



Speaking Style Dependency of Formant Targets

Akiko Amano-Kusumoto, John-Paul Hosom, Alexander Kain

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

{akusumoto, hosom, kain}@cslu.ogi.edu

Abstract

Previous work on formant targets has assumed that these targets are independent of the speaking style. In this paper, we estimate consonant and vowel targets in a database of “clear” and “conversational” speech, using both style-independent and style-dependent models. The test-set errors and clustering of the estimated target values indicate that for this corpus, formant targets depend on the speaking style. Vowel classification accuracy was then tested on estimated target values and compared with classification based on observed formant values. Token-based style-independent classification shows greater accuracy for conversational speech (82.19%) than observed-value classification (73.97%).

Index Terms: formant target, clear speech, formant contour model.

1. Introduction

Clear (CLR) speech has been studied because of increased intelligibility and distinct acoustic characteristics as compared with conversational (CNV) speech [1]. Kain et al. [2] investigated which acoustic feature(s) contribute to increased intelligibility of CLR speech. By modifying the feature combination of spectrum and phoneme duration, the intelligibility of CNV speech was improved (degree of contribution 56%), while individual features, spectrum or phoneme duration alone, did not always significantly improve intelligibility. It is necessary to further understand the relationship between the formant contours (representing both spectrum and duration) and speech intelligibility.

It is known that the degree of formant undershoot varies based on speaking rate, vowel duration, speaking style, and word stress [3]. In CNV speech production, formant undershoot is more dynamic than in CLR speech [4]. In this paper, we examine whether the speaker sets the target of the articulators (tongue, lips and jaw), which are reflected in formant contours, differently in CNV and CLR speech.

Our previous work showed that a formant contour model fitted well to words with a limited context (*/wVl/* and */tVl/* words) [5]. The model was developed with a linear combination of style-independent target values and coarticulation functions. The assumption was that the speaker aims for the same target position regardless of the speaking style. However, it may be true that the speaker aims for a different target position in different speaking styles. Therefore, the style-independent targets might have caused an increased error. In order to investigate the style dependency of the target position, in the current work we estimate style-independent formant targets as well as style-dependent formant targets. The mean error rate and clustering of targets is examined to compare the performance from the

two methods. Finally, as an application of the formant contour model, vowels are classified using the Mahalanobis distance with both style-independent and style-dependent token-based targets, and results are compared with classification based on observed steady-state formant values.

2. Formant Contour Model

As described in [5], the equation of the formant contour model is given as

$$\hat{\mathbf{F}}(t) = d_1(t; s_1, p_1) \cdot (\mathbf{T}_{C_1} - \mathbf{T}_V) + \mathbf{T}_V + d_2(t; -s_2, p_2) \cdot (\mathbf{T}_{C_2} - \mathbf{T}_V) \quad (1)$$

where $\hat{\mathbf{F}}(t)$ is the estimated formant contour of a CVC word as a function of time t , as in a study by Niu and van Santen [6]. \mathbf{T}_{C_1} , \mathbf{T}_V , and \mathbf{T}_{C_2} are the target formant vectors of the pre-vocalic consonant (C_1), vowel (V), and postvocalic consonant (C_2), respectively. The target formant vectors consist of the first four formant values (4×1 dimension). The function $d(t; s, p)$ represents the degree of coarticulation of C_1 and C_2 , and is proportional to the differences in target formant values. The sigmoid function is used for the coarticulation function:

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

which is characterized with two coefficients, s (slope) and p (slope position).

Constraints for the parameters are the following:

- $d_1(t; s_1, p_1) + d_2(t; -s_2, p_2) \leq 1.0$.
- $s \geq 0$ in Equation 2.
- p_1 (or p_2) in Equation 2 may range from middle of C_1 (or V) to the middle of V (or C_2).
- $200 < F1 < 900, 400 < F2 < 3000, 900 < F3 < 3700, 2500 < F4 < 5000$.
- $F2 - F1 > 200, F3 - F4 > 200, F4 - F3 > 200$.

3. Experiment 1

The corpus consists of 242 consonant-vowel-consonant (CVC) meaningful English words, with 23 initial and final consonants and 8 vowels. The CVC words were spoken by a male speaker with two speaking styles (CNV and CLR style). The total of 968 (242 words \times 2 repetitions \times 2 speaking styles) tokens are fitted to the formant contour model with both style-independent and style-dependent target estimation.

This work was supported by NSF Grant IIS-0915754

3.1. Parameter Estimation: Style-independent targets

We estimate model parameters (s_1, p_1, s_2, p_2) per token, and formant target values \mathbf{T} for each phoneme for both speaking styles. (This is called context-independent and style-independent target estimation.) In our corpus, 484 samples per speaking style are available for training and testing. Each sample is tested with a jackknife procedure, with 218 (or 217) randomly selected training samples from each speaking style and 24 (or 25) test samples in 20 test groups, total of 436 in training and 48 test sets. Parameters are initialized with $s = 0.7$, $p =$ corresponding phoneme boundaries. Formant target values are initialized with values provided by Allen et al. [7]

The error function to estimate parameters (s_1, p_1, s_2, p_2) is

$$Errr1^{(k)} = \sqrt{\frac{\sum_{t=T_1^{(k)}}^{T_2^{(k)}} \left| \mathbf{w} \left(\widehat{\mathbf{F}}(t)^{(k)} - \mathbf{F}(t)^{(k)} \right) \right|^2}{N^{(k)}}} \quad (3)$$

where $\mathbf{F}(t)^{(k)}$ and $\widehat{\mathbf{F}}(t)^{(k)}$ are the observed and estimated formant contours for the k -th word, converted to Bark scale. $N^{(k)}$ is the number of frames from $T_1^{(k)}$ to $T_2^{(k)}$. The contribution to the error from F1 and F2 are weighted more than F3 and F4, represented as $\mathbf{w} = [1 \quad 1 \quad 0.5 \quad 0.1]$. $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the middle of C_1 (or C_2) when C_1 (or C_2) is an approximant. Otherwise, $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the C_1V (or VC_2) boundary. Thus, $Errr1$ is calculated only when formant values are available.

Formant targets are estimated by minimizing

$$Errr2 = \sum_{style=1}^S \sum_{k:phn \in k}^K Errr1_{style}^{(k)} \quad (4)$$

$Errr1^{(k)}$ is summed over all the words (k) that have a particular phoneme (phn) and over all speaking styles. K is the number of words ($K = 218$) used in the training set. K has been reduced to 218 to allow direct comparison of $Errr2$ with $Errr3$ (below). S is the number of speaking styles ($S = 2$). Each parameter is estimated using a steepest ascent hill-climbing approach, as described in [5].

3.2. Parameter Estimation: Style-dependent targets

We estimate model parameters (s_1, p_1, s_2, p_2) in $d(t; s, p)$ per token, and formant target values \mathbf{T} for each phoneme in each speaking style. (This is called context-independent and style-dependent target estimation, where one set of formant values is estimated per phoneme for each speaking style.) Each sample is tested once with a jackknife procedure, with 48 (or 49) test samples and the remaining 436 (or 435) training samples from each speaking style in 10 different groups.

The same error function ($Errr1$ in Equation 3) is used to estimate parameters (s_1, p_1, s_2, p_2) , while $Errr2$ (Equation 4) is changed to

$$Errr3_{style} = \sum_{k:phn \in k}^K Errr1_{style}^{(k)} \quad (5)$$

$Errr1^{(k)}$ is summed over all the words (k) that have a particular phoneme (phn) for each speaking style. K is the number of words ($K = 436$) used in a training set. The initial values and the steepest ascent hill-climbing approach for the optimization method are the same as style-independent target estimation.

	Training set		Test set	
	CNV	CLR	CNV	CLR
Style-Indep.	0.2740	0.2975	0.3076	0.3104
Tgt Estimation	(0.1246)	(0.1091)	(0.1572)	(0.1164)
Style-Dep.	0.2592	0.2747	0.2780	0.2851
Tgt Estimation	(0.1242)	(0.1041)	(0.1362)	(0.1107)

Table 1: Mean error $E_{s,target}$ (in Bark) in training and test sets.

3.3. Goodness of fit

The root mean square error $E_{s,target}^{(k)}$ (Bark) is calculated for each word k as in Equation 3. The subscript $s, target$ indicates that the error depends on the variables s and $target$, while the value p for slope position is adjusted to the best fit in each token. Mean $E_{s,target}^{(k)}$ values over the samples in the training and test sets are reported for each style in Table 1. The error difference between training set and test set is relatively small, indicating little over-fitting to the training set. The error rate was successfully reduced with style-dependent target estimation both in training and test sets.

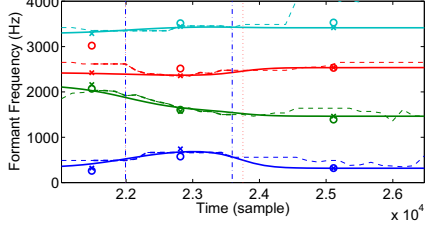
Figures 1a and 1b represent the modeled formant contour of the word “yes” ($j \varepsilon s$) (solid lines) as well as observed contour (dotted lines) in both CNV and CLR styles. The estimated target values (crosses) are the result of style-dependent targets for each speaking style. For the unvoiced consonant /s/ where formant values are not available, the virtual formant target values are estimated reasonably well. Figures 1c and 1d show the corresponding coarticulation functions ($d_1(t_1; s_1, p_1)$ and $d_2(t_2; s_2, p_2)$). The slope of the coarticulation functions show how fast the contour moves from one target to the other. The $E_{s,target}^{(k)}$ values were 0.1964 (Figure 1a) and 0.1803 Bark (Figure 1b) in CNV and CLR styles, respectively.

The test-set results were submitted to three way analysis of variance (ANOVA) (2 methods \times 2 speaking styles \times 8 vowel identities). The main effect of the method was significant ($F(1, 1926) = 16.62$, $p < 0.0001$) as well as the vowel identity ($F(7, 1926) = 5.91$, $p < 0.0001$). No significance was observed due to the speaking style ($p > 0.05$). These results show that the model fitted better with style-dependent target estimation, but fitted equally well to the CNV and CLR data. Post-hoc analysis (HSD) showed that the error rate on words with the vowel /u/ was significantly different from vowels (/i:/, /e/, /æ/ and /ʌ/), at $\alpha = 0.0018$ level, and that no other combinations were significantly different. This is due to the fact that the formant contour of /u/ has high variability because of the neighboring consonants.

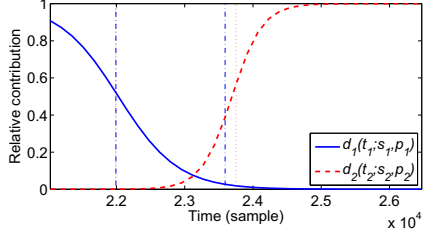
3.4. Estimated formant target values

The parameter estimation process resulted in 20 sets of formant target values with the style-independent target estimation, and 10 sets for two speaking styles with the style-dependent target estimation. Figure 3 shows estimated style-dependent target values for \mathbf{T}_V , \mathbf{T}_{C_1} and \mathbf{T}_{C_2} from the 10 different training sets, while only mean values are shown for the style-independent targets. Selected consonants (/t/, /ɹ/, /l/, /j/ and /w/) are shown in Figures 3b and 3c.

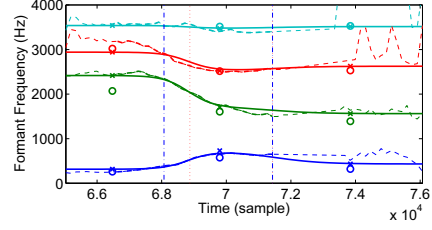
The estimated vowel target values (\mathbf{T}_V) are tightly clustered for each speaking style with different data partitions, which supports the hypothesis that style-dependent targets exist. The estimated vowel target values from style-independent estimation (shown in black squares in Figure 3a) are close to the CLR tar-



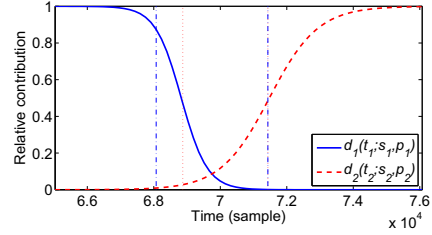
(a) Modeled formant contour (solid lines) and observed contour of CNV style (dotted lines). The $E_{s,target}^{(k)}$ was measured 0.1964 Bark.



(c) Coarticulation functions $d(t; s, p)$ for CNV.



(b) Modeled formant contour (solid lines) and observed contour of CLR style (dotted lines). The $E_{s,target}^{(k)}$ was measured 0.1803 Bark.



(d) Coarticulation functions $d(t; s, p)$ for CLR.

Figure 1: The results of formant contour model for the word “yes (*lj ε s*)”. Circles (same in (a) and (b)) are the the initial values, while crosses are estimated values with style-dependent estimation (different in (a) and (b)). Blue vertical dash-dot lines show the phoneme boundaries, while vertical red dashed lines represent p_1 (left) and p_2 (right).

	/p/	/t/	/k/	/b/	/d/	/g/
Style dep.	1.27	0.21	0.11	0.29	0.14	0.11
Style indep.	1.48	1.34	0.43	0.51	0.33	0.40
	/ɹ/	/l/	/j/	/w/	/m/	/n/
Style dep.	0.08	0.05	0.05	0.05	0.18	0.12
Style indep.	0.17	0.10	0.12	0.15	0.95	0.36

Table 2: Average distance from estimated target values to centroid, in Bark, for twelve consonants and both speaking styles.

gets (filled blue triangles) from the style-dependent estimation.

For the voiced consonants (\mathbf{T}_{C_1} and \mathbf{T}_{C_2}) /w/, /j/, /ɹ/ and /l/, where formant values are available, the estimated target values are tightly clustered, similar to the vowels. The consonant /ɹ/ in the C_1 position, in particular, shows two distinct target classes for CNV and CLR speaking styles. For the voiced stop consonants /b/, /d/, /g/ (not shown in figures), even though formant values are not available, the estimated targets are closely clustered, while the unvoiced consonants (only /t/ is shown in Figures 3b and 3c) show more scattered estimated values. Table 2 indicates how well various consonants are clustered according to style-independent or style-dependent estimation. This table shows the average distance from each estimated target to the centroid for that target. (There is one centroid for style-independent estimation and two centroids for style-dependent estimation.) It can be seen that style-dependent estimation always yields better clustering than style-independent estimation.

Previously, the formant contour model was fitted to words with a limited context (/wVl/ and /tVl/) [5]. Results showed that the estimated target values tended to be located at the most extreme F1 and F2 values. Unlike this previous study in which we had limited phoneme contexts, the current style-independent target estimation yields higher error because of the different formant contour shapes that depend on the phonetic context. Figure 2 shows observed formant contours for C_1 approximants

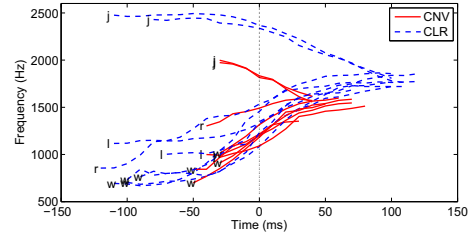


Figure 2: Observed F2 contours for C_1 = approximant, V =/ε/ in both CNV (red solid lines) and CLR (blue dashed lines) speech. All contours are centered at the $C_1 - V$ boundary (0 ms).

and the vowel /ε/, for both CLR and CNV speech styles. The style-independent estimation yields a large error for /j ε/, for example, because a high /ε/ target can not accurately model the lower observed F2 contour.

In summary, based on the lower test-set error from model fitting and more closely clustered estimated consonant target values in the style-dependent model, we conclude that formant targets are style dependent.

4. Experiment 2

As an application of the formant contour model, Experiment 2 is to determine whether vowel classification results can be improved by using model target values instead of observed steady-state formant (SS) values. The hypothesis is that, because the targets represent the vowels independent of phonetic context and formant undershoot, target classification will have higher accuracy than SS classification. (The observed SS values are extracted from the middle position of the vowel.) The target values are evaluated with both style-independent and style-dependent models, since style-independent estimation has the advantage over style-dependent estimation of not requiring prior knowledge about the speaking style.

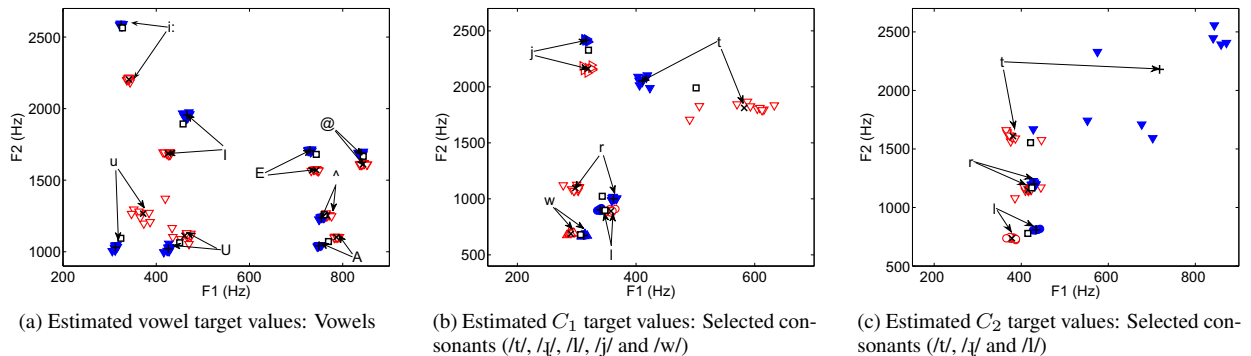


Figure 3: Estimated style-dependent formant target values in F1–F2 space for CNV style (open red) and CLR style (filled blue). The means of each phoneme from style-dependent target estimation are shown in black crosses, while the means from style-independent target estimation are shown in black squares.

	Style-indep.		Style-dep.	
	CNV	CLR	CNV	CLR
Observed data	73.97	88.43	76.47	91.93
Token-based target	82.19	86.13	76.05	88.23

Table 3: Mean percent correct rate in vowel classification experiment.

4.1. Token-based target estimation

Using the formant contour model (described in Section 2), the token-based vowel targets are estimated while neighboring consonants (C_1 and C_2) are set with global target values. (Previously-estimated target values are referred to as global targets.) $Errr1_{style}^{(k)}$ in Equation 3 is minimized per token by adjusting s_1, p_1, s_2, p_2 and \mathbf{T}_V with global target values \mathbf{T}_{C_1} and \mathbf{T}_{C_2} . The outcome of this procedure is a style-independent (or style-dependent) token-based target value for each vowel, as well as s_1, p_1, s_2, p_2 coefficients. A style-independent token-based target means that the global target values \mathbf{T}_{C_1} and \mathbf{T}_{C_2} were obtained from previous style-independent target estimation. The mean error rates for the style-independent token-based target values are 0.1838 (CNV) and 0.2132 (CLR), while style-dependent token-based target values are 0.1770 (CNV) and 0.2009 (CLR). If token-based target values are sufficiently close to the global target (\mathbf{T}_V) values, we expect that the vowel classification results will be better with the token-based target values than with observed formant SS values.

4.2. Vowel classification results

Similar to the jackknife procedure described in Section 3, training and test sessions are completed in 20 different groups for style-independent targets, and 10 groups for style-dependent targets. During the training, one mean (μ) and covariance matrix (S) for the style-independent method and two sets of μ and S are computed for the style-dependent method.

In the test set, the Mahalanobis distance, $D(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$, is measured with μ and S from the training set for all vowels. The vowel identity for each token is assigned based on the minimum Mahalanobis distance ($D(x)$). The vowel classification results are shown in Table 3. A two-sample t -test showed that the CNV vowel classification results were significantly improved with token-based targets, compared with classification based on the observed values

($p = 0.0004$), while none of other comparisons were significant at the $\alpha = 0.0125$ level. The estimated per-token formant targets therefore have potential for improving vowel classification without prior knowledge about the speaking style, which may be applicable to automatic speech recognition (ASR) systems.

5. Conclusions

In this study, we examined whether style-dependent targets would yield better model results as compared with previously presented style-independent targets. The mean test-set error rate was reduced significantly with style-dependent target estimation, while both CNV and CLR formant contours were fitted equally well. The style-dependent consonant targets were better clustered than style-independent targets, further supporting the conclusion that formant targets depend on the speaking style.

As an application of the formant contour model, vowel identity was classified based on either the observed steady-state values or the token-based formant target values. The style-independent vowel classification of CNV speech, using the Mahalanobis distance measure, was improved by a relative 32% by using token-based formant targets.

6. References

- [1] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29: 434–446, 1986.
- [2] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.
- [3] B. Lindblom. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35(11):1773–1781, 1963.
- [4] S. J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55, 1994.
- [5] A. Amano-Kusumoto and J.-P. Hosom. Effect of speaking style and speaking rate on formant contours. In *Proceedings of ICASSP*, pages 4202–4205, 2010.
- [6] 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 Int’l Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2003.
- [7] J. Allen, M. S. Hunnicutt, and D. Klatt, editors. *From text to speech: The MITalk system*. Cambridge University Press, 1987.