Phonetic Optimization: Compromise in Speech Production

EDWARD FLEMMING

The goal of this paper is to motivate a model of phonetic realization which bears substantial similarities to optimality theoretic (OT) phonology (Prince and Smolensky, 1993) in that the phonetic output is optimized with respect to conflicting, violable constraints. However, we will see that the mechanism for the resolution of constraint conflicts must employ numerically weighted constraints rather than the strict constraint domination assumed in OT phonology.

This approach can be contrasted to models of phonetic realization that operate primarily in terms of phonetic realization rules (e.g. Pierrehumbert, 1980), and represents a shift in perspective parallel to that made in phonology in the move from input-oriented re-write rules to output-oriented constraints (Prince and Smolensky, 1993:1-6). Rather than directly specifying the realization of particular phonological configurations, constraints are imposed on that realization, and the actual realization is selected so as to best satisfy the full range of constraints.

The use of constraints in models of phonetic realization is not new[1]. However, this paper will focus on a relatively unexplored possibility raised by such models, namely that phonetic constraints might conflict. It will be argued that compromise between conflicting requirements plays an important role in shaping phonetic realizations.

If constraints are permitted to conflict, then output forms cannot satisfy all constraints, so the constraints must be violable and there must be some method for determining which output best satisfies them. We will see evidence that in cases of conflict, the outcome is a compromise between the conflicting requirements, and that this process of compromise can be modeled as the selection of the candidate output which minimizes the summed violations of numerically weighted constraints.

This type of optimizational model has been employed in phonetics before - e.g. Lindblom (1986) and ten Bosch et al (1987) attempt to derive the distribution of vowels in inventories of different sizes as the result of optimization with respect to conflicting constraints on perceptual distinctiveness of vowel contrasts and minimization of effort. The present proposal is also related to work on motor control which hypothesizes that speech movements, as well as other movements, are optimized with respect to various constraints, such as effort minimization (e.g. Nelson, 1983; Jordan, 1990).

We will motivate a model that incorporates constraint conflict, and conflict resolution based on numerical weighting of constraints through two case studies of phonetic compromise:

(i) A study of formant transitions in CV sequences, showing that properties of second formant transitions can be analyzed as the result of a compromise between the demands of the consonant, the vowel, and minimization of movement rate.

(ii) Compensatory relationships between the durations of segments in the same constituent (e.g. syllable or foot), analyzed as the result of compromise between preferred segment durations and preferred durations for the prosodic constituents containing them.

1 Formant Transitions in CV Sequences

In this section we will outline an optimizational model of aspects of second formant transitions in CV sequences. Adjacent consonants and vowels exhibit mutual assimilation in second formant frequency

[1]For example, Cohn (1990) proposes phonetic constraints in addition to realization rules. In Keating's (1990) 'window' model of coarticulation, the phonetic targets assigned to segments are ranges, or windows, on various dimensions rather than point values. These windows could be regarded as phonetic constraints, since the actual realization is constrained to pass through them, in additions to satisfying other requirements such as a smoothness constraint. Perhaps most similar to the present proposal is Byrd's (1996) window-based model of consonant timing in which timing relations in a particular cluster are determined by the joint action of competing constraints ('influencers'), however Byrd proposes that constraints, and constraint combination, should be probabilistic.

(F2) - i.e. F2 at consonant release varies in the direction of F2 in the following vowel, and F2 in the vowel in turn varies in the direction of F2 at the release of the consonant. These partial assimilations can be viewed as a compromise between achieving the F2 targets for the consonant and vowel, and a preference to avoid fast movements between the two (hence a preference to minimize the difference between the two).

First we will lay out established empirical results regarding this mutual assimilation, and then show that these results can be derived from a simple optimizational model.

1.1 Targets and Interpolation

The first two formants, or resonances of the vocal tract, are the primary determinants of vowel quality - different vowels have different formant frequencies. Formant transitions are the formant movements between consonants and vowels - the formants change as the articulators move from the position for the consonant to the position for the vowel. The formant frequencies at the consonant edge are cues to the place of articulation distinctions among consonants. We will restrict attention to the second formant (F2) which corresponds to the front-back dimension in vowels, and is important in distinguishing place of articulation in consonants.

A first pass at modeling formant transitions is to posit a simple 'targets and interpolation' model. In this type of model of phonetic realization, targets on a number of dimensions are assigned to segments, then the realization is derived by interpolating between targets (e.g. Pierrehumbert 1980). In the case of formant transitions we could assign formant targets for the consonant offset and for the vowel center, then interpolate between them to derive the formant transitions. This is illustrated schematically in figure 1, where $F2_C$ is the frequency of F2 at the release of the consonant, and $F2_V$ is the frequency of F2 at the steady state of the vowel.

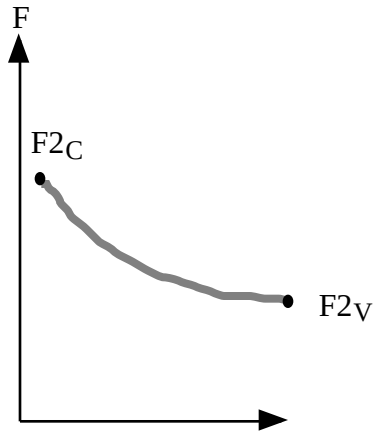

Figure 1. Schematic representation of a 'targets and interpolation' model of formant transitions

The problem with this simple model is that it is well known that $F2_C$ is not invariant for a given consonant, it varies as a function of the following vowel. Similarly vowel formants vary according to their consonant context. So our theory of formant transitions must be able to account for these contextual effects. There are fairly simple and accurate empirical generalizations about exactly how $F2_C$ depends on $F2_V$, and how $F2_V$ depends on $F2_C$. We will outline these relationships, and then show that they follow from a simple optimizational model in which these values are the result of a compromise between minimizing deviation from consonant and vowel targets, and minimizing rate of movement between them.

1.2 Locus Equations

A large number of studies have shown that, for obstruent consonants, $F2_C$ varies linearly with $F2_V$ (Lindblom, 1963; Krull, 1987, 1988; Sussman, 1989, 1991; Sussman, Hoemeke and Ahmed, 1993; Fowler, 1994; Crowther, 1994, etc.). That is, the second formant at the consonant assimilates to that of the vowel. This is illustrated for the english voiced velar /g/ in figure 2. The measurements are from one speaker, reading /gVt/ syllables where V is each of /i, I, eI, œ, ø, A, O, u/.

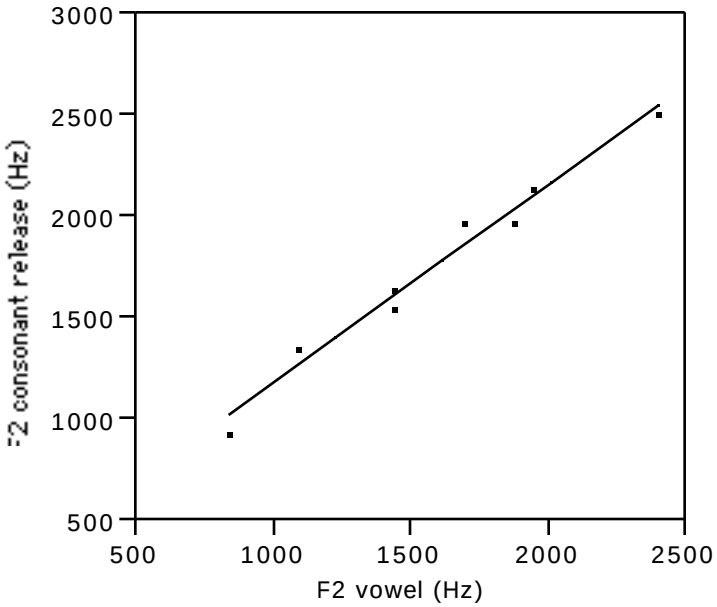

Figure 2. Plot of F2 measured at consonant release against F2 at the steady state, or stationary point, of the vowel, with regression line ($r^2 = 0.98$).

Since the relationship between $F2_C$ and $F2_V$ for a given consonant is linear, it can be expressed as an equation of the following form, known as a locus equation:

(1) $$F2_C = k_1 F2_V + c_1$$

Where k_1 and c_1, the slope and intercept of the line, are fixed for a given consonant[2].

This relationship can be expressed in an alternative form, due to Klatt (1987):

(2) $$F2_C = k_1(F2_V - F2_L) + F2_L$$

The formulation in (2) is equivalent to (1) where $F2L = c_1/(1-k_1)$. The interpretation of (2) is that there is a target or 'locus' for F2 for a given consonant, $F2_L$, but the actual value of F2 at the consonant deviates towards the F2 value in the vowel by a proportion of the difference between consonant locus and F2 in the vowel. That proportion is specified by k_1, the slope parameter - the larger k_1 is, the greater the degree

[2]More accurately, these are fixed for a given consonant in a given style and rate of speech, since, as discussed below (1.4.1), there is evidence that locus equations vary in careful vs. casual speech (Moon and Lindblom 1994) and in citation forms vs. spontaneous speech (Duez 1989). Thus we are not adopting the view that locus equations are invariant properties of particular consonants or places of articulation (cf. Sussman 1989, 1991; Sussman, Hoemeke and Ahmed 1993), they simply summarize empirical observations about the way in which F2 at the release of a consonant varies with F2 at the steady state of the following vowel in a given style of speech, e.g. as elicited by a particular experimental condition.

of accommodation to the vowel. This parameter varies from consonant to consonant, e.g. it is higher for /b/ and /g/ than for /d/.

1.3 Target-Locus Proportionality

Locus equations describe the way in which $F2_C$ varies as a function of $F2_V$, but as mentioned above, $F2_V$ itself varies depending on adjacent consonants. Modeling studies by Lindblom (1963) and Broad and Clermont (1987) have found support for the relationship stated in (3), which is very similar in form to the locus equation.

(3) $$F2_V = k_2(F2_C - F2_T) + F2_T$$

Where $F2_T$ is the F2 target for the vowel. I.e. there is a target, $F2_T$, for the vowel, but the actual F2 of the vowel deviates towards the F2 value of the consonant by a proportion k_2 of the difference between the vowel target and the consonant F2, hence the label 'target-locus proportionality' adopted by Broad and Clermont[3].

So the overall picture obtained from these results is that there are targets for the second formant of each consonant and vowel, $F2_C$ and $F2_L$ respectively, but these targets are systematically 'undershot' with the actual F2 values being displaced towards each other - i.e. there is mutual assimilation in F2 of consonant and vowel. This is illustrated schematically in figure 3.

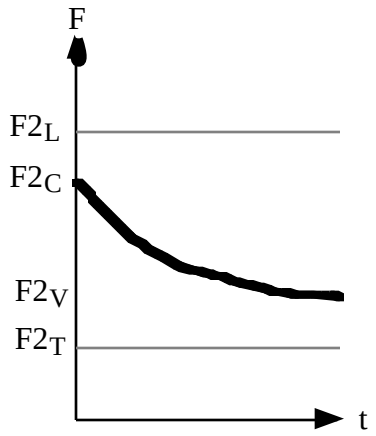

Figure 3. Schematic representation of target undershoot in a CV formant transition.

1.4 An Optimization Model

Having seen how $F2_V$ and $F2_C$ vary systematically with each other, we will now show that these empirically-supported relations can be derived from a model in which $F2_V$ and $F2_C$ for a given CV sequence are selected by optimization with respect to two constraints. Specifically, so as to minimize violation of two basic constraints:

(4) i. Don't deviate from targets.
ii. Don't move quickly.

---

[3]Broad and Clermont (1987) actually propose that deviation of F2 in the vowel from its target is proportional to the difference between $F2_T$ and $F2_L$, not $F2_C$. However, given that the deviation of $F2_C$ from $F2_L$ is proportional to $F2_V - F2_L$, this also implies the proportionality shown in (3).

The second constraint, (4ii), is assumed to be related to effort minimization on the reasonable assumption that faster movements require more expenditure of effort, other things being equal.

Obviously these constraints conflict - achieving all targets may entail rapid articulator movement, avoiding rapid articulator movement may result in failure to reach targets. However it is not appropriate to cast them as ranked constraints since neither is completely dominant. If one were completely dominant then targets would either always be achieved, or would be completely ignored. Instead the constraints are specified as terms of a cost function (5). $F2_V$ and $F2_C$ are selected so as to minimize this cost function.

$$(5) \qquad c = w_c(F2_C - F2_L)^2 + w_v(F2_V - F2_T)^2 + w_e(F2_C - F2_V)^2$$

The targets, $F2_L$ and $F2_T$ are fixed for each consonant and vowel, respectively; w_c, w_v, and w_e are positive weights.

The first two terms of the cost function implement constraint (4i), 'don't deviate from targets'. These terms impose a cost for deviating from targets equal to the square of the difference between the achieved value and the target. Separate terms are included for consonant and vowel targets so that the costs of deviation from each target can be given different weights, w_c and w_v respectively.

The effort terms penalizes larger F2 transitions between C and V. This involves a number of simplifications to keep all constraints in the acoustic domain although effort minimization should properly be specified in articulatory terms. First, change in F2 is only an approximate index of the distance moved by the articulators. Second, articulator velocity obviously depends on the duration of movement - for present purposes, we will simply assume a fixed vowel duration. Finally, different articulators presumably require different amounts of effort to move at a given speed. This variation will be incorporated into the model as differences in the effort weight factor, w_e, but ideally should be derived from some effort metric.

Finally, it is redundant to specify all three weight factors - one weight could be set to a value of one since only the ratios of the weights are relevant to the outcome. All three are specified since this makes the solutions to the optimization easier to interpret.

$F2_V$ and $F2_C$ are selected so as to minimize cost as specified by the cost function (5) - i.e. so as to best satisfy the conflicting constraints on them. We can see that the constraints conflict whenever the targets for consonant and vowel differ, since the first two constraint terms will be minimized when $F2_C$ and $F2_V$ are equal to their target values, whereas the effort constraint is minimized when they are equal to each other.

Graphically, the values of $F2_V$ and $F2_C$ that minimize cost can be identified from the low point in a plot of cost against $F2_V$ and $F2_C$, as shown in figure 4.

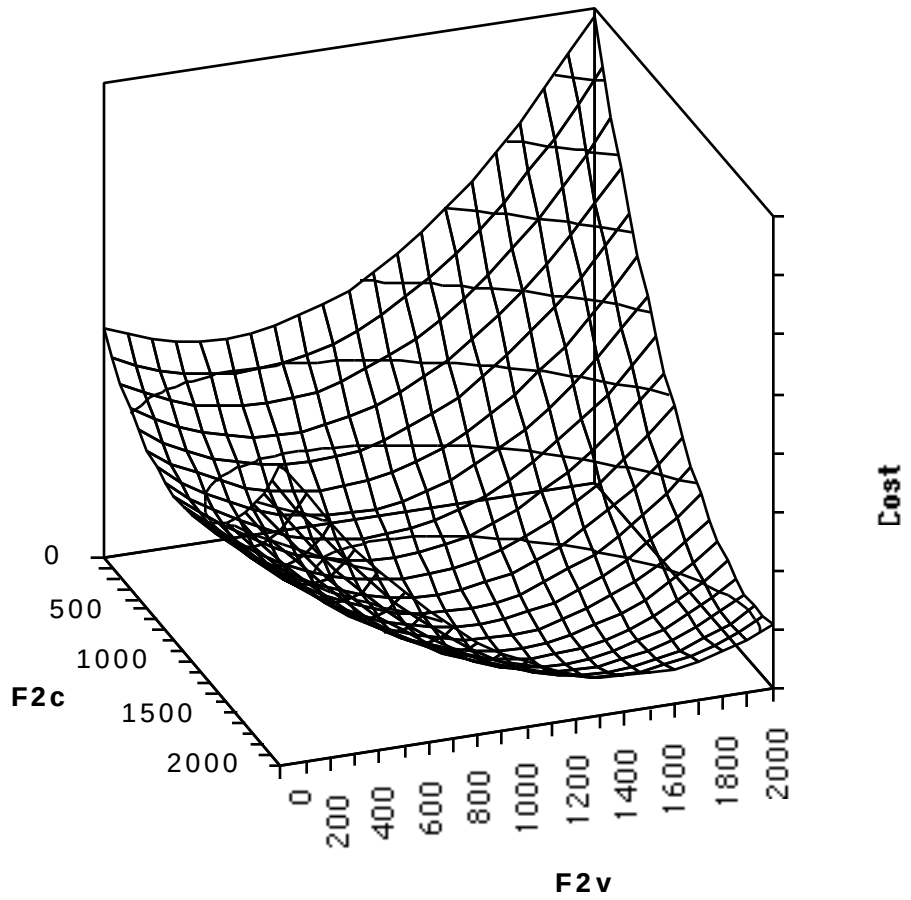

Figure 4. Plot of cost for values of $F2_V$ and $F2_C$, with $F2_L$ = 1700 Hz, $F2_T$ = 1000 Hz, and all weights set to 1.

Analytically, we can find the location of this minimum by finding where the gradients along the $F2_V$ and $F2_C$ dimensions are zero. The gradients are given by the partial derivatives along each dimension. The resulting equation for the minimum along the $F2_C$ dimension is given in (6), and has the form of a locus equation as in (2) above.

(6) $$\frac{\partial c}{\partial F2_C} = 0 \text{ when } \qquad F2_C = \frac{w_e}{w_c+w_e}(F2_V - F2_L) + F2_L$$

Taking the minimum along the $F2_V$ dimension derives target-locus proportionality as in (3) above:

(7) $$\frac{\partial c}{\partial F2_V} = 0 \text{ when } \qquad F2_V = \frac{w_e}{w_v+w_e}(F2_C - F2_T) + F2_T$$

So the optimization model can derive the linear relationship between $F2_C$ and $F2_V$ and the target-locus proportionality pattern for vowels using simple, output-oriented constraints.

The actual solutions for $F2_C$ and $F2_V$, obtained by substituting (6) into (7), are given in (8) and (9).

(8) $F2_C = -u_c(F2_L - F2_T) + F2_L$ where $u_c = \frac{w_e w_v}{w_e w_c + w_v w_c + w_e w_v}$

(9) $F2_V = u_v(F2_L - F2_T) + F2_T$ where $u_v = \frac{w_e w_c}{w_e w_c + w_v w_c + w_e w_v}$

These equations state that $F2_C$ and $F2_V$ undershoot their respective targets by a proportion of the distance between consonant locus and vowel target. The proportion depends on the relative weights of the terms of the cost function. In effect, the interval between $F2_L$ and $F2_T$ is divided into three parts: consonant undershoot, vowel undershoot, and transition in proportions $w_e w_v$: $w_e w_c$: $w_v w_c$. So the more heavily weighted effort avoidance is, the more important it is to have a small transition, and thus more undershoot of vowel and consonant targets results. This shortfall is distributed between consonant and vowel according to the relative weights of the consonant and vowel terms.

1.4.1 The Role of Effort in the Model

The equations derived from the optimization model in (6) and (7), above, have the forms of a locus equation and target-locus proportionality equation respectively, but the scaling factors are functions of the effort weight, w_e, and thus will only be constants if w_e is fixed for a given consonant.

We have said that w_e varies as a function of the consonant since, in this model, it incorporates inherent differences in the effort involved in moving different articulators, but it is also natural to assume that w_e should be varied to model variation between careful and casual speech. This would predict more vowel undershoot and steeper locus equation slopes in casual speech where effort avoidance is more important, i.e. w_e is higher.

There is evidence that these predictions are correct - the locus equation and target-locus proportionality relationships are are not invariant across speaking styles. For example, Lindblom and Moon (1994) demonstrate that vowel undershoot is reduced in careful speech, even when differences in vowel duration are taken into account. Duez (1989) shows, in a study of French, that locus equations tend to be steeper in spontaneous speech than in citation forms of isolated words.

1.4.2 Comparison with OT Phonology

In spite of considerable superficial differences in formulation, there are substantial similarities between the proposed optimizational model of phonetic realization and OT phonology. Both models operate in terms of conflicting, violable constraints which apply to the output. In both cases, the outputs are selected so as to best satisfy these conflicting constraints. In each case, this allows complex mappings from input to output to be analyzed as the result of interactions between simple constraints.

The phonetic constraints look different from familiar OT constraints, because they are stated as mathematical expressions. However, an OT constraint can be regarded as essentially a function from phonological forms to marks of violation (Prince and Smolensky, 1993:68f.), and this is the role of these expressions also. In this case they map phonetic realizations onto marks of violation, expressed as real numbers (e.g. the square of the deviation from a target), and thus are simply non-binary constraints (Prince and Smolensky, 1993:72f.).

The real difference between the two models lies in their modes of constraint interaction, i.e. how the relative harmony of candidates with respect to a full set of constraints is determined from their evaluations with respect to individual constraints. In the phonetic model proposed here, the overall evaluation of a candidate is expressed as a single number (its 'cost'), which is the weighted sum of the costs assigned by each constraint (cf. 5 above). In OT phonology, on the other hand, constraint interaction is governed by a strict dominance hierarchy of constraints.

There are two basic reasons why the constraint-weighting approach is more suitable in the present context. First, it lends itself to the modeling of compromise between conflicting, scalar constraints. We have seen that this is crucial in the analysis of formant transitions: actual F2 values are a compromise between the requirements of a target, and of effort minimization. If achieving targets were strictly dominant, targets would always be hit, and if effort minimization were strictly dominant, assimilation would be total. With summed constraint violations, a large violation of either constraint results in a high cost over all, so it is best to partially violate both[4], with the higher-weighted constraint being violated less.

Compromise between conflicting requirements can be modeled with strict constraint dominance if each requirement is decomposed into a set of ranked sub-constraints which can then be interleaved in the constraint ranking. This strategy is adopted by Prince and Smolensky (1993) in their analysis of the compromise in syllabification between the requirements that nuclei be maximally sonorous and that syllable margins be minimally sonorous. However, this approach does not naturally generalize to the present case where there is a trade-off involving two continuous values, $F2_C$ and $F2_V$. The constraints that are expressed here as simple terms of the cost function would have to be decomposed into a great many sub-constraints (essentially quantizing the F2 dimension).

The second difference between the numerical weighting approach and strict constraint dominance lies in the fact that the latter does not allow for additive effects. That is, violation of multiple lower-ranked constraints can never outweigh a single violation of a higher-ranked constraint. If constraint violations are summed, lower-weighted constraints can 'gang-up' to out-weigh higher weighted constraints. I.e. the sum of the violations of lower-weighted constraints may add up to more than the cost of violating a higher-weighted constraint. Again, this effect is essential in the model of formant transitions: Less vowel undershoot and less consonant undershoot can together make up for more effort.

There is a device for modeling additive effects within the framework of strict constraint domination, namely local conjunction (Smolensky, 1995). Local conjunction takes two constraints and forms a conjoined constraint which is violated if both of the base constraints are violated within some local domain. This conjoined constraint can be ranked higher than constraints which outrank both base constraints, thus in effect allowing violations of two lower-ranked constraints to outweigh violation of a higher-ranked constraint. Again, conjoined constraints would have to be proliferated to account for all of the acceptable trade-offs in the model of formant transitions, but it is interesting to note that the devices

[4]Note that the quadratic form of the constraints is important in deriving this result - squaring the deviations means that large violations of a constraint are liable to incur very high costs, even if the constraint has a fairly low weight.

of constraint decomposition and local conjunction narrow the differences between a system that operates in terms of strict constraint dominance, and one that sums numerically weighted constraint violations.

1.4.3 Extending the Model: The Role of Contrast

Finally, before moving on to another application of optimizational models in phonetic realization, we will briefly consider one interesting direction in which the model of formant transitions appears to need development. As developed so far, the model still relies on the assignment of F2 targets to segments, which effectively constitutes the use of phonetic realization rules. These rules are potentially very simple because contextual variation is governed by constraints, but a fully constraint-based approach to phonetic realization would derive the targets themselves from the interaction of constraints, particularly since there is almost certainly language-specific variation in targets (Disner, 1984). As noted above, optimizational models of the distribution of vowel targets have already been proposed by Lindblom (1986) and ten Bosch et al (1987). These models posit two basic constraints, one requiring that contrasting vowels be maximally distinct, and another requiring that effort be minimized by avoiding extreme articulatory configurations. For a vowel inventory of a given size, vowels targets are selected so as to achieve an optimal compromise between these conflicting requirements.

We shall see here that there is some evidence that constraints on the distinctiveness of contrasts operate to regulate contextual variation in vowels also, and thus should be incorporated into the model of formant transitions in place of constraints requiring the achievement of specified targets. Specifically, there is preliminary evidence that the degree of target undershoot permitted is more restricted where it would have more impact on the distinctiveness of a contrast. Manuel (1990) found evidence that vowel-to-vowel coarticulation is more limited in more crowded vowel inventories, where neigboring vowel phonemes are less distinct acoustically. Coarticulatory variation in vowel quality was measured in the related languages Shona (5 vowels), Ndebele (5 vowelsl), and Sotho (7 vowels). The results showed the low vowel varied less across vowel contexts in Sotho than in the 5-vowel languages, consistent the hypothesis that contextual variation is subject to a distinctiveness constraint.

A pilot study of vowel assimilation to consonants provides clearer evidence of a effect of contrast in limiting contextual variation. The study examines undershoot of long /u/ in the context of coronal stops. The vowel /u/ has a low F2 while coronal stops have a relatively high F2 locus, and so raise the F2 of adjacent vowels. Thus substantial undershoot of /u/ can be observed between coronal stops. Four languages were studied: Two in which /u/ contrasts with front rounded /y/ (Finnish and German) and two in which there are no front rounded vowels (English and Farsi). The contrast between /u/ and /y/ is less distinct than that between /u/ and /i/ since it generally involves smaller differences in F2 and F3, and thus will be more adversely affected if the F2 of /u/ is raised.

The target value for F2 of /u/ in each language was estimated by measuring the minimum value of F2 in the vowel adjacent to a laryngeal (/h/ or glottal stop), e.g. in a word like ‘who’. This is a reasonable estimate of the target value since the laryngeal should have no influence on vowel formants. The effect of the coronal context was determined by measuring the minimum value of F2 in /u/ between two coronals, as in English ‘toot’. Undershoot was then measured as the difference between the F2 in the coronal context, and F2 in the laryngeal context. The measurements in (10), below, are averaged over five repetitions of each word in a carrier phrase, from one speaker of each language. The measurements made adjacent to a laryngeal are labelled ‘h_’, the measurements made between coronal stops are labelled ‘t_t’. The difference between these values is the measure of undershoot, and is reported in the third row of (10). The amount of undershoot is substantial in English and Farsi, while the coronals have almost no fronting effect on /u/ in Finnish and German.

(10)

language	English		Farsi		Finnish		German	
context	h_	t_t	h_	t_t	h_	t_t	h_	t_t
mean F2 (Hz)	869	1195	818	1041	531	574	725	763
‘undershoot’	326		223		43		38	

These results suggest a striking effect of the system of contrasts: Contextual fronting of /u/ is severely restricted where it would endanger a contrast with /y/. In Farsi and English, the nearest front vowel is /i/, so there is more room for variation while maintaining sufficient distinctiveness.

This pattern could be accounted for if constraints on the distinctiveness of contrasts, along the lines proposed by Lindblom (1986) or ten Hout et al (1987), are included in the model of formant transitions so they can interact with the velocity minimization constraint that favors assimilation of vowels to neighboring consonants. However further investigation is required to determine whether either of the specific formulations of distinctiveness constraints proposed by these researchers yields appropriate results for vowel undershoot. For example, Lindblom proposes to penalize the reciprocal of the summed squares of the auditory distances between pairs of vowels. A cost term of this kind will not derive the linear target-locus proportionality described above (1.3) (the same applies to the distinctiveness constraint proposed by ten Hout et al). This is not necessarily incorrect - the fits obtained by Lindblom (1963) and Broad and Clermont (1987) are not as striking as the fits of locus equations - but the precise predictions of a contrast-based model remain to be tested.

2. Other Applications: Duration Compensation

We have seen that mutual assimilation between consonants and vowels in CV sequences is one instance of phonetic compromise that can be perspicuously analyzed in terms of optimization with respect to conflicting constraints. This type of contextual assimilation, or coarticulation, is widespread as would be expected based on the proposed model: The dispreference for high articulator velocity leads to conflicts between the requirements of adjacent segments wherever those requirements differ. E.g. we find contextual partial nasalization of vowels adjacent to nasals, vowel-to-vowel coarticulation across consonants, etc. However we will now look at a rather different case of compromise in phonetic realization: The phenomenon of duration compensation, resulting from conflicting requirements on the duration of segments and of the constituents that contain them.

There is a range of evidence for compensatory relationships between the durations of segments within the same constituent. Striking examples are provided by quantity contrasts in the Scandinavian languages, Swedish, Icelandic, Norwegian in which vowel and following consonant duration are in a compensatory relationship - short vowels occur before long consonants or consonant clusters, and long vowels occur in open syllables or before short consonants only, so schematically, VCC contrasts with VVC. A standard way of conceptualizing this distribution of length is to suppose that the vowel and consonant durations co-vary to try to keep the duration of a larger constituent such as the syllable or foot relatively constant (e.g. Lehiste, 1970), i.e. vowel and consonant durations compensate for each other.

Similar effects are observed at the foot level in Estonian where, in a disyllabic word, the duration of the second vowel is inversely proportional to the duration of the first (Lehiste, 1970). This pattern is also observed in Finnish dialects (Lehtonen, 1970). A compensatory effect that results in less obvious variation in duration is the cross-linguistically common pattern of closed syllable shortening where a vowel is shorter in a closed syllable than in an open syllable (Maddieson, 1985). This latter effect clearly operates below the level of contrastive duration.

A pilot study of English provides evidence of similar compensatory effects at this more subtle phonetic level. The study examined words of the form /tVC/ where V was one of /œ, A, aI/ and C was drawn from /b, d, p, t/. The words were spoken in a carrier phrase by one speaker of American English.

Broadly, the results show that a longer vowel implies a shorter coda consonant, and vice versa. This compensatory pattern is most obvious in words of the form /tœC/, as shown in figure 4. Where coda consonant is longer, as shown by the bottom panel, mean vowel duration is shorter, as shown by the corresponding column in the top panel. Note however, that the magnitude of the differences in vowel and consonant durations are not equal - i.e. a fixed syllable duration is not maintained, because compensation is not exact.

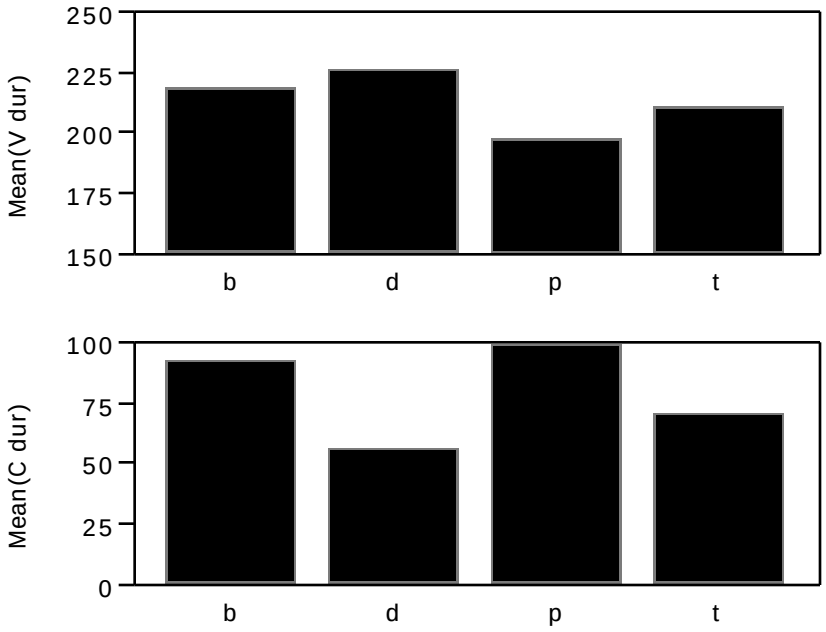

Figure 4. Mean C and V durations by coda consonant for words of the form /tœC/.

The overall patterns of duration are more complex than would be predicted by compensation alone. Not all of the vowels show compensation patterns as clear as figure 4, for example /I/ varies very little in duration across coda consonants, for this speaker. However, a number of patterns indicative of duration compensation are significant in the full data set[5]:

(i) Vowels are shorter before labial stops than before coronal stops, and labial stops are longer than coronal stops.

(ii) Vowels are shorter before voiceless stops than before voiced stops, and voiceless stops are longer than voiced stops.

Both results fit the generalization that vowels are shorter when the coda stop is longer.

The results concerning vowel durations replicate findings of larger studies, e.g. Lehiste and Peterson (1960). In particular the vowel duration difference due to differences in voicing of following stops is extremely well documented. The concomitant consonant duration results are also consistent with previous results, e.g. Byrd's (1993) study of stop durations in the TIMIT corpus yielded the same relative durations, although the magnitude of the difference between coronals and labials was much smaller than found for this speaker.

---

[5]Effects on vowel duration were evaluated by calculating a full-factorial ANOVA of vowel duration with vowel, consonant place, and consonant voicing as factors. All main effects were significant (see table below). There were significant interactions between vowel and place and vowel and voicing, indicating that consonants did not have the same effects on the durations of all vowels.

	d.f.	F Ratio	Prob>F
vowel	3	153	p< .0001
voice	1	48	p< .0001
place	1	17	p< .001
vowel*voice	3	13	p< .0001
vowel*place	3	4.2	p< .01
voice*place	1	0.8	p=0.38
vowel*voice*place	3	0.3	p=0.84

However, an important result of the present study which has been less well documented in the literature is that shortening is reciprocal; i.e. not only are vowels generally shorter before longer stops, but consonants are shorter after long vowels. This is shown in figure 5 where duration of all consonants is plotted by vowel type: 'long' /œ, A, aI/ versus 'short' /I/ (the long vowels were all of similar duration in this study). The difference in mean consonant duration (90 ms with short vowels, 76 ms with long vowels) is significant ($p<0.05$).

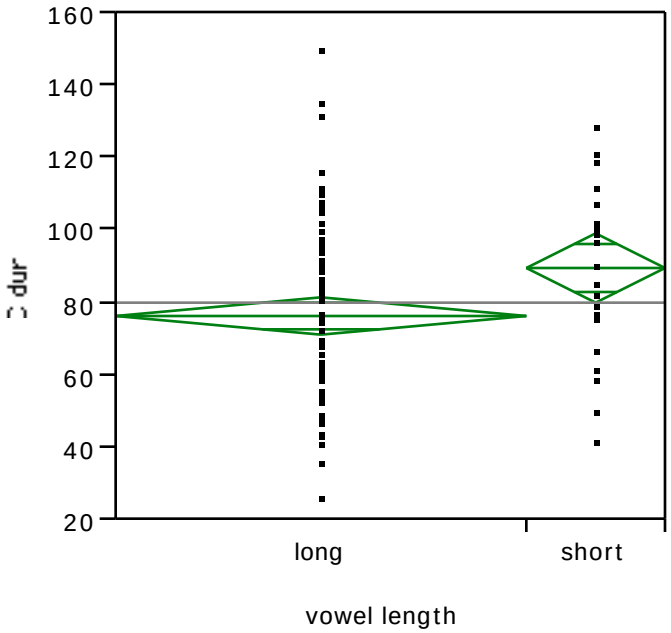

Fig.5. Consonant duration by vowel length (long /œ, A, aI/ vs. short /I/).
Diamonds indicate the means and the 95% confidence intervals, width indicates sample size.

In summary, the results show that, within a monosyllabic word, vowels are shorter when the coda consonant is longer, and coda consonants are shorter when the vowel is longer. I.e. there appears to be low-level duration compensation in English.

2.1 Models of Duration

We will now turn to an optimizational model that can account for the qualitative properties of duration compensation as the result of a compromise between requirements on the duration of segments, and of the larger constituents that contain them.

Existing duration models can be divided into two basic classes, 'top-down' and 'bottom-up' models. Top-down models of duration are designed to account for compensation effects: Duration is assigned to higher units (e.g. syllable, foot, or word), then subdivided amongs the segments that make up these units (e.g. Kohler, 1986; Campbell, 1992). The problem faced by this type of model is that the duration assigned to a higher unit has to depend on the number of segments in that unit (e.g. Kohler's 'complexity' factor). This is because, as noted above, duration compensation effects are often not total. E.g. [p] in 'tap' is about 20ms longer than the [t] in 'tat', but the vowel is only about 6ms shorter than in 'tat'. So there is some shortening in response to a longer consonant, but overall the length of the syllable still increases. As a result, the duration of higher units such as syllables is not fixed any more than the duration of segments.

In a 'bottom-up' model, duration is assigned segment by segment by context-sensitive duration rules (e.g. Klatt 1979). The problem faced by this approach is incorporating compensatory effects. Formulating compensatory effects in terms of context sensitive duration assignment rules results in complex rules, and misses the basic generalization of compensation.

A more satisfactory alternative is a 'parallel' model which incorporates both top-down and bottom-up effects. This is possible in an optimizational model where durations are assigned to best satisfy multiple constraints on the output. Compensatory effects can be obstained by assigning durational targets to both segments and to larger constituents. Constraints require the output to meet these target durations. For example, the cost function in (11) represents some of the constraints on the realization of a C1VC2 sequence. The terms of the function penalize deviation from target durations for each of the segments ($C1_T$, $C2_T$,V_T), and from the target for the duration of the whole syllable (σ_T).

(11) $c = w_{c1}(C1-C1_T)^2 + w_v(V-V_T)^2 + w_{c2}(C2-C2_T)^2 + w_\sigma((C1+V+C2)-\sigma_T)^2$

These constraints conflict when the target syllable duration differs from the sum of the targets for the individual segments. The optimal result in such case will be a compromise between segmental and higher targets, yielding a pattern of partial compensation. E.g. if C2 is lengthened, V will shorten in compensation to prevent excessive violation of the syllable-level constraint, but the constraint on C2 duration will generally prevent total compensation. So syllables will be longer if they contain more or longer segments, but individual segment durations will also be shorter .

However, it is also apparent that additional constraints will be required to account for the full complexity of duration patterns. For example, in the speech of the subject of the pilot study above, differences in coda stop duration dueto voicing are *over*-compensated by vowel duration - E.g. the difference in vowel duration between 'tap' and 'tab' is greater than the difference in duration between /p/ and /b/. The model in (11) cannot predict over-compensation. Possibly this effect represents the fact that the difference in vowel duration is exploited as a cue to voicing contrasts in a following obstruent (Massaro and Cohen, 1983), so the difference in vowel duration could be the result of a distinctiveness constraint. This proposal fits in with the observation that the magnitude of vowel-shortening before voiceless obstruents is much greater in English than in other languages that have been studied (Chen, 1970; Keating, 1985).

3. Universal Phonetics and Language-Specific Variation

We have seen evidence for the importance of conflict and compromise in phonetic realization, and that we can model these phenomena in terms of optimization with respect to output goals. Compromise is then the natural consequence where constraints conflict. We will close by drawing a final parallel between this model and OT phonology, which suggests that the model can provide the basis for a promising approach to separating the universal from the language-specific in phonetic realization.

The search for universals in phonetics has turned up few, if any, strong universals, while cross-linguistic tendencies are much more easily discerned (e.g. Keating, 1985). For example, as discussed by Keating, vowels are shorter before voiceless obstruents than before voiced consonants in a wide range of languages (Chen, 1970), but the magnitude of the effect is greater in English than in most other languages, and shortening is absent in Polish, Czech and Saudi Arabic. Given this variation, no mechanical, deterministic explanation of the shortening effect is feasible, but the tendency is very widespread and is never reversed, so it would seem to follow from some universal factor.

A similar situation obtains in phonology. For example, there is a well-established cross-linguistic preference for syllables to have onsets, as shown by the facts that many languages require all syllables to have onsets, no language requires onsetless syllables, and intervocalic consonants are preferentially syllabified into onset. However, this generalization is only a tendency - not all languages require syllables to have onsets (English is an obvious example). Optimality Theory offers an illuminating account of this type of situation: It posits a universal set of constraints, such as the requirement that syllables should have onsets, but these constraints are violable and may conflict. In cases of conflict, the outcome is determined by the relative importance of the constraints, expressed by the constraint ranking. The ranking of constraints is language specific, so cross-linguistic tendencies are a result of the fact that all languages are subject to the same constraints, while variation results from the fact that languages resolve conflicts between constraints in different ways. For example, languages which permit onsetless

syllables rank universal constraints against epenthesizing a consonant or deleting a vowel higher than the requirement that syllables have onsets.

A parallel account of cross-linguistic tendencies in phonetics is available in the optimizational model of phonetic realization: We can hypothesize that the constraints on phonetic realization are universal, but their weighting is partly language-specific. The effect of a constraint is modulated by interaction with other, possibly conflicting, constraints. Thus constraints give rise to tendencies rather than hard universals. For example, the tendency to shorten vowels before voiceless obstruents mentioned above is plausibly due to compensatory duration effects along the lines described in section 2: Vowels are shorter before voiceless obstruents than before voiced obstruents to compensate for the greater duration of voiceless obstruents (Maddieson 1997). It was suggested above that compensation is motivated by constraints enforcing duration targets for prosodic constituents. These constraints interact with other constraints. For example, Keating (1985) suggests that the shortening effect may be reduced in Czech because vowel duration is contrastive in that language. This account could be formalized by positing a distinctiveness constraint on vowel duration contrasts. In a smilar vein, it was suggested above that the greater shortening effect observed in English might serve to increase the distinctiveness of voicing contrasts in following obstruents.

In general, a model based on violable constraints appears highly suitable as the basis for an account of cross-linguistic phonetics because it can accomodate tendencies, and because it offers an account of when and how those tendencies might be violated - i.e. where they are outweighed by other universal preferences.

References

Bosch, L.F.M. ten, L.J. Bonder, and L.C.W. Pols (1987) 'Static and dynamic structure of vowel systems'. *Proceedings of the 11th international congress of phonetic sciences*, Vol.1, 235-238.

Broad, D.J. and F. Clermont (1987). 'A methodology for modeling vowel formant contours in CVC context'. *Journal of the Acoustical Society of America* 81, 155-165.

Byrd, Dani (1993). '54000 Stops'. *UCLA Working Papers in Phonetics* 83, 97-115.

Byrd, Dani (1996). 'A phase window framework for articulatory timing'. *Phonology* 13, 139-169.

Campbell, Nick (1992). 'Multi-level timing in speech'. ATR Technical Report.

Chen, Matthew (1970). 'Vowel length variation as a function of the voicing of consonant environment'. *Phonetica* 22, 129-59.

Cohn, Abigail (1990). *Phonetic and Phonological Rules of Nasalization*. Ph.D. dissertation, UCLA. Distributed as*UCLA Working Papers in Phonetics* 76

Crowther, C.S. (1994). 'Modelling coarticulation and place of articulation using locus equations'. *UCLA Working Papers in Phonetics* 88, 127-148.

Disner, Sandra (1983). *Vowel Quality: The Relation between Universal and Language-Specific Factors* . Ph.D. dissertation, UCLA. Distributed as*UCLA Working Papers in Phonetics* 58.

Duez, Danielle (1989). 'Second formant locus-nucleus patterns in spontaneous speech: Some preliminary results on French'. *PERILUS* 10, 109-114.

Fowler, C.A. (1994). 'Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation'. *Perception and Psychophysics* 55, 597-610.

Jordan, M.I. (1990). 'Motor learning and the degrees of freedom problem'. M. Jeannerod (ed.) *Attention and Performance XIII: Motor Representation and Control*, Lawrence Erlbaum, Hillsdale NJ, 796-836.

Keating, Patricia A. (1985). 'Universal phonetics and the organization of grammars'. Victoria A. Fromkin (ed.) *Phonetic Linguistics*. Academic Press, New York, 115-32.

Keating, Patricia A. (1990). 'The window model of coarticulation: articulatory evidence'. John Kingston and Mary E. Beckman (eds) *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, Cambridge, 451-470.

Klatt, Dennis H. (1979). 'Synthesis by rule of segmental durations in Englsih sentences'. Björn Lindblom and Sven Öhman (eds) *Frontiers in Speech Communication Research*, Academic Press, New York, 287-300.

Klatt, Dennis H. (1987). 'Review of text-to-speech conversion for English'. *Journal of the Acoustical Society of America* 82, 737-793.

Kohler, Klaus J. (1986). 'Invariance and variability in speech timing: From utterance to segment in German'. J.S. Perkell and D.H. Klatt (eds) *Invariance and Variability in Speech Processes*, LEA, Hillsdale, NJ, pp. 268-289.

Lehiste, Ilse (1970). *Suprasegmentals*. MIT Press, Cambridge.

Lehiste, Ilse, and Gordon Peterson (1960). 'Duration of syllable nuclei in English'. *Journal of the Acoustical Society of America* 32.

Lehtonen, J. (1970). 'Aspects of quantity in standard Finnish'. *Studia Philologica Jyväskyläensia* 6.

Lindblom, Björn (1963). 'Spectrographic study of vowel reduction'. *Journal of the Acoustical Society of America* 35, 1773-1781.

Lindblom, Björn (1968). 'Phonetic universals in vowel systems'. J.J. Ohala and J.J. Jaeger (eds) *Experimental Phonology*. Academic Press.

Maddieson, Ian (1985). 'Phonetic cues to syllabification'. Victoria A. Fromkin (ed.) *Phonetic Linguistics*. Academic Press, New York, 203-221.

Maddieson, Ian (1997). 'Phonetic Universals'. W.J. Hardcastle and J. Laver (eds) *The Handbook of Phonetic Sciences*, Blackwell.

Manuel, Sharon Y. (1990). 'The role of contrast in limiting vowel-to-vowel coarticulation in different languages'.*Journal of the Acoustical Society of America* 88, 1286-1298.

Massaro, Dominic W., and Michael M. Cohen (1983). 'Consonant/vowel ratios: An improbable cue in speech'. *Perception and Psychophysics* 33, 501-5.

Moon, S-J, and B. Lindblom (1994). 'Interaction between duration, context, and specking style in English stressed vowels'. *Journal of the Acoustical Society of America* 96, 40-55.

Nelson, W. (1983). 'Physical principles for economies of skilled movements'. *Biological Cybernetics* 46, 135-147.

Pierrehumbert, Janet B. (1980)*The Phonology and Phonetics of English Intonation.* Ph.D. dissertation, MIT.

Prince, Alan, and Paul Smolensky (1993) *Optimality TheoryI*. Ms., Rutgers and University of Colorado.

Sussman, H.M. (1989). 'Neural coding of relational invariance in speech: Human language analogs to the barn owl'. *Psychological Review* 96, 631-642.

Sussman, H.M. (1991). 'The representation of stop consonants in three-dimensional space'. *Phonetica* 48, 18-31.

Sussman, H.M., K. Hoemeke, and F. Ahmed (1993). 'A cross-linguistic investigation of locus equations as a source of relational invariance for stop place categorization'. *Journal of the Acoustical Society of America* 94, 1256-1268.

Edward Flemming
Dept. of Linguistics
Building 460
Stanford, CA 94305-2150
flemming@csli.stanford.edu