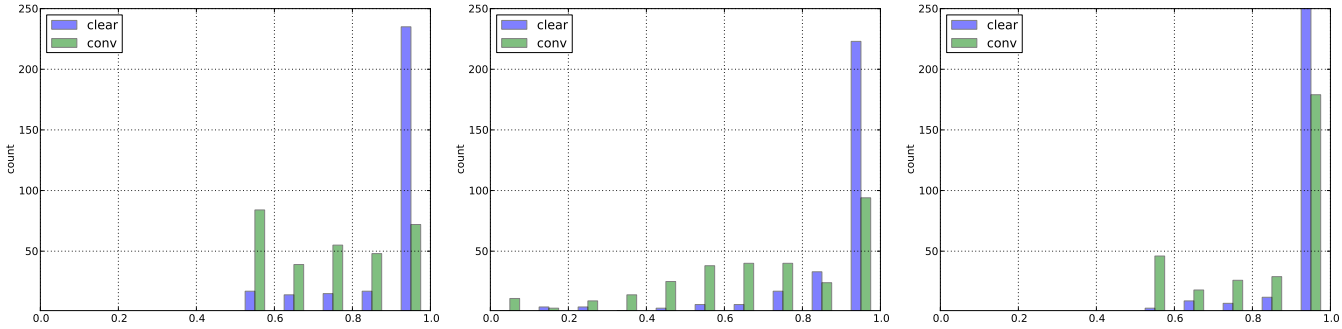
The average proportion of words heard correctly were as follows: natural 89.7% (σ=4.1), vocoded 78.8% (σ=5.6), and model 76.0% (σ=5.7), combining styles. We performed a planned one-sample two-tailed t-test comparing the vocoded and model conditions, yielding a value of 2.2 with 17 degrees of freedom, significant at 0.05. Measuring the effect size using Cohen's d function [21] yielded a value of 0.6, considered a “moderate” effect size. Considering speech style separately, we obtained for CLR speech: natural 94.1% (σ=3.7), vocoded 88.9% (σ=5.4), and model 84.8% (σ=4.8), while for CNV speech we obtained: natural 85.5% (σ=7.6), vocoded 68.8% (σ=7.3), and model 66.8% (σ=9.3). The CLR speech benefit is apparent when comparing the CNV and CLR natural conditions.



(a) Vowels (with observed data) (b) Approximants (with observed data) (c) Nasals

(d) Unvoiced fricatives (e) Voiced fricatives (f) Unvoiced and voiced stops

Fig. 4: Formant targets (with observed data where available) and iso-contours based on the global minimum of E_{sweep}



(a) $d_{C_l}(t_l)$ (b) $\max d_v(t)$ (c) $d_{C_r}(t_r)$

Fig. 5: Histograms of coarticulation function values

Moreover, we observe that CNV speech appears to be relatively more difficult to resynthesize with our chosen vocoder. However, for CNV speech, the model condition is trailing the vocoded condition by only 2.0 percentage points, versus 4.1 for CLR speech.

7. CONCLUSIONS

In this study, we presented a new data-driven methodology to estimate vowel and consonant targets for one speaker. We identified the existence of local optima in the error for single tokens. This finding highlights the source of difficulties in finding consistent formant targets in our previous work. Our new approach estimates global phoneme formant targets robustly using fully automatic estimation methods and highlights statistically significant differences in coarticulation parameters between CLR and CNV speech. An intelligibility test validated that resynthesized CVC words using modeled formant trajectories were nearly as intelligible as those using observed formant trajectories. While the model condition did not perform as well as the vocoded condition, the difference of 3% can be considered small, especially in light of the compactness of the model parameter set versus the raw formant trajectories, with a compression ratio of approximately 1:12 for our corpus. We speculate that some phoneme targets have ranges or multiple targets; e. g. it is possible that targets are different in the onset vs coda of a stressed syllable.

Our findings demonstrate fully automatic estimation of phoneme formant targets and provides evidence that targets are consistent between speech styles. Future work will apply our model to more speakers.

8. REFERENCES

[1] B. Lindblom, "Spectrographic study of vowel reduction," *The Journal of the Acoustical Society of America*, vol. 35, no. 5, pp. 1773–1781, 1963.

[2] D. J. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," *Journal of the Acoustical Society of America*, vol. 81, no. 1, pp. 155–165, 1987.

[3] D. J. Broad and F. Clermont, "Target-locus scaling methods for modeling families of formant transitions," *Journal of Phonetics*, vol. 38, no. 3, pp. 337 – 359, 2010.

[4] B. Lindblom and H. M. Sussman, "Dissecting coarticulation: How locus equations happen," *Journal of Phonetics*, vol. 40, no. 1, pp. 1 – 19, 2012.

[5] R. Sproat, ed., *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*. Kluwer, 1998.

[6] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 769–773, July 1955.

[7] P. C. Delattre, "From acoustic cues to distinctive features," *Phonetica*, vol. 18, pp. 198–230, 1968.

[8] S. E. Öhman, "Numerical model of coarticulation.," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.

[9] B. O. Bush, J.-P. Hosom, A. Kain, A. Amano-Kusumoto, "Using a genetic algorithm to estimate parameters of a coarticulation model," in *Interspeech*, pp. 2677–2680, 2011.

[10] X. Niu and J. P. H. van Santen, "A formant-trajectory model and its usage in comparing coarticulatory effects in dysarthric and normal speech," in *Proc. of the Third International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, (Firenze, Italy), Dec. 2003.

[11] A. Amano-Kusumoto and J.-P. Hosom, "Effect of speaking style and speaking rate on formant contours," in *ICASSP*, pp. 4202–4205, 2010.

[12] A. Amano-Kusumoto, J.-P. Hosom, and A. Kain, "Speaking style dependency of formant targets," in *Proc. of InterSpeech*, 2010.

[13] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech.," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, 1985.

[14] K. S. Helfer, "Auditory and auditory-visual recognition of clear and conversational speech by older adults," *Journal of the American Academy of Audiology*, vol. 9, pp. 234–242, 1998.

[15] K. Sjölander and J. Beskow, "WaveSurfer — an open source speech tool," in *Proc. of ICSLP*, pp. 464–467, 2000.

[16] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *The Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S55–S55, 1987.

[17] S. H. Ferguson and D. Kewley-Port, "Talker Differences in Clear and Conversational Speech: Acoustic Characteristics of Vowels," *J Speech Lang Hear Res*, vol. 50, no. 5, pp. 1241–1255, 2007.

[18] J. Allen, M. S. Hunnicutt, and D. Klatt, eds., *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[19] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels," *Journal of the Acoustical Society of America*, vol. 97, pp. 3099–3111, May 1995.

[20] B. Moore, *An Introduction to the Psychology of Hearing*. Acad. Press, 2003.

[21] J. Cohen, *Statistical Power 2nd Ed*. Taylor & Francis Group, 1988.