Quantitative approaches to problems in linguistics

Studies in honour of Phil Rose

Cathryn Donohue, Shunichi Ishihara, William Steed (editors)

(pp. 155-169)
Linear scaling effects of phonetic context on vowel formants: A tutorial

Frantz Clermont
Linear-scaling effects of phonetic context on vowel formants: A tutorial*

Frantz Clermont

J.P. French Associates & University of York,
Forensic Speech & Acoustics Laboratory, U.K.
e-mail: akustikfonetiks@yahoo.com.au

Systematic effects of phonetic context on vowel-formant transitions are uncovered using the hypothesis (Broad & Clermont 2002) that, for a given speaker and a given formant, the relative spacing between vowels is invariant, and unaffected by consonantal context or by location within a syllable. The assumed invariance implies that, within and between contexts, inter-vowel spacings will be geometrically similar to and, therefore, linearly-scaled copies of one another. Here we explain the linear-scaling approach and describe the sequence of steps leading to scale factors, with which one can evaluate the strength of a context or compare different context effects. This is illustrated using second formant-frequency data (Al-Bannai 1992) obtained from 5 Arabic vowels in 13 fricative contexts.

1. Introduction

In many applications of acoustic-phonetic analysis of spoken language, it is common practice to select centre frames of vowel segments where phonetic context is presumed to have very little effect on formant frequencies. However, previous studies of coarticulation (Delattre et al. 1955; Lindblom 1963; Stevens & House 1963; Öhman 1966; Broad & Fertig 1970; Broad & Clermont 1987, 2002, 2010; Yang et al. 2000; Lindblom et al. 2006) have given consistent evidence that the formant transitions at all instances through a vowel segment, depend not only on the current vowel, but also on its preceding and following contexts.

An underestimated implication of phonetic-context dependency is that the representative target of a speaker's vowel may not be assumed even at those frames where formant movements appear relatively stationary. This is part of the problem posed by the phonetic interpretation of formant transitions, and by context effects as a non-negligible source of within-speaker variability. Is there a systematic behaviour in such effects and, if so, how can it be quantified?

Our approach to uncovering systematicity in context effects is to view the formant values of a speaker's vowel set as a family tied by linear scaling across frames and contexts (Broad & Clermont 2002). Section 2 motivates this shift in perspective and explains the scaling approach. Section 3 describes the formant-frequency data from 5 Arabic vowels in 13 fricative contexts (Al-Bannai 1992), which are employed for this tutorial. Section 4 describes the sequence of steps leading to scaling relationships. Section 5 summarises this work, and Section 6 looks ahead at wider validations and deeper evaluations of the scaling approach.

* This paper is based on work reported by Clermont (2010, 2009), and by Broad & Clermont (2002).
2. Phonetic context effects: A linear-scaling approach

Figure 1 illustrates what happens to a formant frequency (e.g. F2) for a set of monophthongal vowels in some fixed consonant–vowel–consonant (CVC) context. Perhaps the more common perspective arises from looking at the frame-to-frame evolution of the F2 transitions one vowel at a time. The initial and final consonants cause each transition to bend towards their respective loci at left and at right, respectively. The transition shapes are generally similar to each other, with an upward or downward translation depending on the vowel.

A different perspective emerges if the set of transitions is viewed as a whole starting from a relatively compressed pattern at the initial-consonant boundary, moving to a more widely spaced pattern near the syllable centre and then becoming more compressed again towards the final-consonant boundary. It is this variation in inter-vowel spacing from frame to frame (and from context to context) that is in focus here, and the hypothesis for it is more easily stated using a new concept, that of the vowel-formant ensemble.

![Figure 1: Hypothetical family of F2 transitions for some vowels in some consonant-vowel-consonant context. The abscissa represents time in terms of the frame number “n”. The vertical line illustrates a vowel-formant ensemble (VFE) at a given frame.](image)

2.1 Vowel-formant ensemble (VFE)

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

As suggested by the vertical line in the figure, we can also look at how the different vowels are distributed for some fixed frame (relative time position in the syllable). Unfortunately, we have had no readily understood term for this idea of how the formant varies on the “vowel axis”. In the absence of a ready-made term for this, Broad & Clermont (2002) introduced the concept of the vowel-formant ensemble (VFE), which comprises the set of formant frequencies realised for the set of vowels for the given frame and context.

The vertical line in figure 1 therefore marks a VFE, which carries its own inter-vowel spacing at the frame and for the context selected. This new perspective allows us to pose the question of whether and how a speaker’s VFEs relate to one another across different frames and contexts.
2.2 Linear-scaling hypothesis

The top two panels of figure 2 show families of hypothetical transitions for the same formant and the same set of vowels but in two different CVC contexts. The transitions have different shapes and different displacements in the two contexts. However, these complexities in the individual transitions give way to a simpler pattern when our point of view shifts from the time axis to the “vowel axis”.

To illustrate this shift, a frame is selected from each context, and its corresponding VFE is marked at that frame. These ensembles from the 2 contexts are transferred to the bottom panel, which has the same vertical scale (representing formant frequency) as the top two plots.

![Figure 2](image)

**Figure 2**: Top panels show families of hypothetical F2 transitions for a set of vowels in 2 hypothetical contexts A and B. From each context a VFE is selected at some frame and transferred in the bottom panel. VFE similarity is portrayed by the relative invariance of inter-vowel spacing within the 2 VFEs.

The horizontal arrangement of the ensembles in the bottom panel is arbitrary; it is planned simply to fit the picture to reasonable proportions. The fact that the vowel placements subdivide the 2 ensembles in the same proportions, i.e., that the ensembles are geometrically similar to one another, is shown by the fact that the lines connecting identical vowels in the 2 ensembles all intersect at the same point. (The location of this point is however meaningless, as can be seen by how it could be moved around by adjusting the arbitrary horizontal placement of the 2 ensembles.)

The arbitrary selection of the ensemble from Context A and of the ensemble from Context B illustrates our hypothesis (Broad & Clermont 2002), namely that all pairs of a speaker’s VFEs for a given formant will be similar to each other, i.e., all these VFEs will be linearly scaled copies of one another across contexts and frames. In other words, the relative spacing within VFEs will be mostly invariant with respect to context or to position within a syllable, and the scale factors will afford two systematic interpretations of context effects: (1) linear compression or expansion of VFEs; and (2) displacement of VFEs up and down in frequency.
3. Speech material and formant measurements

To illustrate the various stages of the linear-scaling approach, we use formant-frequency data from Al-Bannai’s acoustic-phonetic study of fricative-vowel coarticulation in spoken Arabic.

Five male speakers of the Gulf Arabic Dialect produced 5 long vowels in context of 13 fricatives in initial position and a fixed plosive in final position. This syllable design affords the possibility of studying effects of the fricatives on the vowels, separately from the effects of the final context. Table 1 details the phonetic components of the syllable corpus.

<table>
<thead>
<tr>
<th>Initial C</th>
<th>Place</th>
<th>Manner</th>
<th>Emphatic(^1)</th>
<th>Medial V</th>
<th>Final C</th>
</tr>
</thead>
<tbody>
<tr>
<td>/f/</td>
<td>labio-dental</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/d, t, t, t/</td>
</tr>
<tr>
<td>/θ/</td>
<td>inter-dental</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, t, t, t/</td>
</tr>
<tr>
<td>/ð/</td>
<td>inter-dental</td>
<td>voiced</td>
<td>✓</td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, t, t, t/</td>
</tr>
<tr>
<td>/s/</td>
<td>alveolar</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, d, d, d, d/</td>
</tr>
<tr>
<td>/z/</td>
<td>alveolar</td>
<td>voiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, d, d, t, t/</td>
</tr>
<tr>
<td>/ʃ/</td>
<td>palato-alveolar</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/d, t, d, d/</td>
</tr>
<tr>
<td>/χ/</td>
<td>uvular</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, t, t, t/</td>
</tr>
<tr>
<td>/s/</td>
<td>uvular</td>
<td>voiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, t, t, t/</td>
</tr>
<tr>
<td>/h/</td>
<td>pharyngeal</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/d, d, t, d/</td>
</tr>
<tr>
<td>/ʁ/</td>
<td>pharyngeal</td>
<td>voiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/d, d, d, d/</td>
</tr>
<tr>
<td>/h/</td>
<td>glottal</td>
<td>unvoiced</td>
<td></td>
<td>/i:, e:, a:, o:, u:/</td>
<td>/t, t, t, t/</td>
</tr>
</tbody>
</table>

\(^1\)Arab philologists and grammarians designate “emphasis as a consonantal rather than a vocalic element” (Al-Ani 1970) in the Arabic alphabet, while acknowledging that the set of such consonants is not necessarily the same across dialects. In addition to a primary coronal articulation, Arabic emphatics involve a secondary articulation, which is realised by retracting and lowering the tongue towards the pharyngeal wall. Al-Ani’s X-ray findings indeed point to a pharyngealised instead of a velarised classification of the secondary articulation. A well-known acoustic attribute of emphatics is their strong coarticulatory influence, which causes significant lowering of the F2 at the onglide of or even throughout neighbouring vowels (Bin-Muqbil 2006).

Table 1: Phonetic particulars of the CVC corpus employed by Al-Bannai.

Every vowel’s F1 and F2 transitions were marked at onglide and at “steady-state”. For each formant and each of these 2 frames, there were 65 values corresponding to 13 initial fricatives and 5 medial vowels, which Al-Bannai averaged over the 5 random tokens generated and over the 5 speakers’ data. He argued that this depth of averaging helped to achieve statistical stability and representativeness of the context effects manifested in his data. Thus, his paper provides only the token- and speaker-averaged values for F1 and F2.

4. VFE scaling: Procedures and results

The F2 data made available by Al-Bannai will serve as test examples for illustrating the VFE scaling steps — Construction of VFEs (section 4.1); vertical alignment of VFEs (section 4.2);
construction of the mean VFE (section 4.3); similarity of VFEs (section 4.4); calculation of VFE scales (section 4.5); and graphical summaries (section 4.6) of the scaling relationships.

4.1 Construction of VFEs

The sequence chart displayed in figure 3(a) gives an immediate view of the wide range of effects caused by the 13 initial fricatives on every vowel’s F2 at onglide. The 2 emphatic fricatives exert the greatest influence on all vowels’ onglides by lowering their F2s to within a very small range. The least influential fricatives are seen at the extremes of the chart, namely the labio-dental /f/ and the glottal /h/. Figure 3(b) shows that context effects are still noticeable but relatively weaker at steady-state. However, the 2 emphatic fricatives continue to exert a significant influence by lowering every vowel’s F2.

Figure 3: Top panels show sequence charts of Arabic vowel F2s (token- and speaker-averaged) in context of 13 initial fricatives: (a) at onglide and (b) at steady-state. Bottom panels (c) and (d) display VFE representations of the data from the top panels.
The horizontal perspective conveyed by the solid lines in figures 3(a) and 3(b) is de-emphasised in figures 3(c) and 3(d) by replacing them with dotted lines. Vowel-to-vowel spacings are connected vertically in order to produce a family of onglide and steady-state VFEs, whose 26 members differ mainly in size and vertical displacement.

4.2 Vertical alignment of VFEs

The vertical displacement of VFEs can be interpreted as a consequence of the pulling effect of each fricative’s locus on their centroids. To illustrate this effect, figures 4(a) and 4(b) display the vowel-averaged F2s obtained for each of the onglide and steady-state VFEs, respectively.

Figure 4(a) depicts a clear pattern of context-dependent effects. The lowest mean-F2 values (circa 800 Hz) for the 2 VFEs in emphatic-fricative context confirm the strongest influence noted earlier for these consonants. The mean-F2 values for the other VFEs span an appreciable range between 1356 Hz and 1863 Hz.

![Figure 4](image)

**Figure 4:** Top panels (a) and (b) display centroids of the onglide and steady-state VFEs, respectively. Bottom panels (c) and (d) display vertically-aligned versions of the onglide and steady-state VFEs shown in the bottom panels of figure 3.
The mean F2s for steady-state VFEs follow a generally similar pattern. However, the pull at steady-state towards the fricatives’ loci is expectedly less, and the displacement effect is indeed smaller. The mean-F2 values for 11 fricatives hover around 1644 Hz in figure 4(b), but the lowest mean-F2 values (circa 1400 Hz) are still obtained for the 2 VFEs in emphatic context. The strong influence of these fricatives persists as far as the steady-state regions.

In order to proceed with the assumed similarity of VFEs and their linear scaling, the mean-F2 values are subtracted from the raw VFEs shown in figures 3(c) and 3(d) to yield VFEs centered at zero. The vertical alignment of the centered VFEs displayed in figures 4(c) and 4(d) leaves relative size as the remaining contrast.

4.3 Construction of the mean VFE

The assumed invariance of inter-vowel spacing with respect to consonantal context and to position within a syllable, implies that all onglide and steady-state VFEs should be similar to one another and, by extention, they should also be similar to the grand-mean VFE. Figure 5 schematises the averaging operation leading to this VFE. It involves both the 13 onglide VFEs recalled in figure 5(a) and the 13 steady-state VFEs recalled in figure 5(b).

![Figure 5](image-url)

Figure 5: At left, the vertically-aligned VFEs are recalled in panels (a) and (b) for onglides and steady-states, respectively. The grand-mean VFE is shown at the far right.
The grand-mean VFE evidently embodies features that are representative of all of the data at hand. For this reason, it affords a key role in seeking relative measures of ensemble-to-ensemble similarity within and between all of the available contexts. In the next section, we use the mean VFE to check for consistent similarity as a pre-scaling precaution.

### 4.4 Similarity of VFEs: A pre-scaling diagnostic

Is there evidence that the 13 onglide and 13 steady-state VFEs are indeed geometrically similar to one another? It is essential to determine this in the affirmative if the linear-scaling step is to be carried out confidently. One approach to detecting departure from similarity is to look at the strength of correlations between the grand-mean VFE and each of the 26 VFEs.

![Figure 6: Distribution of the 13 onglide VFEs versus the grand-mean VFE. Every panel includes a correlation coefficient at bottom right.](image)

The scattergrams displayed in figure 6 give correlation coefficients for 11 onglide VFEs that are very convincing with values ranging between 0.9318 and 0.9994. By contrast, the coefficients of 0.7758 and 0.7341 obtained for the 2 VFEs in emphatic-fricative context
indicate some departure from similarity and therefore suggest the exclusion of these VFEs from the onglide family. However, the substantial reduction of all vowels’ onglide F2 observed in emphatic-fricative context and the consequential degradation of regularity in inter-vowel spacing provide a reasonable basis for interpreting the weaker correlations as tolerable if not encouraging.

Similarity should be even more consistent amongst the steady-state VFEs, since these are located further from the initial fricatives’ loci. Indeed, all 13 scattergrams plotted in figure 7 show tight clusterings of the 5 vowels about the diagonal line. Strong correlations ranging between 0.9922 and 0.9996 confirm this observation. In other words, similarity holds up amongst the steady-state VFEs without reservation. The non-problematic nature of the steady-state VFEs in emphatic-fricative context is also reassuring vis-à-vis our earlier argumentation for retaining their relatively less similar counterparts at onglide.

**Figure 7**: Distribution of the 13 steady-state VFEs versus the grand-mean VFE. Every panel includes a correlation coefficient at bottom right.

We have established that there is a strong trend of similarity within and between the families of onglide and steady-state VFEs. The way ahead is now cleared for the numerical implementation of linear scaling described in the next section.
4.5 Calculation of VFE scales

Analogously to the scattergram representation explored above, the scaling procedure also implicates one or more VFEs and the grand-mean VFE. Instead of computing a correlation, a least-squares line is fit through the scatter of vowels obtained by plotting the individual VFE of interest against the grand-mean VFE. The line’s slope is the estimate for the so-called *VFE scale*, which describes a proportion with respect to the grand-mean VFE.

![Diagram showing calculation of VFE scales](image)

**Figure 8:** The top panel displays the vertically-aligned VFEs at onglide — 2 members (alveolar /z/ and glottal /h/) of this family are highlighted, with the grand-mean VFE at the far right. The bottom left and right panels give scatterplots of the 2 selected VFEs versus the grand-mean VFE, separately. Line fits and slopes are superimposed.
Figure 8 illustrates the procedure outlined above and contrasts the results for two VFEs selected from the onglide family. The line fit for the alveolar /z/ context yields a slope of 0.49, while the fit for the glottal /h/ context yields a slope that is larger by a factor close to 2.5. This confirms the geometric observation that the latter VFE is more expanded than the former, but it also points to the following systematic effect — the larger the VFE scale, the weaker the context effect; and the smaller the VFE scale, the stronger the context effect! The scales obtained for the 13 onglide and the 13 steady-state VFEs are summarised in the next section.

4.6 Similarity plots

In order to bring together all scaling relations yielded by the linear-regression procedure illustrated above, each fitted VFE is positioned along the abscissa according to its scale, resulting in a plot that organises the 26 fitted VFEs from left to right in increasing order of the scales. For clarity’s sake, the plot is split in two parts — figure 9(a) for the 13 onglide VFEs, and figure 9(b) for the 13 steady-state VFEs.

![Figure 9: Similarity plots summarising linear-regression fits to scaling relations amongst the 26 VFEs, split into: (a) onglides and (b) steady-states. Fitted VFEs are positioned according to their respective scale. Context effects grow weaker with larger VFE scales.](image-url)
Perhaps the most striking pattern is the pencil of vowel-specific lines, which is formed by the juxtaposition of the fitted VFEs from *most compressed at left* towards *most expanded at right*. The pattern confirms the geometric similarity relation by the fact that each VFE subdivides the pencil of lines into the same proportions and, for this reason, Broad & Clermont (2002) justly referred to this graphical representation as a *similarity plot*.

The similarity plots in figure 9 afford the possibility of visualising the 13 fricative contexts in order of strength along the VFE scale. The ordering is more evident in figure 9(a) where the onglide VFEs are easily distinguished owing to the wide spread and range of their scales from 0.15 to 1.33.

Since the scale measures VFE size relative to that of the grand-mean VFE, it becomes possible to interpret its numerical value as a quantitative measure of context strength. A scale value > 1 indicates a relatively more expanded VFE and a relatively weaker context effect, the opposite for a value < 1, and a very strong effect of context for a value approaching 0. Accordingly, figure 9(a) shows (*emphatic dental*) and (*emphatic alveolar*) as the *most influential* contexts, and (*glottal*) and (*labio-dental*) as the *least influential* contexts as far as onglide VFEs are concerned. Between these extremes, intermediate effects of context include those from (*non-emphatic dentals, alveolars, pharyngeals*), (*palato-alveolar*) and (*uvulars*), in this order of decreasing strength.

By comparison with the wide-ranging scales observed in figure 9(a) for the onglide VFEs, figure 9(b) shows a tighter pattern of scales well above the reference scale of 1.0, which explains the concentration of the steady-state VFEs in that upper region. However, even within the smaller range of their scales (1.16 to 1.42), these VFEs display some noticeable differences in size that warrant further scrutiny.

**Figure 10:** Sequence chart of the two sets of VFE scales (onglide and steady-state), labelled with the 13 fricative contexts along the abscissa. The two sets are positively correlated with each other, except for the 2 uvular scales highlighted.
In order to enhance our perspective on the repercussion of fricative contexts at steady-state, the 13 scales for onglide VFEs and those for steady-state VFEs are thus superimposed in figure 10. The two sets of scales follow the same pattern of variation across contexts (except for the 2 uvulars), with a positive correlation near 0.9. At steady-state location, for example, the emphatic fricatives continue to cause the greatest compression of VFEs, while the labiodental and glottal fricatives continue to cause the least compression. The correlated pattern for the 2 sets of scales lends support to the established notion that context effects persist as far as the central regions of vowel nuclei often assumed to carry immunity.

5. Summary
Our first aim here has been to familiarise the reader with the new approach proposed by Broad & Clermont (2002) for uncovering systematic effects of phonetic context on vowel-formant transitions. Central to the approach are the concept of VFEs and the geometric similarity of their inter-vowel spacings within and between contexts. From the assumed similarity follows linear scaling as a practical tool, which requires only the elementary operations of taking averages and fitting lines to data.

Our second aim has been to describe this tool through a logical sequence of steps, which gradually highlighted the revealing power of linear scaling. Sections 4.1, 4.2, 4.3 and 4.5 laid out the steps for calculating VFE scales as quantitative measures of context effects. The similarity plots in figure 9 (Section 4.6) provided a graphical means of displaying (i) the invariance in relative positioning of the 5 Arabic vowels across the 13 contexts; and (ii) the relative strength of these contexts at onglides and steady-states of the vowel nuclei.

6. Ways forward
The examples offered in this tutorial show that F2s of Arabic vowels in fricative contexts exhibit a linear-scaling structure that is consistent with Broad & Clermont’s (2002) original finding of VFE similarity amongst F2s of Australian English vowels in plosive contexts. While it can be argued that the strong sensitivity of F2 to phonetic context makes it an ideal parameter for evaluating the linear-scaling approach, there is no a priori reason that precludes applicability of the approach to the lower formant F1 or even the higher formant F3. The key question is whether VFE similarity will hold up for these formants across frames and contexts. It will be instructive to elucidate this question, especially vis-à-vis the higher formants, which tend to carry more speaker-specific information than the lower formants (Peterson 1959; Liljencrants 1971; Kuwabara & Takagi 1991; Mokhtari & Clermont 1994; Mokhtari 1998; Clermont & Mokhtari 1998; Rose & Clermont 2001; Clermont et al. 2008).

Al-Bannai’s two levels of averaging (over tokens and speakers) of his F2 data made it possible to focus on the description and the illustration of the linear-scaling steps, while leaving unattended issues of variability and robustness. However, a rigorous evaluation of VFE-scaling for context effects will require: (i) dealing with the available data on a per-speaker basis in order to avoid compounding errors due to speaker differences in vocal-tract structure; (ii) retaining the token-to-token variation in inter-vowel spacing as a baseline measure for assessing the goodness of the line fits; and (iii) eventually gaining a working knowledge of the mean-VFE’s statistical fluctuations. The robustness of this ensemble will expectedly improve as additional tokens, frames and contexts are recruited in the averaging
process, but it is not yet known how many of each of these will suffice to prevent erosion of the usefulness of the mean VFE as a reference ensemble.

In addition to the statistical analyses outlined above, one productive step towards wider validations of the linear-scaling approach will be to look beyond Al-Bannai’s corpus of Arabic syllables. Their pre-determined phonetic content and their controlled elicitation do not reflect many practical situations, in which vowels and consonantal contexts occur in unpredictable combinations and frequencies. Uncontrolled speech samples will therefore necessitate new strategies in order to satisfy “context and phonetic comparability” as a requirement for being able to seek scaling relationships amongst VFEs. This is reminiscent of Rose’s (2002) requirement of “situational and linguistic comparability” of forensic speech samples for being able to achieve successful identification amongst speakers.

References


