Journal of Phonetics 47 (2014) 47–80

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Technical Article

A method for analyzing the coarticulated CV and VC components of vowel-formant trajectories in CVC syllables

David J. Broad [a,*], Frantz Clermont [b,1]

[a] 2638 State Street, Unit 12, Santa Barbara, CA 93105-3577, USA
[b] J. P. French Associates & University of York, Forensic Speech and Acoustics Laboratory 86, The Mount, York Y024 1AR, United Kingdom

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form
18 September 2014
Accepted 25 September 2014
Available online 31 October 2014

Keywords:
Coarticulation
Context
Consonant locus
Vowel target
Linear scaling
Formant dynamics

ABSTRACT

To better understand the dynamic structure of vowels in CVC′ contexts one must account for the temporally-overlapping effects of the initial and final consonants C and C′. Here we present a linear-decomposition (LD) method for analyzing these effects as perturbations of the vowel-formant trajectories from their targets. The perturbations are modeled by the superposition of their CV and VC′ components, which are scaled by the differences between the vowel targets and their respective consonant loci. We use a dataset of second-formant frequencies (*F*2) from bVd, dVd, and gVd syllables containing seven vowels to illustrate how to estimate each element of the model by taking advantage of its additive structure and scaling properties. The model represents a family of formant trajectories unified by its scaling relationships, and the LD method that follows from it reveals how contextual effects combine and change over the duration of a vowel. Formulas for implementing the method are presented in appendices along with the two speakers' *F*2 datasets employed in this study.

1. Introduction

Articulatory movements and the formant patterns that result from them embody dynamic characteristics that challenge us to find principled ways to connect discrete phonetic units with the physical behavior of the speech mechanism. This is a methodological paper in which we demonstrate how datasets of vowel formants in CVC contexts can be analyzed to reveal how the temporally-overlapping effects from the initial and final consonants combine over the duration of the vowel. This is also a theoretical paper because the method depends on a mathematical model for describing these effects as shifts of the formant away from the vowel target, shifts that are composed of additive components from the CV and VC′ combinations. Here “additive” means that these components are superposed and combined by simple arithmetic addition.

Fig. 1 illustrates what this means for *F*2 from a high-front vowel with a target equal to 2000 Hz in a CVC′ context. (The prime (′) on the second C is to distinguish the final consonant C′ from the initial one C.) In the figure, time runs from left to right in each panel. The first panel shows the vowel target as a horizontal line. The next two panels show the CV and VC′ components of the formant's shift from the vowel target, which are plotted relative to zero. The second plot portrays the initial CV component as a decaying persisting effect, while the third plot depicts the final VC′ component as an anticipatory effect growing downward. The last panel shows how the first three add together by the layering of the CV component (horizontal hatching) under the target line and layering the VC′ component (vertical hatching) under that. The red contour (arrowed) underlying these is the resulting *F*2 trajectory.

The right panel of Fig. 1 reveals the structure of the syllable nucleus. The bottom curve for the *F*2 trajectory represents the vowel's acoustically observable trajectory, while the hatched areas and target line above it are the internal structures that shape it and locate it on the formant-frequency scale.

Abbreviations: ANOVA, analysis of variance; B84, Broad (1984); BC84, Broad and Clermont (1984); BC87, Broad and Clermont (1987); BC02, Broad and Clermont (2002); BC10, Broad and Clermont (2010); CAbS, Cepstral Analysis-by-Synthesis; df, degrees of freedom; LD, linear decomposition; LE, locus equation; LPC, linear predictive coding; RMS, root mean square; VFE, vowel-formant ensemble.

* Corresponding author. Tel.: +1 805 687 7157.
E-mail addresses: djbroad@silcom.com (D.J. Broad), akustikfonetiks@yahoo.com.au (F. Clermont).
[1] Tel.: +44 1904 634 821.

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.09.003

Nomenclature

$a_{CC'}(n)$ ensemble scale of VFE(C,C′,*n*), i.e., its scale relative to that of the mean VFE
$a_C(n)$, $a_{C'}(n)$ $a_{CC'}(n)$ for near-boundary frames treated as single-sided CV and VC′ transitions
a_C, $a_{C'}$ $a_C(n)$ and $a_{C'}(n)$ as continuous variables (not only values for frames *n*)
a_T target scale, i.e., the scale of VFE(C,C′,*n*) relative to the target ensemble
β_C, $\beta'_{C'}$ reciprocal exponential time constants for consonants C and C′
C, V, C′ identifiers for specific initial consonants C, vowels V, and final consonants C′
[d] as a superscript on *F*: designates a formant value from the data
D duration of a vowel
F-statistic statistic for testing the significance of effects in an ANOVA (in context this should not be confused with a formant.)
*F*1, *F*2, *F*3 frequencies of the first, second, and third formants
$F_{CVC'}(n)$ formant trajectory of vowel V in the context of consonants C and C′ at frame *n*
$F_{C\bullet C'}(n)$ $F_{CVC'}(n)$ averaged over the vowels V
• indicates averaging over a variable, as in $F_{C\bullet C'}(n)$ above
$F_{C\bullet}(n)$ $F_{C\bullet C'}(n)$ for near-CV boundary frames treated as a single-sided CV transition
$F_{\bullet C'}(n)$ $F_{C\bullet C'}(n)$ for near-VC′ boundary frames treated as a single-sided VC′ transition
$F_{C\bullet}$, $F_{\bullet C'}$ $F_{C\bullet}(n)$ and $F_{\bullet C'}(n)$ as continuous variables (not only values for frames *n*)
$F_{\bullet V\bullet}(\bullet)$ $F_{CVC'}(n)$ averaged over consonants C and C′ and frames *n*
$F...(\bullet)$ $F_{CVC'}(n)$ averaged over consonants C and C′, vowels V, and frames *n*
$G_C(n)$, $G'_{C'}(n)$ transition-shape functions for initial consonant C and final consonant C′
$J_{CC'}(n)$ additive displacement of VFE(C,C′,*n*) in formant frequency
k expected ratio of RMS error of an exponential model to that of an unmodified linear-decomposition model
$K_{CC'}(n)$ scale of VFE(C,C′,*n*) relative to the target ensemble
$K_C(n)$, $K'_{C'}(n)$ $K_{CC'}(n)$ for peripheral frames to approximate CV and VC′ transitions
κ_C, $\kappa'_{C'}$ exponential scale factors for consonants C and C′
L_C, $L'_{C'}$ loci of consonants C and C′
N number of vowel frames in a CVC′ syllable nucleus
$N_{\mathbf{C}}$, $N_{\mathbf{V}}$, $N_{\mathbf{C}'}$ number of initial consonants, vowels, and final consonants in a dataset
N_{data} number of data points in a dataset
N_P number of peripheral frames used to estimate consonant loci and vowel targets
$N_{par,exp}$ number of parameters in an exponential model
$N_{par,LD}$ number of parameters in a linear-decomposition model
n discrete time frame with integer values
$S^{[\mathbf{x}]}$, $SS^{[\mathbf{y}]}$ shorthand notation for summations of type **x** and type **y**
s_{est} expected error for an exponential model
$s_{expon.model}$ RMS error for an exponential model
$s_{LD.model}$ RMS error for an unmodified linear-decomposition model
t physical time variable measured, e.g., in milliseconds
T_V, $T_\bullet$ target of vowel V and the inter-vowel average of T_V
$(T_V - L_C)G_C(n)$ CV component of a formant trajectory
$(T_V - L'_{C'})G'_{C'}(n)$ VC′ component of a formant trajectory
VFE(C,C′,*n*) vowel-formant ensemble (VFE) for the context CVC′ and time frame *n*

The analysis of a CVC′ dataset to produce a diagram such as the rightmost plot in the figure requires a number of steps involving the estimation of the consonant loci and vowel targets, followed by the determination of the initial-CV and final-VC′ components of the formant's shifts away from the vowel target. To make the method available for use, its steps must be described in some detail. The utility of these steps may not be obvious at first, but some idea of the outcome can be seen by turning to Figs. 11(b) and 21(a) below to see diagrams like Fig. 1 that are derived from data.

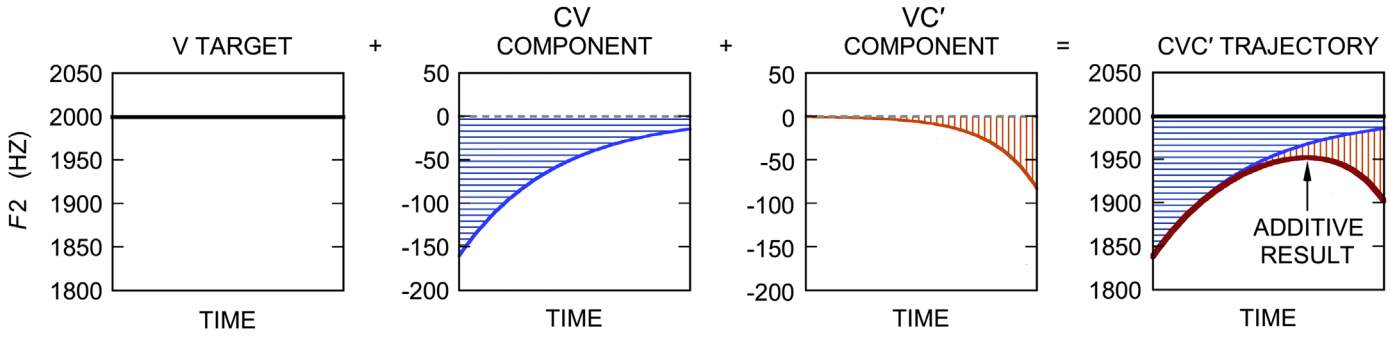

Fig. 1. Illustration of how the additive CV and VC′ components combine to shift *F*2 from its vowel target to form a CVC′ trajectory. A 2000 Hz target is represented by the horizontal line in the first plot. The second plot is for the CV component of the formant trajectory's shift from the target, shown with blue hatching. The third plot is for the VC′ component of the shift, shown with orange hatching. The rightmost plot shows how the first three plots combine to produce the CVC′ *F*2 trajectory. Addition of the components is represented by the stacking of the second plot under the target line and of the third plot under that. The red curve (arrowed) below the others is the *F*2 trajectory.

The model schematized in Fig. 1 draws on many decades of efforts by phoneticians. As cited by Heffner (1960), Menzerath and de Lacerda (1933) introduced the term *coarticulation* to refer to parts of two articulations being produced simultaneously. For example, an articulator not needed for a current sound can be released to anticipate upcoming ones. Öhman (1967) incorporated this concept into his vocal-tract model for coarticulation in VCV utterances with his *coarticulation function*, which characterized each consonant's constraints on the vocal-tract shape. Constraints on articulatory dynamics also imply that articulatory actions can overlap in time (Lindblom, 1963a, 1963b) through superposition (as also suggested by Stevens, House, & Paul, 1966).

For vowels in CVC context, Lindblom (1963a, 1963b) obtained evidence for this idea in the form of significant contextual effects at the vowel center, manifested as stronger vowel reduction for shorter vowel durations. His inference was that a CVC syllable will be spoken faster not by making the articulators move faster but by means of articulatory actions with fixed time courses that are triggered by commands issued at a faster rate. A new command issued while previous actions are still in progress will initiate a new action that will combine with these instead of replacing them. Hence a command to move toward the final consonant issued before the vowel is fully achieved will begin to steer the articulators away from the vowel and so reduce its realization.

In another study that made significant strides toward a quantitative characterization of coarticulation and articulatory dynamics, Houde (1968) analyzed extensive measurements of cineradiographic records of tongue-body movements in $V_1CV_2CV_1$ utterances taken at 100 frames per second. He discovered that the vowel-to-vowel component of these movements had similarly-shaped time courses for all the vowel pairs transitioning with a given consonant. This observation led him to an economic numerical characterization of his data by linear scaling of these consonant-specific trajectory shapes, or *transition functions*, a term we adapt below for our scalable consonant-specific time functions.

In Section 2 of this paper we present the key developments that set the stage for our method, which we call *linear-decomposition* (LD). We first introduce some preliminary concepts (e.g., vowel-formant ensemble (VFE), ensemble scale, and VFE similarity within and between contexts), and then describe the model on which the method is based. Consonant loci and vowel targets parallel earlier counterparts, but are redefined operationally in terms of how they are estimated. Target-locus differences then become scale factors that determine the ranges of the consonantal components of the shifts away from the vowel target illustrated in Fig. 1. These scale factors are applied to *transition-shape* time functions associated with each initial or final consonant. We also describe the dataset of *F*2 measurements from two speakers' productions of CVd syllables, which we use to illustrate the steps in the method.

Section 3 develops the linear-decomposition method for single CVC′ contexts. In a single-context dataset, the vowels are all preceded by the same initial consonant and followed by the same final consonant. Each step is illustrated with a geometric interpretation of the method's formalism.

Section 4 extends the single-context method to joint analyses of multiple contexts. In a dataset containing multiple contexts, each vowel occurs in the context of more than one initial consonant and/or more than one final one. In a joint analysis, such a dataset is analyzed as a whole − not one context at a time as in the single-context version of the method. In Section 5 the transition-shape time functions are fit to exponential functions with their asymptotes identified with the vowel targets as proposed by Lindblom (1963a, 1963b). These exponential transition-shape functions greatly ameliorate inconsistencies between a context's outcomes from its single-context analysis and its participation in a joint analysis in the company of other contexts.

Section 6 discusses various aspects of the model and LD method, including the requirements for a dataset to be amenable to linear-decomposition analysis, the comparative advantages of the single- and multiple-context analyses, a critique of error-reduction methods, an outline of connections between linear-decomposition and locus equations, and an observation on the desirability and difficulty of taking vowel duration into account. Section 7 summarizes the paper.

To maintain an uninterrupted flow in the logical development, the body of the paper is limited to the mathematical relations that contribute conceptually to the ideas and methods. Mathematics directed only toward technical details or the practicalities of computation is relegated to Appendices A–C. Appendix D lists the *F*2 data for our two speakers' bVd, dVd, and gVd contexts. These data will enable readers to check their implementations against results in the main text.

2. Background, preliminary concepts, and vowel formant dataset

In this section we explain our conceptual framework for characterizing dynamic effects of consonantal contexts on vowel formants. The model on which our method is based is the result of a series of developments (Broad & Fertig, 1970; Broad, 1984; Broad & Clermont, 1984, 1987, 2002, 2010; cited below as BF70, B84, BC84, BC87, BC02, and BC10, respectively), in which we progressively incorporated the additivity of consonantal effects, the concept of consonant-specific transition shapes, and the scaling of these shapes by differences between vowel targets and consonant loci. These properties are recalled and discussed in Sections 2.1 through 2.5 with a view towards explaining them and adapting them for use in the linear-decomposition method. Section 2.6 describes our vowel formant dataset.

2.1. Additivity of effects from initial and final consonants

The additivity of consonantal effects shown in Fig. 1 was shown to be a real effect by Broad and Fertig (1970) who reported on spectrographic measurements of one male speaker's *F*1, *F*2, and *F*3 of the American English vowel /ɪ/ in CɪC′ contexts with all combinations of 24 initial Cs and 24 final C′s in a corpus of 576 syllables, each repeated three times, with formants measured at 11 equally-spaced time frames through the vowel. Fig. 2 shows the *F*-statistics obtained for the *F*2 data from two-way analyses of

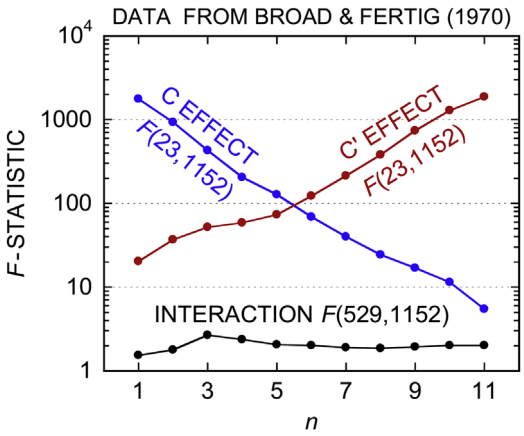

Fig. 2. *F*-statistics from analyses of variance (ANOVAs) for *F*2 obtained separately at 11 time frames for the American English vowel /ɪ/ produced by one male speaker, each repeated three times, in 576 CɪC′ contexts: 24 initial contexts in all combinations with 24 final ones. The *F*-statistics are plotted against the time frame *n*, with a separate ANOVA for each of the 11 time frames. The *F*-statistics for the main C and C′ effects are *F*(23,1152) and those for the interaction are *F*(529,1152).

variance (ANOVA) done at each frame. The *F*(23,1152) for the main effects from both the initial and final consonants is highly significant ($p<0.005$) through the duration of the vowel. The initial-consonant effects decrease with time and the final-consonant effects increase to form an X pattern on the plot. The interaction *F*(529,1152) at the bottom of the plot is much smaller in magnitude, but also highly significant ($p<0.005$). The large values of the main-effect *F*(23,1152) arise from the sensitivity of the *F*-statistic to even modest excesses in the differences among the category means. Similar X patterns have been reported by Tabata and Sakai (1973, 1976) for Japanese VCV utterances.

BF70's results demonstrate that context effects are effectively additive over the duration of the vowel. Focusing on the single vowel /ɪ/ in various contexts permitted BF70's two-way ANOVAs to bring out the dominance of the main effects for the initial and final consonants. As illustrated in Fig. 2, this additivity of consonantal effects is a significant step in our understanding of how vowels are coarticulated with their surrounding consonants. One could exploit this by repeating the analysis for other vowels, but this would yield only a number of different vowel-specific additive models with no clear way of relating them one to another.

A more comprehensive model that covers a variety of vowels in a variety of CVC′ contexts needs to incorporate some other properties along with additivity. The model presented below will have as its goal the accurate representation of the formant trajectories in CVC′ datasets. As will be seen, the accuracy of these representations compares favorably with a speaker's ability to repeat his or her own utterances and with a listener's ability to perceive differences in formant frequencies. As a consequence of its accuracy, the model's properties can reasonably be interpreted as properties of the data, i.e., the model can reasonably be said to explain the simultaneous C and C′ effects on the production of vowels in CVC′ contexts.

2.2. *Per-consonant similarity and target-locus scaling*

One desirable property of a descriptive model would be for the time courses of its additive components to display some kind of regularity. The simplest possibility would be for all the components from either an initial or a final consonant to share some universal shape for these time courses. However, in B84 it had been found that the additive components for the initial consonants in the BF70 model differed in shape from consonant to consonant and so this idea of a universal shape had to be rejected. In its place, B84 suggested the next-simplest hypothesis of *per-consonant similarity* in which the additive components from any given initial or final consonant would have similarly-shaped time courses for their transitions with the different vowels. Per-consonant similarity was verified on some small datasets in BC84 and BC87 and so was incorporated into their model. With per-consonant similarity, a consonant's additive component of the formant trajectory for any given vowel would be a scaled copy of its component shape or *transition-shape function*, as defined below.

It remained to be seen whether there was a simple way to scale these consonant-specific shapes without having to use an idiosyncratic scale factor for each consonant–vowel combination. The phonetically-motivated hypothesis adopted was that the scales should be directly proportional to the differences between vowel targets and consonant loci. This idea of *target-locus scaling* was also supported in the BC84 and BC87 studies and incorporated into their model.

These studies also developed some methods for estimating the consonant loci and vowel targets. In BC10 we took advantage of the simpler structure of single-sided VC′ syllables to refine these methods by concentrating on the loci and targets without having to deal with additivity at the same time. Now that they have been developed, we bring them back to deal with an additive model for vowels in CVC′ contexts. For this, we use BC02's simple reformulation of the BC84/BC87 models.

2.3. The additive model

We retain the additive structure of the BC87 model, but here we proceed with BC02's streamlined form of its elements:

$$F_{CVC'}(n) = T_V + (T_V - L_C)G_C(n) + (T_V - L'_{C'})G'_{C'}(n), \quad 1 \le n \le N \tag{1}$$

Here $F_{CVC'}(n)$ is the model's formant at time frame n for a syllable with initial consonant C, vowel V, and final consonant C′. It is also referred to as the *formant trajectory*. Each element on the right side of the equation is tied to one of the syllable's phonetic units: each vowel V has its own target T_V, each initial consonant C has its own locus L_C and *transition-shape function* $G_C(n)$, and each final consonant C′ has its own locus $L'_{C'}$ and transition-shape function $G'_{C'}(n)$.

The term $(T_V - L_C)G_C(n)$ is then the CV component of the trajectory's shift away from the vowel target, while the term $(T_V - L'_{C'})G'_{C'}(n)$ is its corresponding VC′ component.

Eq. (1) formalizes the idea illustrated in Fig. 1, where the first panel corresponds to the vowel target which is the first term on the right side of Eq. (1), the second panel of the figure corresponds to the CV component which is the second term of the equation, and the third panel for the VC′ component corresponds to the equation's third term. The superposition of these three terms in the fourth panel of Fig. 1 corresponds to the formant trajectory, which is the left side of Eq. (1).

In Eq. (1), the non-numerical subscripts C, V, and C′ serve as tags for the consonants and vowels. Time is expressed in terms of the discrete frame counter n running from 1 to the number N of frames, which is the same for all the syllables in a dataset.

Eq. (1) provides one way to relate the discrete phonetic units C, V, and C′ to the dynamically changing formant trajectory within a CVC′ syllable nucleus by describing the contribution from each unit as it is distributed over the duration of the vowel.

Because each of the model's elements is to be derived from data, it is primarily a descriptive model with no preconceived constraints on numerical values for its constants or on the forms of its transition-shape functions. Yet it is still a predictive model to the extent that it predicts that a CVC′ dataset can reasonably be described in terms of its additive structure, the per-consonant similarity of the time courses of its contextual transition-shape functions, and the linear scaling of these functions by target-locus differences.

2.4. Formants as linear functions of the vowel targets

In working toward the present method, in BC02 it was observed that Eq. (1) implies that the vowel formant is a linear function of the vowel target, independent of the details of context or temporal placement in the syllable. To see this, we follow BC02 in rewriting Eq. (1) by separating quantities that are factors of T_V from those that are not:

$$F_{CVC'}(n) = K_{CC'}(n)T_V + J_{CC'}(n), \quad 1 \le n \le N \tag{2}$$

where

$$K_{CC'}(n) = 1 + G_C(n) + G'_{C'}(n), \quad 1 \le n \le N \tag{3}$$

$$J_{CC'}(n) = -L_C G_C(n) - L'_{C'} G'_{C'}(n), \quad 1 \le n \le N \tag{4}$$

When C, C′, and n are fixed, Eq. (2) shows the formant as a linear function of the vowel target with its slope equal to $K_{CC'}(n)$.

2.4.1. A shift in perspective: vowel-formant ensembles (VFEs) and their similarity

In fixing a context and time frame and varying only the vowel, we are in effect looking at a “vowel axis” (BC02, BC10). This observation led to our concept of the *vowel-formant ensemble* or VFE by which we mean the arrangement of formants for the different vowels at a fixed time frame within a fixed context.

The idea of the VFE is illustrated in Fig. 3, which shows *F*2 trajectories synthesized from Eq. (1) with idealized transition-shape functions. In the figure, *F*2 for five vowels preceded by consonant C and followed by consonant C′ are plotted against the time-frame counter n. The vertical box around the formants at $n = 10$ marks the VFE for this frame in this context. Each time frame in each

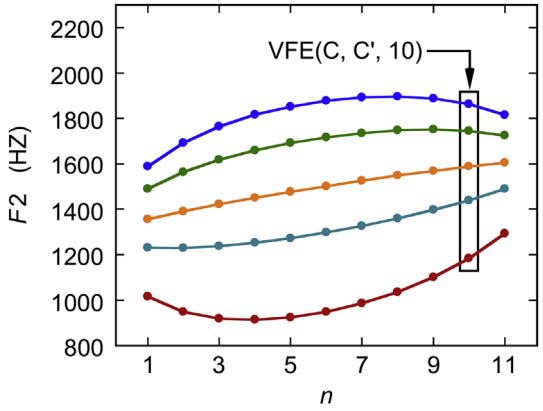

Fig. 3. The concept of the vowel-formant ensemble (VFE). Stylized *F*2 trajectories for five vowels in the context of a preceding consonant C and a following consonant C′ are plotted against n, a frame counter running from 1 to 11. The box around the formants at frame $n = 10$ marks the VFE for this context and this time frame and is denoted VFE(C,C′,10).

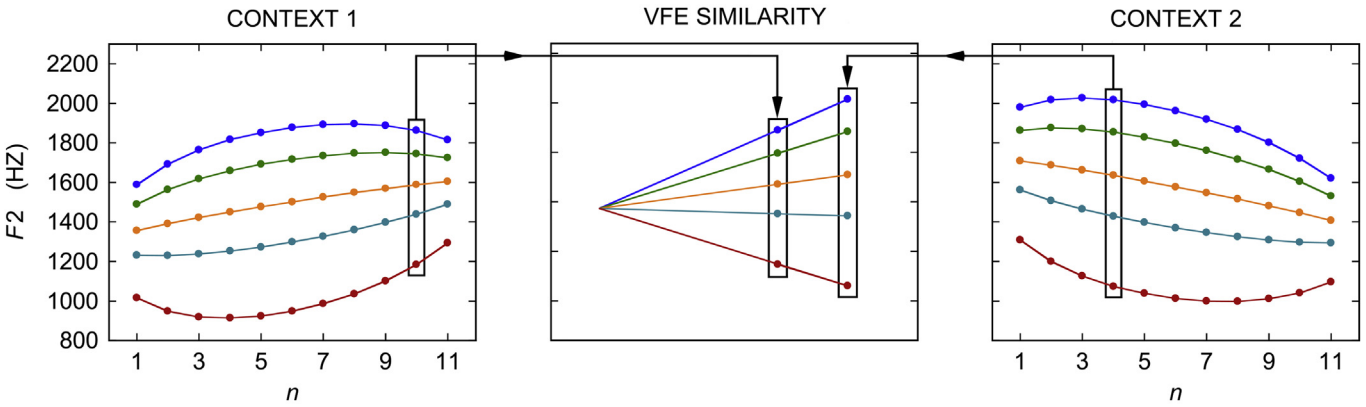

Fig. 4. Geometric similarity of vowel-formant ensembles (VFEs). In the left and right plots, *F*2 is plotted against the time frame *n* for the same 5 vowels from Context 1 in the left plot and Context 2 in the right one. As indicated by the arrowed lines, the VFEs from frame 10 of Context 1 and frame 4 of Context 2 are copied into the center plot. The two VFEs can thus be compared by drawing a line between each vowel's two *F*2s and extending it leftward to intersect the lines from the other vowels. Because the two VFEs form the sides of two sets of triangles with the same three angles, the VFEs must be geometrically similar. Hence they must have the same proportionate spacing of *F*2 among the 5 vowels.

context will have its own VFE, which for context CVC′ and frame *n* will be denoted with the shorthand notation VFE(C,C′,*n*). Hence the VFE marked in the figure is VFE(C,C′,10).

VFEs provide an interesting corollary to the fact that Eq. (1)'s formants are linear functions of the vowel target, namely that all the VFEs characterized by Eq. (1) for a given set of vowels will be geometrically similar to one another. This is illustrated in Fig. 4, which shows *F*2 for the vowels from Fig. 3 in two contexts with differing initial and final consonants. *F*2 for the first context in the left panel replicates the data from Fig. 3 and the *F*2 trajectories for the other context are shown in the right panel. The VFE for time frame 10 for the first context is again enclosed in a box, as is the VFE for time frame 4 of the other context. These VFEs are copied into the center plot as shown by the arrowed lines. For each vowel in this center plot, a line is drawn through its two *F*2 points and extended leftward to an intersection shared by the lines from all the vowels. The two VFEs have exactly the same angles with the sheaf of lines radiating from the left and so are sides of a set of similar triangles. The two VFEs are therefore geometrically similar and the relative spacing of their vowel formants is maintained.

Eq. (2) characterizes VFE similarity in more formal terms. In geometry, shapes are similar if they retain the same proportions when they are uniformly stretched or squeezed or if they are moved around. For our purposes, the uniform stretching or squeezing is embodied in Eq. (2) by its slope $K_{CC'}(n)$, which is the scale factor of VFE(C,C′,*n*) relative to the VFE with the vowel targets as its elements. (We call this VFE the *target ensemble*.) In the same way, the displacement of a VFE up or down the formant-frequency scale is embodied by $J_{CC'}(n)$, which is the additive term in Eq. (2). It therefore characterizes the placement of VFE(C,C′,*n*) on the formant-frequency scale.

In BC02 we used Eqs. (1) and (2) to predict VFE similarity in a dataset of a speaker's *F*1 and *F*2 across several CVd contexts. This prediction was borne out with some precision and was relatively easy to obtain because it depended only on the linear structure of Eq. (1) and did not require the estimation of any of its elements, such as consonant loci or vowel targets.

The linear relation in Eq. (2) implies (BC02, BC10) that context and temporal location in a syllable can affect a vowel's formant entirely through their effects on the VFE to which it belongs. Such effects are limited to some combination of translation of the VFE as a whole up or down in formant frequency (by the term $J_{CC'}(n)$) and linear scaling as either a linear compression or linear expansion (by the factor $K_{CC'}(n)$).

Note that VFE similarity is implicit in the linear relations between vowel formants from onset and center frames in CVC contexts that are consistently reported in the locus-equation literature (e.g., Krull, 1987, 1989; Sussman, McCaffrey, & Matthews, 1991; Sussman, Fruchter, Hilbert, & Sirosh, 1998; Lindblom & Sussman, 2012). The study by Nearey and Shammass (1987) reports another early example of linear relations for both *F*2 and *F*3 of Canadian English vowels in CVd contexts, which are clearly interpretable in terms of VFE similarity.

As described in Sections 2.5.1 and 2.5.2, the concept of VFE similarity led to BC10's method for finding consonant loci and vowel targets in single-sided CV or VC′ syllables, a method that has opened the way to the linear-decomposition method presented in Sections 3 and 4 for analyzing CVC syllables.

2.4.2. *Ensemble scales*

The *ensemble scale* of VFE(C,C′,*n*) is its scale relative to that of the *mean VFE*, the elements of which are the vowel-by-vowel means of the formant taken over all C, C′, and *n* in a dataset. For the vowel V, this per-vowel mean is denoted $F_{\cdot V \cdot}(\bullet)$, which is the average of $F_{CVC'}(n)$ over all the C, C′, and values of *n*. (The dot "•" in a variable position is Scheffé's (1959) economic notation for averaging over that variable.) Within the model, the mean VFE will be similar to all its other VFEs. A first link between the model and data stems from the calculation of the mean VFE from the data. The ensemble scale $a_{CC'}(n)$ for frame *n* from the context CVC′ is determined as the slope of the linear relation between $F_{CVC'}(n)$ and its per-vowel average $F_{\cdot V \cdot}(\bullet)$.

This is illustrated in Fig. 5 for VFE(C,C′,10) from the frame-10 formants marked in Fig. 3. The ensemble scale therefore serves as a data-referenced scale of a VFE. Because each ensemble scale is referenced to the same standard (the mean VFE), a dataset's

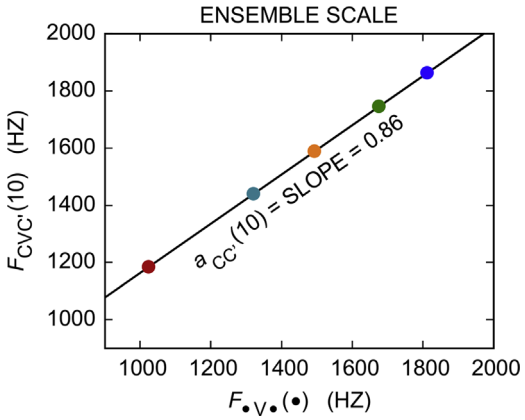

Fig. 5. Estimation of the ensemble scale $a_{CC'}(10)$ for VFE(C,C′,10) as the slope of the linear relation between $F_{CVC'}(10)$ and $F_{\cdot V \cdot}(\bullet)$ for the frame-10 formants from Fig. 3.

ensemble scales will all be directly comparable to one another. Hence they provide a first step toward a unifying framework for a dataset.

Note that because Fig. 3 represents a single-context dataset, the per-vowel average over "all" its C and C′ will be $F_{\cdot V \cdot}(\bullet) = F_{CVC'}(\bullet)$. However, for consistency with the definition, the more general form $F_{\cdot V \cdot}(\bullet)$ is used for the *x*-axis in Fig. 5. The notation is still valid because "all" the C and C′ are the *only* C and C′ in this single-context dataset.

$K_{CC'}(n)$ from Eq. (2) is the slope of the linear relation between a VFE's formants and the corresponding targets from the target ensemble. Hence, like the ensemble scale, $K_{CC'}(n)$ is also a measure of a VFE's scale − but a model-referenced one that uses the target ensemble as its standard. Within the model, the target and mean ensembles are similar to each other and, because these are the standards to which $a_{CC'}(n)$ and $K_{CC'}(n)$ are referenced, this means that these two measures of VFE scales differ only by a scale factor. For this factor, we define the *target scale* a_T to be the ensemble scale of the target ensemble, i.e., the scale of the target ensemble relative to the mean ensemble (BC02). $K_{CC'}(n)$ is therefore related to $a_{CC'}(n)$ as

$$K_{CC'}(n) = a_{CC'}(n)/a_T \tag{5}$$

Eq. (5) is a first step in relating an element of the model, $K_{CC'}(n)$, to an element derived from the data, $a_{CC'}(n)$.

By now we have accumulated a number of different types of linear scaling, each with its own meaning: (a) target-locus scaling in which the scale factor is the a target-locus difference, e.g., $T_V - L_C$, (b) the scaling implied by Eq. (2) for the linear relation between a VFE and the target ensemble with $K_{CC'}(n)$ as its scale factor, (c) the ensemble scale $a_{CC'}(n)$, which is the scale factor for a VFE relative to the mean VFE, and (d) the target scale a_T which is the target ensemble's scale relative to that of the mean VFE. For understanding the model and the linear-decomposition method it is important to keep the meanings of these different instances of linear scaling in mind.

2.5. Consonant loci and vowel targets: How our usages relate to earlier concepts

The scaling of context effects by target-locus differences in Eq. (1) carries with it a view of loci and targets that parallels some earlier ideas but also departs from them in some details. The main property on which our concepts agree with earlier ones is that the loci and targets are properties of the consonants and vowels that are not necessarily actually observed values. This idea stems from Potter, Kopp, and Green's (1947) idea of the *hub*, which was a "zone of influence" for a consonant or vowel which could be either "visible or hidden" (Kopp & Green, 1946, p. 81). Since then the term "locus" has become the recognized term for a consonant-specific value just as "target" has become one for vowels.

2.5.1. Consonant loci

In BC87 we developed the idea of the locus as an axis of symmetry for the pattern of formant trajectories in a family of vowels transitioning with a given consonant. The locus had originally been defined as "a place on the frequency scale at which a transition begins or to which it may be assumed to point" (Delattre, Liberman, & Cooper, 1955, footnote 3, p. 769). Our concept and theirs are both embraced by the diagram in Fig. 6, which shows a stylized *F*2 for a family of vowel formants transitioning from a hypothetical consonant. The solid lines represent the observable formant and the dashed lines show their continuation back to a consonantal locus, where the trajectories all intersect. For both concepts the locus becomes obvious from the multi-vowel pattern: the extrapolation to an intersection, and the horizontal axis around which the trajectories are distributed. In BC87 and BC10 we showed that there is a straightforward way to find the optimum location for this stationary axis.

As can be seen from Fig. 6, the two locus concepts are consistent with one another and differ only in which aspect of the diagram to take as a defining characteristic: the backward extrapolation of the trajectories to a common intersection or the axis around which they are distributed. Delattre et al. (1955, p. 771) use a similar diagram to illustrate the stimuli they used for one of their perception experiments, though they stated no observations on its symmetry.

One problem in using the "pointing" concept to estimate a locus is that extrapolation of a formant track can be noisy. Another is that with only one vowel, one would not know how far to extend the extrapolation. Schouten and Pols (1979a, 1979b, 1981)

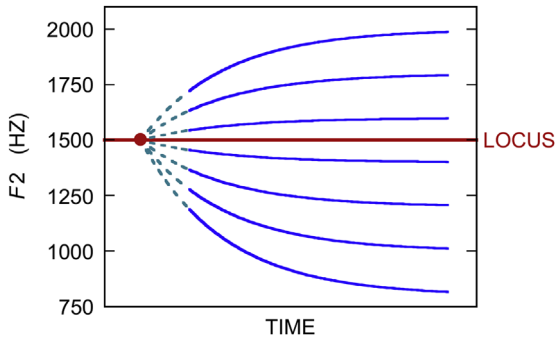

Fig. 6. Stylized *F2* transitions from a consonant into 7 vowels. Observable formants are shown as solid blue lines, unobservable extrapolations in dashed aqua lines. The consonant locus is represented both by the red horizontal line, which is the axis around which the transitions are distributed, and by the red dot, which is the point where the extrapolations intersect this axis.

addressed both these problems by using four vowels in an analog of the *F*1–*F*2 plane. For each vowel, a small range of possible extrapolations was created which yielded a region of overlap for the four vowels, a region they called a "locus area." Their use of multiple vowels for locus estimation anticipated our multi-vowel method for estimating the locus as a horizontal axis.

Sussman et al. (1991) argued that a consonant locus will correspond to the point on a locus-equation line where the onset and center formants are equal, i.e., a point where a hypothetical vowel would have a constant trajectory. Such a trajectory would correspond to the horizontal red axis in Fig. 6, so the locus concept advanced by Sussman, et al. is consistent with ours. In Section 6.4.1 we illustrate how a consonant locus can function as an organizing element for a set of two-frame transitions from a locus-equation dataset.

2.5.2. *Vowel targets*

Lindblom (1963a, 1963b) found an elegant way to define vowel targets as formant values that are the asymptotes of exponential curves, most notably in his fitting of exponentials to the consonant-specific relations he observed between vowel reduction and vowel duration. He also succeeded in fitting exponential time functions to some formant transitions (Lindblom, 1963a). It would be hard to imagine a more natural explanation for his exponential vowel-reduction effect than exponentially-shaped transitions with consonant-specific time constants.

In BC87 we also looked at exponentials to represent transition shapes with their asymptotes identified with the vowel targets. In the same paper (BC87, Eq. (29)), we developed another approach to the vowel target which did not depend on transitions having any preconceived functional forms, but instead depended only on the scaling properties of a model such as Eq. (1) simplified to represent only single-sided CV and VC′ syllables (see Appendix A):

$$F_{CV}(n) = L_C + (T_V - L_C)K_C(n) \tag{6}$$

In BC10 we elaborated on the idea embodied in Eq. (1) that loci and targets anchor scaling relationships for formant transitions. This idea is the basis of the method described below for estimating the loci and targets for Eq. (1): indeed, it is remarkable that their function in Eq. (1) also leads to practical methods for estimating their values. Turned the other way around, the loci and targets are more than the outcomes of certain numerical exercises: they help to characterize observed formant trajectories as members of a coherent family linked by these scaling relationships.

Implicit in Eq. (1) is its treatment of the vowel target T_V as a property inherent to the vowel, which remains fixed across contexts. In BC10 we referred to this characteristic of the target as "portability" from one context to another. This a key feature of the linear-decomposition method, and in the model it connects the different contexts through their shared vowel targets.

2.6. *Vowel formant dataset*

Before turning our attention from properties of the Eq. (1) model to the linear-decomposition method and its application to real data, we next introduce our vowel material and the acoustic analyses we performed to obtain our formant measurements.

The data are from two adult male speakers of Australian English who produced CVd syllables with C = /b, d, g/ in all combinations with the non-diphthongized vowels /ɪ, ɛ, æ, a, ɒ, ʌ, ɜ/. Key words for these vowels are "bid", "bed", "bad", "bod", "bard", "bud", and "bird". Five readings of the syllables from each speaker were digitally recorded at one sitting with 16-bit samples at a rate of 10,000 samples per second. For statistical stability, the data used here are the averages of these measurements over the five repetitions.

Vowels were segmented by visual inspection of the acoustic signal and the segmentation points were used to automatically locate 13 equally-spaced 256-sample frames. The first sample of the first frame coincides with the initial segmentation point and the last sample of the 13th frame coincides with the final segmentation point. To avoid unreliable measurements at the vowel boundaries, the first and 13th frames are ignored, leaving the internal 11 frames available for analysis.

Initial estimates of the first four formants were obtained from 14th-order LPC-pole analyses (Markel & Grey, 1976) from which formant peaks were tracked using a Cepstral Analysis-by-Synthesis (CAbS) method. The CAbS tracker (Clermont, 1991, 1992) exploits the index-weighted cepstral distance (Yegnanarayana & Reddy, 1979) for its sensitivity to spectral slope, and incorporates a dynamic-programming procedure that takes inter-frame spectral continuity into account. The CAbS tracker has since been improved

using vowels produced in a wide range of phonetic contexts and in different dialects of English (Clermont & Mokhtari, 1998; Clermont, Harrison, & French, 2007; Clermont, French, Harrison, & Simpson, 2008). This work was briefly described in BC10, and the formant-patterns tracked for the present study are tabulated in Appendix D.

3. The linear-decomposition method I: Single-context case

The way is now clear for bringing together the properties outlined above to use in our linear-decomposition method. In this section we will consider only the single contexts of our CVC′ dataset, where each vowel is preceded by the same initial consonant and followed by the same final one. Fig. 7 shows the method's steps for this case. These steps will be described using Speaker 1's bVd context for illustration. His ensemble scales $a_{bd}(n)$ and inter-vowel mean trajectory $F_{b\bullet d}(n)$ are described in Section 3.1. Section 3.2 shows how the peripheral frames from these two time functions are used to estimate the model's loci and targets. Section 3.3 describes how these values are used together with the formant data to obtain the transition-shape functions $G_b(n)$ and $G'_d(n)$. At this point all the elements of Eq. (1) will have been determined for Speaker 1's bVd dataset. Section 3.4 shows how these elements combine to model the dataset's formant trajectories and enable us to visualize their structures. Section 3.5 describes the linear-decomposition results for Speaker 1's dVd and gVd contexts, while Section 3.6 describes the parallel analyses of Speaker 2's bVd, dVd, and gVd contexts.

3.1. Ensemble scales and mean formant trajectories: The initial steps

Fig. 8(a) shows Speaker 1's *F*2 data for the seven vowels from his bVd context. Fig. 8(b) shows the $a_{bd}(n)$ time-function of ensemble scales that results from applying the method illustrated in Fig. 5 to all the frames of Fig. 8(a). From its maximum at frame 4, the function drops to smaller values as it approaches either of the vowel boundaries. This behavior tracks the corresponding changes in the inter-vowel spacing of *F*2, which can be seen in Fig. 8(a).

The translation effect of context and frame placement on a VFE is characterized from the data by the inter-vowel mean $F_{C\bullet C'}(n)$ of the formant for context CVC′. Fig. 8(c) shows the time function $F_{b\bullet d}(n)$ for the formants from Fig. 8(a). It closely resembles the trajectory in Fig. 8(a) for the vowel /ɜ/ and its steady upward trend follows the overall pattern that is clear in that figure. Two quartets of peripheral frames are enclosed in boxes in each of the three parts of Fig. 8. These frames will be used in the estimation of the consonant loci, the target scale, and the mean vowel target.

Even if implementing all the steps in the linear-decomposition method seems to be overly burdensome, it may be worthwhile to carry the method through this stage of calculating the ensemble-scale and inter-vowel-mean-formant time functions. These can already yield interesting information about the dynamics of vowels in different contexts.

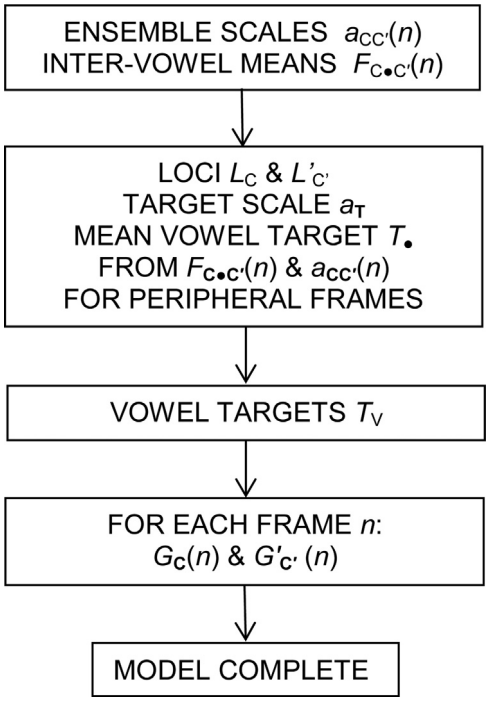

Fig. 7. Flow diagram showing the steps in the linear-decomposition analysis for a single context.

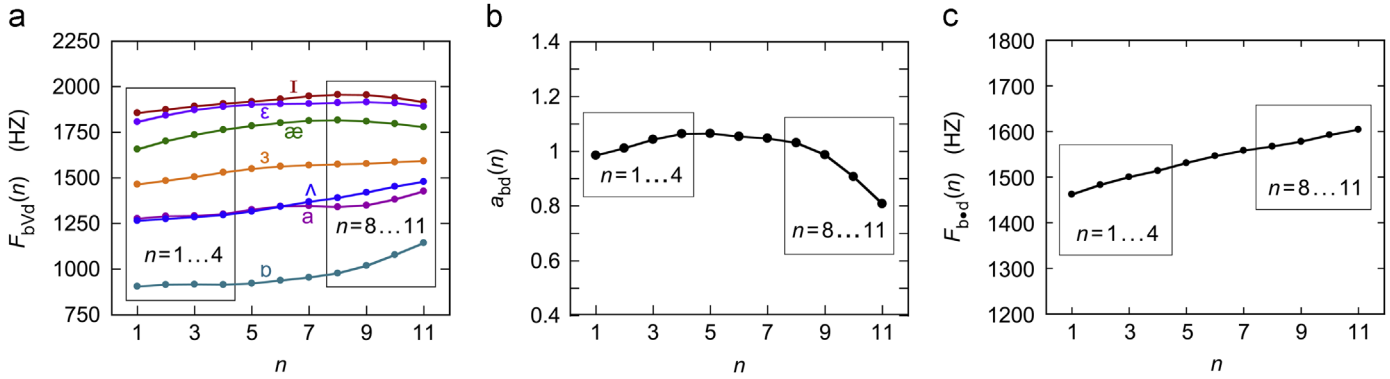

Fig. 8. Initial steps for Speaker 1's bVd context. (a) *F*2 trajectories measured for his vowels in bVd context. (b) The ensemble scales $a_{bd}(n)$ derived from these trajectories. (c) The inter-vowel mean $F_{b•d}(n)$ of the trajectories. The peripheral quartets of frames (for $n=1, 2, 3, 4$ and $n=8, 9, 10, 11$) in the three plots are enclosed in boxes.

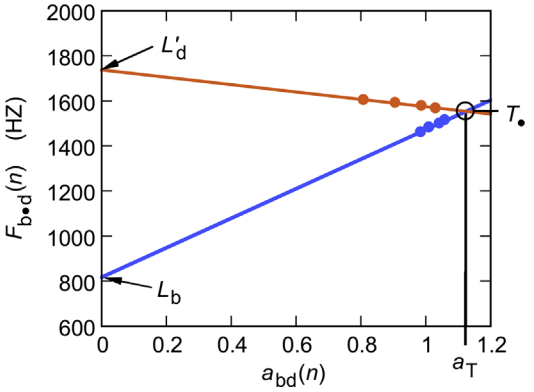

Fig. 9. Speaker 1's $F_{b•d}(n)$ plotted against $a_{bd}(n)$ showing the estimation of the loci L_b and L'_d, the target scale a_T, and the mean vowel target $T_•$. The quartets of data points are from the boxed intervals in Fig. 8: frames 1–4 for L_b (plotted in blue) and 8–11 for L'_d (plotted in orange).

3.2. Estimation of consonant loci and vowel targets

As we showed in BC10, the translation and scaling effects can be brought together to estimate consonant loci and vowel targets in single-sided CV and VC′ contexts. We now treat the peripheral frames such as those marked in Fig. 8 as quasi-single-sided contexts, in this case as bV and Vd contexts. (Note that the number of time-frames through the vowel nuclei must be large enough to allow at least a few to be close to the consonant boundaries.) Estimating the loci involves linear relationships between the scaling effect as embodied in the ensemble scale $a_{CC'}(n)$ and the translation effect as embodied in the inter-vowel-mean formant $F_{C•C'}(n)$ (see BC10 and Appendix A). In BC10 we showed that the locus for a C or C′ in a single-sided CV or VC′ context can be estimated as the *y*-axis intercept of such a linear relation. As shown in Fig. 9, these relations for the peripheral-frame quartets from Fig. 8 have the *y*-axis intercepts $L_b = 821$ Hz and $L'_d = 1738$ Hz, which are the initial-/b/ and final-/d/ loci.

In BC10 we also pointed out that in a diagram like Fig. 9 with only two contexts (bV and Vd in the present case) the intersection between the two lines will mark the point where the ensemble scale equals the target scale a_T and the inter-vowel mean formant equals the mean vowel target $T_•$. (See Appendix A). This follows from BC10's criterion of *inter-contextual consistency* for defining the vowel targets, which means the targets should be the same for any context. The fact that the intersection between the lines in Fig. 9 satisfies this criterion follows from the reasoning following Eq. (A.9) in Appendix A. Here we have $a_T = 1.1235$ and $T_• = 1554$ Hz.

Thus Fig. 9 yields the values of four critical parameters: L_b, L'_d, a_T, and $T_•$. Formulas for computing these parameters are given in Appendix B, Sections B.1.1 and B.1.2. It should still be noted that although the figure shows these four parameters together, fitting the lines to the peripheral-frame data is the first operation. The *y*-axis intercepts of these immediately provide values for the consonant loci. It is only after the lines are constructed that the target scale and inter-vowel mean target can be found from their intersection.

With $T_•$ and a_T in hand, we obtain the individual vowel targets as (BC10):

$$T_V = T_• + a_T[F_{•V•}(•) - F_{•••}(•)] \tag{7}$$

Table 1 lists the vowel targets from Eq. (7). These targets seem reasonable in terms of phonetic expectation. This may seem surprising for estimates that are derived from only near-boundary time frames. We might expect that excluding the central frames ($n=5, 6, 7$) from the estimates should lead to less reasonable targets. This seemingly odd situation does not arise from any intention of ignoring the central frames, but arises instead from the method's need to have formants that can reasonably be interpreted as being from single-sided CV and VC′ syllables, bV and Vd syllables in the present case. Hence in Fig. 9, the two lines are indeed fit to data from these quasi-bV and quasi-Vd syllables, but once the lines are determined, the data points no longer matter and it is the

Table 1
Speaker 1's vowel targets for his bVd context from Eq. (7).

Vowel	$T_{ɪ}$ (Hz)	$T_{ɛ}$ (Hz)	$T_{æ}$ (Hz)	$T_{ɑ}$ (Hz)	$T_{ɒ}$ (Hz)	$T_{ʌ}$ (Hz)	$T_{ɜ}$ (Hz)
Target	1978	1944	1811	1323	915	1345	1560

lines themselves and their intersection that determine the target scale and mean vowel target. Even with this explanation it may still be at least a little surprising that this approach to target estimation should work so well.

With the loci and targets in place we can move to the final step in the linear-decomposition method, the determination of the transition-shape functions $G_b(n)$ and $G'_d(n)$. In dealing with only a single CVC′ context, this step can be illustrated with a geometric interpretation.

3.3. Estimation of the transition-shape functions $G_C(n)$ and $G'_{C'}(n)$

Estimation of the transition-shape functions is the last stage in finding all the elements of Eq. (1), which describes how effects from preceding and following consonants combine to shift the formant away from its vowel target. Determining the values of the transition-shape functions will now characterize how these shifts are to be allocated between the effects from the two consonants. At any given time frame, each vowel will have its own tradeoff for allocating the consonantal effects and so $G_C(n)$ and $G'_{C'}(n)$ will be estimated by finding the best agreement among the tradeoffs for the different vowels.

To illustrate this concept with a simple example, consider how the tradeoffs work for frame $n=4$ of the two vowels /ɪ/ and /ɒ/, which are the two that differ the most in $F2$. The loci L_b and L'_d and the targets $T_{ɪ}$ and $T_{ɒ}$ have already been estimated and the data values for the corresponding formants $F_{bɪd}(4)$ and $F_{bɒd}(4)$ are available from Table D1 in Appendix D. From these various values we write two instantiations of Eq. (1): one using the vowel /ɪ/ and the other using the vowel /ɒ/. After the substitutions, calculation of target-locus differences, and some rearrangements of terms, we obtain the following two equations in the unknown $G_b(4)$ and $G'_d(4)$:

$$\text{Using /ɪ/:} \quad G'_d(4) = -0.29955 - 4.8304\,G_b(4) \tag{8a}$$

$$\text{Using /ɒ/:} \quad G'_d(4) = 0.001593 + 0.11976\,G_b(4) \tag{8b}$$

As plotted in Fig. 10(a), these relations represent two linear vowel-dependent tradeoffs between $G_b(4)$ and $G'_d(4)$. Their intersection is the solution of Eqs. (8a) and (8b):

$$G_b(4) = -0.0608 \;\; G'_d(4) = -0.0057 \tag{9}$$

This exercise illustrates why the linear-decomposition method requires a dataset with at least two vowels: with only a single vowel all the points along its tradeoff line would have equal standing and choosing one could only be arbitrary. Hence it is differences among the tradeoff lines for the various vowels that make the method possible.

The example just worked for frame 4 could be repeated for all the other time frames, but rather than use only two well-separated vowels which serve for illustration, it is best to take all the vowels into account. This will require us to find a compromise among the tradeoffs, a compromise that minimizes the modeling error in fitting Eq. (1) to the data. This involves some simple mathematics, as detailed in Section B.1.3 of Appendix B.

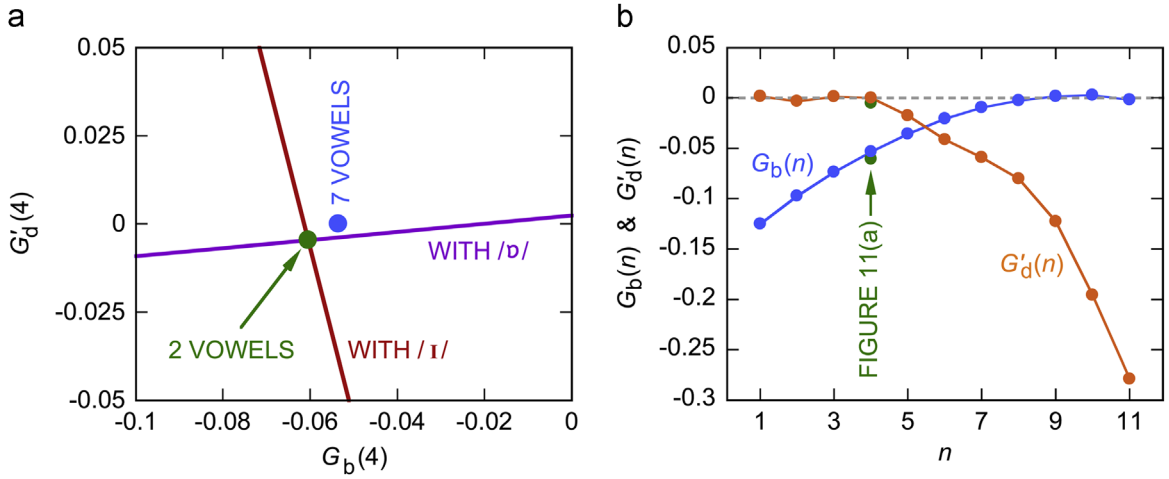

Fig. 10. (a) Linear tradeoffs between Speaker 1's $G'_d(4)$ and $G_b(4)$. The one plotted in red is for the vowel /ɪ/ and the one in purple is for the vowel /ɒ/. The intersection (green dot) is where the two tradeoffs agree and so is an estimate for $G_b(4)$ and $G'_d(4)$. The blue dot is the solution from a least-squares fit to Eq. (1) taking all 7 vowels into account. (b) Speaker 1's $G_b(n)$ and $G'_d(n)$ from the least-squares 7-vowel solutions plotted against the frame n. The green data points at $n=4$ represent the 2-vowel solution from (a).

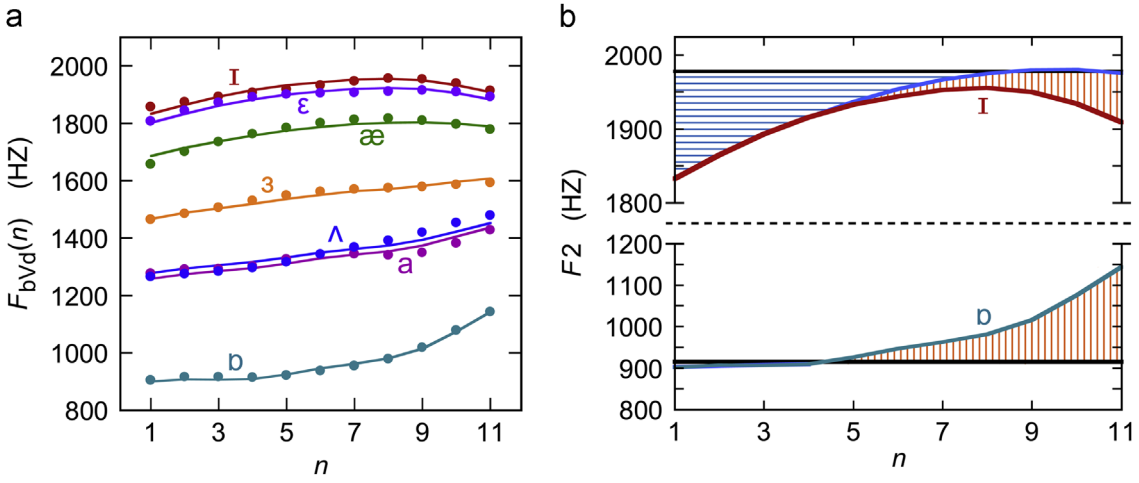

Fig. 11. (a) Speaker 1's *F*2 data (plot symbols only) shown with the *F*2 trajectories generated by Eq. (1) (contours without plot symbols). (b) How the additive components from Eq. (1) combine with the vowel targets to model Speaker 1's *F*2 trajectories for the vowels /ɪ/ (top plot, red curve) and /ɒ/ (bottom plot, aqua curve) in bVd context. The horizontal black lines represent the vowel targets and hatched areas show the components from the two consonants: horizontal blue hatching for the bV component and vertical orange hatching for the Vd component.

The $G_b(n)$ and $G_d'(n)$ time functions which result from applying the linear-decomposition method using data and targets from all seven vowels to all 11 frames are shown in Fig. 10(b). The results for frame 4 from Fig. 10(a) are the two green data points in Fig. 10(b). These differ only slightly from their 7-vowel counterparts.

The initial-/b/ transition-shape function $G_b(n)$ displays tangible persistent effects through frame 7, while the final-/d/ function $G_d'(n)$ shows anticipatory effects from frame 5 onward. The first 4 frames of $G_d'(n)$ and the final 4 frames of $G_b(n)$ are close to zero and so indicate that effects from the trans-vowel consonant become negligible for the two pairs of peripheral time-frame quartets.

3.4. Reconstruction of the formant trajectories from the linear-decomposition model

Together with the previously obtained loci and targets, the transition-shape functions in Fig. 10(b) complete the linear decomposition of Speaker 1's bVd context. With all the model's elements now in hand for this context, we can use Eq. (1) to reconstruct its *F*2 trajectories. The outcomes for all Speaker 1's vowels are shown in Fig. 11(a) where the model *F*2s are shown as contours without plot symbols and the original *F*2 data from Fig. 8(a) are shown as plot symbols without connecting lines. The model tracks the data quite well, with an RMS error of only 12 Hz.

Because the model is a good description of the data, it seems reasonable to think that it is also a good description of the properties of the data. This can be seen in Fig. 11(b) which shows diagrams of how the formant trajectories are structured in terms of Eq. (1) for the vowels /ɪ/ and /ɒ/. As in the conceptual diagram in Fig. 1, the targets are represented by horizontal black lines, the additive bV component by horizontal blue hatching, and that from the Vd component with vertical orange hatching. The /ɪ/ trajectory is the red curve underlying its vowel target and contextual components while the /ɒ/ trajectory is the aqua curve mostly overlying its target line and contextual components.

The differences between these trajectories arise entirely from their different vowel targets: the loci L_b and L_d' are the same for the different vowels in the context, as are the transition-shape functions $G_b(n)$ and $G_d'(n)$. The /ɪ/ target is well above both consonant loci so that the target-locus differences for the initial and final contexts are both positive. Because the transition-shape functions are negative, their products with the target-locus differences are also negative and so pull the formant trajectory down from the /ɪ/ target. This is shown in the figure by these components lying under the target line. At 915 Hz, on the other hand, the /ɒ/ target is still above the /b/ locus of 821 Hz, but by only 94 Hz. Hence its bV component is still convex upward, though its small scale renders it nearly invisible in the plot. Because the /ɒ/ target is smaller than the /d/ locus, its target-locus difference is negative and so its product with the negative $G_d'(n)$ is positive. This positive Vd component pulls the /ɒ/ trajectory upward from its target toward the /d/ locus, as shown by its lying atop the Vd component.

Because the consonant-specific /b/ and /d/ loci and transition-shape functions remain the same, the vowel target is the only model element that changes between the two vowels. As the vowel target changes, the trajectories shift systematically in their shapes and ranges. All the vowel targets are larger than the /b/ locus, so all the onset transitions are convex upward but become shallower as the target drops. The target-locus difference drops along with it, thus decreasing the scale of the bV component.

These observations illustrate how the linear-decomposition method provides a direct view of the way individual vowel trajectories are structured and of how these structures change systematically from one vowel to another.

3.5. Results for the other single contexts (dVd, gVd)

Fig. 12(a) shows $G_d(n)$ and $G_d'(n)$ for Speaker 1's symmetric dVd context. The peripheral frames display a pattern comparable to that for the bVd context, but the central frames display a large mirror-image distortion, in which $G_d(n)$ takes an exaggerated local rise

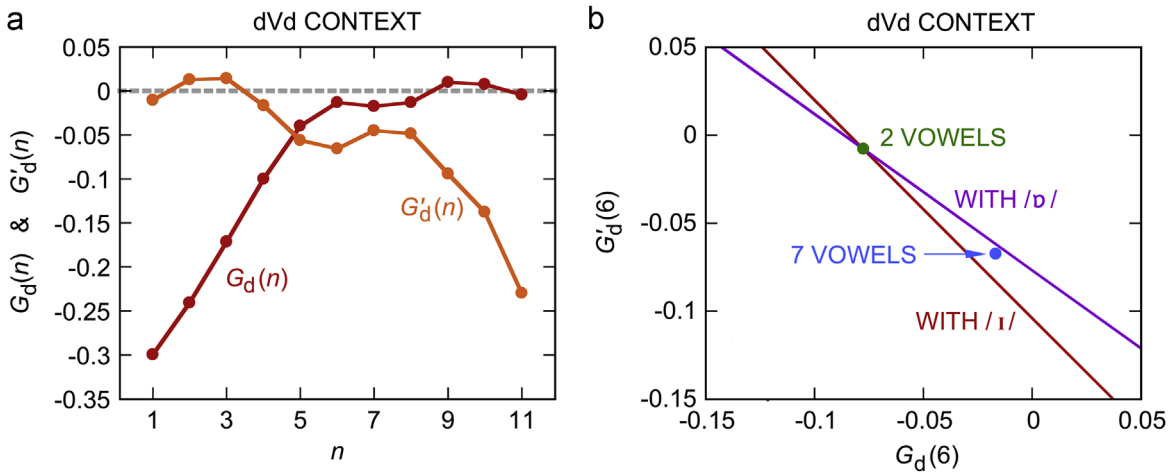

Fig. 12. (a) Transition-shape functions for Speaker 1's dVd context. (b) Graphical 2-vowel estimates of $G_d(6)$ and $G'_d(6)$ showing their sensitivity to the small angle between the lines analogous to Eqs. (8a) and (8b) using the targets and data for /ɪ/ (red line) and /ɒ/ (purple line). Their intersection (green dot) is the 2-vowel solution while the light blue dot is the linear least-squares estimate taking all 7 vowels into account.

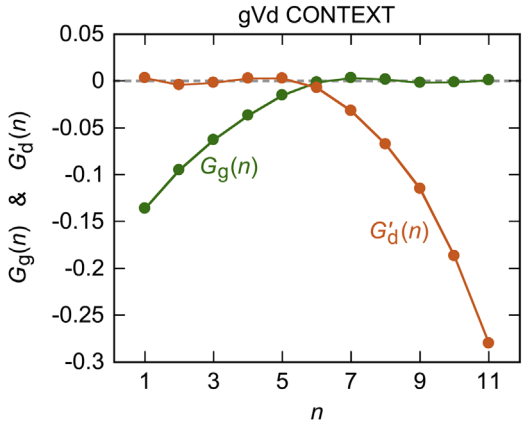

Fig. 13. Transition-shape functions $G_g(n)$ and $G'_d(n)$ for Speaker 1's gVd context.

topping out at $n=6$ after which it connects with the convergence toward zero. $G'_d(n)$ takes an exaggerated local drop bottoming out at $n=6$ after which it connects with the expected drop toward the boundary with the final /d/.

This odd behavior appears to arise from the small difference of only 76 Hz between the two /d/ loci, $L_d=1589$ Hz and $L'_d=1665$ Hz. As shown in Fig. 12(b), the angle between the lines for the geometric two-vowel estimation of $G_d(6)$ and $G'_d(6)$ becomes quite small and the location of their intersection becomes sensitive to small variations in the loci, targets, and *F*2 data. This becomes evident from the distinct locations of the data points for $G_d(6)$ and $G'_d(6)$ for the 2- and 7-vowel solutions. With the 7-vowel estimate falling between the two lines in the figure, it is easy to imagine how even small shifts in the lines would shift the intersection to a range of points along the general linear trend. Here the two estimates use the same /d/ loci and the same data and targets for the vowels /ɪ/ and /ɒ/ and differ only in their use of two or seven vowels. As will be shown in Section 4.5, the outcome for the dVd context can be improved by including it in a joint analysis with one or more other contexts which include at least one with an initial-C locus well separated from the two /d/ loci.

Fig. 13 shows the transition-shape functions for the gVd context. These are close to zero for frames distant from their associated consonants and only at frame 6 do both functions contribute to the resultant and this only minimally. Hence the gV component is virtually finished by the time the Vd component begins.

With Speaker 1's bVd, dVd, and gVd contexts now analyzed, we turn to the analyses for Speaker 2.

3.6. Results for Speaker 2

Fig. 14 shows the transition-shape functions for Speaker 2's bVd, dVd, and gVd contexts. The outcomes for both his bVd and gVd contexts resemble those from Speaker 1, particularly in terms of their timings for the persisting and anticipatory effects. Speaker 2's outcome for the dVd context is even more problematic than Speaker 1's: his $G_d(n)$ and $G'_d(n)$ look even wilder than Speaker 1's and virtually impossible to interpret. This result again seems to arise from closely-spaced /d/ loci: $L_d=1854$ Hz and $L'_d=1741$ Hz. As with Speaker 1, an approach to ameliorating this outcome is described in Section 4.5. Except for minor details, the results for the two speakers' single-context analyses are in good agreement.

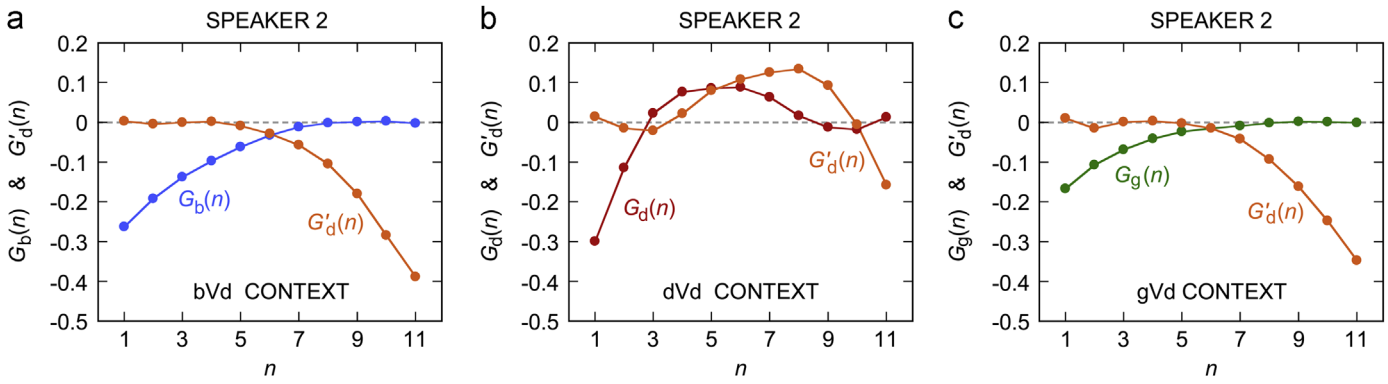

Fig. 14. Speaker 2's transition-shape functions. (a) For his bVd context. (b) For his dVd context. (c) For his gVd context.

To better observe differences and convergences among the outcomes for the various analyses, the loci and targets for these and later analyses are summarized in Section 5.6.

4. The linear-decomposition method II: Multiple-context case

We next consider joint analyses of multiple contexts using Speaker 1's full dataset with the three contexts bVd, dVd, and gVd for illustration. As shown in Fig. 15, the method for this case differs from that for single contexts most obviously in its estimates of the target scale and mean vowel target which, instead of their direct geometric estimation for the single-context case, must now be found by searching for a target scale that leads to the best fit between Eq. (1) and the data.

As will be seen in Section 4.2, locus estimation still uses *y*-axis intercepts, but now intercepts shared by lines for the different trans-vowel consonants. Section 4.3 deals with adapting the estimation of the target scale, mean vowel target, and individual vowel targets to multiple-context datasets. The method for the transition-shape functions is described in Section 4.4. The outcomes for the joint linear-decomposition analyses of the bVd, dVd, and gVd contexts for both speakers are described in Section 4.5.

4.1. Ensemble scales and inter-vowel mean trajectories: The initial steps

The ensemble scales are still obtained from the slopes of lines fit to the relations between formants for each VFE in a multi-context dataset and the mean VFE. The one difference to note in the multiple-context case is that the elements of the mean VFE are now obtained by averaging each vowel's formant over both frames and contexts.

The inter-vowel mean trajectories are still computed by averaging a context's trajectories over all its vowels. The one difference from the single-context case is that now a mean trajectory must be computed for each context, i.e., for each combination of initial and final consonants in the dataset.

4.2. Locus estimation

The method for locus estimation for the single-context case has to be elaborated for it to apply to consonants that are paired with more than one trans-vowel consonant. The locus will still be obtained from the linear relations between peripheral frames of $F_{C\bullet C'}(n)$ and $a_{CC'}(n)$, but now with one such linear relation for each trans-vowel consonant. Lines fit to the data in this case will now have to be constrained to have the same *y*-axis intercept which, as in the single-context case, will be identified with the consonant locus. The formulas for the locus and the different line slopes are given in Appendix B, Section B.2.1.

Each of our speaker's initial consonants, /b/, /d/, and /g/, occur only with the final consonant /d/ and so their loci are still determined by the method used for single contexts.

His final /d/ occurs with the three initial consonants, so his final-/d/ locus must be found as a *y*-axis intercept shared by three lines. These are shown in Fig. 16 where the intercept yields Speaker 1's final-/d/ locus as $L'_d = 1663$ Hz.

With their common intersection at the *y*-axis, the lines must diverge away from it so there is no question of obtaining the target scale and mean vowel target from an intersection as in the single-context case. For this we turn to another adaptation of BC10's use of target-locus-scaling.

4.3. Vowel target estimation

For the single-context case, the target scale and mean vowel target were found as the intersection between two lines, as illustrated in Fig. 9. Fig. 17 shows an analogous plot for the inter-vowel mean formant plotted against the ensemble scale for peripheral frames from the bVd, dVd, and gVd contexts. It includes the data from Fig. 16, with their three lines now closer together due to the expanded range of the *y*-axis, and the three lines used to determine the loci for the initial /b/, /d/, and /g/. Instead of two lines, as in Fig. 9, there are now six lines with multiple distinct intersections. To deal with this, we still use the idea behind the

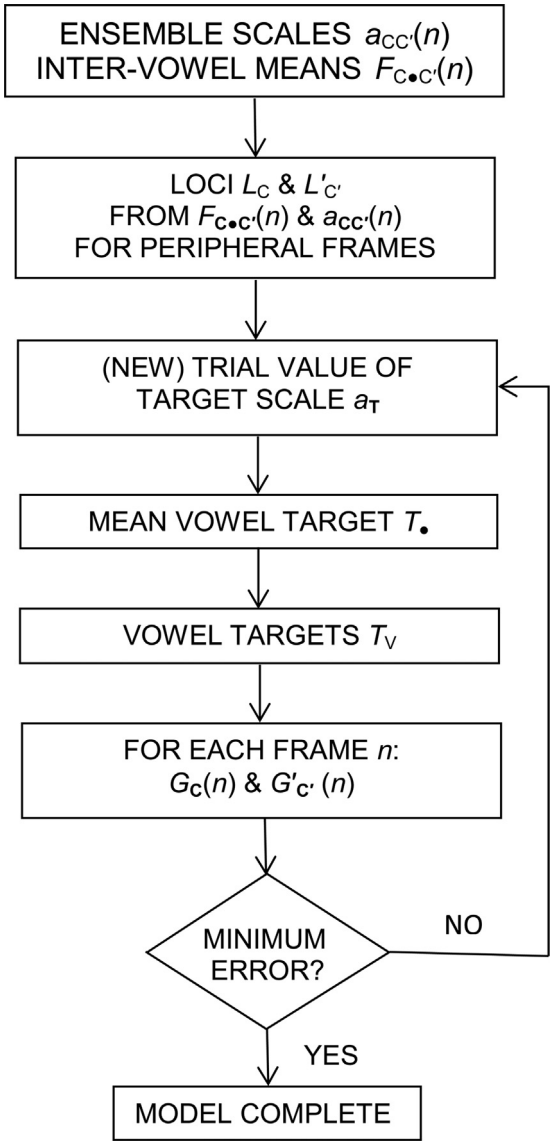

Fig. 15. Flow diagram for the joint linear decomposition of multiple contexts.

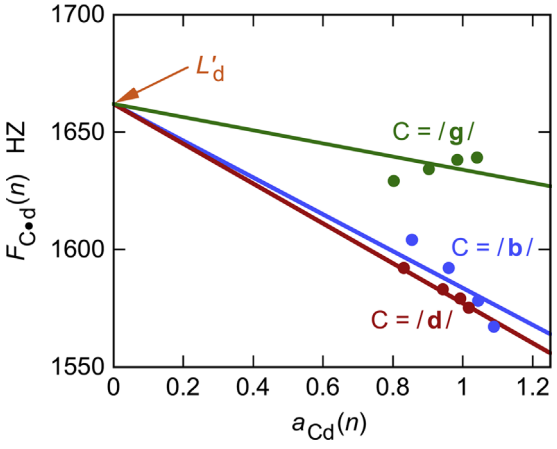

Fig. 16. Estimation of the final-/d/ locus L'_d for Speaker 1's 3-context dataset. $F_{C\bullet d}(n)$ is plotted against $a_{Cd}(n)$ for the final four frames ($n=8, 9, 10, 11$) of his bVd, dVd and gVd contexts.

single-context case, but now adapt it to the task of finding a compromise "near intersection", a point on the plot that represents the target scale and mean vowel target that minimizes the RMS error in fitting the model to the data.

A solution could be found by a two-dimensional search of (a_T,$T_\bullet$) pairs. However, BC10 showed that the problem can be reduced to a one-dimensional search of target scales for single-sided CV or VC′ contexts. This is made possible by a formula for the best $T_\bullet$

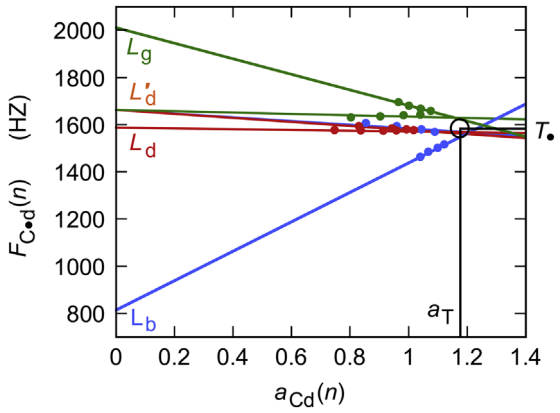

Fig. 17. Estimation of the target scale and mean vowel target for the joint analysis of Speaker 1's bVd, dVd, and gVd contexts. The inter-vowel mean formant $F_{C\bullet d}(n)$ is plotted against the ensemble scale $a_{Cd}(n)$ for the peripheral frames. The first four frames ($n=1, 2, 3, 4$) are used for the initial-consonant loci L_b, L_d, and L_g. The final four frames ($n=8, 9, 10, 11$) are used for the final-/d/ locus L'_d and reproduce the data and lines from Fig. 16. The circled right-angle between the black lines marks the optimum positions for the target scale a_T and mean vowel target $T_\bullet$.

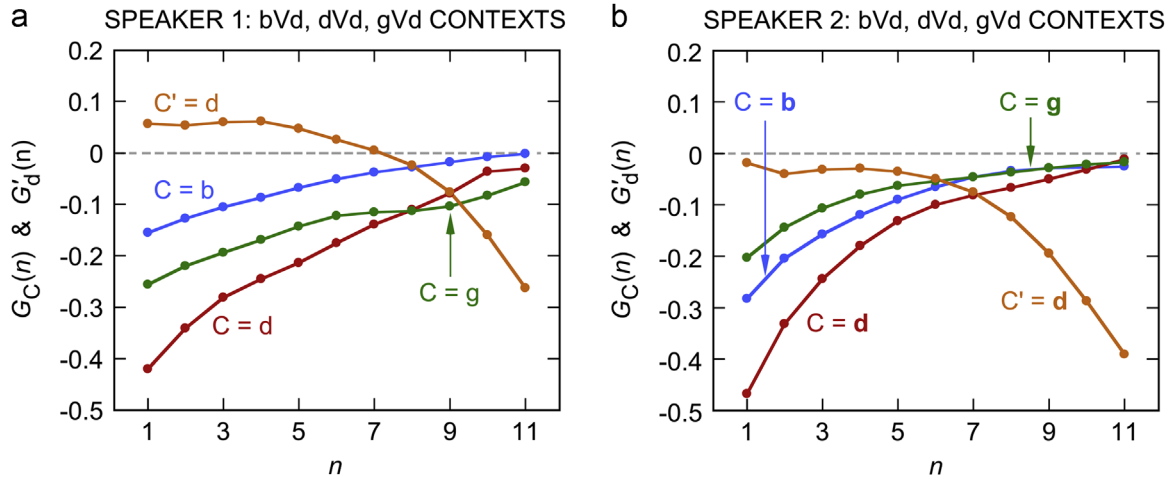

Fig. 18. Transition-shape functions for the bVd, dVd, and gVd contexts resulting from joint analyses. (a) For Speaker 1. (b) For Speaker 2.

that corresponds to any trial value of a_T. This formula is adapted to two-sided CVC′ contexts in Section B.2.2 of Appendix B. The targets for the individual vowels are still obtained from Eq. (7) using this trial a_T and its implied $T_\bullet$. As was the case in BC10, the best value for a_T must be determined computationally by a search of trial values for the one that gives the best fit to Eq. (1) through the subsequent determination of trial vowel targets and trial values for the $G_C(n)$ and $G'_{C'}(n)$. To anticipate part of the outcome, the optimum a_T and $T_\bullet$ for Speaker 1 that result from this search are marked by the right-angle intersection between the two black lines in Fig. 17 in the area of confluence among the six lines.

4.4. Estimation of $G_C(n)$ and $G'_{C'}(n)$

As with the single-context case, the transition-shape functions are estimated one frame at a time with no a priori constraints for frame-to-frame continuity. With transition-shape functions to obtain for more than one context at a time, the explicit formulas for the single-context case no longer apply and a more general form of linear algebra must be used. As with the single-context case, the method still depends on linear relationships among the $G_C(n)$ and $G'_{C'}(n)$ as conditioned by the values of the consonant loci and trial values of the vowel targets. If we let N_C be the number of initial consonants and $N_{C'}$ be the number of final ones, there will be $N_C+N_{C'}$ linear equations in the $N_C+N_{C'}$ unknown $G_C(n)$ and $G'_{C'}(n)$, with one set of these equations to be solved for each frame n. The computing formulas for these equations are in Section B.2.3 of Appendix B and their solutions are readily obtained by well-known standard methods (see, e.g., Lipschutz and Lipson (2013), Section 3.8, pp. 73–78).

With all the trial elements of Eq. (1) in hand, the trial model is complete. To complete the final model, it only remains to do a simple search for the trial value of the target scale that yields the trial model with formants having the smallest RMS modeling error. Once this is found, the model is complete.

4.5. Results of joint analyses of the full dataset

Fig. 18 shows the transition-shape functions that result from applying the method for multiple contexts to the data for our two speakers' CVd data with C=/b, d, g/. Note that the final-/d/ transition-shape function is shared by the by the three contexts.

For Speaker 1, the early frames for $G'_d(n)$ are displaced upward and further away from zero than in the outcomes for the single-context analyses of his bVd and gVd contexts. However, in terms of how this function changes over time, it resembles those from the single bVd and gVd analyses in which the anticipatory drop from near-constancy sets in only near the vowel center.

His initial-consonant transition-shape functions for /b/ and /g/ differ significantly from those for their single-context analyses. They no longer converge on zero or any other constant value near the vowel center, but instead approach zero gradually over the duration of the vowel, a behavior now shared by the initial /d/.

For Speaker 2 the first frame of $G'_d(n)$ starts close to zero, but the other early frames hover slightly below it. As with Speaker 1, its first substantial downward shift occurs about midway through the vowel. Also in parallel with Speaker 1, Speaker 2's initial-consonant functions approach zero in a regular rise throughout the duration of the vowel instead of peaking near the vowel center as in the single-context analyses. For Speaker 1 the initial-/g/ function is displaced substantially downward from its position in his single-context analysis.

For both speakers, embedding the dVd context in a joint analysis with others avoids the problem of the irregular outcomes for $G_d(n)$ and $G'_d(n)$ observed in their single-context analyses. Including contexts with more stable single-context analyses has kept the dVd context from wandering too far from what we see as a more reasonable outcome.

For interpretation we are left with the question of how long the effects of the initial consonants persist: only to about midway in the vowel as in their single-context analyses for the bVd and gVd contexts or practically to the vowel boundaries as in their joint analyses? This question will be taken up in Section 5.

5. Exponential transition-shape functions

The outcomes for the bVd and gVd contexts in the multiple-context analyses are inconsistent with those for their single-context analyses, most notably in the timings of their persisting effects from the initial /b/ or /g/: In single-context analyses these come close to zero about midway through the vowel while in their multiple-context analyses they approach zero more gradually over the duration of the vowel. A further complication for the multiple-context cases is that their $G_C(n)$ and $G'_d(n)$ only hover in the neighborhood of zero instead of convincingly closing in on it for frames near the boundaries with their trans-vowel consonants.

In this section we attempt to resolve these concerns by representing the $G_C(n)$ and $G'_{C'}(n)$ with exponential functions of the following form, where β_C and $\beta'_{C'}$ are reciprocal time constants and κ_C and $\kappa'_{C'}$ are scale factors:

$$G_C(n) = \kappa_C \exp(\beta_C n) \tag{10}$$

$$G'_{C'}(n) = \kappa'_{C'} \exp(\beta'_{C'} n) \tag{11}$$

5.1. Fitting the model with exponential transition-shape functions to data

Implementation of the model follows the steps shown in the flow diagram in Fig. 19. This now includes a search for the optimum target scale for both the single- and multiple-context cases. This was necessary for the multiple-context case just described, but not for the single-context case for which the target scale was found by direct construction. When we wondered whether searching for an optimum in the single-context case would yield the same answer as the construction did, we were surprised to find that every trial target scale yielded the same error. This meant that the construction in Section 3.2 was not only possible, but necessary. The new need for a search, even for a single-context, arises as follows when the shape functions are represented by exponentials.

When these functions are represented by exponentials with asymptotes equal to zero, as in Eqs. (10) and (11), different values of the target scale a_T imply different placements of these zero asymptotes in relation to the $G_C(n)$ and $G'_{C'}(n)$ from the underlying linear-decomposition analysis. This means that a search for an optimum a_T will now make sense, because each trial value will yield different transition-shape functions that will differ in how well they fit zero-asymptote exponentials. These in turn will result in different errors in fitting Eq. (1). Hence for the models with exponential transition-shape functions, a search for the optimum target scale is both possible and necessary for both the single- and multiple-context cases.

To fit each transition-shape function to an exponential, we search a range of values for its reciprocal time constant, β_C or $\beta'_{C'}$. For each of these, a simple formula from Appendix C yields the optimum value of the scale factor, κ_C or $\kappa'_{C'}$. Once the best exponentials are found, the trial model is complete and its RMS modeling error can be found. The final model is then the one with the trial target scale that leads to the smallest modeling error.

5.2. The speakers' exponential transition-shape functions

Fig. 20 shows the exponential $G_C(n)$ and $G'_d(n)$ for both speakers' single-context analyses of the bVd and gVd contexts and their joint analyses of the bVd, dVd, and gVd contexts. (No single-context analyses for the dVd context are shown, because for neither speaker are $G_d(n)$ and $G'_d(n)$ suitable for exponential fits.) All the curves go asymptotically toward zero for frames more removed in time from their associated consonants. The exponentials (curves without data points) track the $G_C(n)$ and $G'_d(n)$ from linear decomposition (data points without connecting lines) reasonably well.

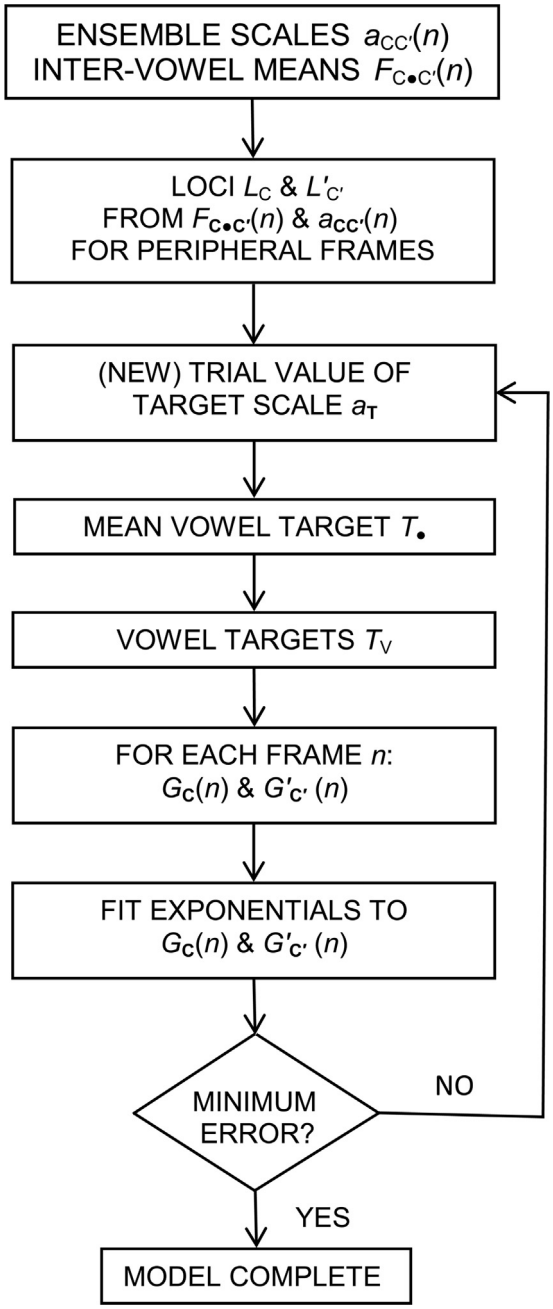

Fig. 19. Flow diagram for models with exponential time functions fit to the transition-shape functions. Here both single and multiple contexts require a search for an optimum target scale, as indicated by the feedback loop.

Speaker 1's exponential $G'_d(n)$ transition-shape function is nearly the same across the three analyses and is in agreement with his basic linear-decomposition analyses by its remaining close to zero through frame 4. Its drop away from zero becomes noticeably steep from frame 5 onward. His exponential $G_b(n)$ for the bVd context is also consistent between its single-context analysis and its appearance in the joint analysis. As was the case for his single- and multiple-context linear-decomposition analyses, the $G_g(n)$ for his gVd context in the joint exponential analysis is still displaced considerably downward from its position in its single-context exponential analysis. At the same time, however, these differently positioned transition-shape functions share a gradual upward trend toward zero across the duration of the vowel, in agreement with their outcomes in unmodified LD analyses for multiple-context datasets, but at variance with their outcomes in single-context analyses.

The exponential transition-shape functions from Speaker 2's joint analysis are in good agreement with all those from his single-context bVd and gVd analyses. As with Speaker 1, his final-/d/ transition-shape function hovers near zero through frame 4, then becomes noticeably steep from frame 5 onward, while his initial-consonant transition-shape functions converge more slowly on zero across the duration of the vowel.

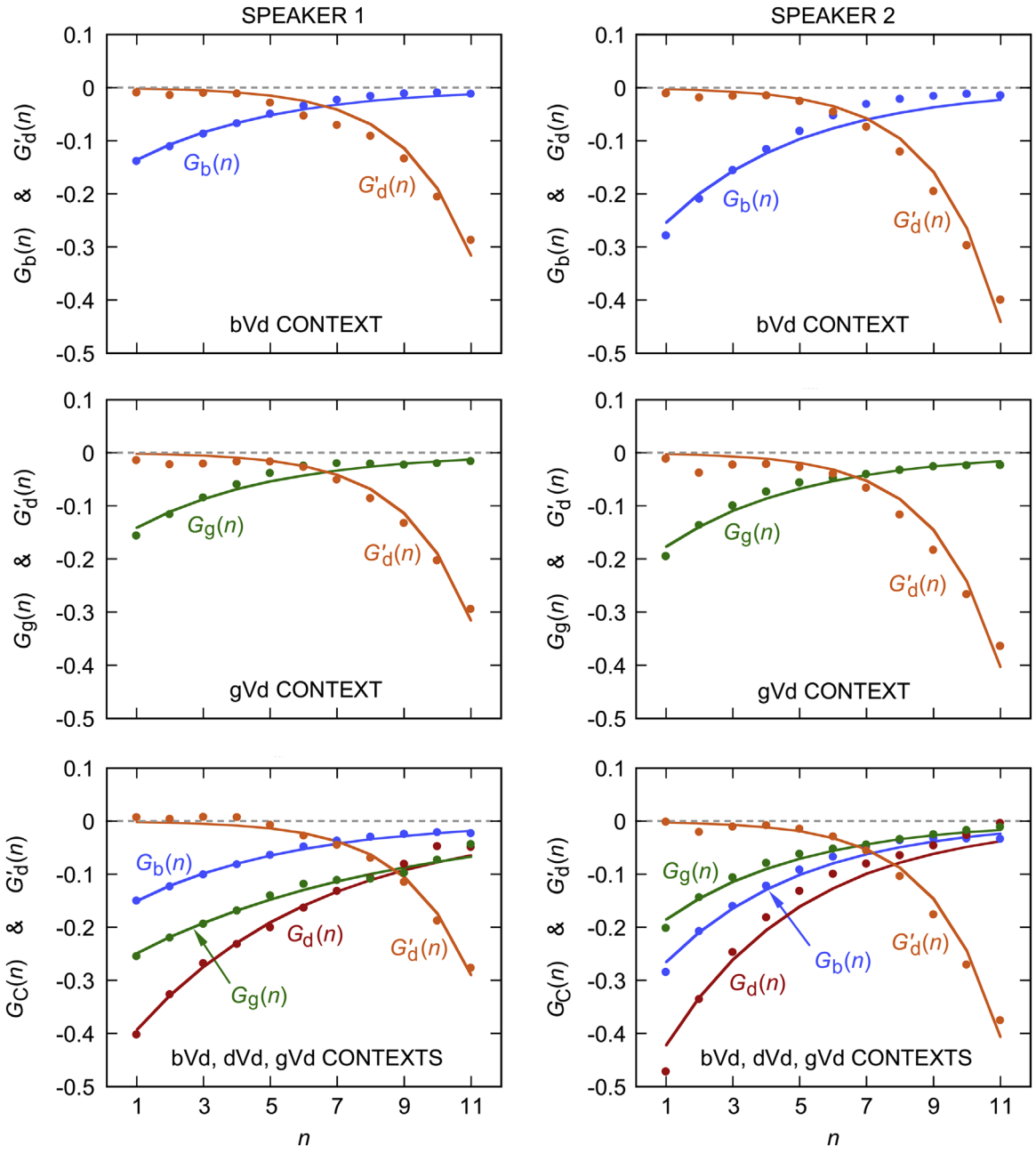

Fig. 20. Exponential transition-shape functions for Speaker 1's (left column) and Speaker 2's (right column) single-context analyses for the bVd (top row) and gVd (middle row) contexts and their joint analyses for the bVd, dVd, and gVd contexts (bottom row). Dots are for the $G_C(n)$ and $G'_d(n)$ from the unmodified LD analyses while the curves are their best-fit exponentials.

5.3. *Parameters of the exponential transition-shape functions*

The outcomes among analyses can be compared more analytically by looking at the parameters for the exponential fits. These are shown in Table 2. The scale factors κ_C and κ'_d are all negative because the $G_C(n)$ and $G'_d(n)$ in Fig. 20 are also all negative. The initial-consonant reciprocal time constants are all negative because their functions decay toward zero as frames increase from the beginning of the vowel. Conversely, all the final-/d/ reciprocal time constants are positive because their functions expand exponentially downward for frames advancing from the vowel onset.

As seen in the last two columns of Table 2 the κ'_d for the final /d/ are much smaller in magnitude than the κ_C for the initial consonants because the final-/d/ function starts from near zero and ends with a larger magnitude while the initial-C functions start with larger magnitudes. The reciprocal time constants for the final /d/ are substantially larger in magnitude than those for the initial consonants.

Aside from the different ranges for Speaker 1's gVd context, the overall result is one of consistency between the single- and multiple-context methods. The more gradual approach to zero for the initial-consonant transition-shape functions, seen above for the joint linear-decomposition analyses of three contexts, appears to be vindicated by the exponential models. Even though $G'_d(n)$ has a

Table 2
Parameters for the exponential fits in the various analyses. The scale factors are κ_C and $\kappa'_{C'}$ while β_C and $\beta'_{C'}$ are the corresponding reciprocal time constants.

Speaker	Context(s)	κ_b	β_b	κ_d	β_d	κ_g	β_g	κ'_d	β'_d
1	bVd	−0.190	−0.239					−0.00664	0.344
	gVd					−0.213	−0.310	−0.00318	0.413
	bVd, dVd, gVd	−0.211	−0.215	−0.501	−0.195	−0.277	−0.138	−0.00258	0.437
2	bVd	−0.391	−0.293					−0.00672	0.375
	gVd					−0.247	−0.303	−0.00454	0.400
	bVd, dVd, gVd	−0.371	−0.293	−0.617	−0.301	−0.253	−0.261	−0.00830	0.352

shorter time constant, its apparently slow in departure from zero arises from its very small scale factor κ'_d, which is less than 0.01 in magnitude.

The initial-consonant reciprocal time constants are smaller in magnitude than those for the final /d/. In terms of time constants, then, the initial consonants are relatively "slow" and the final /d/ is relatively "fast".

5.4. How the trajectories are structured with exponential transition-shape functions

As observed in connection with Fig. 11(b), the good fits between the model and the data suggest that the model is providing us with a good description of how the formant data are structured. Fig. 21(a) illustrates how Speaker 1's *F*2 bVd trajectories are structured in terms of the exponential version of Eq. (1) for the vowels /ɪ, æ, ɜ, ʌ, ɒ/. (The /ɛ/ and /a/ are not shown because they are not very distinct from the /ɪ/ and /ʌ/.) As in Fig. 1, each panel shows the vowel target as a horizontal line, the bV component of the trajectory as an area with horizontal hatching, and the Vd component as one with vertical hatching. Each resultant trajectory is shown as a curve with the same color as used for the same vowel in Figs. 8(a) and 11(a).

Here the additivity of the components can be shown by pure stacking only for /bɪd/ and /bæd/ for which the vowel targets are both larger than the initial-/b/ and final-/d/ loci. For /bɜd/, /bʌd/, and /bɒd/, on the other hand, the vowel targets are less than the final-/d/ locus which makes their dV components positive, while their bV components remain negative. For these syllables the final-/d/ component now pulls the formant trajectory up while the initial-/b/ component continues to pull it down. This can be seen most clearly in the plots for /bɜd/ and /bʌd/ where the horizontal hatching for the bV component and the vertical hatching for the Vd component overlap in the small area under the target line and above the arc of the trajectory. In all five plots the construction starts from the vowel target, adds the bV component to it (which in these cases means moving downward from the target to the arc marking the boundary of the horizontally hatched area) and from that boundary then adds the Vd component, which is a downward move for /bɪd/and /bæd/ and an upward one for /bɜd/, /bʌd/ and bɒd/, all to arrive at the bVd trajectory.

Because the consonant-specific /b/ and /d/ loci and transition-shape functions remain the same, the vowel target is the only model element that changes from one panel to another. As the vowel target changes, the trajectories shift systematically in their shapes and ranges. All the vowel targets are larger than the /b/ locus, so all the onset transitions are convex upward but become shallower as the target drops. The difference between the vowel target and the /b/ locus drops along with it, decreasing the scale of the bV component.

Because the final-/d/ locus is smaller than the /ɪ/ and /æ/ targets, their Vd transitions are, like the bV transitions, convex upward. On the other hand, the targets for the vowels /ɜ, ʌ, ɒ/ are all smaller than the /d/ locus so their Vd transitions become convex downward, though for the /ɜ/ the target is about midway between the initial-/b/ and final-/d/ loci so the resultant trajectory exhibits only the slightest sinuosity as it moves steadily upward away from the /b/ locus and toward the /d/ locus. The Vd transition becomes more distinctly convex downward for the /ʌ/ and /ɒ/ as their targets become more distant from the final-/d/ locus.

Not only can we now see how the individual vowel trajectories are structured, but we can also see a larger picture of how these structures change systematically from one vowel to another.

It is perhaps a little too easy to think of the /bɪd/ trajectory as typical: with its wide-ranging /bɪ/ and /ɪd/ components, it conveys a mental impression that suggests that there should be an extremum in the trajectory near the vowel center, an extremum that might serve as a proxy for the vowel target. Indeed, /ɪ/ and /æ/ have maxima somewhat past the vowel center, but the other three vowels exhibit mostly monotone upward trends which for /ɜ/ and /ʌ/ do pass through their targets near mid-vowel, while the /ɒ/ trajectory hovers near its target for the first 4 or 5 frames. Hence the trajectory for each vowel may have its own closest approach to its target, but this will also depend on the context.

In Fig. 21(a) we have seen how the vowel trajectories are structured according to the relations between the vowel targets and consonant loci, relations that shift systematically with the vowel targets. The details of the five trajectories in Fig. 21(a) should not distract us from how Eq. (1) represents a coherent family of trajectories, all related by scaling relationships that are anchored by their loci and targets. In this sense, the panels in Fig. 21(a) are samples of members of this family, samples that necessarily come from real vowels, as these were the means by which the elements of Eq. (1) could be determined. Now that these elements are in hand, however, we are in a position to illustrate the family it represents. Fig. 21(b) shows the model trajectories for the five vowels from Fig. 21(a) embedded with members of Speaker 1's bVd family, here generated with targets between 800 and 2050 Hz in steps of 50 Hz. This shows the model's representation of the context as a coherent family of curves, all having their loci and transition-shape

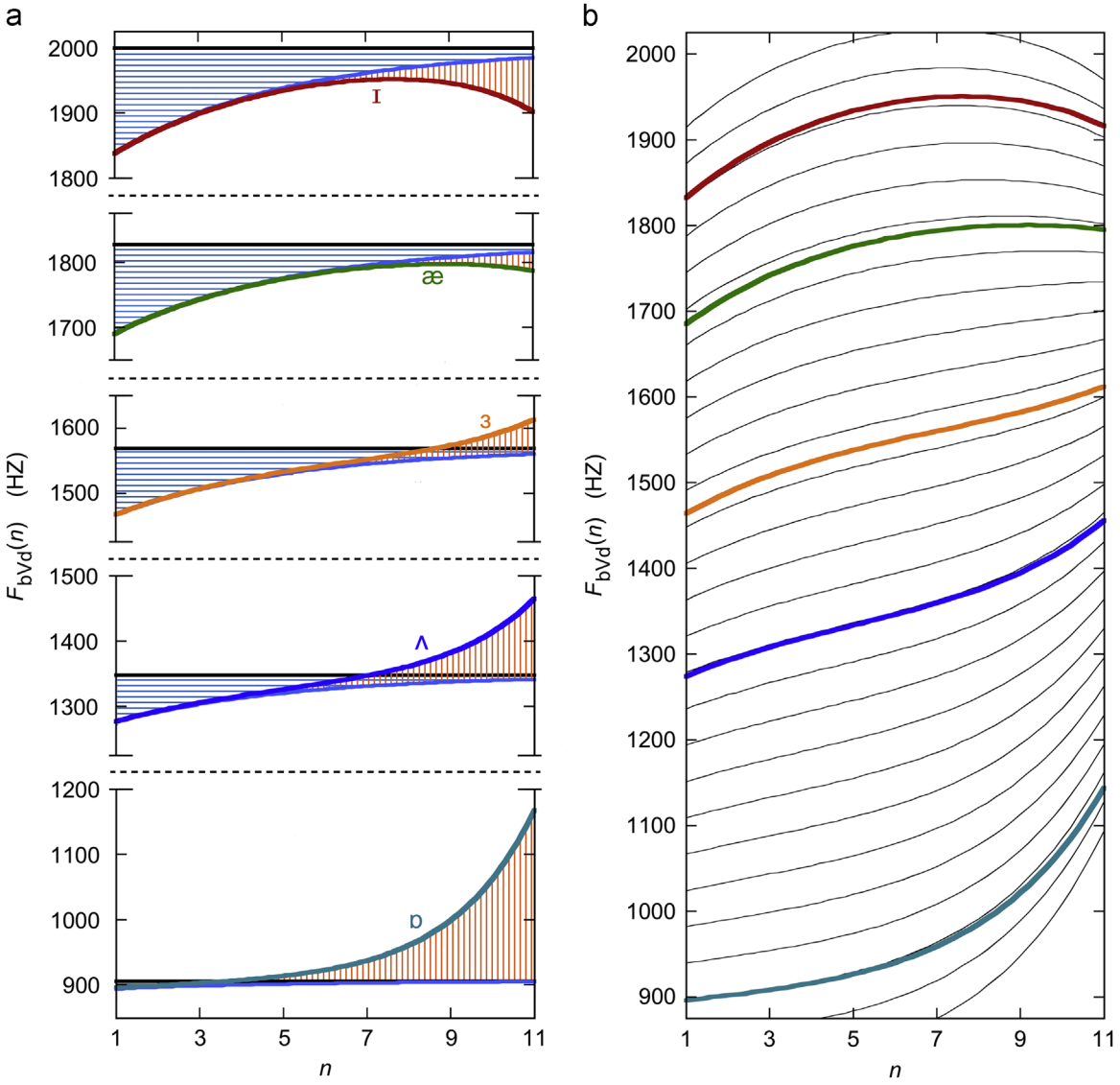

Fig. 21. (a) Five plots separated by dashed lines showing how the additive exponential components combine with the vowel targets to model Speaker 1's *F*2 trajectories for vowels in bVd context. From top to bottom the plots are for the vowels /ɪ/, /æ, /ɜ/, /ʌ/, and /ɒ/. In each plot, the horizontal black line represents the vowel target and hatched areas show the contextual components: horizontal blue hatching for the bV component and vertical orange hatching for the Vd component. (b) *F*2 trajectories from Eq. (1) for the same vowels shown embedded with Speaker 1's family of trajectories in the context, with vowel targets between 800 and 2050 Hz in steps of 50 Hz. In both parts of the figure the vowel colors are as in Figs. 8(a) and 11(a).

functions in common and varying only in their vowel targets. Hence the model not only yields good representations of the individual trajectories, it also shows them as samples from a continuum of trajectories from the same family.

5.5. Analysis of the modeling errors

How reasonable are the models with exponential transition-shape functions? Here these models will be referred to as exponential models. To check on the quality of their fits we compare their errors to those expected on the basis of their numbers of parameters. The exponential models have many fewer parameters than the unmodified LD models because they involve only two parameters for each transition-shape function versus the 11 for the raw functions from their underlying linear-decomposition models. Because of their smaller numbers of parameters, the exponential models are expected to have larger errors.

We can estimate these from an argument based on the numbers of degrees of freedom for both the data and the models under the hypothesis that shifts in the errors are all attributable to the differences in the number of modeling parameters. The data have as many degrees of freedom N_{data} as there are data points to be fit. Let $N_{par,LD}$ be the number of parameters in a linear-decomposition model, and $N_{par,exp}$ be their number in a model with exponential transition-shape functions. Also let $s_{LD.model}$ be the RMS error from a linear-decomposition model and s_{est} be the RMS error estimated for its associated exponential model under the degrees-of-freedom

hypothesis. The expected error based only on the degrees of freedom will then be

$$s_{\text{est}} = \sqrt{\frac{N_{\text{data}} - N_{\text{par,exp}}}{N_{\text{data}} - N_{\text{par,LD}}}}\, s_{\text{LD.model}} = k\, s_{\text{LD.model}} \tag{12}$$

Here the constant k is the value of the square root. The modeling errors for the exponential models are summarized in Table 3. In five of the six cases, the actual errors $s_{\text{exp.model}}$ for the exponential models are smaller than expected. In the other two cases the expected and actual errors are equal (to the nearest 1 Hz). Even though the modeling errors for the exponential models are larger than those for their underlying linear-decomposition models, these are no larger than expected on the basis of their smaller number of parameters. The exponential fits therefore come at no intrinsic cost to the quality of the models.

Some of the errors for the raw linear-decomposition models may appear unrealistically small, particularly those for Speaker 1's single-context analyses. These small errors arise from two factors: (1) the dataset's makeup of averages of five repetitions and the consequent reduction in random variations by a factor of about the square root of five and (2) the large number of parameters in Eq. (1) [consonant loci and vowel targets together with the 11 frames of each $G_{\text{C}}(n)$ and $G'_{\text{C}'}(n)$] in comparison with the number of data points being fit.

The exponential models provide an economic representation of the data with no intrinsic harm to the quality of the fits. Fig. 20 shows that these models go a long way toward resolving the discrepancies between outcomes for single- and multiple-context analyses. The exponential models also realize the phonetic expectation that a consonant's effects on a vowel should diminish toward zero for frames more temporally removed from it.

The errors in Table 3 compare favorably with the speakers' variations among the five tokens of each syllable type. This means that the model formant typically differs from a speaker's inter-repetition mean by less than his own tokens do. These errors also compare favorably with a listener's ability to detect changes in *F*2 (Flanagan, 1955).

5.6. Variations in the consonant loci and vowel targets for all analyses

Table 4 lists the consonant loci from the various analyses of the two speakers' datasets. The initial-consonant loci for both speakers are quite consistent across their different analyses. Speaker 1's final-/d/ locus varies over a range of 144 Hz for his different analyses. While this is not an especially large variation, it is a larger range than for any other locus in the table. Speaker 2's final-/d/ locus displays a much smaller range of 44 Hz for the different cases. For Speaker 1 the initial-consonant loci differ between their single- and multiple-context cases by at most 7 Hz, while the corresponding maximum difference for Speaker 2 is 13 Hz.

Table 5 lists the vowel targets from the various analyses, together with the corresponding mean vowel targets $T_{\cdot}$ and target scales a_{T}. Now exponential models are included in the list because their values differ from those of the corresponding raw linear-

Table 3
Errors for exponential models compared with those from the basic linear-decomposition models. N_{data} is the number of data points being fit. $N_{\text{par,LD}}$ is the number of parameters in a basic linear-decomposition model. $N_{\text{par,exp}}$ is the number of parameters in an exponential model. The RMS error for a basic linear-decomposition model is $s_{\text{LD.model}}$ and $s_{\text{exp.model}}$ is the RMS error for an exponential model. The constant k is the expected ratio of $s_{\text{exp.model}}$ to $s_{\text{LD.model}}$ from Eq. (12). The expected error estimate from Eq. (12) is s_{est}.

Speaker	Context(s)	N_{data}	$N_{\text{par,LD}}$	$N_{\text{par,exp}}$	$s_{\text{LD.model}}$	k	s_{est}	$s_{\text{exp.model}}$
1	bVd	77	31	13	12	1.18	14	13
	dVd	77	31		13			
	gVd	77	31	13	15	1.18	18	16
	bVd, dVd, gVd	231	55	19	19	1.10	21	21
2	bVd	77	31	13	21	1.18	25	22
	dVd	77	31		23			
	gVd	77	31	13	35	1.18	41	37
	bVd, dVd, gVd	231	55	19	33	1.10	36	34

Table 4
Consonant loci for the two speakers' single- and multiple-context analyses. Loci for exponential models are not shown because once the loci are determined in a linear-decomposition model, they remain the same for exponential models.

Speaker	Context(s)	Analysis type	L_{b} (Hz)	L_{d} (Hz)	L_{g} (Hz)	L'_{d} (Hz)
1	bVd	Single	821			1738
	dVd	Single		1589		1665
	gVd	Single			2012	1594
	bVd, dVd, gVd	Multiple	814	1589	2011	1663
2	bVd	Single	1410			1766
	dVd	Single		1854		1741
	gVd	Single			3100	1722
	bVd, dVd, gVd	Multiple	1408	1854	3087	1743

Table 5
Vowel targets for the two speakers' single- and multiple-context analyses.Under Type, LD signifies a basic linear-decomposition analysis and EXP signifies an analysis with exponential transition-shape functions. The target scales are a_T and the mean vowel targets are $T_{\bullet}$. For neither speaker were the transition-shape functions for the dVd context suitable for exponential fits.

Speaker	Context(s)	Type	a_T	$T_{ɪ}$ (Hz)	$T_{ɛ}$ (Hz)	$T_{æ}$ (Hz)	$T_{ɜ}$ (Hz)	$T_{ʌ}$ (Hz)	$T_{ɑ}$ (Hz)	$T_{ɒ}$ (Hz)	$T_{\bullet}$ (Hz)
1	bVd	LD	1.1235	1978	1944	1811	1323	915	1345	1560	1554
	bVd	EXP	1.1752	2012	1976	1837	1327	900	1350	1575	1568
	dVd	LD	1.1670	1984	1893	1817	1337	988	1408	1561	1570
	dVd	EXP									
	gVd	LD	1.1032	2067	1991	1880	1411	1044	1487	1623	1643
	gVd	EXP	1.1540	2078	1998	1882	1392	1008	1472	1614	1635
	bVd, dVd, gVd	LD	1.1772	2021	1951	1840	1341	951	1400	1575	1583
	bVd, dVd, gVd	EXP	1.2652	2056	1982	1862	1326	906	1389	1577	1585
2	bVd	LD	1.2018	2213	2102	1918	1210	917	1279	1481	1589
	bVd	EXP	1.2810	2255	2137	1941	1186	873	1260	1475	1590
	dVd	LD	0.9734	2152	2027	1854	1337	1170	1421	1567	1647
	dVd	EXP									
	gVd	LD	1.1385	2347	2121	1935	1257	1011	1395	1531	1657
	gVd	EXP	1.1960	2339	2102	1907	1195	936	1340	1482	1614
	bVd, dVd, gVd	LD	1.2894	2297	2117	1904	1167	895	1281	1469	1590
	bVd, dVd, gVd	EXP	1.3100	2302	2119	1903	1154	878	1270	1460	1584

decomposition analyses. Target values for any given vowel differ from one analysis to another with per-vowel minimum-to-maximum ranges between 63 and 144 Hz for Speaker 1 and between 73 and 297 Hz for Speaker 2. While these are not remarkably consistent, within each analysis the vowels consistently follow the same ordering in *F2*: $T_{ɪ} > T_{ɛ} > T_{æ} > T_{ɜ} > T_{ʌ} > T_{ɑ} > T_{ɒ}$.

6. Discussion

Here we discuss various aspects of the model and LD method. In Section 6.1 we describe the requirements for datasets to be treatable by the LD method; Section 6.2 compares the relative benefits of single- and multiple-context analyses; Section 6.3 discusses error-reduction methods and their pitfalls; Section 6.4 describes connections between linear-decomposition and locus equations; the challenge of taking vowel duration into account is discussed in Section 6.5.

6.1. Dataset requirements and caveats

In order to be analyzed a dataset must have at least two vowels and its formant data must have time frames closely enough spaced to allow good resolution of the formant dynamics. This is essential for the early steps that use peripheral frames to estimate the consonant loci and vowel targets. Hence the method is not applicable to a dataset consisting of only a single syllable or one with measurements from only a few frames.

The vowels should be true monophthongs. For example, the English /e/ and /o/ cannot be used because they are typically realized as the glides [e^{ɪ}] and [o^{ʊ}]. In our dataset, the Australian /i/ and /u/ are excluded because, as Bernard (1970, p. 114) has observed, they are "less happily" produced as "one-target sounds." These would be examples of vowel-inherent dynamics as described by Nearey (2013). If syllable nuclei with inherent dynamics are included in a dataset, the linear-decomposition method will be unable to distinguish these dynamics from those due to context.

The notation and formalism used here are designed for datasets in which all combinations of the initial and final contexts occur. This will be true of nearly all the datasets found in the literature. Clearly, any single-context dataset satisfies this requirement, because "all" the combinations consist of the single C and C′ pair.

Exceptions are possible for multiple-context datasets. For example, Fig. 4 has two contexts sharing the same vowels but sharing neither their initial nor their final consonants. Here, even the calculation for the mean VFE presents difficulties. Let the syllable types be denoted as C1VC′1 and C2VC′2, where C1 differs from C2 and C′1 differs from C′2. Then with only these two contexts the elements of the mean VFE will be $0.5(F_{C1VC2}(\bullet)+F_{C'1VC'2}(\bullet))$. Our expression $F_{\bullet V\bullet}(\bullet)$ for the mean VFE would become meaningless because it assumes that the dataset would also contain data from contexts C1VC′2 and C2VC′1, combinations absent in the example from Fig. 4.

The model is easy to adapt to datasets such as these, though notation and computations become more cumbersome. For notation, it would become necessary to treat contexts as single entities. Hence a CVC′ syllable would be said to have a context which could be denoted as $\underline{CC'}$. The formant would then be denoted somewhat less intuitively as $F_{\underline{CC'}V}(n)$ instead of $F_{CVC'}(n)$. Its advantage would be that the corresponding expression $F_{\bullet V}(\bullet)$ for the elements of the mean VFE would be valid and all the implied modifications to the structuring of computations would work for all multiple-context datasets sharing the same vowels. The practical implication for coding would be that C and C′ pairs would have to be dealt with as compound units, not as separate C and C′ units. We have not derived the formulas for this and so we cannot say how practical such adaptations might be.

With this caveat, we again note that researchers will usually not use datasets with such disparate contexts.

Faster processor speeds and larger-capacity memories also open possibilities for dealing with larger datasets. For research in phonetics the limiting factor might no longer lie with computing power but with possible unreliability of automated segmentation and formant-tracking methods. This would call for some human supervision. Another limiting factor would be the endurance of speakers. Thoughtful experimental design would be needed to avoid combinatorial explosion. For example, applying the linear-decomposition method to a corpus with 7 vowels (as here) with all combinations of 10 initial and 10 final consonants would consist of 700 syllables. Obtaining 5 repetitions of each syllable (as here) would then yield a corpus of 3500 syllables. One would have to listen to the recordings to verify that they reasonably represent the scripts. Limits on attention could become a challenge for speakers and listeners alike. How many such datasets would be humanly feasible?

Our model's additivity of C and C′ effects might ameliorate combinatorial explosion by treating these effects separately instead of in combination. The method would have to be adapted to handle an “additive” dataset (see B84), with, for example, CVd syllables augmented with mirror-image dVC′ syllables. The sizes of “multiplicative” datasets with all combinations of C and C′ would grow with the products of their numbers while the sizes of additive ones would grow only with their sums.

6.2. Single- and multiple-context analyses

The methods for both single and multiple contexts are based on the fact that for each n Eq. (1) is linear in its transition-shape functions. Once consonant loci and vowel targets are in place, therefore, the only remaining unknowns are these $G_C(n)$ and $G'_{C'}(n)$. Eq. (1) is then easily solved using the method of least squares (Hamming, 1986, pp. 435–442).

Though the methods for single and multiple contexts are both based on this observation, their implementations differ. Single-context datasets cannot be analyzed as special cases of the multiple-context method by setting its numbers of initial and final consonants both equal to 1. When this is tried, the multiple-context method's search for an optimum target scale never converges because each of its trial values yields the same modeling error. (It should be possible to prove this mathematically but we have not yet been able to do this.) Hence the phonetically-motivated deterministic method for finding the target scale and mean vowel target in the single-context method (see Fig. 9) has not only been useful for illustration, it is necessary for the single-context case.

When multiple contexts are available, the multiple-context version of the linear-decomposition method provides a unified picture of the contexts as a family of formant trajectories tied together through an analysis in which (1) all the ensemble scales are referred to the same mean VFE, (2) the vowel targets are shared by all the contexts, and (3) any initial or final consonant shared by contexts in the dataset will have the same locus and transition-shape function. With such shared benchmarks, the results from a multiple-context analysis permit direct comparisons among the contexts, comparisons that can be problematic between single-context analyses with their context-local benchmarks.

Outcomes from the single-context linear-decomposition analyses for bVd and gVd contexts conflict with those from multiple-context analyses in which they are included: in the single-context analyses the initial-consonant transition-shapes go nearly to zero just past the midpoint of the vowel and their final-/d/ transition shapes begin to depart from zero just before the midpoint. When they appear in multiple-context analyses, on the other hand, the CV transition-shape functions display more gradual shifts toward zero that span the duration of the vowel. Fitting exponential time-functions to the transition-shape functions largely resolves these discrepancies in favor of the more gradual convergence on zero for the initial-consonant transition-shape functions. It is not just the exponential representation that accomplishes this, but the new target scales that also shift the transition-shape functions in the underlying LD analyses. In addition to its resolution of most of the discrepancies between single- and joint-method outcomes, the models with exponential transition-shape functions also have the advantage of using a much smaller number of parameters than the basic linear-decomposition models do.

It should be kept in mind that the exponential fits are an added stage of the overall analysis and not an alternative to the linear-decomposition method. Indeed, without a prior linear decomposition there would be no transition-shape functions to fit.

It is also important to remember that not all transition-shape functions will be reasonably represented by exponentials. It is therefore advisable to examine the outcomes and their RMS errors to check that exponential fits have introduced little or no tangible distortions.

6.3. Numerical methods for reducing the modeling error

It can be tempting to treat the analyses developed so far as a way to obtain only initial estimates for the elements of Eq. (1) and to use these as a starting point for reducing the fitting error by some kind of numerical method. For example, once the initial $G_C(n)$ and $G'_{C'}(n)$ are in place, Eq. (1) becomes linear in its loci and targets. One could then find new loci and targets that minimize the error with respect to these $G_C(n)$ and $G'_{C'}(n)$, after which a new LD analysis could find new $G_C(n)$ and $G'_{C'}(n)$. This pair of refinements might be repeated as many times as desired. We have found, however, that the modeling error can begin to increase with only a few iterations of this procedure: the best Ls and Ts for a given set of $G_C(n)$ and $G'_{C'}(n)$ are not necessarily better than the ones used to find these functions: the procedure retains no memory of what worked best in the past, and so has no basis for working progressively for better results.

We have also tried some variations on the method of steepest descent in which the error does progressively fall, but sometimes at the cost of phonetically absurd results, such as loci that grow to very large magnitudes (e.g., 10^5 Hz) and corresponding transition-shape functions with tiny magnitudes. The problem with purely numerical refinements is that their only criterion is error reduction

without regard to any phonetic consideration. Our priority in the LD analysis is not to reduce errors at any cost, but to reveal in a phonetically meaningful way how vowels are coarticulated with the consonants that precede and follow them in the speech stream.

The linear-decomposition method initializes loci and targets by treating peripheral frames as if they are from single-sided CV and VC′ syllables, even though Eq. (1) allows for effects from the trans-vowel consonant to span the duration of the vowel. In principle it would therefore be desirable for the locus and target estimates to be informed by data from all the available frames, but so far we have not found a way to do this that also yields phonetically convincing outcomes. For the present, the use of peripheral frames for loci and targets serves operationally as a phonetic constraint on the LD method through its recognition that a consonant's effect must be strongest on the frames closest to it.

6.4. The locus equation (LE) revisited

6.4.1. The underlying phenomenon

The locus equation (LE) is the linear relationship now widely observed between *F*2 for vowels in consonantal context as measured at their onset and center time frames. The extensive literature on locus equations has shown that there is much to be learned about vowels transitioning from different consonants by using datasets with only these two time frames. The slopes and intercepts of these relationships have been effective parameters for describing this empirical observation. The linear representation of how onset and center frames are related is the descriptive model at the heart of locus-equation studies. Though it is descriptive, the locus equation model is also predictive to the extent that it predicts that the relations between the formants from onset and center time-frames will be well described by straight lines.

As with the locus equation, our Eq. (1) is also a descriptive model and is also based on empirical observations yet, again like the LE, it is predictive to the extent that it predicts that data can be well described by it and, that if this holds, that the properties of the model are good approximations to properties of the data.

Because the LD method requires data from multiple frames within one or more CVC′ contexts and locus equations involve only two frames, it is only in the early stages of the LD method that the methods become comparable, specifically the LD method's determination of ensemble scales. Fig. 22 illustrates this point in its two plots that show two ways of exhibiting Speaker 1's *F*2 from the onset and center frames for his seven vowels in the bVd context.

Fig. 22(a) shows these data in a locus-equation format with the onset-frame *F*2 plotted against its center-frame counterpart. The dotted lines from the data points represent projections of the onset *F*2s onto the *y*-axis and of the center *F*2s onto the *x*-axis. The two sets of projected values are the vowel-formant ensembles (VFEs) for these two frames. Because they are drawn from frames 1 and 6 (onset and center, respectively) in the speaker's bVd context, these are denoted VFE(b,d,1) and VFE(b,d,6).

Each of these VFEs has an ensemble scale which is obtained from the slope of a line fit to its relation to the mean VFE, the elements of which in this case are the averages of the *F*2 pairs for the different vowels. Note that the ensemble-scale calculation is done with respect to the mean VFE, not to the center-frame VFE as in the locus-equation plot in Fig. 22(a).

Fig. 22(b) shows the two VFEs plotted against their ensemble scales. The *x*-axis starts at zero where the /b/ locus is plotted. (A zero ensemble scale would correspond to a VFE with each vowel having the same *F*2.) The pencil of lines radiating from it are best fits to the seven vowels' data pairs. Though the locus seems unrealistically low, it helps us visualize the relative spreads of the *F*2 values in the two VFEs and their relative placements on the *F*2 axis.

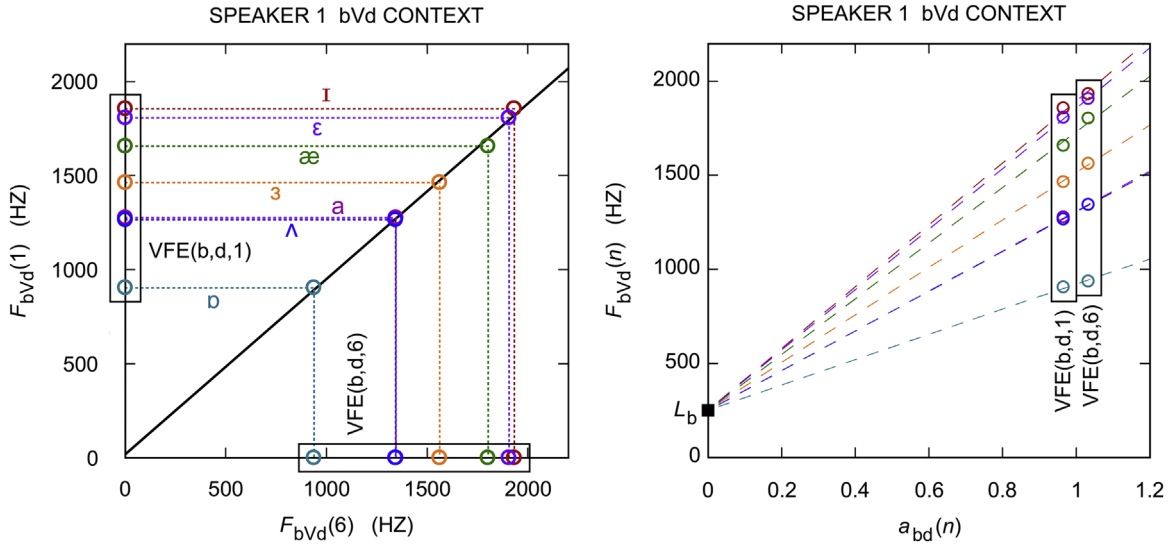

Fig. 22. Speaker 1's initial- and center-frame *F*2 from his bVd context displayed in two ways: (a) In locus-equation format with the onset-frame ($n=1$) data plotted against the center-frame ($n=6$) data. Colors for the 7 vowels match those from Figs. 8(a) and 11(a). The line fit to the data is the locus-equation relationship. Dotted lines from the data points connect to the vowel-formant ensembles (VFEs) for the two frames, VFE(b,d,1) along the *y*-axis and VFE(b,d,6) along the *x*-axis. (b) The two VFEs plotted against their ensemble scales $a_{bd}(n)$. The locus obtained from these is plotted at the *y*-axis ($a_{bd}=0$) and lines radiating from it are best fits to the vowel-specific *F*2 pairs.

The two plots in Fig. 22 carry the same information, but displayed in different ways. A link between the two is that the linearity of the LE plot (a corollary of the geometric similarity between the onset and the center VFEs) is what guarantees the good fits between the seven lines and data pairs in Fig. 22(b). As long as the consonant locus had no clear function, there was no reason to keep track of it in locus-equation studies. In Fig. 22(b) it takes on the role of a focal point for the connections between the elements of the two VFEs. In this way, Fig. 22(b) provides a phonetically-motivated perspective on the information that an LE dataset has to offer.

6.4.2. Limitations of the current practice of LEs

Locus equations as they are currently implemented have some limitations. One is that they are context-local in the sense that each one is limited to two frames from the same context. As a consequence, cross-context comparisons have to be made on the basis of LE plots each of which is referred to a different standard: the context's own center-frame data. Center-frame data can vary systematically between contexts and so must be seen as context-local standards for the scaling of initial-frame data, scaling as measured by the LE slope. Such variations among local standards will weaken comparisons among contexts. This weakness could easily be avoided by referring all a dataset's VFEs to a common standard. If the mean VFE were adopted as such a standard, comparisons across multiple frames and contexts would become more straightforward and meaningful. In particular, this could immediately yield information on how center frames are affected by different contexts. By drawing on all the data in a given dataset, the mean VFE becomes statistically more stable than a VFE for any individual center frame, which will have its own random variations in addition to its inherent differences from other contexts.

Leaving the question of statistical stability aside, in BC10 we showed that there is a formal relationship between the locus equation and ensemble scales. This relation is quite simple: the LE slope is the ratio of the ensemble scale of the onset-frame VFE to that of the center-frame VFE. In BC10 we further argued that this use of only the ratio of ensemble scales instead of the two scales themselves entails the loss of potentially valuable information.

To illustrate this point, we showed how more could be learned from the LE data in the thoughtful article by Duez (1992), which included useful tables of formants and vowel durations from French syllables uttered both in spontaneous speech and read from lists of the same syllables in citation form after their transcription from the spontaneous modality. We found that the *F*2 in vowels from both speech modalities could be represented by a common set of consonant-place loci and vowel targets. Differences between the modalities were confined to differences in the ranges of the transition shapes. Hence there was no evidence that speakers used more careful vowel targets in reading words from lists, though in the reading modality their vowels came closer to realizing the vowel targets. Furthermore, we were able to show a relation between vowel duration and vowel reduction, though this result had to be taken with some caution because it was obtained by pooling the data from the study's five speakers. From that result we predicted that vowel reduction would also show up as an effect in the pattern of LE slopes, a prediction borne out by Duez' data.

Here Fig. 22(b) shows how treating the onset- and center-frame VFEs as separate entities yields a view of the data which makes immediate sense in terms of how the vowel formants are organized in relation to the consonant locus.

6.4.3. Toward more effective LE analyses

We suggest that using ensemble scales based on the mean VFE as a standard reference would be better practice for the analysis of two-frame data from multiple contexts in locus-equation studies. Cross-context comparisons would then be on a stronger foundation.

We also suggest that plotting the VFE pairs against their ensemble scales and relating these to the consonant locus (as in Fig. 22 (b)) would aid in a better understanding of what the data are telling us. This would not necessarily replace an LE plot, as in Fig. 22(a), but could serve as a useful adjunct to one.

Although there is more that can be learned from two-frame data, the fact remains that much more is to be learned from multiple-frame data that capture dynamic variations in vowel formants. As long as we had few methods for going beyond only looking at plots of dynamic variations in the raw and perceiving only the possible existence of regularities, there may have appeared to be little point in gathering data from which one could obtain only such subjective impressions. With better analytic tools, such as the linear-decomposition method, we can move beyond this to achieve a better understanding of how these dynamics can be described in a systematic way.

6.5. Taking vowel duration into account

Some inaccuracies in the model no doubt arise from variations in vowel duration. The mean duration of Speaker 1's four shortest vowels (/ɪ, ɛ, a, ʌ/) is 136 ms while the mean duration of his three longest vowels (/æ, ɒ, ɜ/) is 220 ms. The corresponding durations for Speaker 2 are 162 ms and 269 ms. These differences in duration are substantial. Because our dataset's discrete time-frame variable *n* is normalized with respect to the vowel duration, Eq. (1) and the linear-decomposition method in their present forms cannot take duration or a physical time scale (e.g., one using units of milliseconds instead of duration-normalized frames) into account.

Lindblom and his colleagues (Lindblom, 1963a, 1963b; Lindblom, Mauk, & Moon, 2006; Lindblom & Sussman, 2012) have made compelling cases for the idea that phonetic gestures are governed by a set of articulatory programs with fixed time courses. This idea would explain Lindblom's discovery (1963a, 1963b) of the relation between vowel reduction at syllable centers and shorter vowel durations. Our duration-normalized time scale puts this relation beyond the direct reach of both Eq. (1) and the linear-decomposition method in their present forms.

The obvious way to take duration into account within the model would be to use a physical time scale. As we have proposed earlier (BC87), this would involve the use of a physical time variable t to rewrite Eq. (1) as

$$F_{\mathrm{CVC'}}(t) = T_{\mathrm{V}} + (T_{\mathrm{V}} - L_{\mathrm{C}})G_{\mathrm{C}}(t) + \left(T_{\mathrm{V}} - L'_{\mathrm{C'}}\right)G'_{\mathrm{C'}}(D-t) \quad 0 \le t \le D \tag{13}$$

where D is the vowel duration. In BC87 we had a somewhat cumbersome way to implement a model with this treatment of the time-axis, and, while it yielded somewhat smaller modeling errors for $F2$ than the analog of Eq. (1) did with its duration-normalized time scale, it was difficult to tell whether these were real improvements or artifacts of the complex fitting procedure.

We think it should be possible to fit Eq. (13) to a real-time dataset even though it is not yet clear how to accomplish this. The root of the problem seems to be the question of which samples from different utterances are comparable. With a duration-normalized time scale it is assumed that the nth frame from each vowel token is comparable to the nth frames from all other vowel tokens. With a dual treatment of a real time scale such as that in Eq. (13), direct comparability across utterances would require the presence of syllables sharing the same values of both t and $D-t$. Such examples would be rare and perhaps nonexistent. Intervals near the vowels' boundaries might still be used to estimate loci and targets, but the problem of solving Eq. (13) for its $G_{\mathrm{C}}(t)$ and $G'_{\mathrm{C'}}(D-t)$ would remain.

A solution might open the way to a unified characterization of coarticulated vowels with varying durations, including long and short vowels, vowels from datasets with syllables spoken at different rates, and random variations between tokens. Adapting the linear-decomposition analysis to accommodate a physical time scale might also provide a definitive way to decide the extent to which vowel reduction arises from physiological constraints (Lindblom, 1963a, 1963b) or from intentional differences in a speaker's manner of speech (Van Son & Pols, 1992; Van Son, 1993).

Although it may be difficult to incorporate duration directly into the linear-decomposition method, duration effects can still be studied by analyzing datasets structured to have the same syllables uttered at different rates. Separate linear-decomposition analyses could then be applied to each dataset with results compared across speaking rates.

Such datasets have previously been used to good effect. Lindblom (1963a, 1963b) elicited CVC syllables with different vowel durations by using four different carrier phrases with different positions for the syllable in the phrase and with the syllable being either stressed or unstressed. Crowther (1994) also used differently structured carrier phrases to elicit syllables with different durations. Duez (1992) obtained syllables with different vowel durations by first recording speakers' spontaneous speech, isolating certain syllables from it, and then having the speakers read lists of the selected syllables at a slow rate in citation form. These studies all took duration into account without having to fit their data to a model such as Eq. (13).

7. Summary

Effects of nearby sounds have long been recognized to have overlapping effects on the production of a current sound (Heffner, 1960; Stevens & House, 1963; Lindblom, 1963a, 1963b; Stevens et al., 1966; Öhman, 1966; Houde, 1968; Broad & Fertig, 1970). Even with this realization, however, the challenge remained of finding a way to identify these overlapping effects. The linear-decomposition method achieves just this for vowels in CVC′ contexts. It is intended as a means for advancing our understanding of the dynamics and phonetic complexity of vowels in CVC′ contexts.

The method's success in achieving accurate characterizations of our datasets shows that it could be used to analyze other datasets in a variety of contexts and speech-modality settings. It also supports the validity of the model in Eq. (1) as a general description of vowels in context, and the idea that properties of the model are good approximations to properties of the data. These properties include the additivity of CV and VC′ components of the formant trajectories for any given vowel and the scaling of these components by target-locus differences applied to consonant-specific transition-shape functions. Hence the model provides some valuable insight into how vowels are coarticulated with their surrounding consonants. Even without applying the method to further datasets this insight alone advances our basic understanding of how coarticulation works and lends quantitative support to Menzerath and le Lacerda's (1933) early idea of articulatory actions being able to overlap in time.

Where possible, the concepts related to the model and the steps in the method have been illustrated with visual interpretations. To minimize the use of mathematics in the main text, computational formulas and more detailed discussions of the model's properties are relegated to appendices. The method is presented in enough detail for readers to program their own working versions. Their results can be checked against those presented here by using the formant data listed in Appendix D. It is our hope that the method will be a useful tool for analyzing the behavior of vowels in consonantal context.

Appendix A. Linear relation between $a_{\mathrm{CC'}}(n)$ and $F_{\mathrm{C\bullet C'}}(n)$ for peripheral frames

The following will be developed for frames near the CV boundary and will apply with appropriate adaptations to those near the VC′ boundary. The early frames (from $n=1$ to $n=N_{\mathrm{P}}$, where the subscript P stands for "peripheral") are assumed to be reasonably approximated as a single-sided CV transition for which we assume $G'_{\mathrm{C'}}(n)\approx 0$ in Eq. (1) from the main text:

$$F_{\mathrm{CVC'}}(n) \approx F_{\mathrm{CV}}(n) = T_{\mathrm{V}} + (T_{\mathrm{V}} - L_{\mathrm{C}})G_{\mathrm{C}}(n) \quad (1 \le n \le N_P) \tag{A.1}$$

Averaging over vowels gives the inter-vowel mean formant as

$$F_{\mathrm{C\bullet}}(n) = T_{\bullet} + (T_{\bullet} - L_{\mathrm{C}})G_{\mathrm{C}}(n) \quad (1 \le n \le N_{\mathrm{P}}) \tag{A.2}$$

$T_{\bullet}$ and L_C are both constants, so Eq. (A.2) represents $F_{C\bullet}(n)$ as a linear function of $G_C(n)$, with slope $T_{\bullet}-L_C$ and y-axis intercept $T_{\bullet}$. Rearrange Eq. (2) from the main text as

$$K_{CC'}(n) = G_C(n) + G'_{C'}(n) + 1 \leftrightarrow G_C(n) + G'_{C'}(n) = K_{CC'}(n) - 1 \tag{A.3}$$

For the initial frames we assume $G'_{C'}(n)$ to be approximately zero so that Eq. (A.3) becomes

$$K_{CC'}(n) \approx K_C(n) = G_C(n) + 1 \leftrightarrow G_C(n) = K_C(n) - 1 \quad (1 \le n \le N_P) \tag{A.4}$$

Substitution of Eq. (A.4) into Eq. (A.2), adding and subtracting L_C, and regrouping terms yields

$$F_{C\bullet} = L_C + (T_{\bullet} - L_C)K_C \quad (1 \le n \le N_P) \tag{A.5}$$

In Eq. (A.5) the time variable n is suppressed in order to explore the relationship between the quantities K_C and $F_{C\bullet}$ for values not necessarily realized for any time frame n. Note that $F_{C\bullet} = L_C$ when $K_C = 0$ and $F_{C\bullet} = T_{\bullet}$ when $K_C = 1$.

We also adapt the ensemble scale $a_{CC'}(n)$ to near-boundary frames:

$$a_{CC'}(n) \approx a_C(n) \quad (1 \le n \le N_P) \tag{A.6}$$

Eq. (5) from the main text adapts in a parallel fashion:

$$K_C(n) = a_C(n)/a_T \tag{A.7}$$

Now Eq. (A.5) can be rewritten as

$$F_{C\bullet}(n) = L_C + (T_{\bullet} - L_C)a_C(n)/a_T \quad (1 \le n \le N_P) \tag{A.8}$$

Both $F_{C\bullet}(n)$ and $a_C(n)$ are available from the data as $F_{C\bullet}(n) \approx F_{C\bullet C'}(n)$ and $a_C(n) \approx a_{CC'}(n)$ for the N_P frames near the CV boundary. Hence the initial-C locus is obtained as the y-axis intercept of Eq. (A.8).

A relation similar to Eq. (A.8) is obtained in a similar fashion for the final frames near the VC′ boundary:

$$F_{\bullet C'}(n) = L'_{C'} + (T_{\bullet} - L'_{C'})a_{C'}(n)/a_T \quad (N - N_P + 1 \le n \le N) \tag{A.9}$$

A line fit to this relation will yield the C′ locus as its y-axis intercept. To look at the linear relation itself, it is appropriate to consider a_C, $a_{C'}$, $F_{C\bullet}$ and $F_{\bullet C'}$ as continuous variables independent of their values at particular frames n: hence the lines fit to Eqs. (A.8) and (A.9) will be continuous functions of a_C and $a_{C'}$. Suppression of the time variable n will now let us focus on Eq. (9)'s linear relationship for values of these variables not represented by any actually realized values, in particular values for a_C and $a_{C'}$ equal to 0 and 1.

The intersection of the lines from Eqs. (A.8) and (A.9) is the one point where $a_C = a_{C'}$ yields equal values for $F_{C\bullet}$ and $F_{\bullet C'}$, and hence consistent values for assigning the mean vowel target as $T_{\bullet} = F_{C\bullet} = F_{\bullet C'}$. This follows directly from the observation noted above in connection with Eq. (A.5) that this is the point where $K_C = K'_{C'} = 1$. Hence this assignment for $a_C = a_{C'} = a_T$ makes Eq. (A.7) consistent between the CV and VC′ contexts.

Appendix B. Computation methods for the linear-decomposition method

For the purposes of computation we use numerical counters for the consonants and vowels. Hence in this Appendix the notations from the main text are modified by replacing the C, V, and C′ in subscripts with the corresponding numerical counters c, v, and c'. For example, in the datasets used in the main text the vowel variable V ranges over the vowels /ɪ, ɛ, æ, a, ɒ, ʌ, ɜ/, while here its corresponding counting variable v ranges between 1 and 7. Also, let $N_{\mathbf{C}}$ be the number of initial consonants, $N_{\mathbf{V}}$ the number of vowels, and $N_{\mathbf{C}'}$ the number of final consonants.

B.1. For single contexts

This is the case in which the dataset consists of formants from a set of vowels all occurring in a single fixed CVC′ context, the case covered in Section 3 of the main text.

B.1.1. Consonant loci

Let N be the total number of frames measured for each syllable and let N_P be the number of peripheral frames to be used for locus and target estimation. The locus L_c for the initial consonant C will be the y-axis intercept of a line fit to the relation between $F_{c\bullet c'}(n)$ and $a_{cc'}(n)$ for $n = 1, 2, \ldots, N_P$:

$$F^{[d]}_{c\bullet c'}(n) \approx L_c + b_c a_{cc'}(n) \quad 1 \le n \le N_P \tag{B.1}$$

Here b_c is the slope of the linear relation. The [d] superscript on the formant expression identifies this as a formant from the data and the approximately-equals sign is a reminder that relative to the data (as distinct from the model) the relation is an approximation. The goal is to find the L_c and b_c that minimize the mean squared error in Eq. (B.1). Define the following shorthand notations for certain

summations:

$$S^{[\mathbf{A}]} = \sum_{n=1}^{N_P} a_{cc'}(n) \quad S^{[\mathbf{B}]} = \sum_{n=1}^{N_P} a_{cc'}^2(n)$$
$$S^{[\mathbf{C}]} = \sum_{n=1}^{N_P} F_{c\bullet c'}^{[d]}(n) \quad S^{[\mathbf{D}]} = \sum_{n=1}^{N_P} a_{cc'}(n) F_{c\bullet c'}^{[d]}(n) \tag{B.2}$$

The boldface superscripts are tags for identifying summation types. Then the two normal equations formed from these summations are

$$N_P L_c + S^{[\mathbf{A}]} b_c = S^{[\mathbf{C}]} \tag{B.3}$$

$$S^{[\mathbf{A}]} L_c + S^{[\mathbf{B}]} b_c = S^{[\mathbf{D}]} \tag{B.4}$$

These are linear in the unknown L_c and b_c and are easily solved for the locus L_c as

$$L_c = \frac{S^{[\mathbf{C}]} S^{[\mathbf{B}]} - S^{[\mathbf{D}]} S^{[\mathbf{A}]}}{N_P S^{[\mathbf{B}]} - S^{[\mathbf{A}]^2}} \tag{B.5}$$

The slope b_c will be useful later for determining the target scale and mean vowel target. The solution of Eqs. (B.3) and (B.4) for it is:

$$b_c = \frac{N_P S^{[\mathbf{D}]} - S^{[\mathbf{A}]} S^{[\mathbf{C}]}}{N_P S^{[\mathbf{B}]} - S^{[\mathbf{A}]^2}} \tag{B.6}$$

Finding the final-consonant locus $L_{c'}^{'}$ involves modifying Eq. (B.1) to apply only to the final N_P frames: $n = N - N_P + 1, N - N_P + 2, \ldots, N - 1, N$:

$$F_{c\bullet c'}^{[d]}(n) \approx L_{c'}^{'} + b_{c'}^{'} a_{cc'}(n) \quad N - N_P + 1 \le n \le N \tag{B.7}$$

We adapt the summations from Eq. (B.2) to apply to the final N_P frames as:

$$S^{[\mathbf{A}']} = \sum_{n=N-N_P+1}^{N} a_{cc'}(n) \quad S^{[\mathbf{B}']} = \sum_{n=N-N_P+1}^{N} a_{cc'}^2(n)$$
$$S^{[\mathbf{C}']} = \sum_{n=N-N_P+1}^{N} F_{c\bullet c'}^{[d]}(n) \quad S^{[\mathbf{D}']} = \sum_{n=N-N_P+1}^{N} a_{cc'}(n) F_{c\bullet c'}^{[d]}(n) \tag{B.8}$$

Following the reasoning for obtaining the initial-consonant locus and its associated slope, the two normal equations are as follows:

$$N_P L_{c'}^{'} + S^{[\mathbf{A}']} b_{c'}^{'} = S^{[\mathbf{C}']} \tag{B.9}$$

$$S^{[\mathbf{A}']} L_{c'}^{'} + S^{[\mathbf{B}']} b_{c'}^{'} = S^{[\mathbf{D}']} \tag{B.10}$$

The solutions to Eqs. (B.9) and (B.10) are as follows:

$$L_{c'}^{'} = \frac{S^{[\mathbf{C}']} S^{[\mathbf{B}']} - S^{[\mathbf{A}']} S^{[\mathbf{D}']}}{N_P S^{[\mathbf{B}']} - S^{[\mathbf{A}']^2}} \tag{B.11}$$

$$b_{c'}^{'} = \frac{N_P S^{[\mathbf{D}']} - S^{[\mathbf{A}']} S^{[\mathbf{C}']}}{N_P S^{[\mathbf{B}']} - S^{[\mathbf{A}']^2}} \tag{B.12}$$

B.1.2. Vowel targets

With the solutions for L_c, $L_{c'}^{'}$, b_c, and $b_{c'}^{'}$ in hand, Eqs. (B.1) and (B.7) become a pair of simultaneous linear equations in $a_{cc'}$ and $F_{c\bullet c'}^{[d]}$. The time variable n is ignored here because we are interested in these as continuous variables with values not necessarily realized at any time frame. As illustrated in Fig. 9 in the main text and by the reasoning in the last paragraph of Appendix A, the target scale and mean vowel target will be the solution to this pair of equations with $T_{\bullet}$ substituted in place of $F_{c\bullet c'}^{[d]}(n)$ and a_T in place of $a_{cc'}(n)$:

$$T_{\bullet} = L_c + b_c a_T \leftrightarrow T_{\bullet} - b_c a_T = L_c$$
$$T_{\bullet} = L_{c'}^{'} + b_{c'}^{'} a_T \leftrightarrow T_{\bullet} - b_{c'}^{'} a_T = L_{c'}^{'} \tag{B.13}$$

The solution to the two equations on the right in the unknown $T_{\bullet}$ and a_T is easily obtained by determinants as:

$$T_{\bullet} = \frac{b_c L'_{c'} - b_{c'}^{'} L_c}{b_c - b_{c'}^{'}},$$

$$a_T = \frac{L'_{c'} - L_c}{b_c - b'_{c'}} \tag{B.14}$$

The constants in Eq. (B.13) are evaluated using Eqs. (B.5), (B.6), (B.11), and (B.12). The individual vowel targets are found by substituting these $T_{\bullet}$ and a_T into Eq. (7) in the main text.

B.1.3. The $G_c(n)$ and $G'_{c'}(n)$ transition-shape functions

Once the loci and targets have been obtained, each instantiation of the main text's Eq. (1) with the initial and final consonant loci and one of the vowel targets becomes a linear relation in the unknown $G_c(n)$ and $G'_{c'}(n)$ with definite numerical values for the constant term and coefficients on its right side, and a definite formant value from the data for the left side. This allows us to use linear-least-squares optimization to find, for each time-frame n, the $G_c(n)$ and $G'_{c'}(n)$ which minimize the mean-squared error between the two sides of the equation. This will involve some summations that range over the vowels. We define these summations as follows:

$$S^{[1]} = \sum_{v=1}^{N_V} (T_v - L_c)^2 \quad S^{[2]} = \sum_{v=1}^{N_V} (T_v - L_c)(T_v - L'_{c'})$$
$$S^{[3]}(n) = \sum_{v=1}^{N_V} (T_v - L_c)\left(F^{[d]}_{cvc'}(n) - T_v\right) \tag{B.15}$$

$$S^{[1']} = \sum_{v=1}^{N_V} (T_v - L_c)(T_v - L'_{c'}), \quad S^{[2']} = \sum_{v=1}^{N_V} (T_v - L'_{c'})^2$$
$$S^{[3']}(n) = \sum_{v=1}^{N_V} (T_v - L'_{c'})\left(F^{[d]}_{cvc'}(n) - T_v\right) \tag{B.16}$$

Again, the square-bracketed boldface superscripts are tags for identifying the summation types. With the summations from Eqs. (B.15) and (B.16), we form the following two linear equations in $G_c(n)$ and $G'_{c'}(n)$:

$$S^{[1]} G_c(n) + S^{[2]} G'_{c'}(n) = S^{[3]}(n)$$
$$S^{[1']} G_c(n) + S^{[2']} G'_{c'}(n) = S^{[3']}(n) \tag{B.17}$$

The summations in Eqs. (B.15) and (B.16) can all be directly calculated from the loci, targets, and formant data and so represent definite numerical coefficients of the $G_c(n)$ and $G'_{c'}(n)$ in Eq. (B.17). Hence this pair of equations can be solved directly by using determinants as

$$G_c(n) = \frac{S^{[3]}(n) S^{[2']} - S^{[3']}(n) S^{[2]}}{S^{[1]} S^{[2']} - S^{[1']} S^{[2]}}$$
$$G'_{c'}(n) = \frac{S^{[1]} S^{[3']}(n) - S^{[1']} S^{[3]}(n)}{S^{[1]} S^{[2']} - S^{[1']} S^{[2]}} \tag{B.18}$$

These are the desired $G_c(n)$ and $G'_{c'}(n)$. Note that the calculation must be carried out once for each frame n. For example, the data for Fig. 10(b) in the main text were obtained by using these formulas 11 times for the 11 values of n.

B.2. Multiple-context case

B.2.1. Consonant loci

A locus will still be estimated as a y-axis intercept, but now one shared by a set of lines fit to the relation between $F_{c\bullet c'}(n)$ and $a_{cc'}(n)$ for peripheral-frame data, with one line for each trans-vowel consonant. Let us take the case of a locus L_c for an initial consonant C. The lines will have a form analogous to Eq. (B.1), but now with a separate slope $B'_{cc'}$ for the line associated with the trans-vowel consonant C'. These lines, one for each final consonant C′ will be of the form:

$$F^{[d]}_{c\bullet c'}(n) \approx L_c + B'_{cc'} a_{cc'}(n), \quad 1 \le n \le N_P \tag{B.19}$$

(The upper-case "B" is used for the slope $B'_{cc'}$ in Eq. (B.19) to avoid ambiguity with the slope $b'_{c'}$ from the single-context case.)

Define shorthand expressions for certain summations as follows:

$$S^{[\mathbf{AA}]}_{cc'} = \sum_{n=1}^{N_P} a_{cc'}(n) \quad S^{[\mathbf{BB}]}_{cc'} = \sum_{n=1}^{N_P} a^2_{cc'}(n)$$
$$S^{[\mathbf{CC}]}_{c} = \sum_{c'=1}^{N_{C'}} \sum_{n=1}^{N_P} F^{[d]}_{c\bullet c'}(n) \quad S^{[\mathbf{DD}]}_{cc'} = \sum_{n=1}^{N_P} a_{cc'}(n) F^{[d]}_{c\bullet c'}(n) \tag{B.20}$$

These summations over n are of exactly the same form as those in Eqs. (B.1) and (B.2), but now there is one form specific to each trans-vowel consonant C′ for the summations of types **AA**, **BB**, and **DD**, while the summation of type **CC** now involves a summation over all the final consonants C′. One equation in the unknown locus L_c and the $N_{\mathbf{C}'}$ unknown line slopes $B'_{cc'}$ is

$$N_{\mathbf{C}'} N_P L_c + \sum_{c'=1}^{N_{C'}} S^{[\mathbf{AA}]}_{cc'} B'_{cc'} = S^{[\mathbf{CC}]}_{c} \tag{B.21}$$

$N_{\mathbf{C}'}$ more equations, one for each final consonant C′ will be of the form:

$$S_{cc'}^{[\mathbf{AA}]}L_c + S_{cc'}^{[\mathbf{BB}]}B'_{cc'} = S_{cc'}^{[\mathbf{DD}]} \tag{B.22}$$

Together, Eqs. (B.21) and (B.22) represent $\boldsymbol{N_{C'}}+1$ normal equations in the unknown locus L_c and the $N_{\mathbf{C}'}$ unknown line slopes $B'_{cc'}$. These are readily solved for these unknowns by means of standard methods (see, e.g., Lipschutz and Lipson, 2013, Section 3.8, pp. 73–76).

From the forms of Eqs. (B.21) and (B.22) it is readily seen that the single-context case in the preceding subsection is a special case of the multiple-context case with the number of trans-vowel consonants set equal to 1.

Similar forms apply for a final-consonant locus $L'_{c'}$, with $L'_{c'}$, taking the place of L_c, $N_{\mathbf{C}}$ that of $N_{\mathbf{C}'}$, and $B_{cc'}$ that of $B'_{cc'}$. In other words, the roles of primed and unprimed variables are exchanged. In addition, the summations over frames n must now range over the final N_{P} frames, i.e., between $N-N_{\mathrm{P}}+1$ and N, instead of between 1 and N_{P} as was the case for Eq. (B.20). New summations to serve as coefficients in the normal equations will be of the form:

$$\begin{aligned} S_{cc'}^{[\mathbf{AA}']} &= \sum_{n=N-N_{\mathrm{P}}+1}^{N} a_{cc'}(n), \quad S_{cc'}^{[\mathbf{BB}']} = \sum_{n=N-N_{\mathrm{P}}+1}^{N} a_{cc'}^2(n) \\ S_{c'}^{[\mathbf{CC}']} &= \sum_{c=1}^{N_{\mathbf{C}}} \sum_{n=N-N_{\mathrm{P}}+1}^{N} F_{c\bullet c'}^{[d]}(n), \quad S_{cc'}^{[\mathbf{DD}']} = \sum_{n=N-N_{\mathrm{P}}+1}^{N} a_{cc'}(n)F_{c\bullet c'}^{[d]}(n) \end{aligned} \tag{B.23}$$

One equation in the unknown locus $L'_{c'}$ and the $N_{\mathbf{c}}$ unknown line slopes $B_{cc'}$ is

$$N_{\mathbf{C}}N_{\mathrm{P}}L'_{c'} + \sum_{c=1}^{N_{\mathbf{C}}} S_{cc'}^{[\mathbf{AA}']}B_{cc'} = S_{c'}^{[\mathbf{CC}']} \tag{B.24}$$

$N_{\mathbf{C}}$ more equations, one for each initial consonant C will be of the form:

$$S_{cc'}^{[\mathbf{AA}']}L'_{c'} + S_{cc'}^{[\mathbf{BB}']}B_{cc'} = S_{cc'}^{[\mathbf{DD}']} \tag{B.25}$$

Together, Eqs. (B.24) and (B.25) are a system of $1+N_{\mathbf{C}}$ equations in the unknown $L'_{c'}$ and the $N_{\mathbf{C}}$ slopes $B_{cc'}$ of the $N_{\mathbf{C}}$ lines for the $N_{\mathbf{C}}$ initial consonants. As with the system defined by Eqs. (B.21) and (B.22), this system is readily solved using standard methods.

B.2.2. Vowel targets

We adapt BC10's Eq. (8) for the mean vowel target $T_{\bullet}$ to apply only to the peripheral frames. For a trial value of the target scale a_{T}, we use Eq. (5) from the main text to convert the ensemble scale $a_{cc'}(n)$ to $K_{cc'}(n)$ as

$$K_{cc'}(n) = a_{cc'}(n)/a_{\mathrm{T}} \tag{B.26}$$

Then let $U_{cc'}(n)$ and $U'_{cc'}(n)$ be

$$\begin{aligned} U_{cc'}(n) &= K_{cc'}(n)\left[F_{c\bullet c'}^{[d]}(n) - L_c(1-K_{cc'}(n))\right] \\ U'_{cc'}(n) &= K_{cc'}(n)\left[F_{c\bullet c'}^{[d]}(n) - L'_{c'}(1-K_{cc'}(n))\right] \end{aligned} \tag{B.27}$$

$U'_{cc'}(n)$ differs from $U_{cc'}(n)$ only in its use of $L'_{c'}$ instead of L_c. The mean vowel target is then

$$T_{\bullet} = \frac{\sum_{c=1}^{N_{\mathbf{C}}}\sum_{c'=1}^{N_{\mathbf{C}'}}\left[\sum_{n=1}^{N_{\mathrm{P}}} U_{cc'}(n) + \sum_{n=N-N_{\mathrm{P}}+1}^{N} U'_{cc'}(n)\right]}{\sum_{c=1}^{N_{\mathbf{C}}}\sum_{c'=1}^{N_{\mathbf{C}'}}\left[\sum_{n=1}^{N_{P}} K_{cc'}^2(n) + \sum_{n=N-N_{\mathrm{P}}+1}^{N} K_{cc'}^2(n)\right]} \tag{B.28}$$

The square-bracketed expressions in the numerator and denominator both have two terms, the first ranging over the first N_{P} frames and the second over the last N_{P}. The individual vowel targets are then obtained by Eq. (7) from the main text.

B.2.3. The $G_C(n)$ and $G'_{C'}(n)$ transition-shape functions

The treatment of multiple CVC′ contexts is just an elaboration of the single-context case. The parallel between the two cases is immediately evident in the definition of a set of shorthand notations now defined by embedding the summations in Eqs. (B.15) and (B.16) in more elaborate summations which must now be indexed for the initial and final consonants C and C′:

$$\begin{aligned} SS_c^{[\mathbf{1}]} &= N_{\mathbf{C}'}\sum_{v=1}^{N_{\mathbf{V}}}(T_v - L_c)^2, \quad SS_{cc'}^{[\mathbf{2}]} = \sum_{v=1}^{N_{\mathbf{V}}}(T_v - L_c)\left(T_v - L'_{c'}\right) \\ SS_c^{[\mathbf{3}]}(n) &= \sum_{c'=1}^{N_{\mathbf{C}'}}\sum_{v=1}^{N_{\mathbf{V}}}(T_v - L_c)\left(F_{cvc'}^{[d]}(n) - T_v\right) \end{aligned} \tag{B.29}$$

$$\begin{aligned} SS_{cc'}^{[\mathbf{1}']} &= \sum_{v=1}^{N_{\mathbf{V}}}(T_v - L_c)\left(T_v - L'_{c'}\right), \quad SS_{c'}^{[\mathbf{2}']} = N_{\mathbf{C}}\sum_{v=1}^{N_{\mathbf{V}}}\left(T_v - L'_{c'}\right)^2 \\ SS_{c'}^{[\mathbf{3}']}(n) &= \sum_{c=1}^{N_{\mathbf{C}}}\sum_{v=1}^{N_{\mathbf{V}}}\left(T_v - L'_{c'}\right)\left(F_{cvc'}^{[d]}(n) - T_v\right) \end{aligned} \tag{B.30}$$

The *SS* notation is meant to parallel the single-*S* notation used above for Eqs. (B.15) and (B.16). Note that the inner summations over the vowels V are identical to those in the corresponding expressions in Eqs. (B.15) and (B.16). For each initial consonant C there will be one equation that is linear in its transition-shape function $G_c(n)$ and in all $N_{\mathbf{C}'}$ of the $G'_{c'}(n)$ functions for the $N_{\mathbf{C}'}$ final consonants C′:

$$SS_c^{[1]}G_c(n)+\sum_{c'=1}^{N_{\mathbf{C}'}}SS_{cc'}^{[2]}G'_{c'}(n)=SS_c^{[3]}(n) \tag{B.31}$$

The first term is for the $G_c(n)$ of current interest, while the summation represents $N_{\mathbf{C}'}$ terms for the $G'_{c'}(n)$, with one term for each final consonant C′.

In a similar fashion, each final consonant C′ will enter into one equation that is linear in its transition-shape function $G'_{c'}(n)$ and in all $N_{\mathbf{C}}$ of the $G_c(n)$ for the $N_{\mathbf{C}}$ initial consonants C:

$$\left(\sum_{c=1}^{N_{\mathbf{C}}}SS_{cc'}^{[1']}G_c(n)\right)+SS_{c'}^{[2']}G_{c'}(n)=SS_{c'}^{[3']}(n) \tag{B.32}$$

Now it is the first $N_{\mathbf{C}}$ terms for the $G_c(n)$ that are expressed as a summation and it is $G'_{c'}(n)$ that is a single term for the C′ of current interest.

Together, Eqs. (B.31) and (B.32) are the forms for $N_{\mathbf{C}}+N_{\mathbf{C}'}$ linear equations in the $N_{\mathbf{C}}+N_{\mathbf{C}'}$ unknown $G_c(n)$ and $G'_{c'}(n)$. These are solved by readily available standard methods (see, e.g., Lipschutz & Lipson (2013), Section 3.8, pp. 73–78). As with the single-context case, the calculation must be done separately for each time-frame *n*.

Appendix C. The best exponential scale factor for a given reciprocal time constant

In this appendix there are no summations ranging over consonants or vowels, so there is no need for the numerical analogs (*c*, *c*′, and v) in the exposition, but they will still be useful for writing a program to implement the method. The best value for the scale factor κ_{C} for fitting an exponential to $G_{\mathrm{C}}(n)$ with a given value of the reciprocal time constant β_{C} is found as the solution to a simple linear least-squares problem. Eq. (9) from the main text gives the exponential approximation as:

$$G_{\mathrm{C}}(n)\approx\kappa_{\mathrm{C}}\exp\left(\beta_{\mathrm{C}}n\right) \tag{C.1}$$

Table D1
*F*2 data in Hertz for Speaker 1's bVd, dVd and gVd contexts.

C	*n*	ɪ	ɛ	æ	a	ɒ	ʌ	ɜ
b	1	1856	1807	1657	1277	905	1265	1464
	2	1873	1843	1701	1290	915	1275	1484
	3	1892	1872	1735	1292	916	1284	1506
	4	1906	1890	1763	1302	914	1296	1530
	5	1918	1901	1784	1326	921	1316	1549
	6	1932	1905	1801	1344	937	1343	1562
	7	1947	1907	1813	1345	954	1368	1570
	8	1956	1911	1816	1341	978	1390	1574
	9	1954	1915	1809	1349	1019	1419	1578
	10	1939	1909	1796	1382	1078	1453	1585
	11	1914	1892	1778	1427	1143	1479	1593
d	1	1900	1787	1744	1420	1200	1441	1543
	2	1919	1813	1762	1405	1139	1428	1546
	3	1933	1840	1780	1396	1086	1417	1551
	4	1940	1858	1795	1387	1059	1416	1557
	5	1947	1868	1805	1377	1050	1422	1563
	6	1951	1876	1810	1365	1040	1425	1568
	7	1953	1880	1810	1353	1027	1426	1573
	8	1950	1881	1806	1346	1023	1438	1580
	9	1943	1877	1800	1349	1035	1457	1589
	10	1924	1867	1793	1367	1061	1472	1597
	11	1896	1850	1775	1392	1147	1479	1604
g	1	2077	1987	1889	1530	1164	1553	1655
	2	2075	1989	1879	1508	1130	1533	1633
	3	2073	1994	1874	1489	1097	1511	1626
	4	2067	1999	1874	1465	1068	1496	1629
	5	2062	1997	1876	1435	1049	1489	1635
	6	2057	1987	1878	1410	1045	1490	1637
	7	2048	1974	1876	1396	1061	1493	1636
	8	2031	1961	1871	1390	1090	1498	1634
	9	2007	1946	1859	1393	1124	1505	1633
	10	1971	1918	1838	1408	1160	1516	1630
	11	1923	1881	1812	1432	1202	1530	1623

Table D2
F2 data in Hertz for Speaker 2's bVd, dVd and gVd contexts.

C	*n*	ɪ	ɛ	æ	a	ɒ	ʌ	ɜ
b	1	2065	1892	1740	1294	1043	1309	1446
	2	2096	1952	1791	1279	1010	1312	1445
	3	2121	2005	1832	1258	978	1315	1440
	4	2139	2044	1861	1243	954	1315	1441
	5	2150	2069	1884	1240	941	1312	1458
	6	2154	2081	1907	1240	941	1310	1483
	7	2150	2080	1929	1241	960	1309	1508
	8	2134	2067	1936	1254	1006	1317	1533
	9	2109	2045	1919	1282	1082	1345	1563
	10	2077	2012	1887	1328	1179	1396	1597
	11	2042	1973	1849	1393	1273	1448	1621
d	1	2112	1978	1828	1488	1403	1562	1583
	2	2143	2009	1824	1385	1288	1505	1550
	3	2174	2030	1820	1321	1181	1448	1535
	4	2206	2043	1828	1291	1106	1396	1532
	5	2228	2055	1844	1273	1060	1355	1535
	6	2227	2066	1859	1265	1038	1331	1543
	7	2208	2075	1877	1276	1037	1326	1553
	8	2184	2072	1896	1295	1062	1343	1563
	9	2152	2054	1903	1308	1110	1379	1578
	10	2110	2025	1890	1327	1177	1427	1601
	11	2063	1989	1862	1362	1250	1470	1617
g	1	2466	2256	2193	1584	1299	1792	1687
	2	2388	2212	2110	1461	1207	1676	1632
	3	2398	2180	2028	1379	1139	1567	1596
	4	2402	2154	1967	1338	1099	1477	1571
	5	2394	2138	1932	1317	1078	1417	1561
	6	2372	2135	1914	1308	1069	1384	1569
	7	2339	2125	1902	1307	1072	1368	1586
	8	2292	2097	1893	1308	1092	1372	1599
	9	2233	2062	1891	1314	1129	1399	1612
	10	2169	2029	1886	1338	1192	1440	1627
	11	2109	1992	1870	1381	1278	1475	1638

The error for fitting frame n will be

$$e(n) \approx \kappa_C \exp(\beta_C n) - G_C(n) \tag{C.2}$$

The total squared error will then be:

$$E^2 = \sum_{n=1}^{N} e^2(n) = \sum_{n=1}^{N} \left[\kappa_C \exp(\beta_C n) - G_C(n)\right]^2 \tag{C.3}$$

The derivative of E^2 with respect to κ_C is

$$\frac{dE^2}{d\kappa_C} = 2 \sum_{n=1}^{N} \left[\kappa_C \exp(\beta_C n) - G_C(n)\right] \exp(\beta_C n) \tag{C.4}$$

Setting the derivative to zero, canceling the factor 2, and slightly rearranging the equation yields the following "normal equation" (Hamming, 1986, pp. 435–442):

$$\kappa_C \sum_{n=1}^{N} \exp^2(\beta_C n) - \sum_{n=1}^{N} G_C(n) \exp(\beta_C n) = 0 \tag{C.5}$$

Solving for κ_C yields the desired optimum as

$$\kappa_C = \frac{\sum_{n=1}^{N} G_C(n) \exp(\beta_C n)}{\sum_{n=1}^{N} \exp^2(\beta_C n)} \tag{C.6}$$

For obtaining the optimum $\kappa'_{C'}$ for a final consonant C′ from a given value of $\beta'_{C'}$ in Eq. (10) from the main text, one modifies each of the above equations by replacing $G_C(n)$ with $G'_{C'}(n)$ β_C with $\beta'_{C'}$, and κ_C with $\kappa'_{C'}$.

The solution for this linear-least-squares problem can be shown here in some detail because it involves only the single unknown κ_C. The normal equations in Appendix B are derived in a similar fashion, only with a larger number of unknown variables that yield an equal number of simultaneous linear normal equations. Hence the derivations for those normal equations would take up a lot of space. But this brief exercise is an example of how all the normal equations can be derived. Indeed, all linear-least squares problems are essentially the same: they begin by writing a formula similar to Eq. (C.3) for the total squared error. This formula is then

differentiated with respect to each of the unknowns. The resulting derivatives, one for each unknown, will be the linear normal equations in the desired unknowns. These are solved by standard methods of linear algebra.

Appendix D. *F*2 data for Speakers 1 and 2

Table D1 lists the *F*2 data for Speaker 1's bVd, dVd and gVd contexts, while Table D2 lists Speaker 2's *F*2 data for the same contexts. These can be used to test readers' implementations against results in the main text.

References

Bernard, J. R. L. (1970). Toward the acoustic specification of Australian English. *Zeitschrift für Phonetik, Sprachwissenschaft, und Kommunikationsforschung*, *23*, 113–128.

Broad, D. J. (1984). Vowels in context: Dynamics, statistics, and recognition. Technical memorandum CSRE-84-08: Center for Speech Research and Education, Voice Control Systems, Goleta CA [reprinted in W. A. Lea (Ed.) (1989). Towards robustness in speech recognition. Apple Valley MN: Speech Science Publications (pp. 373–408)].

Broad, D. J., & Clermont, F. (1984). A superposition model for coarticulation in certain CVC utterances. *Journal of the Acoustical Society of America*, *76*, S14–S15.

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

Broad, D. J., & Clermont, F. (2002). Linear scaling of vowel formant ensembles (VFEs) in consonantal contexts. *Speech Communication*, *37*, 175–195.

Broad, D. J., & Clermont, F. (2010). Target-locus scaling methods for modeling families of formant trajectories. *Journal of Phonetics*, *38*, 337–359.

Broad, D. J., & Fertig, R. (1970). Formant-frequency trajectories in selected CVC syllable nuclei. *Journal of the Acoustical Society of America*, *47*, 1572–1582.

Clermont, F. (1991). *Formant-contour models of diphthongs: A study in acoustic phonetics and computer modeling of speech* (Ph.D. thesis). Australian National University.

Clermont, F. (1992).Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching. In *Proceedings of the IVth Australian international conference on speech science & technology* (pp. 48–53). Brisbane.

Clermont, F., French, J. P., Harrison, P. T., & Simpson, S. (2008). Population data for English spoken in England: A modest first step. In *Abstract proceedings of the XVIIth annual conference of the International Association for Forensic Phonetics & Acoustics (IAFPA)*, Lausanne, Switzerland.

Clermont, F., Harrison, P. T., & French, J. P. (2007) Formant-pattern estimation guided by cepstral compatibility. In *Abstract proceedings of the XVIth annual conference of the International Association for Forensic Phonetics & Acoustics (IAFPA)*, Plymouth, U.K.

Clermont, F., & Mokhtari, P. (1998) Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency. In *Proceedings of the 5th international conference on spoken language processing* (Vol. 2, pp. 527–530), Sydney.

Crowther, C. S. (1994). *Modeling coarticulation and place of articulation using locus equations*. UCLA working papers in phonetics (Vol. 88, pp. 127–148).

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*, 769–773.

Duez, D. (1992). Second formant locus-nucleus patterns: An investigation of spontaneous French speech. *Speech Communication*, *11*, 417–427.

Flanagan, J. (1955). A difference limen for vowel formant frequency. *Journal of the Acoustical Society of America*, *27*, 613–617.

Hamming, R. W (1986). *Numerical methods for scientists and engineers* (2nd ed.). New York: Dover.

Heffner, R.-M. S. (1960). *General phonetics* (3rd printing). Madison, Wisconsin: University of Wisconsin Press (1st printing, 1950).

Houde, R. A. (1968). *A study of tongue body motion during selected speech sounds. SCRL monograph no. 2*. Speech Communications Research Laboratory, Santa Barbara, CA.

Kopp, G. A., & Green, H. C. (1946). Basic phonetic properties of visible speech. *Journal of the Acoustical Society of America*, *18*, 74–89.

Krull, D. (1987). *Second formant locus patterns as a measure of consonant–vowel coarticulation*. Phonetic experimental research. Institute of Linguistics, University of Stockholm, PERILUS V (pp. 43–61).

Krull, D. (1989). *Consonant–vowel coarticulation in reference words* (Vol. 30, pp. 101–105). *Quarterly progress and status report*. KTH Department for Speech, Music, and Hearing, STL-QPSR.

Lindblom, B. (1963a). *On vowel reduction* (Report 29). Stockholm: Royal Institute of Technology (Section 4.3).

Lindblom, B. (1963b). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Lindblom, B., Mauk, K., & Moon, S-J. (2006). Dynamic specification in the production of speech and sign. In P. Divenyi, S. Greenberg, & G. Meyer (Eds.), *Dynamics of speech production and perception (NATO science series, Series I: Life and behavioral sciences)*, Vol. 374 (pp. 7–20). Amsterdam: IOS Press.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*, 1–19.

Lipschutz, S., & Lipson, M. L. (2013). *Linear algebra (5th ed.). (Schaum's Outlines)*. New York: McGraw-Hill.

Markel, J. D., & Grey, A. H., Jr. (1976). *Linear prediction of speech*. New York: Springer.

Menzerath, P., & le Lacerda, A. (1933). *Koartikulation, Steuerung und Lautabgrenzung*. Berlin: Dümmler.

Nearey, T. M. (2013). Vowel inherent spectral change in the vowels of North American English. In G. S. Morrison, & P. F. Assmann (Eds.), *Vowel inherent spectral change* (pp. 49–85). Berlin: Springer Verlag.

Nearey, T. M., & Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, *15*, 17–24.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, *41*, 310–320.

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). *Visible speech*. New York: D. van Nostrand.

Scheffé, H. (1959). *The analysis of variance* (p. 56). New York: Wiley.

Schouten, M. E. H., & Pols, L. C. W. (1979a). Vowel segments in consonantal contexts: A spectral study of coarticulation—Part I. *Journal of Phonetics*, *7*, 1–23.

Schouten, M. E. H., & Pols, L. C. W. (1979b). CV- and VC-transitions: A spectral study of coarticulation II. *Journal of Phonetics*, *7*, 205–224.

Schouten, M. E. H., & Pols, L. C. W. (1981). Consonant loci: A spectral study of coarticulation III. *Journal of Phonetics*, *9*, 225–231.

Stevens, K. N., & House, A. S. (1963). Perturbation of articulations by consonantal context. *Journal of Speech and Hearing Research*, *6*, 111–128.

Stevens, K. N., House, A. S., & Paul, A. P. (1966). An acoustic description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. *Journal of the Acoustical Society of America*, *40*, 123–132.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*, 241–299.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as sources of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, *90*, 1309–1325.

Tabata, K., & Sakai, T. (1973). Multivariate statistical analysis of Japanese VCV utterances. *Studia Phonologica*, *7*, 31–54.

Tabata, K., & Sakai, T. (1976). Evaluation of the speaker-factor in Japanese VCV utterances. *Studia Phonologica*, *10*, 60–70.

Van Son, R. J. J. H. (1993). *Spectro-temporal features of vowel segments* (*Studies in language and language use, 3*). Amsterdam: IFOTT.

Van Son, R. J. J. H., & Pols, L. C. W. (1992). Formant movements of Dutch vowels in a text, read at fast and normal rate. *Journal of the Acoustical Society of America*, *92*, 121–127.

Yegnanarayana, B., & Reddy, D. R. (1979) A distance measure based on the derivative of the linear prediction phase spectrum. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing,* Washington DC (pp. 744–747).