VOWEL TARGETS AND CONSONANT LOCI FROM SCALING PROPERTIES OF FORMANT TRANSITIONS

David J. Broad[1] and Frantz Clermont[2]

[1]2638 State St., Unit 12, Santa Barbara CA 93105, USA
[2]JP French Associates & University of York, United Kingdom
djbroad@silcom.com akustikfonetiks@yahoo.com.au

ABSTRACT

From the assumptions that vowel-to-consonant formant transitions in any given VC context are (1) similar in shape and (2) scaled in proportion to the difference between the vowel target and consonant locus, we show that the implied set of scaling relationships leads to a method for estimating the loci and targets from formant data.

Keywords: vowel, formant, locus, target, scaling

1. INTRODUCTION

As observed more than fifty years ago [3], formant transitions exhibit no invariant properties for either vowels or their consonantal contexts. It has therefore been more promising to think of consonant loci [4] and vowel targets [5] as values that stand in certain relationships to these transitions.

In [1] this idea was carried forward by looking at consonant loci as axes of symmetry for families of formant transitions and at vowel targets as optimum elements for scaling the transitions. This approach worked, but required separate iterative procedures for the individual loci and targets.

Here we show how consonant loci and vowel targets can be found more efficiently in an approach where scaling relationships among families of formant transitions play a unifying role.

Section 2 develops the method's basic concepts around a simple model for formant transitions while Section 3 describes the method and illustrates it with formant data. Section 4 is a summary.

2. CONCEPTUAL FOUNDATIONS

2.1. Phonetically-motivated model

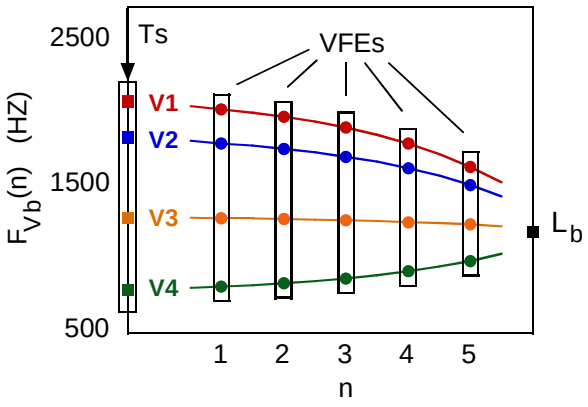
We start from the model in [1] which has consonant-specific transition shapes that are scaled in proportion to the target-locus distance:

$$(1) \quad F_{VC}(n) = L_C + (T_V - L_C)K_C(n)$$

$F_{VC}(n)$ is the formant transition from the vowel V into the consonant C, expressed as a function of the time-frame number n. L_C is the locus for consonant C, T_V is the target for vowel V, and $K_C(n)$ is the transition shape specific to the consonant C.

Figure 1 illustrates Eq. (1) with some idealized F_2 trajectories which trend forward toward a /b/-like locus and backward toward the vowel targets.

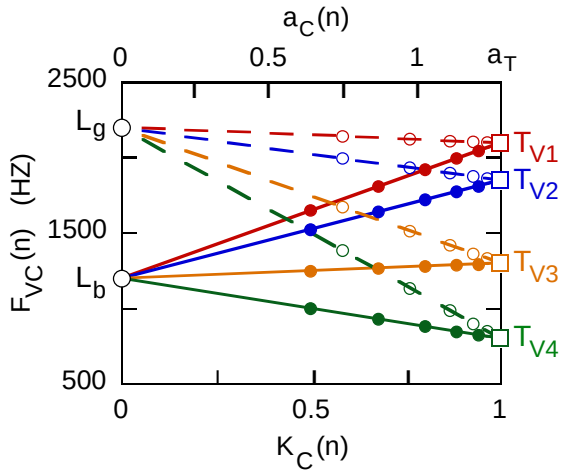
Figure 1: Idealized second-formant $F_{Vb}(n)$ transitions for four vowels. Rectangles enclose the vowel-formant ensembles (VFEs) for each frame n and the target ensemble at the left edge [see Sec. 2.3].



2.2. A new independent variable

In Eq. (1) it is natural to think of the frame number as the independent variable and to view each transition in Fig. 1 as a function of n. However, we can also treat the whole function $K_C(n)$ as the independent variable. As shown in Fig. 2 with the Vb-like transitions from Fig. 1 and with four Vg-like ones, convergent curved trajectories as in Fig. 1 become straight divergent lines which take on the values of the loci at $K_C(n) = 0$ and those of the targets at $K_C(n) = 1$.

Figure 2: Formant frequency plotted as a function of $K_C(n)$ for transitions between four vowels and /b/- and /g/-like consonants according to Eq. (1). The top scale is the ensemble scale, explained in Sec. 2.4.



If we knew the form of $K_C(n)$ we could get the *L*s and *T*s directly from a plot such as Fig. 2. However, $K_C(n)$ is not yet known so we next develop a data-referenced surrogate for it.

2.3 Vowel-formant ensembles (VFEs)

Each vertical rectangle in Fig. 1 encloses the formant values for the different vowels at a particular time frame. Such a set of values is defined [2] as a *vowel-formant ensemble*, or VFE. Each frame *n* in each context C will have its own VFE. Here we also define the *target ensemble* as the VFE made up of the vowel targets, as illustrated at the left edge of Fig. 1.

The salient property of Eq. (1) from the perspective of VFEs is that for a fixed frame in a fixed context the formant values for the vowels are a linear function of the vowel targets. As pointed out in [2], this implies that each VFE will be geometrically similar to the target ensemble. Thus all VFEs from Eq. (1) will be scaled copies of one another. We now turn to the scaling of VFEs.

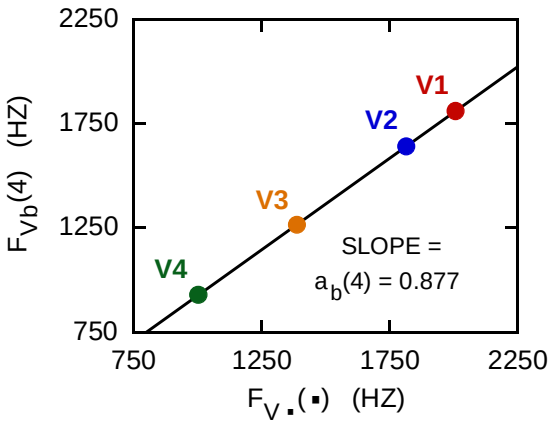
2.4. Scaling *relative to* the mean VFE

To connect the model to data, the following will assume a database of VC syllables with all combinations of the Cs and Vs equally represented.

The *ensemble scale* $a_C(n)$ of the VFE for frame *n* and consonant C is defined [2] as the slope of a line fit to a plot of that VFE's formants against those of the mean VFE, i.e., the VFE made up of the intra-vowel averages $F_{V\bullet}(\bullet)$ of the formant over all the frames and contexts in the database. (Here and below a dot "•" symbolizes averaging over a variable.)

Figure 3 shows such a plot for $n = 4$ from Fig. 1. The slope 0.877 is the ensemble scale $a_b(4)$ for this VFE. The ensemble scale for each frame in each context can be determined by this type of slope calculation and the resulting values for $a_C(n)$ are the first step in linking the model to data. The next step is to link the $a_C(n)$ to the $K_C(n)$ in Eq. (1).

Figure 3: The VFE for $n = 4$ from Fig. 1 versus the mean VFE. The slope is this VFE's ensemble scale.



2.5. Scaling *relative to* the target VFE

In Eq. (1) $K_C(n)$ is the scale of the VFE for consonant C and frame *n* relative to that of the target ensemble. If we now define the *target scale* a_T as the ensemble scale of the target ensemble, the scalings among the mean VFE, the target ensemble, and the VFE for consonant C and frame *n* are related as:

$$(2) \quad K_C(n) = a_C(n) / a_T$$

Figure 2 shows this relation by its two horizontal axes: As the bottom axis $K_C(n)$ runs from 0 to 1 the top axis $a_C(n)$ runs from 0 to a_T.

Our method's building blocks are now in place.

3. DETERMINING LOCI AND TARGETS

3.1. Rationale

Substitution of Eq. (2) into Eq. (1) yields

$$(3) \quad F_{VC}(n) = L_C + (T_V - L_C)\, a_C(n) / a_T$$

In Eq. (3) the formant transition becomes a linear function of $a_C(n)$ which intersects the consonant locus at $a_C(n) = 0$ and the vowel target at $a_C(n) = a_T$.

Lines fit to actual data will not in practice intersect precisely as they do for the idealized case shown in Fig. 2. Therefore we must find a set of intersections for the loci and targets that will yield the best fit to the data.

The method for this will next be illustrated with a database structured as specified in Sec. 2.4.

3.2. Data for illustration

Our illustration uses the VC data from [1] which consist of three repetitions of each combination of the eight vowels /ɪ, ɛ, æ, ɑ, ɔ, ʊ, ʌ, ɝ/ and the three consonants /b, d, g/. VCs were favored over CVs because the CVs tend to be diphthongized in American English.

Vowel boundaries were hand-marked at the voice onset and at the sudden decrease of amplitude for the consonant closure on a visual display of the waveform.

The first three formants were estimated for 11 equally-spaced frames in the vowel segments by hand-editing an LPC analysis to check for missed or inserted formants. The results shown here are based on averages of the three repetitions.

3.3. Step 1: consonant loci

To find the locus for consonant C, we average Eq. (3) over the vowels to get its inter-vowel mean transition:

$$(4) \quad F_{\bullet C}(n) = L_C + (T_\bullet - L_C)\, a_C(n) / a_T$$

where the dot "•" in the place of the "V" from Eq. (3) now indicates averaging over vowels. $T_\bullet$ is the average of the vowel targets.

A consonant's mean transition smoothes statistical idiosyncrasies of individual vowel transitions. In addition, Eq. (4) shows that the *y*-axis intercept of a line fit to consonant C's $F_{\bullet C}(n)$ data plotted against the ensemble scale $a_C(n)$ should yield a unique estimate for the C locus.

Figure 4 shows such a plot for our data with C = /b, d, g/. The loci determined by its *y*-axis intercepts are listed in Table 1. These seem to be in line with phonetic expectation.

Figure 4: Inter-vowel mean F_2 for Vb, Vd, and Vg contexts versus the ensemble scale. Loci are the *y*-axis intercepts of the lines fit to the data. Inset: the near-intersection of the lines on an expanded scale.

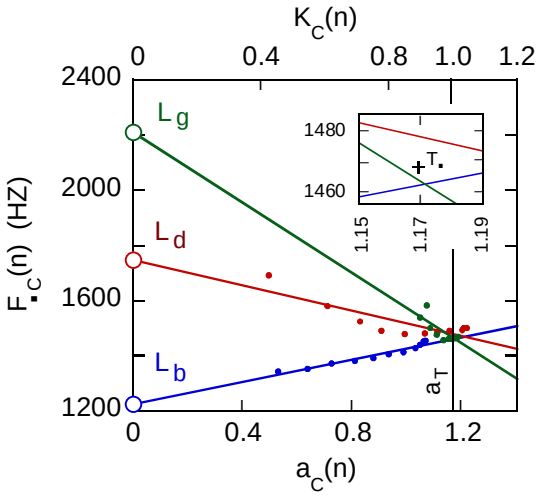


Table 1. Consonant loci (in Hertz) via Eq. (4).

L_b	L_d	L_g
1225	1747	2210

3.4. Step 2: vowel targets

The inset to Fig. 4 shows that the three lines only come close to intersecting near an ensemble scale of 1.17. To get a geometric diagram like Fig. 2, we need lines which will still pass through the loci at the left but which will pass exactly through a common point at the right. This point will fix the target scale (a_T) and mean vowel target ($T_\bullet$). Next we find the a_T and $T_\bullet$ that best fit the data.

The problem is nonlinear because the unknowns a_T and $T_\bullet$ in Eq. (4) occur as a ratio. However, any trial value for a_T implies a set of $K_C(n)$ values via Eq. (2). From these the $T_\bullet$ that minimizes the rms error in fitting Eq. (4) can be calculated as:

$$(5) \quad T_\bullet = \frac{\sum_{c=1}^{N_C}\sum_{n=1}^{N} K_c(n)\left[F_{\bullet c}(n) - L_c\left(1 - K_c(n)\right)\right]}{\sum_{c=1}^{N_C}\sum_{n=1}^{N} K_c^2(n)}$$

where N is the number of frames, N_C is the number of consonants, and c is a consonant index $c = 1, 2, \ldots, N_C$.

Figure 5: The rms error in fitting Eq. (4) plotted as a function of the target scale a_T.

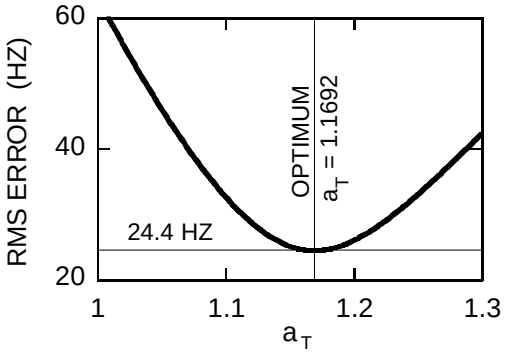


Figure 5 shows the rms error for a range of trial values of a_T, with a minimum of 24.4 Hz at a_T = 1.1692 for which $T_\bullet$ = 1467 Hz. These are the coordinates of the cross in the inset to Fig. 4.

We can now use these values for $T_\bullet$ and a_T to obtain the individual vowel targets. For this, we start by subtracting Eq. (4) from Eq. (3) to eliminate L_C and begin to isolate T_V:

$$(6) \quad F_{VC}(n) - F_{\bullet C}(n) = (T_V - T_\bullet)\, a_C(n) / a_T$$

The remaining dependency on consonants and frames is removed by averaging both sides of Eq. (6) over all Cs and all values of n. From the additional fact that $a_\bullet(\bullet)$ is the ensemble scale of the mean VFE and thus equal to unity, the vowel targets can then be expressed as:

$$(7) \quad T_V = T_\bullet + a_T\,[F_{V\bullet}(\bullet) - F_{\bullet\bullet}(\bullet)]$$

The targets, loci, and target scale just found yield an rms error of 43 Hz in fitting Eq. (3) to the VC data. Table 2 lists the vowel targets and Fig. 6 shows the fit to the full dataset. As with the loci, the vowel targets seem to agree with expectation.

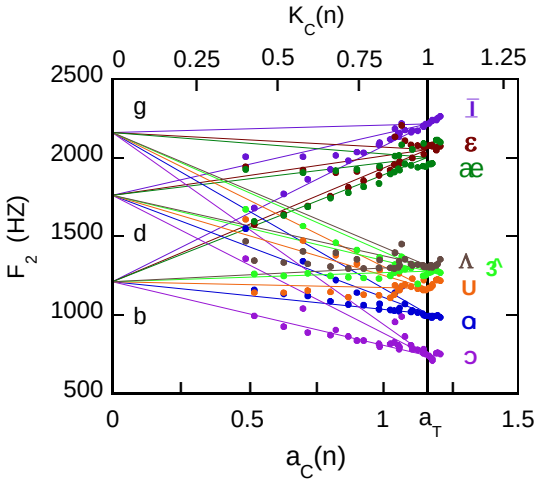
Table 2: Vowel targets (in Hertz) via Eq. (7).

$T_{ɪ}$	$T_{ɛ}$	$T_{æ}$	$T_{ɑ}$	$T_{ɔ}$	$T_{ʊ}$	$T_{ʌ}$	$T_{ɝ}$
2212	2046	1998	1006	731	1161	1276	1308

4. SUMMARY AND CONCLUSION

We have explicated vowel targets and consonant loci in terms of scaling properties of formant transitions in VC context. This conceptualization

Figure 6: F_2 data plotted against the ensemble scale. The consonant loci lie along the y-axis and the vowel targets along the vertical line at $a_C(n) = a_T$.



leads to our method which uses only the simple operations of averaging, fitting lines to data, and a single one-dimensional numerical minimization. (For a step-by-step example, see the supplementary document WorkedExample.pdf.)

The method also works for CV syllables, but CVCs will additionally involve a superposition model [1] for the interaction between initial and final Cs.

The method yields a sort of locus-target nomogram (Fig. 6) which exhibits the scaling relationships among the transitions. In model form (Fig. 2) the nomogram helps to pose and solve the problem of estimating the loci and targets from formant data.

5. REFERENCES

[1] Broad, D. J., Clermont, F. (1987). A methodology for modeling vowel formant contours in CVC context, *J. Acoust. Soc. Am.* 81, 155-165.

[2] Broad, D. J., Clermont, F. (2002). Linear scaling of vowel formant ensembles (VFEs), *Speech Communication* 37, 175-195.

[3] Cooper, F. S., Delattre, P., Liberman, A. M., Borst, J. M., Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds, *J. Acoust. Soc. Am.* 24, 597-606.

[4] Delattre, P., Liberman, A. M., Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants, *J. Acoust. Soc. Am.* 27, 769-773.

[5] Lindblom, B. (1963). Spectrographic study of vowel reduction, *J. Acoust. Soc. Am.* 35, 1773-1781.