ELSEVIER

Speech Communication 37 (2002) 175–195

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Linear scaling of vowel-formant ensembles (VFEs) in consonantal contexts

David J. Broad [a,*], Frantz Clermont [b]

[a] 2638 State St., Unit 12, Santa Barbara, CA 93105, USA

[b] School of Computer Science, University College (ADFA), University of New South Wales, Canberra, 2600 ACT, Australia

Received 10 March 2000; received in revised form 20 December 2000; accepted 11 January 2001

Abstract

There are familiar terms such as "contour" and "trajectory" to refer to a vowel formant frequency as a function defined on the time axis, but there is no readily understood term for the analogous idea of how a formant behaves on the "vowel axis". For this we introduce the concept of a vowel-formant ensemble (VFE) as the set of values realized for a given formant (e.g., F_2) in going from vowel to vowel among a speaker's vowel phonemes for a fixed time frame in a fixed CVC context. The VFE affords a simple description of our development: we observe that D.J. Broad and F. Clermont's [J. Acoust. Soc. Am. 81 (1987) 155] formant-contour model is a linear function of its vowel target and that as a consequence all its VFEs for a given speaker and formant number are linearly scaled copies of one another. Are VFEs in actual speech also linearly scaled? To show how this question can be addressed, we use F_1 and F_2 data on one male speaker's productions of 7 Australian English vowels in 7 CVd contexts, with each CVd repeated 5 times. Our hypothesized scaling relation gives a remarkably good fit to these data, with a residual rms error of only about 14 Hz for either formant after discounting random variations among repetitions. The linear scaling implies a type of normalization for context which shrinks the intra-vowel scatter in the F_1F_2 plane. VFE scaling is also a new tool which should be useful for showing how contextual effects vary over the duration of the syllable's vocalic nucleus. © 2002 Elsevier Science B.V. All rights reserved.

Résumé

"Contour" et "trajectoire" sont devenus des termes familiers qui, pour toute voyelle, servent à décrire, sur l'axe des temps, l'évolution des fréquences propres à chacun des formants. Par contre, il y aurait lieu d'établir des vocables analogues permettant de préciser le profil de ces fréquences sur "l'axe des voyelles". On introduit, donc, le concept de *vowel-formant ensemble* (et l'acronyme VFE qui en découle) afin de pouvoir regrouper, de voyelle à voyelle, les fréquences d'un formant (e.g., F_2) qui sont obtenues, à un instant fixe de l'axe des temps, pour le même locuteur et dans le même contexte syllabique CVC. Notons que le concept de VFE contient à lui seul toute la démarche adoptée ici, à savoir que notre modélisation précédente des trajectoires des formants (Broad et Clermont, J. Acoust. Soc. Am. 81, 1987, 155–165) repose sur une fonction linéaire de la cible des voyelles et, de ce fait, suggère l'hypothèse que des relations linéaires devraient aussi servir à caractériser les VFEs propres à un locuteur et chacun des formants à la fois. De telles relations sont-elles vérifiables sur des échantillons réels de parole? On aborde cette question pour les fréquences des formants F_1 et F_2 de 7 voyelles de l'anglais australien qui ont été prononcées par un locuteur masculin, 5 fois de

[*] Corresponding author. Tel.: +1-805-687-7157.
E-mail addresses: djbroad@silcom.com (D.J. Broad), frantz@cs.adfa.edu.au (F. Clermont).

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(01)00010-3

suite, dans 7 contextes syllabiques du type CVd. Nonobstant les variations aléatoires inhérentes aux 5 répétitions, l'application de notre hypothèse aux voyelles en question engendre un écart quadratique moyen qui ne dépasse pas la valeur, remarquablement faible, de 14 Hz pour F_1 et F_2. Les relations linéaires ainsi obtenues se prêtent à une normalisation par rapport au facteur contexte, que l'on démontre par une réduction de la dispersion intra-voyelle dans l'espace planaire F_1F_2. Les relations dérivées du concept de VFE constituent également un nouvel outil devant permettre la mise en évidence des effets de différents contextes au travers des noyaux vocaliques de syllabes.

Keywords: Vowel; Coarticulation; Context; Formant; Scaling

1. Introduction

1.1. Vowels in CVC context

A goal in the study of speech communication is to better understand the link between the continuous stream of physical events in speech and the corresponding discrete sequence of phonetic units. One of the problems we face in establishing this link at the acoustic level is that formant contours for vowels in context consist mainly of transitions and have steady states that are only fleetingly realized. At any instant the formants are functions not only of the current vowel, but also of its preceding and following contexts.

Fig. 1 illustrates what happens to a formant (such as F_2) for a set of monophthongal vowels in some fixed consonant–vowel–consonant (CVC) context (bVd, for example). Just as the contour for each individual vowel makes its transition from the initial consonant to the syllable center and then its transition to the following consonant, the set of contours taken as a whole starts from a relatively compressed pattern at the initial-consonant boundary, moves to a more widely spaced pattern in the syllable center and then becomes more compressed again as it approaches the final consonant.

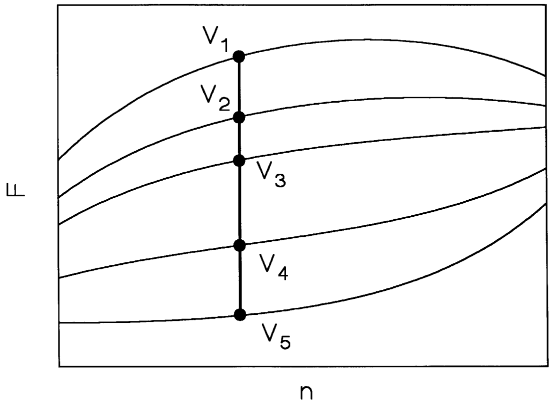

Fig. 1. Hypothetical family of formant (F) contours for a set of vowels in some context. The abscissa represents time in terms of the frame number *n*. The vertical line marks a VFE, which is the set of formant frequencies realized for the different vowels at some fixed frame in the context.

It is this systematic variation in the intervowel spacing for vowels in CVC context that is the topic of this paper and our hypothesis for it will be more easily stated if we first define a new concept, that of the vowel-formant ensemble.

1.2. Concept of the vowel-formant ensemble (VFE)

As just described, Fig. 1 shows variations of a given formant along the two dimensions of time and vowel category for a fixed CVC context. To look at the variations with time while keeping the vowel fixed amounts to selecting a vowel in the figure and following the course of its formant trajectory from beginning to end. For such variation along the time axis we have readily understood terms such as "contour", "trajectory" and "transition", terms for which the notion of the time axis is already implicit.

But as illustrated by the vertical line in the figure, we can also look at how the different vowels are distributed for some fixed frame (relative time position in the syllable). Unfortunately, we have no readily understood term for this idea of how the formant varies on the "vowel axis". In the absence of a ready-made term for this, we now introduce the concept of the vowel-formant ensemble (VFE) by which we mean the set of formant frequencies realized for the set of vowels for the given frame and context. In the figure the vertical slice represents a vowel-formant ensemble.

In this paper our focus will be on the vowel-formant ensemble rather than on individual vowel contours, particularly on how the ensembles for different time frames and contexts are scaled in relation to one another.

1.3. The hypothesis

The hypothesis we explore is the simplest scaling relation we can imagine, namely, that for a fixed formant (such as F_2) all a speaker's VFEs will be linearly scaled copies of one another across CVC contexts and relative time frames in the syllable, i.e., that all these VFEs will be geometrically similar to one another.

Our approach is to first motivate the hypothesis by showing how our earlier time-domain model for formant contours (Broad and Clermont, 1987; cited below as BC87) predicts the linear scaling of VFEs and then to show how this prediction can be tested.

The hypothesis itself is illustrated in Fig. 2 where the top two panels show families of contours for the same formant and the same set of vowels but in two different CVC contexts. A frame is selected from each context and its corresponding vowel-formant ensemble is marked by a vertical line. These ensembles from the two contexts are transferred to the bottom panel, which has the same vertical scale (representing formant frequency) as the top two plots. The horizontal arrangement of the ensembles in the bottom panel is arbitrary, and is planned simply to fit the picture to reasonable proportions. The fact that the vowel placements subdivide the two ensembles in the same proportions, i.e., that the ensembles are geometrically similar to each other, is shown by

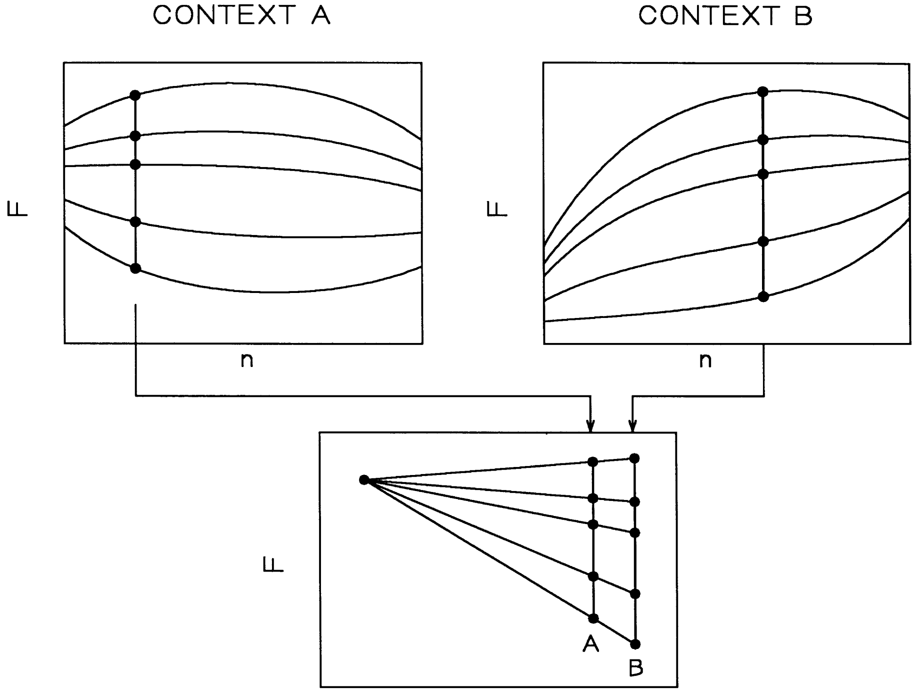

Fig. 2. Similarity of vowel-formant ensembles. The two top panels show the families of formant contours for the same hypothetical set of vowels in two contexts. From each context a VFE is selected and transferred to the same vertical (formant) scale in the bottom panel, where the horizontal arrangement of the ensembles is arbitrary. The similarity of the ensembles is illustrated by the common intersection of all the lines connecting identical vowels in the two ensembles.

the fact that the lines connecting identical vowels in the two ensembles all intersect at the same point. (The location of this point in and of itself is meaningless, as can be seen by how it could be moved around by adjusting the arbitrary horizontal placement of the two ensembles.) That the ensemble from Context A and the one from Context B are selected arbitrarily illustrates our hypothesis: that *all* pairs of a speaker's vowel-formant ensembles for a given formant will be similar to each other, i.e., all these VFEs will be linearly scaled copies of one another across contexts and time frames.

The vowel-formant contours in the different contexts in Fig. 2 have different shapes and different displacements. As the diagram suggests, however, these complexities in the individual contours may give way to a simpler pattern when our point of view shifts from the time axis to the vowel-formant ensemble.

1.4. Overview of study

In Section 2 we start from one of the time-domain models we developed in BC87. It embodies the simple properties of (1) additivity of effects from initial and final consonants, (2) per-consonant similarity of transition shapes, and (3) scaling of transitions by the differences between vowel targets and consonant loci. The model involves parameters and functions which are indexed by the vowel and by the initial and final consonants. The only element of the model that depends on the vowel category is the vowel target, which enters the model only as a first-order factor. These properties imply that within a given formant number (such as F_2) the vowel-formant ensemble from any frame and any context in the model is a linearly scaled version of the model's target ensemble and that as a result all the VFEs for the given formant in the model are linearly scaled versions of one another.

We finish Section 2 by showing how this theoretical result can be connected to data through some simple numerical operations. F_1 and F_2 data on a single speaker's vowels spoken in different CVd contexts are used in Section 3 to illustrate the linear scaling relation and test it statistically using interrepetition variation as a baseline. For this example dataset the hypothesized linear scaling of its VFEs cannot be confirmed in the strictest statistical sense, but its implied rms departure from linearity is only about 14 Hz for each formant. Therefore the F_1 and F_2 VFEs for this dataset are only nearly linearly scaled. If not strictly true, then our hypothesis of linear scaling will be seen to be a remarkably good approximation for this dataset.

In Section 3 we also adapt the linear scaling relation to a form of normalization of vowel-formant ensembles for context and show how vowels become better separated on an F_1F_2 plot. Limitations and applications of the scaling relation are discussed in Section 4 while our conclusions are summarized in Section 5.

It is noteworthy that none of the data operations for analyzing the linear scaling relation require the estimation of any of the parameters or functions that make up the original formant-contour model. This follows from the fact that the result arises from the linear structure of the model and not from any specific implementation of it. As a consequence, the linear scaling of VFEs can be studied as a phenomenon in its own right without reference to the model which predicted it.

2. Development of the scaling hypothesis

In this section we derive the linear scaling of vowel-formant ensembles as a prediction from one of our models in BC87. In Section 2.1 we introduce the model in its time-axis formulation and discuss some of its properties and their roots in earlier work. In particular, we note the linear structure of the model which leads to its prediction of the linear scaling of VFEs. In Section 2.2 a "vowel-axis" reformulation of the model provides an explicit characterization of the linear scaling relation in terms of the parameters and functions of the model. Although this might make it appear that one must first fit vowel data to the model before testing it for the similarity of its VFEs, we show in Section 2.3 that this intermediate step is unnecessary and that the scaling relation can be tested directly on data by fitting a least-squares line to a plot of each VFE against the average ensemble.

We will use such linear fits in Section 3 to exhibit and test the predicted linear scaling on some actual formant measurements.

2.1. A time-axis model for formant contours in CVC context

From a sound spectrogram it is obvious that contextual effects on a vowel in CVC context include at least the transitional intervals leading inward from the consonantal boundaries toward the vowel center. Lindblom (1963) showed that these effects extend at least to the vowel center and show up as a tendency for the phonetic value of the vowel to be undershot. Later, Broad and Fertig (1970) showed by means of two-way analyses of variance on measurements taken at 11 equally spaced frames through a vowel in CVC context that the main effects from both preceding and following consonants on the first three formants were highly significant throughout the vowel's duration, reaching even to the boundaries with the opposite consonants. Öhman (1966) has demonstrated significant effects from transconsonantal vowels in VCV context, showing that contextual effects are not limited to just a vowel's immediate context. More recently, Yang et al. (2000) have found significant information relevant to the phonetic classification of a typical vowel frame to be distributed over an even longer interval extending as far as 200 ms in either direction. In the present paper we make no attempt at characterizing effects from such an extended range and confine ourselves to the effects of the consonants that immediately precede and follow a vowel in CVC context.

In BC87 we studied a series of models for vowel formant contours in CVC$'$ context. (The prime ($'$) on the second consonant C$'$ is to distinguish it from the first one. In the following the prime will also denote constants and functions associated with the second consonant.) Our point of departure in this paper is BC87's formant-contour model for the linear superposition of CV and VC$'$ transitions with per-consonant similarity and target-locus scaling. These properties will be discussed below along with the model itself. The speaker and formant number are fixed for any given implementation of the model and so there will be one model for a speaker's F_1 and another for his or her F_2, etc. A model of this type can be written in the form

$$F_{\mathrm{CVC}'}(n) = T_{\mathrm{V}} + (T_{\mathrm{V}} - L_{\mathrm{C}})G_{\mathrm{C}}(n) + \left(T_{\mathrm{V}} - L'_{\mathrm{C}'}\right)G'_{\mathrm{C}'}(n), \qquad (1)$$

where $F_{\mathrm{CVC}'}(n)$ is the formant frequency realized at frame n $(1 \leqslant n \leqslant N)$ for the vowel V in the context of consonants C and C$'$, T_{V} is the vowel target, L_{C} and $L'_{\mathrm{C}'}$ are the consonant loci, and $G_{\mathrm{C}}(n)$ and $G'_{\mathrm{C}'}(n)$ are time functions that characterize the CV and VC$'$ transition shapes. In BC87's notation, these functions are of the forms $\mathrm{G}_{\mathrm{C}}(n) = \mu_{\mathrm{C}} f^*_{\mathrm{C}}(n) - 1$ and $\mathrm{G}'_{\mathrm{C}'}(n) = \mu'_{\mathrm{C}'} \mathrm{g}^*_{\mathrm{C}'}(n) - 1$ (see their Eqs. (1), (27) and (28)). Note that the subscripts C, V and C$'$ are indices ranging over the phonetic categories of the initial consonant, vowel and final consonant, respectively.

2.1.1. Additivity of effects from C and C$'$

The three terms in Eq. (1) correspond, respectively, to the vowel, the transition from the initial consonant, and the transition into the final one. This superposition of transitions shown here mathematically as the addition of overlapping time functions was hinted at in the observation by Stevens et al. (1966) that "The articulatory processes during the vocalic portion of a consonant–vowel–consonant (CVC) syllable consist of a superposition of several events...". Houde (1968) gave substance to this idea by using the linear superposition of transition functions (with only two fixed forms) to represent his X-ray measurements of tongue movements in VCV$'$CV utterances. Later Broad and Fertig (1970) found that linear superposition of C and C$'$ effects gave a good approximation to their data on one speaker's F_1, F_2 and F_3 for the vowel /ɪ/ in all 576 combinations of 24 initial contexts with 24 final ones. Additivity of C and C$'$ effects was later incorporated into the superposition model of Broad and Clermont (1984) which evolved into the formant models developed in BC87, one of which is Eq. (1) above.

The model's additivity is illustrated in Fig. 3 which shows the two consonant transitions

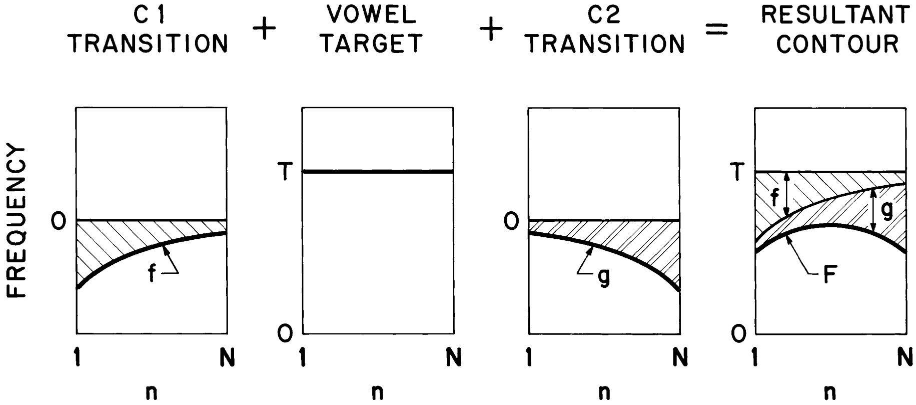

Fig. 3. Additivity of effects from initial and final consonants. The resultant contour on the right is the superposition of the intial transition function *f* and the final transition function *g* with the vowel target. Reproduced with permission from (Broad and Clermont, 1987).

superimposed on the static vowel target. The initial transition function decreases in size with time, but is still nonzero at the final-consonant boundary. Likewise, the final transition function grows from a nonzero value at the left boundary to its maximum size at the right boundary. The resultant contour makes its closest approach to the target near the vowel center where the decreasing contribution from the initial consonant is overtaken by the growing contribution from the final one.

The additivity of C and C′ effects is a component of the model's linear structure which leads to the linear scaling of VFE's.

2.1.2. The vowel target

The vowel target is the only vowel-specific element in Eq. (1) and it occurs only as a first-order factor in each additive term. As will be seen, the linear scaling of VFEs in this model follows from this simple fact. The idea of the target itself is drawn from the target concept of Lindblom (1963) which he defined as the asymptote of a decaying exponential function of vowel duration. With this single parameter he could give a unified description of a vowel's variable realizations for different contexts and durations.

In Eq. (1) we have no exponential functions to provide asymptotic targets. Nevertheless in BC87 we found that unique vowel targets could be obtained for Eq. (1) as values that minimized its rms errors once the consonantal characteristics had been determined by other means. The optimization was not for the individual vowel targets but for a single parameter which in our present terminology corresponds to the scale of the target ensemble. This target concept is more general than that of the target-as-asymptote because it entails no assumption about exact functional forms. It still retains Lindblom's idea of an abstract feature that unifies the description of a vowel's realizations over a range of contexts.

We estimate no vowel targets in this paper and so this specific concept of the target will only be implicit in what follows. Instead the significance of the target will be as a formal element of Eq. (1).

2.1.3. The consonant locus

In a similar fashion it is desirable to be able to associate a characteristic value with the preceding or following consonant. Delattre et al. (1952) observed that the family of vowel formant transitions with a given consonant appeared to be pointing toward a common point of intersection just beyond the vowel boundary, an appearance echoed above in Figs. 1 and 2. Delattre et al. associated this apparent point of intersection with their concept of the consonant *locus*. In BC87 we drew upon this concept but avoided the difficulties of estimating projected points of intersection by redefining the locus as the more easily computed axis of symmetry for the family of transitions in a single-sided CV or VC context.

This idea of the locus as an axis of symmetry for a set of formant contours is implicit in Eq. (1) but, as with the vowel targets, we will not be estimating any consonant loci and the focus will instead be on the locus as a formal element in the model.

2.1.4. Transition shapes

Particular functional forms for the transition shapes have been suggested in earlier works. Stevens and House (1963) used parabolic contours while the exponential duration effect found by Lindblom (1963) strongly suggested exponential transition shapes. The model developed by Houde (1968) for tongue movements used a pair of transition functions which were completely defined by data but which could also be accurately represented by a simple integral formula. Van Bergem (1994) even found straight lines to give good fits to formant transitions in the short-duration unstressed Dutch schwa, which he interpreted as a "targetless" vowel. Broad and Fertig (1970) took the approach of not adopting any particular functional form for the transitions but expressed them numerically in terms of formant data. BC87 did consider exponential transitions but also advanced the less restrictive hypothesis of *per-consonant similarity* which makes no assumptions about the exact forms of the transition shapes but assumes only that the transitions with any consonant in a given position (initial or final) are just scaled versions of some consonant-specific contour shape determined from data. Per-consonant similarity combines simplicity with the flexibility needed to handle different transition shapes being associated with different consonants.

Per-consonant similarity of the CV and VC′ transitions is embodied in Eq. (1) through the fact that the transition-shape functions $G_{\mathrm{C}}(n)$ and $G'_{\mathrm{C}'}(n)$ are indexed only by C and C′.

The model developed by Öhman (1967) for the dynamics of vocal-tract shapes contains a consonant-specific coarticulation function which is a conceptual precursor to the idea of per-consonant similarity in the above model. Like our transition-shape functions, its form is data-driven and not dependent on any particular mathematical form. The vowel enters both our model and Öhman's only in the form of its target and both models assign differences in formant transition shapes (ours) or in articulatory constraints (his) to the consonant.

2.1.5. Target-locus scaling

Stevens and House (1963) and Stevens et al. (1966) observed that the size of a contextual effect grew with the formant-frequency difference between the vowel and the adjacent consonant. The vowel target and consonant locus provide a handy means for expressing this observation quantitatively: the simplest assumption would be for transitions to be scaled in proportion to the difference between the transition's vowel target and consonant locus. Versions of this idea of *target-locus scaling* were incorporated formally into the models of Lindblom (1963); Öhman (1967) and, later, in the form of a separate hypothesis, in BC87. It is incorporated in Eq. (1) through the factors $T_{\mathrm{V}} - L_{\mathrm{C}}$ and $T_{\mathrm{V}} - L'_{\mathrm{C}'}$.

Target-locus scaling is what makes the target a first-order factor in each of the two transition terms and, as will presently be shown, is also what makes Eq. (1) imply the linear scaling of VFEs. This fact also depends on the model's property of per-consonant similarity because target-locus scaling requires the existence of the objects to be scaled, in this case the transition shape functions.

2.1.6. General observations on the model

Note that each parameter or function that makes up the right-hand side of Eq. (1) corresponds to just one phonetic element, the C, V or C′. Hence the model isolates the contributions of the individual sounds and at the same time shows their contributions as being distributed over the duration of the vowel.

Note also that the formalism of Eq. (1) permits a given consonant to affect the formant contour differently depending on whether it occurs in initial or final position. Thus a consonant (such as /d/) can have both different loci for initial and final positions (e.g., L_{d} is not necessarily equal to L'_{d}) and differently shaped transitions (e.g., the final-/d/ transition shape $G'_{\mathrm{d}}(n)$ is not necessarily the mirror image of the initial-/d/ transition shape $G_{\mathrm{d}}(n)$, i.e., is not necessarily equal to $G_{\mathrm{d}}(N-n+1)$).

We will not be using Eq. (1) to represent formant values themselves but will only be using its structure to explore the linear scaling relation among vowel-formant ensembles. For this we will not need any values for the targets or loci but will only need to assume their existence as structural elements of the model. In order to focus on the formal relationships implied by Eq. (1), therefore, we leave aside the questions of how targets and loci might be estimated from data and of how they might be interpreted in terms of articulation or higher-level representations. In a similar fashion our present purposes will require only the *existence* of the transition shapes $G_{\mathrm{C}}(n)$ and $G'_{\mathrm{C'}}(n)$ and so their exact forms need concern us no further. (BC87 describes some methods for estimating the elements of Eq. (1).)

Indeed, with all its special properties – additivity, per-consonant similarity, target-locus scaling, and the dependence of each element on only a single phonetic segment – one might well ask: Besides describing individual formant contours, should the model not also imply some more general characteristic of vowel data? One might well hope for some sort of analogy with the theory of electric circuits where, for example, the mere fact that a circuit is linear implies that its response to a sinusoidal input of a given frequency will also be sinusoidal with the same frequency, independently of the details of the circuit. In a similar way the linear scaling relation about to be derived is a general characteristic of the model which follows from its linear structure, independently of the numerical details of its parameters and functions. Because it depends on these general properties the linear scaling relation will also provide an indirect test of the model's assumptions without our having to go into its details.

2.2. Deriving the linear scaling of VFEs from a "vowel-axis" reformulation

In terms of Eq. (1) the vowel-formant ensemble entails going from one value of the target T_{V} to another. This can be emphasized by a slight rewrite of Eq. (1):

$$F_{\mathrm{CVC'}}(n) = K_{\mathrm{CC'}}(n)T_{\mathrm{V}} + J_{\mathrm{CC'}}(n), \tag{2}$$

where

$$K_{\mathrm{CC'}}(n) = 1 + G_{\mathrm{C}}(n) + G'_{\mathrm{C'}}(n) \tag{3}$$

and

$$J_{\mathrm{CC'}}(n) = -L_{\mathrm{C}}G_{\mathrm{C}}(n) - L'_{\mathrm{C'}}G'_{\mathrm{C'}}(n) \tag{4}$$

are constants for any given vowel-formant ensemble, i.e., for any frame-context combination. $K_{\mathrm{CC'}}(n)$ represents a scale factor for the VFE corresponding to Frame n and Context C_C′. It is independent of the consonant loci and depends only on the transition-shape functions. The loci do enter into $J_{\mathrm{CC'}}(n)$, which represents a translation of the VFE. $K_{\mathrm{CC'}}(n)$ and $J_{\mathrm{CC'}}(n)$ are both independent of the vowel target.

Eq. (2) is just a form of Eq. (1) with the emphasis shifted away from the time-axis representation of individual contours and toward a "vowel-target-axis" representation of the vowel-formant ensemble. Indeed, the representation is extremely simple on the T_{V} axis: the realized formant is just a linear function of the target, i.e., the vowel-formant ensemble for any context-frame pairing is a linearly scaled copy of the ensemble of vowel targets. As explained in connection with Fig. 2, this is the same as saying that any of the model's VFEs will be geometrically similar to the target ensemble.

It follows that the ensembles for all context-frame combinations in this model are similar to each other as well.

We have therefore just shown that our hypothesis is a prediction from the model of Eq. (1): the model's VFEs are linearly scaled copies of the target ensemble and hence linearly scaled copies of each other.

As it stands, the right-hand side of Eq. (2) consists entirely of unknown parameters and functions of the model. Although this might make it appear that testing the predicted scaling relation depends on first fitting the model to data, we show in the next subsection that this intermediate step is not necessary.

2.3. Relation to observables

2.3.1. The ensemble centroid

Our hypothesis concerns the scaling term in Eq. (2) and has nothing to do with the translation term. Our first step will therefore be to eliminate $J_{\mathrm{CC'}}(n)$ by translating each ensemble via a subtraction of its own centroid. The centroid for Frame n of Context C_C′ will be

$$\begin{aligned} A_{\mathrm{CC'}}(n) &= \frac{1}{N_{\mathrm{V}}} \sum_{\forall \mathrm{V}} F_{\mathrm{CVC'}}(n) = F_{\mathrm{C{\cdot}C'}}(n) \\ &= K_{\mathrm{CC'}}(n) T_{\cdot} + J_{\mathrm{CC'}}(n), \end{aligned} \tag{5}$$

where N_{V} is the number of vowels and $T_{\cdot}$ is the intervowel average of the vowel targets. (Following Scheffé (1959) we use the dot $(\cdot)$ in a subscript or argument position to denote averaging over the corresponding index or variable. This avoids proliferating summations whose only purpose is to implement averaging operations.)

Eq. (5) is of the same form as Eq. (2) except that the average target takes the place of the individual vowel target. Taking the difference between Eqs. (2) and (5) results in the desired form from which the translation term $J_{\mathrm{CC'}}(n)$ is eliminated:

$$F_{\mathrm{CVC'}}(n) - A_{\mathrm{CC'}}(n) = K_{\mathrm{CC'}}(n)(T_{\mathrm{V}} - T_{\cdot}). \tag{6}$$

2.3.2. The average vowel-formant ensemble

Ideally, we would like to use Eq. (6) to compare each vowel-formant ensemble directly with the target ensemble. But since we do not have the target ensemble, we will instead use the ensemble of per-vowel averages as the basis of comparison. Let the average of Eq. (6) for Vowel V over contexts and frames be

$$\Phi_{\mathrm{V}} = F_{\cdot\mathrm{V}\cdot}(\cdot) - A_{\cdot\cdot}(\cdot) = K_{\cdot\cdot}(\cdot)(T_{\mathrm{V}} - T_{\cdot}). \tag{7}$$

The Greek letter Φ is chosen to suggest kinship with F. The term $F_{\cdot\mathrm{V}\cdot}(\cdot)$ is the average over all frames and contexts of the data for Vowel V while the term $A_{\cdot\cdot}(\cdot)$ is the centroid of the ensemble of $F_{\cdot\mathrm{V}\cdot}(\cdot)$'s. Φ_{V} is therefore the result of translating the mean ensemble by its own centroid. Eq. (7) shows that Φ_{V} is proportional to $T_{\mathrm{V}} - T_{\cdot}$ and so provides an easily computed surrogate for it. Our next step will be to show how the mean ensemble leads to an explicit way to exhibit the linear scaling relation among VFEs using only simple operations on data.

2.3.3. The scaling relation in terms of data

If we solve Eq. (7) for $T_{\mathrm{V}} - T_{\cdot}$ and substitute the result into Eq. (6), we obtain

$$F_{\mathrm{CVC'}}(n) - A_{\mathrm{CC'}}(n) = \frac{K_{\mathrm{CC'}}(n)}{K_{\cdot\cdot}(\cdot)} \Phi_{\mathrm{V}} = a_{\mathrm{CC'}}(n) \Phi_{\mathrm{V}}. \tag{8}$$

We may not be in a position to estimate either $K_{\mathrm{CC'}}(n)$ or $K_{\cdot\cdot}(\cdot)$, but Eq. (8) shows that we should be able to obtain their ratio $a_{\mathrm{CC'}}(n)$ as the slope of a line from the origin fit to a vowel-by-vowel plot of $F_{\mathrm{CVC'}}(n) - A_{\mathrm{CC'}}(n)$ against Φ_{V}. Both these quantities are obtainable directly from formant data and so, therefore, is the slope $a_{\mathrm{CC'}}(n)$, which represents the scaling of the vowel-formant ensemble for Frame n of Context C_C′ with respect to that of the mean ensemble.

Eq. (8) will be our vehicle for testing data on vowels in context for the linear scaling of their vowel-formant ensembles. Our hypothesis is equivalent to the prediction from Eq. (8) that each VFE's plot of $F_{\mathrm{CVC'}}(n) - A_{\mathrm{CC'}}(n)$ against Φ_{V} will fit a straight line from the origin. In Section 3 this will be our link between the hypothesis and data.

2.3.4. The ensemble scale

In Eq. (8) $a_{\mathrm{CC'}}(n)$ is the constant of proportionality between a VFE's scale and that of the mean ensemble. We call $a_{\mathrm{CC'}}(n)$ the *ensemble scale* for the frame and context. In terms of Figs. 1 and 2, the more compact formant spacings among the vowels near the consonant boundaries correspond to more condensed VFEs and hence to smaller values of the ensemble scale while the more dilated spacings near the syllable centers correspond to more elongated VFEs and to larger values of the ensemble scale.

3. Test with data

3.1. The data

3.1.1. Utterances and recording

The data we use are from a General Australian English speaker's productions of $N_{\mathrm{V}} = 7$

undiphthongized vowels in 7 CVd contexts (C = /h, b, d, g, p, t, k /, V = ɪ, ɛ, æ, a, ɒ, ʌ, ɜ/, C′ = /d/). We follow Bernard (1970) for the phonetic symbolization of the vowels. Key words for their pronunciations are, respectively, “hid”, “head”, “had”, “hard”, “hod”, “hudd” and “herd”. Each combination of vowel and context is represented in the dataset by $N_{\mathrm{rep}} = 5$ repetitions.

The vowels included in the dataset are monophthongs in Australian English, but the /i/ and /u/ which one might expect to see are, as noted by Bernard, “less happily so”. Indeed, our speaker diphthongized these vowels and so we did not include them in the following. Exclusion of diphthongs is discussed in Section 4.1.2.

A computer program randomized the ordering of the syllables in the dataset and prompted the speaker by displaying one syllable every 3 s. The speaker was encouraged to take a fresh breath after each display. The recordings thus obtained were found to have no list intonation or breath group modulations, but do exhibit some detectable variations in stress and pitch.

3.1.2. Segmentation and frame placement

The analog recordings were converted to digital form at a sampling frequency of 10 kHz and quantized to 12 bits. The onset and offset of every syllable’s vocalic nucleus were determined iteratively by visual inspection of the waveform and auditory confirmation. For each utterance the left segmentation point for the vowel was placed at the onset of the voiced vocalic interval. For the initial voiced plosives this was typically only a few milliseconds after the transient for the release. For the voiceless initial plosives the placement is at the onset of voicing following the interval of post-release aspiration. Likewise, voicing onset is the criterion used for the boundary with the initial voiceless fricative /h/. The segmentation point with the final /d/ was placed at the last sample that could be reasonably associated with the vowel. This is usually clearly marked by the sudden change in the waveform at the moment of closure.

After segmentation, $N + 2 = 13$ equally spaced analysis frames 25.6 ms in width were positioned within the vowel interval. Because reliable formant measurements at the vowel boundaries were not always possible, the first (Frame 0) and last (Frame 12) of these frames were skipped and formants were tracked for the central $N = 11$ frames. This equal spacing of a fixed number of frames was accomplished by the method described in BC87 in which the size of the frame advance for the analysis is adjusted on the basis of the vowel duration, with the beginning of Frame 0 aligned with the vowel onset and with the end of Frame 12 aligned with the end of the vowel. Durations (including all 13 frames) ranged between 102 and 318 ms. This method of frame placement amounts to normalizing the vowels to a common duration.

3.1.3. Formant measurements

The three lowest formant frequencies were obtained via a 14th-order LPC-autocorrelation analysis of Hamming-windowed frames of 25.6 ms duration. Initial tracking was done by means of peak-picking on the LPC magnitude spectrum using the algorithm described by McCandless (1974). This was followed by a manual verification and correction stage guided by criteria based on bandwidth estimates, formant ranges and continuity.

3.1.4. Averaging over repetitions

Because the elements of the formant-contour model are indexed by phonetic categories, the best it can do for any CVC′ is to represent the statistical mean for repetitions of a speaker’s productions. Details of variations among the speaker’s individual tokens are beyond the reach of such a model and must be characterized statistically. We shall return to this point in Section 3.4 but for the moment the point is that what we fit to the linear scaling relation (Eq. (8)) will not be the data from the individual utterances but the values obtained by averaging each CVC′ combination over the $N_{\mathrm{rep}} = 5$ repetitions. More precisely, if we let $i = 1, 2, \ldots, N_{\mathrm{rep}}$ be an index for repetition number, then we can denote the formant realized at the nth frame for the ith repetition of the sequence CVC′ as $\mathrm{F}_{\mathrm{CVC}'i}(n)$. Averaged over the repetitions this becomes $\mathrm{F}_{\mathrm{CVC}'\cdot}(n)$. In the following it will be this interrepetition average that is used in our fits to Eq. (8). We shall return to the variation among individual repetitions in Section 3.4 for the statistical evaluation of the hypothesis.

3.1.5. A glimpse at the data

Before getting into the abstraction of the linear scaling relation it may be useful to get some idea of what the data look like and at the same time to have a concrete picture of how we move from a set of contours to a set of VFEs. Fig. 4 shows by way of example the family of F_2 contours for the bVd context. As just discussed, each contour in the figure represents the average of the speaker's five repetitions of the syllable. One can see a faint but perceptible tendency for the contours to converge toward a /b/ locus on the left and a more pronounced contraction toward a /d/ locus on the right. In the center the contours are somewhat more widely separated as they approach their vowel targets. Though not as accentuated as the stylization in Fig. 1, this family of actual contours plainly exhibits the existence of regularity in the bVd context. One can also see that not all the VFEs for this context will be exactly linearly scaled copies of one another, most obviously from the contour for /a/ crossing that for /ʌ/ thus reversing these vowels' order within their ensembles. Departures from the scaling relation will be discussed in relation to the statistics of interrepetition variation in Section 3.4.

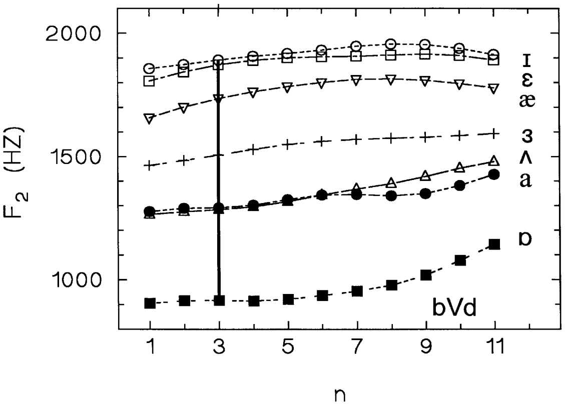

Fig. 4. The family of F_2 contours for vowels in bVd context. Each contour represents the mean of five repetitions of its syllable. The vertical line marks the VFE for Frame 3 in the context.

In anticipation of its serving as our illustration for the scaling relation in Section 3.2, the VFE for Frame 3 of the bVd context is marked by a vertical line.

3.2. Implementing the scaling relation

Fig. 5 illustrates Eq. (8) for the F_2 VFE from Fig. 4 for Frame 3 ($n = 3$) of vowels in bVd context. F_2 values for $F_{\mathrm{bVd}}(3) - A_{\mathrm{bd}}(3)$ are plotted against Φ_{V} for the seven vowels in the database. As just discussed in Section 3.1.4, each data point represents the average of five repetitions. The least-squares line from the origin fits the data with an rms error of 23 Hz. Its slope, 1.06, is then an estimate of the F_2 ensemble scale $a_{\mathrm{bd}}(3)$ for Frame 3 of Context b_d. That $a_{\mathrm{bd}}(3) > 1$ indicates that this ensemble is a slightly enlarged copy of the mean ensemble.

For the VFE corresponding to each combination of context and frame the least-squares calculation for the fit to Eq. (8) will yield a value for the ensemble scale and a record of the errors in fitting the data to the line. In Section 3.3 we give a graphical summary of these calculations while in Section 3.4 we describe a statistical analysis of the fit between the data and the linear scaling.

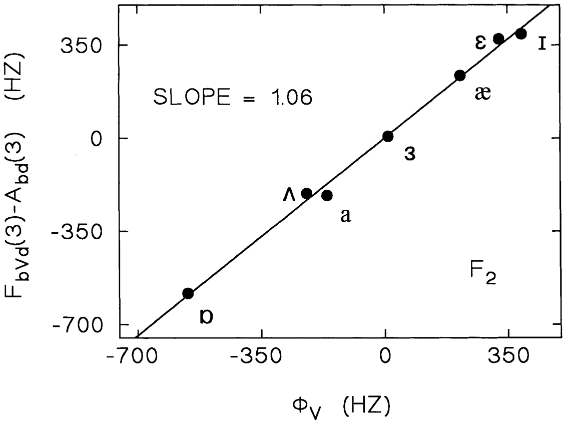

Fig. 5. Using Eq. (8) to fit the scaling relation to data. Second-formant values of $F_{\mathrm{bVd}}(3) - A_{\mathrm{bd}}(3)$ are plotted against Φ_{V} for Frame 3 of an Australian English speaker's productions of 7 vowels in bVd context. A line from the origin fits the data with an rms error of 23 Hz. Its slope of 1.06 is then an estimate of the ensemble scale $a_{\mathrm{bd}}(3)$.

3.3. Summarizing the linear scaling relation

It is neither very practical nor informative to generate plots like Fig. 5 for all the ensembles. Our database would have $NN_{\mathrm{C}}N_{\mathrm{C}'} = 77$ such plots for each formant. (Even this would be an improvement over looking at the $1/2 \times 77 \times 76 = 2926$ plots for all the pairwise comparisons of a formant's VFEs.) It is still desirable to have a more efficient graphical means for summarizing the results of the fits of the type shown in Fig. 5. Once the above calculations have been performed on all of a formant's VFEs we can use the resulting values of the ensemble scale to produce a plot which summarizes the fits to the scaling relation. For this we plot each VFE (still translated by its own centroid as in Fig. 5) against its own ensemble scale. Fig. 6 shows such a summarizing plot for the first-formant ensembles and Fig. 7 shows one for the second-formant ensembles. In these plots each VFE is represented by a set of 7 data points (one for each of its vowels), positioned horizontally according to its ensemble scale. In such a plot a formant's VFEs will be organized left to right in ascending order of their ensemble scales. As in Fig. 5, each data point represents the mean of five repetitions.

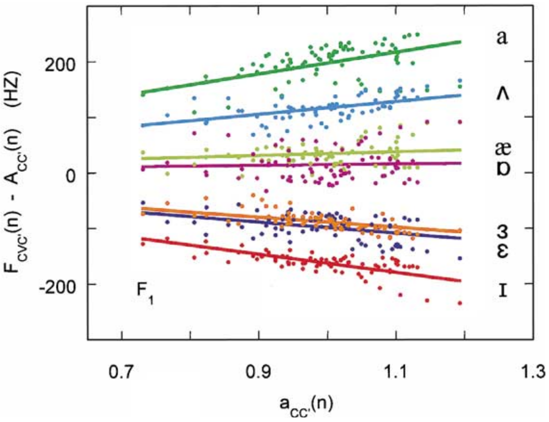

Fig. 6. Similarity of the F_1 vowel-formant ensembles from the database. $F_{\mathrm{CVC}'}(n) - A_{\mathrm{CC}'}(n)$ is plotted against the ensemble scale $a_{\mathrm{CC}'}(n)$ for 11 equally spaced frames from 7 Australian English vowels in 7 CVd contexts.

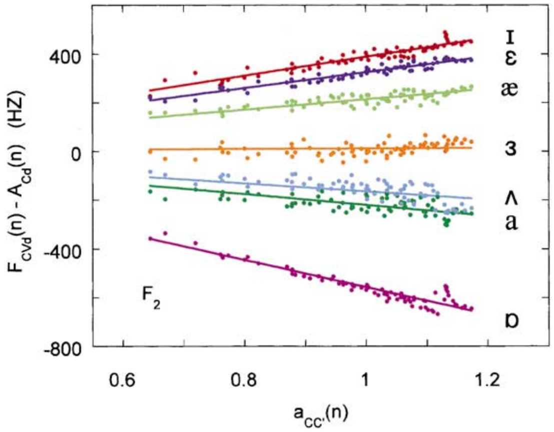

Fig. 7. Similarity of the F_2 vowel-formant ensembles. The plot is of the same type as Fig. 6 and the vowels have the same colors as they do in Fig. 6.

According to Eq. (8) the data for each vowel in Fig. 6 or 7 should be a line from the origin with a slope equal to its value of Φ_{V}. Like Fig. 5, then, Figs. 6 and 7 also represent Eq. (8) but with a reversal of the roles between the slope and the independent variable. In these summarizing plots the vowel-specific lines give a picture of how the data fit the linear scaling relation. These lines also help to visualize the scaling relation as a geometric similarity relation by the fact that they will subdivide any vertical slice into the same proportions. In the following we shall therefore refer to Figs. 6 and 7 as *similarity plots*.

A strong pattern is apparent in the data which seems to support the idea of similarity among vowel-formant ensembles. Through its range of values $a_{\mathrm{CC}'}(n)$ shows the strength of contextual effects on ensemble scaling, which for F_1 spans a factor of about 1.6 and for F_2 one of about 1.8. In light of Eq. (6), Φ_{V} is just the mean value of the formant for the vowel V relative to the formant's grand mean. Hence the pencils of lines in Figs. 6 and 7 are ordered with increasing slopes (Φ_{V}'s) corresponding to increasing formant values. Thus for the F_1 results in Fig. 6 the vowels are ordered from bottom to top roughly by their vertical places of articulation from high to low, while for the F_2 results in Fig. 7 they are ordered roughly by their horizontal places of articulation from back to front.

3.4. Statistical variation

The idea of using variation among repetitions as a criterion for evaluating a model has a certain intuitive appeal which Houde (1968) used to advantage by displaying tongue motions predicted by his model side by side with motions measured from pairs of actual utterances to show that the predictions differed from the real motions by no more than these differed from each other. Later in (Broad and Fertig, 1970) and in BC87 statistics on interrepetition variation were used as a baseline for evaluating models of formant trajectories. As discussed in more detail in (Broad, 1976), this variation tends to be surprisingly small and is comparable in size to the perceptual difference limen for formant frequencies (Flanagan, 1955). We now take up the variation among the individual repetitions of the syllables in our database as a means for evaluating the linear scaling of its VFEs.

Table 1 shows a vowel-by-vowel breakdown of the observed rms error $s_{\mathrm{obs,V}}$ in fitting the scaling relation and of the unbiased estimates of the interrepetition variation $s_{\mathrm{rep,V}}$. Each error in the table is the rms value of the vowel's contributions to the errors in fitting Eq. (8) to the $NN_{\mathrm{C}}N_{\mathrm{C'}}$ distinct ensembles. The table also shows the overall rms values s_{obs} and s_{rep}. The vowels differ substantially in how well they fit the scaling relation. These differences tend to be tracked by differences among the interrepetition variations, but not without exceptions: for example, /ɒ/ has the largest interrepetition variation for F_2 and only the third largest rms error for this formant.

How well do the errors in Table 1 compare with those we would expect on the basis of interrepetition variation? Under the null hypothesis that all the errors are attributable to interrepetition variation, these errors would to a first approximation be expected to be about $s_{\mathrm{rep}}/\sqrt{N_{\mathrm{rep}}}$. However, the errors also depend on the number of vowels, as can be seen from the zero error we would have for the determinate case with $N_{\mathrm{V}} = 2$. (A plot like Fig. 5 with two data points equidistant from their mean would fit a line from the origin perfectly.) This would suggest an educated guess of a further factor of $(N_{\mathrm{V}} - 2)/N_{\mathrm{V}}$ for the variance, which would lead to a rule-of-thumb estimate s_{est} for the rms error expected for the null hypothesis of

$$s_{\mathrm{est}} = s_{\mathrm{rep}}\sqrt{\frac{N_{\mathrm{V}} - 2}{N_{\mathrm{V}}N_{\mathrm{rep}}}}. \tag{9}$$

Because the exact rms error to be associated with the null hypothesis is hard to determine from first principles, however, we did some Monte-Carlo simulations in which artificial data were first computed from Eq. (8) using the same values of Φ_{V} and $a_{\mathrm{CC'}}(n)$ as in the original data. Each of the formant values was then perturbed by the addition of the mean of N_{rep} random gaussian samples with the values of $s_{\mathrm{rep,V}}$ set equal to those in Table 1. The simulations assume utterances to be statistically independent of one another but take account

Table 1
Vowel-by-vowel statistics[a]

Formant	F_1		F_2	
Vowel V	$s_{\mathrm{obs,V}}$ (Hz)	$s_{\mathrm{rep,V}}$ (Hz)	$s_{\mathrm{obs,V}}$ (Hz)	$s_{\mathrm{rep,V}}$ (Hz)
ɪ	15	16	17	33
ɛ	20	31	13	31
æ	20	29	18	31
a	28	49	29	45
ɒ	29	60	25	80
ʌ	17	30	35	61
ɜ	12	19	24	36
rms	21	36	24	48

[a] For vowel V the rms error in fitting the linear scaling relation is $s_{\mathrm{obs,V}}$ and the unbiased estimate of the interrepetition variation is $s_{\mathrm{rep,V}}$.

of dependencies among frames within an utterance by modifying the random perturbations to emulate the interframe covariances of the actual data. We did 100,000 simulations for each formant with 77 VFEs per simulation. These simulations provide the baseline statistics for the null hypothesis.

Table 2 summarizes the results of the simulations and the actually observed rms errors together with the values from Eq. (9) for the educated guesses based on the number of repetitions and the number of vowels. The null-hypothesis values s_{MC} from the Monte-Carlo simulations fall close to those of s_{est} from Eq. (9).

The actual rms spreads s_{obs} of the data about the lines in the similarity plots are 21 Hz for F_1 and 24 Hz for F_2. Of the 100,000 Monte-Carlo simulations for each formant, only 29, or about 0.03%, had rms errors exceeding the actual value for F_1 while 215, or about 0.22%, had rms errors exceeding that for F_2. Therefore we may conclude that the null hypothesis can be rejected at about the 99.98% level for F_1 and at about the 99.7% level for F_2. There is therefore almost certainly some systematic error attributable to the model itself and in the strictest statistical sense the linear scaling relation cannot be confirmed.

But besides knowing that departures from the linear scaling relation are highly significant statistically, we also want to know how large these errors might be. If we assume that errors due to interrepetition variation (with a variance of s^2_{MC} from the Monte-Carlo simulations) are uncorrelated with the errors in the underlying model (with a variance of s^2_{mod}), then these variances would sum to the observed mean squared error s^2_{obs} and the rms error attributable to the underlying model would then be

$$s_{mod} = \sqrt{s^2_{obs} - s^2_{MC}}. \qquad (10)$$

As shown in Table 2, s_{mod} is 14 Hz for both F_1 and F_2. In comparison with the interrepetition variations and even in comparison with our ability to measure formants, we consider these implied levels of accuracy to be quite encouraging. Indeed, it seems remarkable that departures from the scaling relation that are this small should be statistically detectable at all. Therefore, even though the above statistical reasoning shows that the linear scaling relation must be rejected in the strictest sense, the estimated sizes of the systematic errors are numerically quite small and so, in a practical sense the linear scaling can be accepted as a good approximation. Within this data set, then, VFEs are linearly scaled to nearly within the speaker's ability to repeat any given vowel.

Although the F_1 plot looks less convincing than the F_2 one does, the implied levels of accuracy for the underlying models are comparable and the difference in appearance may be attributed in part to contextual effects being weaker for F_1 and in part to its smaller total range.

3.5. Normalizing vowel-formant ensembles for context

A slight rearrangement of Eq. (8) shows how the linear scaling of VFEs leads to a sort of normalization for context:

Table 2
Statistics related to the linear scaling relation[a]

	Rms errors				
Formant	s_{obs} (Hz)	s_{MC} (Hz)	s_{rep} (Hz)	s_{est} (Hz)	s_{mod} (Hz)
F_1	21	15	36	14	14
F_2	24	19	48	18	14

[a] The observed rms spread of the data points about their best-fit lines is s_{obs}. The rms spread expected on the basis of the Monte-Carlo simulations is s_{MC}. The unbiased estimate of the interrepetition variation is s_{rep} and the related rule-of-thumb estimate from Eq. (9) for the rms error estimated from the number of repetitions and the number of vowels is s_{est}. The rms error attributable to the model itself as estimated by Eq. (10) is s_{mod}.

$$\Phi_V = \left[F_{CVC'}(n) - A_{CC'}(n)\right]/a_{CC'}(n). \qquad (11)$$

Eq. (11) shows that as far as the model is concerned the ensemble-local ratio of $F_{CVC'}(n) - A_{CC'}(n)$ to $a_{CC'}(n)$ is the constant Φ_V for each vowel V, i.e., this ratio is invariant over the range of frames and contexts for which the model is defined. This invariance is a property of the coarticulation model and in practice there will be some statistical variation in using Eq. (11) to estimate Φ_V from frame- and context-local data, variation just described in relation to the statistical analysis of the scaling relation itself.

Fig. 8 illustrates the normalization with a plot of F_2 against F_1 superimposed on one of Φ_2 against Φ_1 for the data from Figs. 6 and 7. In accordance with Eq. (7), the Φ scales are just translations of the corresponding F scales by the grand means $A_{..}(\cdot)$. The dots in the plot represent the per-vowel means for both F and Φ. (That these should coincide in the plot is clear from Eq. (7).) Around each vowel's mean are two convex hulls, an outer one enclosing the formant data and an inner one the Φ data.

In going from F to Φ under the normalization the hulls shrink toward their intravowel means. The compression is generally better for the second formant than for the first. Although the Φ hulls are not all completely interior to their corresponding F hulls, the vowels do become better separated, in this instance completely without overlap.

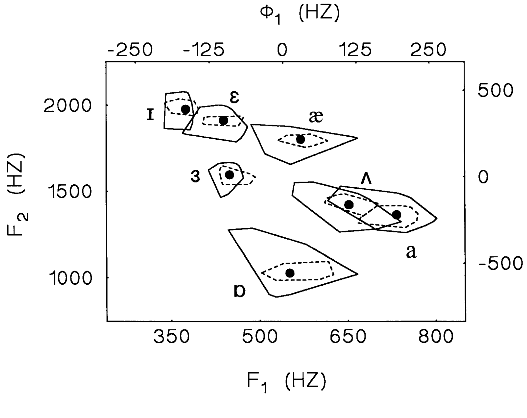

Fig. 8. Normalizing vowel-formant ensembles for context using Eq. (11). Convex hulls enclose each vowel's 77 data points for F_2 plotted against F_1 (solid loops) and for Φ_2 against Φ_1 (dashed loops). Fs and Φs use the same scale sizes and are positioned to make their per-vowel means (dots) coincide. The normalization shrinks the intravowel dispersion and enhances the intervowel separation.

4. Discussion

4.1. Some limitations

4.1.1. Systematic errors in the model

We know ahead of time that Eq. (1) will almost certainly not be exact even for the mean statistical trend of phonetic data. As already shown by Broad and Fertig (1970) there exists a systematic departure from additivity which, though numerically small, is highly significant statistically. This is similar to our scaling result where the departures from Eq. (8) are greater than those attributable to interrepetition variation to a high (>99.5%) level of significance, even though numerically the excesses attributable to inaccuracy in the model itself are quite small (~14 Hz).

For the data presented here we are therefore inclined to accept the linear scaling of VFEs for practical purposes because, with only simple operations on the data and with errors only slightly larger than expected, it captures a strong regularity in the contextual conditioning of vowels. This use of an admittedly inexact model would be for much the same reason we use the acoustic theory of speech production (Fant, 1960) which treats the vocal tract as a linear filter. Even though we know ahead of time that the vocal tract is subject to various nonlinear effects, its response is linear enough for the theory to provide reasonable predictive power and really powerful explanatory power. In the same way, we also expect the additive model of Eq. (1) and its corollary, the linear scaling of VFEs, to be usefully descriptive and to provide a point of departure for studying nonlinear effects.

4.1.2. VFEs and contexts should be phonetically homogeneous

Phonetic homogeneity within the vowel set is implicit in the idea of the vowel-formant ensemble. The set of vowels must be stable across the contexts and frames for which the linear scaling relation is to be studied. Stability is also assumed for the context in the sense that the C and C′ are assumed to maintain their phonetic identities across the range of vowels. Inhomogeneities can arise from allophonic variation or from dynamic variation among frames in the case of a diphthong.

Allophonic variations of a consonant conditioned by the adjacent vowel might cause some departure from uniform scaling of VFEs. In our data, for example, the /g/ in a gVd will be realized with a more palatal allophone when the V is a front vowel and a more velar one when the V is a back vowel. Contrary to what we might have expected, however, we found here as in BC87 for this particular case that lumping all the gVd data together created little problem, possibly because for neither F_1 nor F_2 is there a strong contextual effect from the /g/. This might be explained by the cineradiographic observation by Houde (1968) that /g/ assimilates along a horizontal continuum with the position of closure dictated completely by the horizontal place of articulation of the vowel. A similar continuum of horizontal adjustments probably also holds for /k/. Allophones which represent a more categorical than continuous variation might present more of a problem for the scaling relation. It remains to be seen, for example, whether other English phonemes such as /l/ which have an analogous pattern of assimilation (e.g., velarized in “all” but not in “ill”) will vary more continuously or more categorically. In either case will they still have linearly scaled VFEs?

Allophonic variation must be kept in mind for special treatment if necessary. Though it appears that no distortion of the scaling relation has resulted from treating /g/ and /k/ as single consonantal categories, this is more the result of a special circumstance than a sign that allophonic variation can be ignored in all cases. For the model such cases may mean that the Cs and Vs will sometimes have to refer not to phonemes but to the phonetic categories of their allophones, each with its own distinctive locus or target. In this event the vowels making up an ensemble for a phonemically designated context would be allotted to different allophonic variants of the context, leaving no single variant with a full complement of vowels. For the linear scaling relation such incomplete VFEs would have to be excluded from consideration, unless one had a method for comparing VFEs made up of different sets of vowels, a development beyond the scope of our present effort.

Diphthongs represent another source of phonetic inhomogeneity in VFEs. A diphthong or glide as a compound syllable nucleus would have its own internal dynamics (Clermont, 1991, 1993) which would preclude its different frames from making up a homogeneous set of VFEs. A diphthong’s path of data points in a similarity plot such as Fig. 6 or 7 would cross from the line for its initial vowel to the line for its final one. The present approach therefore requires the database to be restricted to monophthongs. As mentioned in Section 3.1.1, diphthongization was the reason /i/ and /u/ had to be excluded from our vowel set. Residual diphthongization in the vowels retained in the dataset could contribute to the systematic error.

4.1.3. Duration-normalized time scale

Our development uses a time scale normalized for duration. This is implicit in the frame numbering scheme where $n = 1$ corresponds to the first modelled frame and $n = N$ to the last. As discussed in more detail in BC87, the model formulated with a duration-normalized time scale cannot capture systematic effects attributable to the duration itself. As shown by the discovery by Lindblom (1963) of duration-dependent vowel reduction in Swedish vowels, such effects can be present and would be a source of systematic errors for the duration-normalized model. As shown by Van Son and Pols (1992) for Dutch vowels extracted from continuous speech produced at different rates, however, there may sometimes be no such effects in a given data set and the duration-normalized time scale could then become appropriate.

4.1.4. Normalization operates on ensembles, not tokens

Because Eq. (11) for the normalization of vowels for context involves the ensemble-wide variables $a_{CC'}(n)$ and $A_{CC'}(n)$, its implementation requires data on all the vowels in the ensemble and so is inapplicable to vowels taken in isolation from their ensembles. As a result it cannot in its present form be directly applied to recognition which operates on individual input tokens.

4.2. The linear scaling as a tool for acoustic phonetic analysis

4.2.1. Similarity plots as diagnostics

These restrictions aside, similarity plots such as Figs. 6 and 7 provide convenient ways to determine whether the vowel-formant ensembles for a set of CVC′ data are reasonably similar. The range of values for the ensemble scale would be a measure of the strength of contextual effects in the dataset.

We have looked at only a small dataset and must allow for situations in which Eq. (1) is simply wrong. In such a case a similarity plot could be a helpful diagnostic. For example, if one were unsuccessful in fitting Eq. (1) it might still be possible that one has only failed to find good values for the elements of the model. If a plot like Fig. 6 or 7 fails, however, one would then know that *no possible* assignment of values to the elements of Eq. (1) would work and one would be saved from a futile effort.

On the other hand, good linear fits to the similarity plots are only necessary but not sufficient conditions for workable forms of Eq. (1) to exist. To see this, imagine synthesizing some perfect data from Eq. (1) and then randomizing the assignments of frames and contexts to the resulting vowel-formant ensembles: the similarity plots would look the same, but the underlying phonetic order would be destroyed. (Randomizing phonetic labels of frames has been used previously by Yang et al. (2000) to establish a baseline in their study of the mutual-information measure.)

4.2.2. Systematics of the ensemble scale

In constructing similarity plots such as Fig. 6 or 7 we ignore the frames and contexts from which VFE's are taken. Once established, the linear scaling relation permits us to study the systematics of the ensemble scale. Fig. 9 shows $a_{CC'}(n)$ for F_1 plotted as a function of the frame number for the different contexts in the database while Fig. 10 is the corresponding plot for the F_2 data. The trajectories in both plots are all fairly smooth curves. Note also that the curves for both F_1 and F_2 for the different contexts remain distinct even toward the boundary with the final /d/, illustrating the persistence of effects from the initial consonant through the duration of the vowel.

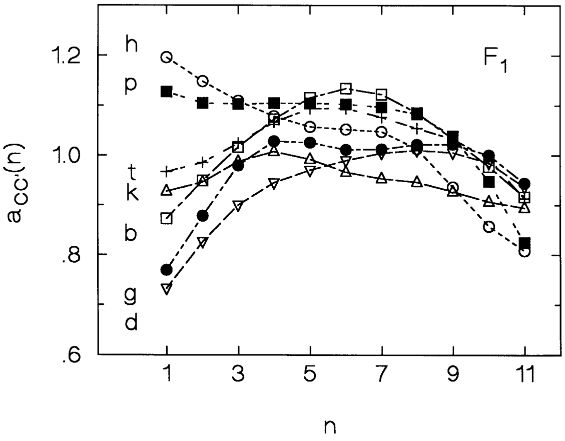

Fig. 9. The ensemble scale $a_{CC'}(n)$ for F_1 as a function of the frame number n for the seven initial consonants used in the study.

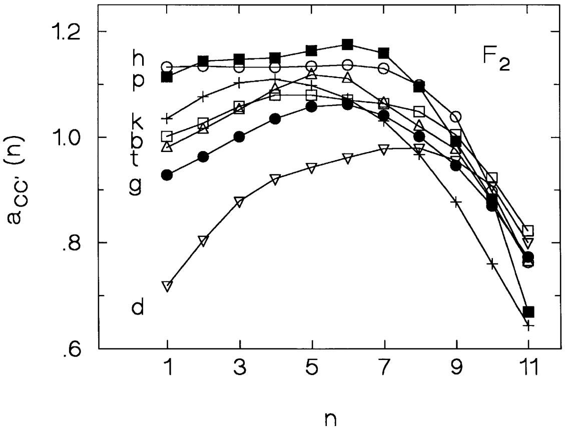

Fig. 10. The ensemble scale $a_{CC'}(n)$ for F_2 as a function of the frame number n for the seven initial consonants used in the study.

For the phonetically symmetric dVd context the ensemble scale is not itself symmetric about the center frame in either of the two plots. In both plots it is larger for the last frame than for the first. It peaks somewhat late in the syllable (at Frame 9 for F_1 and at Frame 8 for F_2) rather than at the syllable center (Frame 6). As noted above in connection with Eq. (1), this asymmetry is allowed for by the formalism of the model. Figs. 9 and 10 suggest that the systematics of VFE scaling might be a useful diagnostic for asymmetries in the effects of a given consonant in initial and final positions.

A smaller value of the ensemble scale $a_{\mathrm{CC'}}(n)$ corresponds to a more compressed VFE and therefore to a stronger contextual effect. Hence the descent to smaller values over the last few frames of all seven curves in Figs. 9 and 10 corresponds to the formant ensembles becoming more compressed as the contours converge on each other as they approach their final-/d/ loci.

The variations among the onsets $(n = 1)$ in Figs. 9 and 10 show some interesting regularities. Onset values of the ensemble scale are larger for the voiceless plosives than for their voiced counterparts for both F_1 and F_2, as shown by the orderings /p/ > /b/, /t/ > /d/ and /k/ > /g/ (the notation is shorthand for $a_{\mathrm{pd}}(1) > a_{\mathrm{bd}}(1)$, etc.). Indeed, for F_1 this ordering applies to the voiced and voiceless sets, as might be expressed by the further shorthand $\{\mathrm{p,t,k}\} > \{\mathrm{b,d,g}\}$. Each voiceless plosive therefore has a less compressed onset VFE than its voiced counterpart, probably because the aspiration interval following the voiceless plosive release would give the articulators time to accomplish much of the transition into the vowel before voicing and measurable formants begin.

In terms of the place of articulation of the initial consonant, Fig. 9 shows that the F_1 ensemble scale onsets have the orderings /h/ > /b/ > /g/ > /d/ and /h/ > /p/ > /t/ > /k/. In both series, the bilabials /b/ and /p/ have the largest onset values for the voiced and voiceless plosives, respectively, but the alveolar /d/ has the smallest onset value in the voiced series while its counterpart /t/ in the voiceless series has the second smallest.

Fig. 10 shows more consistency among the ensemble scale onsets in terms of the place of articulation of the initial consonant. The onset values of the ensemble scale in Fig. 10 exhibit the parallel orderings /b/ > /g/ > /d/ and /p/ > /k/ > /t/. These orderings parallel those for the voiced plosives in Fig. 9, but not, as noted, the voiceless series. These results for F_2 lend themselves to comparison with results from the body of literature on so-called locus equations (see, e.g., Sussman et al., 1991). In terms of the present development (see Appendix A), a locus equation is a plot of a context's onset VFE for F_2 against its center-frame VFE for F_2. Though the locus-equation approach should be equally applicable to other formants it seems to have been confined to F_2. In this literature the slope k of a locus equation is an important variable which, in terms of the present development (see Appendix A), would be the ratio of the corresponding ensemble scales:

$$k_{\mathrm{CC'}} = \frac{a_{\mathrm{CC'}}\ (\mathrm{onset})}{a_{\mathrm{CC'}}\ (\mathrm{center})}. \tag{12}$$

If we take our Frame 1 to be the onset (though the skipped Frame 0 would be better) and Frame 6 to be the center, then Eq. (12) yields the equivalent locus-equation slopes $k_{\mathrm{bd}} = 0.93$, $k_{\mathrm{dd}} = 0.75$ and $k_{\mathrm{gd}} = 0.87$ for the F_2 ensemble scales from Fig. 10. Possibly because our Frame 1 is too late to capture the true onset, our values are somewhat larger than those reported in (Sussman et al., 1991) for vowels in CVt context which for the mean of their 10 male speakers are 0.87, 0.43 and 0.66 for initial /b/, /d/ and /g/, respectively. Both sets of slopes preserve the ordering /b/ > /g/ > /d/ observed above for Fig. 10.

We can hardly leave the topic of locus equations without noting that the consistently reported linearity of locus-equation data (see, e.g., (Löfqvist, 1999) and his list of references) is a special case of the linear scaling of VFEs and so tends to corroborate our hypothesis. We should also note that the term "locus equation" itself arises from the fact that a locus equation's intersection with a 45° line from the origin approximates the initial-consonant locus (Sussman et al., 1991). To be consistent with our Eq. (1) this type of estimate would entail some special condition such as the effects from the second consonant C′ being negligible (as should be reasonable for a single-sided CV context) or the initial and final consonants having equal loci (see Appendix A).

4.2.3. The ensemble scale as a step toward a model

Besides the regularities seen in Figs. 9 and 10, the ensemble scales give us a step toward a model because through Eq. (8) they determine all the scale factors $K_{CC'}(n)$ in terms of their still unknown mean $K_{..}(\cdot)$ and, in light of Eq. (5), the translation terms $J_{CC'}(n)$ in terms of $K_{..}(\cdot)$ and the still unknown average vowel target $T_.$.

This is not to suggest that completing a model from this point forward is a trivial matter. This will depend on the structure of the database. For example, the methods developed in BC87 require some single-sided CV or VC contexts to obtain consonant loci and transition shapes. Hence the database used here would not be adequate for a straightforward application of those methods. We would either have to augment the database with more contexts or develop a more general methodology. Looking at the scaling of VFEs therefore has less demanding database requirements than modelling does. Not only can the scaling be exhibited without the need for an intermediate modelling stage, this can still be done when a modelling stage is not possible.

5. Conclusion

We began by noting that although contextual effects on the formant contour of a single vowel in CVC context will show up most obviously in its transitions, the overall pattern of a context's effects is more clearly exhibited by the family of formant contours for the context's range of vowels. Such a family may have a fairly close intervowel spacing near a consonantal boundary where the contours tend to converge toward a consonant locus and a more dilated one toward the syllable center where they approach their vowel targets.

Our purpose here was to quantify this observation both theoretically in terms of a model for formant contours and practically in terms of formant data. To proceed we needed to augment our terminology in order to describe the behavior of a body of vowels in context. In particular, we needed a "vowel-axis" analog of the time-axis concept of the contour. Given that a contour is the sequence of formant values encountered in going from frame to frame for some fixed vowel in a context, we defined a vowel-formant ensemble (VFE) to be the sequence of formant values encountered in going from vowel to vowel for some fixed frame in the context.

The above observation about a family of contours for a context would then be restated as the context's VFEs being compressed near consonantal boundaries and enlarged near the vowel center. To quantify this observation we first showed how for a given formant number the VFE for any frame in any context in BC87's coarticulation model is just a linear transformation of the model's target ensemble. Hence all the VFEs within the model are just linearly scaled copies of one another across all frames and contexts.

We then showed how this theoretical prediction could be tested directly on formant data without having to first go through an intermediate stage of implementing the model. Using this method we verified the prediction for some data on the first two formant frequencies of a single speaker's vowels in CVd context, with theory and data agreeing to nearly within the speaker's ability to reproduce any given vowel.

This verification of our prediction tends to confirm the hypotheses underlying the model, all the more so because the linear scaling relation is a true prediction from the model and not just a previously known effect in need of an a posteriori explanation. BC87's unification of the hypotheses of additivity, per-consonant similarity, and target-locus scaling into the model in Eq. (1) therefore withstands a test that is all the more significant. Although the linear scaling relation is not as strong a vindication as a full-blown fit to the model would be, it does have the advantage of not depending on the fine tuning of a number of parameters.

Once it is established for a body of data, the scaling relation leads to a form of normalization of VFEs for context. On an F_1F_2 plot the sizes of the vowel clusters are reduced and their separation enhanced. Because the normalization operates on ensembles and not on individual vowel tokens, however, it cannot be readily applied to recognition.

The *ensemble scale* for a combination of frame and context was defined as the scale of its VFE relative to that of the average VFE. It is a new tool with which one can both track the effects of any single context and compare the effects of different contexts.

In addition to its relation to a model for contextual effects and even its prediction from such a model, the linear scaling of VFEs is an interesting phenomenon in its own right which as far as we know has not been previously reported. As a phenomenon in its own right it does not require any of the machinery of the model or knowledge of any of its parameters. To exhibit it requires only the elementary operations of taking averages and fitting lines to data. This means that a database with as few as two VFEs can easily be checked for linear scaling, as can one with other parameters such as cepstral coefficients or vocal-tract area functions.

Appendix A. The "locus equation" as a corollary

A.1. Linearity of a locus equation

To construct a locus equation from the model, let the VFEs for the center and onset frames be associated with x and y, respectively. Eq. (2) gives these as

$$\begin{aligned} x &= F_{\mathrm{CVC}'}\ (\text{center}) \\ &= J_{\mathrm{CC}'}\ (\text{center}) + K_{\mathrm{CC}'}\ (\text{center}) T_{\mathrm{V}} \\ &= J_{\mathrm{c}} + K_{\mathrm{c}} T_{\mathrm{V}}, \end{aligned} \tag{A.1}$$

$$\begin{aligned} y &= F_{\mathrm{CVC}'}\ (\text{onset}) \\ &= \mathrm{J}_{\mathrm{CC}'}\ (\text{onset}) + K_{\mathrm{CC}'}\ (\text{onset}) T_{\mathrm{V}} \\ &= J_{\mathrm{o}} + K_{\mathrm{o}} T_{\mathrm{V}}, \end{aligned} \tag{A.2}$$

where the singly subscripted Js and Ks are shorthand versions for saving on notation. To get the locus equation (i.e., y as a function of x), we solve Eq. (A.1) for the target and substitute the result into Eq. (A.2),

$$y = \left[J_{\mathrm{o}} - J_{\mathrm{c}} \frac{K_{\mathrm{o}}}{K_{\mathrm{c}}} \right] + \frac{K_{\mathrm{o}}}{K_{\mathrm{c}}} x. \tag{A.3}$$

Eq. (A.3) shows that our model implies linear locus equations.

A.2. The slope as a ratio of ensemble scales

The slope of Eq. (A.3) is

$$\begin{aligned} k = \frac{K_{\mathrm{o}}}{K_{\mathrm{c}}} &= \frac{K_{\mathrm{CC}'}\ (\text{onset})/K_{\cdot\cdot}(\cdot)}{K_{\mathrm{CC}'}\ (\text{center})/K_{\cdot\cdot}(\cdot)} \\ &= \frac{a_{\mathrm{CC}'}\ (\text{onset})}{a_{\mathrm{CC}'}\ (\text{center})}, \end{aligned} \tag{A.4}$$

which is Eq. (12) in the main text.

A.3. Assumptions needed for locus estimate

As stated in the locus-equation literature (e.g., Sussman et al., 1991) the intersection of a 45° line with the locus equation gives an estimate of the initial-consonant locus. The 45° line will be

$$y = x. \tag{A.5}$$

It corresponds to a contour for which the onset and center frames share the same formant value. If we substitute Eq. (A.5) into the locus equation (A.3) and solve for x we obtain

$$x = \frac{J_{\mathrm{o}} K_{\mathrm{c}} - J_{\mathrm{c}} K_{\mathrm{o}}}{K_{\mathrm{c}} K_{\mathrm{o}}}. \tag{A.6}$$

Using the defining Eqs. (3) and (4) from the main text and Eqs. (A.1) and (A.2) above to carry out the algebraic details for Eq. (A.6) yields the following formula:

$$x = \frac{-\left(L_{\mathrm{C}} G_{\mathrm{o}} + L'_{\mathrm{C}'} G'_{\mathrm{o}}\right)\left(1 + G_{\mathrm{c}} + G'_{\mathrm{c}}\right) + \left(L_{\mathrm{C}} G_{\mathrm{c}} + L'_{\mathrm{C}'} G'_{\mathrm{c}}\right)\left(1 + G_{\mathrm{o}} + G'_{\mathrm{o}}\right)}{G_{\mathrm{c}} + G'_{\mathrm{c}} - G_{\mathrm{o}} - G'_{\mathrm{o}}}. \tag{A.7}$$

Eq. (A.7) includes the C′ locus $L'_{\mathrm{C}'}$ and C′ transition-shape function G' evaluated at both the onset and center frames (where the subscripts o and c have the same shorthand meanings as in Eqs. (A.1) and (A.2)). However, as can be verified from Eq. (A.7), one can obtain the desired solution $x = L_{\mathrm{C}}$ as an estimate of the initial-consonant locus under either of the two simplifying assumptions mentioned in the main text:

Case 1. Final-consonant effects negligible. In terms of our model, this assumes

$$G'_{\mathrm{o}} \equiv G'_{\mathrm{C}'}\ (\text{onset}) = G'_{\mathrm{C}'}\ (\text{center}) \equiv G'_{\mathrm{c}} = 0. \tag{A.8}$$

This condition might be reasonably met in a single-sided CV context. In fact, the single-sided context is what we used in BC87 for estimating loci. The locus estimate just outlined is equivalent to our earlier method for the special case of $N = 2$ frames (see Section III.B of BC87). When $N > 2$ there will be $N(N-1)/2$ distinct pairings of frames from which the locus might be estimated by the above method but with each pairing yielding a slightly different value. We handled this situation in BC87 by drawing on data from all the available frames to find a locus that minimized the rms error of the model.

Case 2. Initial and final loci the same. In terms of our model this condition assumes

$$L_C = L'_{C'}. \tag{A.9}$$

Although this might have been expected to hold for symmetric contexts such as bVb, in BC87 we found that a consonant's locus was in general not the same for initial and final positions (see Table VI of BC87). This case is therefore probably of limited practical interest because applying the locus estimate would entail imposing equality on the initial and final consonant loci without independently determining how close they really are to each other.

References

Bernard, J.R.L., 1970. Toward the acoustic specification of Australian English. Z. Phonetik Sprachwiss. Komm. Forsch. 23, 113–128.

Broad, D.J., 1976. Toward defining acoustic phonetic equivalence for vowels. Phonetica 33, 401–424.

Broad, D.J., Clermont, F., 1984. A superposition model for coarticulation in certain CVC utterances. J. Acoust. Soc. Am. 76, S14–S15.

Broad, D.J., Clermont, F., 1987. A methodology for modeling vowel formant contours in CVC context. J. Acoust. Soc. Am. 81, 155–165; also cited as BC87.

Broad, D.J., Fertig, R.H., 1970. Formant-frequency trajectories in selected CVC utterances. J. Acoust. Soc. Am. 47, 1572–1582.

Clermont, F., 1991. Formant contour models of diphthongs: a study in acoustic phonetics and computer modelling of speech. Ph.D. dissertation, Australian National University.

Clermont, F., 1993. Spectro-temporal description of diphthongs in F_1–F_2–F_3 space. Speech Communication 13, 377–390.

Delattre, P.C., Liberman, A.M., Cooper, F.S., 1952. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769–773.

Fant, C.G.M., 1960. The Acoustic Theory of Speech Production. Mouton, The Hague.

Flanagan, J., 1955. A difference limen for vowel formant frequency. J. Acoust. Soc. Am. 27, 613–617.

Houde, R.A., 1968. A study of tongue body motion during selected speech sounds. SCRL Monograph No. 2. Speech Communications Research Laboratory, Santa Barbara.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773–1781.

Löfqvist, A., 1999. Interarticulator phasing, locus equations, and degree of coarticulation. J. Acoust. Soc. Am. 106, 2022–2030.

McCandless, S.S., 1974. An algorithm for automatic formant extraction using linear prediction spectra. IEEE Trans. Acoust. Speech Signal Process. ASSP-22, 135–141.

Öhman, S.E.G., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39, 151–168.

Öhman, S.E.G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41, 310–320.

Scheffé, H., 1959. The Analysis of Variance. Wiley, New York, p. 56.

Stevens, K.N., House, A.S., 1963. Perturbation of articulations by consonantal context. J. Speech Hear. Res. 6, 111–128.

Stevens, K.N., House, A.S., Paul, A.P., 1966. Acoustical description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Am. 40, 123–132.

Sussman, H.M., McCaffrey, H.A., Matthews, S.A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Am. 90, 1309–1325.

Van Bergem, D.R., 1994. A model of coarticulatory effects on the schwa. Speech Communication 14, 143–162.

Van Son, R.J.J.H., Pols, L.C.W., 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am. 92, 121–127.

Yang, H.H., Van Vuuren, S., Sharma, S., Hermansky, H., 2000. Relevance of time-frequency features for phonetic and speaker-channel classification. Speech Communication 31, 35–50.