Focalization within Dispersion predicts vowel inventories better*

Roy Becker-Kristal
roybecker@humnet.ucla.edu

ABSTRACT

A revision of the dispersion-focalization theory (DFT) is presented, with focalization based upon the Center-of-Gravity-Effect embedded in the perceptual representation of vowels. The revised model dispenses with scalar parameters, perceptual discontinuities and hybrid energy function, and yet the predictive accuracy of its simulation reported here surpasses that of its predecessor and of other models.

1. INTRODUCTION

Although Lindblom’s theorization that perceptual dispersion plays a key role in the organization of vowel inventories seems unquestionable, only moderate improvement has been achieved in the predictive power of dispersion-based models since the earliest simulations [19]. The crucial prediction of /ieæou/ in 5-vowel inventories was achieved with a practical (though not theoretical) drawback of using whole spectra instead of formant frequencies [20]. Only coarse predictions and anecdotal improvements have been reported in various studies that changed certain aspects of the dispersion model, including the frequency unit, the contribution ratio of the perceptual dimensions to the distance metric, and non-perceptual factors such as articulatory effort or organizational symmetry [3,4,5,14,22]. The long-standing problem of predicting too many high vowels in 7-vowel inventories has been overcome only recently, as a model that ‘reintroduced’ formants (by using noise to cancel the contribution of spectral troughs or by using temporally resolved auditory excitation to boost spectral peaks) predicted a /ieæaoou/ inventory [10]. Prediction of non-peripheral strategies of larger inventories such as series of front-rounded (but also central-unrounded) vowels seems impossible, as the non-peripheral series which is equidistant from the front and back peripheries is that of back-unrounded vowels (with F2≈10Bk).

The Dispersion-Focalization Theory (DFT, [24]) has overcome some of the shortcomings of dispersion-only models and achieved better predictions. DFT is unique in incorporating a bias towards

* This paper has been submitted to the 16th International Congress of Phonetic Sciences. I would like to thank Patricia Keating and Sun-Ah Jun for their reviews and comments. I also thank the many discussants during presentations of earlier versions of this work at the 4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan and at the Phonetics seminar of the Linguistics department at UCLA, in particular Edward Flemming and Colin Wilson. Finally, this work would have never taken off the ground without the accompaniment, encouragement and suggestions given (and challenges posed) to me, during its initial stages, by Bruce Hayes, to whom I am mostly indebted.
‘focalization’, or formant convergence, into its perceptual model. Focalization creates perceptual ‘quantization’, which is further facilitated by the Center-of-Gravity Effect (CoGE) of perceptual integration of formants dispersed less than 3.5Bk apart [9], and is thus preferred in terms of Stevens’ Quantal Theory [25]. Focalization constitutes a second component (E_F) added to dispersion (E_D) in an inventory’s energy function. E_F is based on distances between adjacent formants within vowels, while E_D is based on F_1xF_2’ weighted Euclidean distance between vowel pairs, where F_2’ explicitly implements the CoGE between F_2, F_3 and F_4. Two scalar parameters, λ and α (0≤λ,α≤1), respectively regulate the contribution ratios F_2’/F_1 in the distance metric and E_F/E_D in the energy function.

Inventory predictions were compared with typological data of real language inventories [23]. Best match was achieved with 0.17≤λ≤0.27 and α=0, yielding the following predictions: 3 vowels: /iau/~iao/; 4: /ieau/; 5: /ieaou/~ieaou/; 6: /ieieoou/~ieieoou/; 7: /ieieoou/~ieieoou/. With considerable contribution of E_F (0.3≤α≤0.4) the model manages to predict the emergence of /y/ (with close F_2 and F_3) in stable sub-optimal 7-vowel inventories.

This model, however, has some theoretical drawbacks: First, the perceptual plasticity represented by α and λ has not been motivated. Second, the CoGE is applied selectively, ignoring the formant pairs ‹F_1,F_2› and ‹F_1,F_0›. Finally, the 3.5Bk discrete threshold dependency of the CoGE yields a discontinuous and fluctuating output range for the calculation of F_2’, as demonstrated in Fig. 1. Such discontinuities have been challenged in studies that showed that perceptual shifts upon formant proximity are gradual [1,11,17].

**Figure 1:** The original model’s F_2’ as a function of (a) F_2 (F_3=14.6Bk) and of (b) F_3 (F_2=12.9Bk). F_4=16.2Bk in both.

In addition, certain predictions are problematic, such as ‘point’ /u/ (F_2<600Hz) in small inventories (/u/ is typically somewhat fronted, with 700Hz< F_2<1000Hz, in inventories lacking a vowel between /u/ and /i/) and the emergence of /y/ not as part of a front-rounded series (as in e.g. Germanic and Finno-Ugric languages). Moreover, predictions for larger inventories were not reported, although the authors argued that inventories with 8 and 9 vowels were still viable [23].

The current paper presents a drastically revised DFT model that overcomes the drawbacks mentioned, as described in section 2. Section 3 presents and discusses the results of a Matlab simulation of the revised model.
2. THE REVISED DFT MODEL

2.1. Reinterpreting the Center-of-Gravity Effect

The CoGE is reinterpreted here as a function from an auditory representation of any two adjacent formants, each with its frequency and default weight, to a perceptual representation, where interpolated frequency shifts and weight boosts may occur. This behavior represents bi-directionally amplified neural response to the range between two spectral peaks, due to the ‘smeared’ nature of auditory excitation, resulting in perceptual peaks that are closer to each other than the original auditory peaks. The effect accelerates, from negligible to peak merger, in a gradual, non-linear manner free from discontinuities, as the between-formant interval decreases from above 3.5Bk to below 1.5Bk (see also [17]). The CoGE-based weight-boosting constitutes the focalization bias in the revised model, thus dispensing with \(E_F\) and \(\alpha\). As a coarse approximation, the CoG is always one quarter of the way from the louder formant (which is shifted and boosted moderately) to the softer (which is substantially affected). Fig. 2 illustrates this interpretation of the CoGE as a function of the interval between formant pairs with a louder lower formant and CoG at 9.8Bk.

Figure 2: CoGE converting auditory (dots) to weighted perceptual (variably-sized circles) representation of formant pairs (the lower formant is louder) with CoG at 9.8Bk.

2.2. Acoustic and perceptual representation of vowels

The CoGE function is integrated into the derivation of a vowel percept from its formant frequencies, including the aperture formant (AF, always \(F_1\)), the front cavity formant (FCF, usually \(F_2\)), the back cavity formant (BCF, usually \(F_3\)), \(F_4\) and \(f_0\). It is assumed that cavity affiliation is recovered, so that FCF always determines the vowel’s chromaticity profile, even when FCF is \(F_3\) (higher-fronter vowels [12,18]), as shown in studies of \(F_2^2\) [2,4,8]. The following fixed intensity hierarchies are also assumed: AF>\(f_0\); AF>FCF>BCF>F_4.

In the simulation, vowel qualities were created in acoustic terms directly. A male-speaker’s AFxFCF ‘cardinal vowel’ plane was created with the ranges (in Hz):

\[
1. 230 \leq AF \leq 830 \\
2. 570x2^{600} \leq FCF \leq 1650 + 21x^{\frac{AF-230}{25}} + 18x1.174^{\frac{AF-230}{25}}
\]
Both AF and FCF determined the range of BCF, which was higher for lower and backer vowels (higher AF and lower FCF), and wider for non-peripheral vowels (FCF closer to its mid-range), reflecting the role of BCF in distinguishing unrounded (higher BCF) from somewhat fronter rounded (lower BCF) non-peripheral vowels with the same FCF, e.g. u/u [12,18,26]. FCF and BCF frequencies overlapped at the higher-fronter region of the vowel space. Articulatorily constrained variants of the AFxFCF plane were delineated by 0.25Bk vertical margins and 0.75Bk horizontal margins for moderate dorsum displacement, and horizontally at AF≤720Hz and AF≤610Hz for intermediate and slight jaw-lowering, respectively. These discrete constraints replace the scalar \( \lambda \) parameter, and reflect the dependency of a particular language’s acoustic space on rigid articulatory constraints often related to quantity contrasts, stress prosody and co-articulatory leniency of consonants. In addition, 65 vowel prototypes with full formant specifications were defined and used for categorizing vowel qualities resulting from the simulation process. The AFxFCF plane is shown in Fig. 3.

**Figure 3:** AFxFCF plane for extreme (solid line) and moderate (dotted line) dorsum displacement, boundaries for slight and intermediate jaw displacement conditions (dashed lines), area of FCF~BCF overlap (grayed) and the 65 vowel prototypes.

Within this space, a vowel’s AF, FCF and BCF frequencies were determined by three fraction variables. Actual ‘pinching’ of FCF and BCF, which is impossible in real vowels due to cavity coupling, was resolved whenever needed by dispersing them to a minimal distance (in Hz) close to the frequency of AF [25]. By keeping FCF well above AF, AF~FCF overlap as in pharyngealized back vowels was avoided (secondary articulation was ignored as in all other models). \( F_4 \) was determined by rule 600~700Hz above the higher among FCF and BCF, up to 3400Hz, and \( f_0 \) was fixed at 125Hz. All frequency ranges, and in particular the FCF~BCF overlap region, were kept as close as possible to those demonstrated in [12,18].

With its formant frequencies set, a vowel was assigned a percept, consisting of a sonority profile (SP) and a chromaticity profile (CP), as follows: BCF and \( F_4 \) were integrated into a ‘speaker formant’ (SF) [16,21], one quarter the way from BCF to \( F_4 \). Within the \( f_0 \)~SF auditory range, AF and
FCF respectively determined SP and CP, each of which consisted of a frequency and weight. For SP, the default frequency was AF (in Bk) and the default weight was 0.7, while for CP, the default frequency was FCF (in Bk), and the default weight was 0.14. The CoGE function, applied separately on the pairs \( \langle \text{AF, } f_0 \rangle \), \( \langle \text{FCF, SF} \rangle \) and \( \langle \text{AF, FCF} \rangle \), could boost the weights of SP and CP up to 1.0 and 0.8 respectively (and also shift their frequencies). The weight ranges are assumed to be arbitrary yet fixed idiosyncrasies of the perceptual system that yield, for two vowels with medial weights (0.85 for SP, 0.47 for CP), the CP/SP contribution ratio of 0.3 (=0.47²/0.85²), as in [3].

2.3. Dispersion score

The dispersion score for a given vowel pair is the SPxCP weighted Euclidean distance between the two percepts. The weight of each dimension is the product of the relevant weights from each vowel. Thus, if /\text{a/}\) has SP of 5;0.7 (frequency in Bk;weight) and CP of 10.8;0.16, and /\text{o/}\) has SP of 3.9;0.9 and CP of 6.6;0.58, then the score for this pair is:

\[
\begin{align*}
  d \!\langle \text{a, o}\rangle &= (0.7 \times 0.9 \times (5-3.9)^2 + 0.16 \times 0.58 \times (10.8-6.6)^2)^{\frac{1}{2}} = 1.549 \\
\end{align*}
\]

Rather than fixed or parameterized CP/SP contribution ratio as in [3,24], here this ratio depends on perceptually-transformed spectral properties of the individual vowels. This is a step towards whole-spectrum representation (even if coarse) while retaining testability against descriptive phonetic studies, which typically report only formant frequencies. This current approach is a reasonable compromise, in particular as the prominent role of formants has been acknowledged and implemented in the most recent (and most successful) whole-spectrum based model [10].

Notice that dispersion is interpreted here in terms of reliable distinctiveness, rather than distance. Two vowels can be rather distant from each other, but only moderately dispersed, if their weights are small. Another pair of more proximate vowels might be better dispersed owing to significantly greater weights. Indeed, it is claimed that focal regions can be more populated than others because they render small distances more reliable. Thus, the likelihood of front founded and central unrounded vowels is similar, because the higher reliability weight of CP of the front-rounded vowels (due to FCF~SF proximity) compensates for their shorter distance from front-unrounded vowels. Similarly, AF~f_0 proximity allows finer height contrasts for higher than for lower vowels.

Finally, it is assumed that an inventory’s score is determined by the score of the its least dispersed vowel pair [6,15]. This score is increased marginally by all other vowel pairs, only to the extent needed to distinguish different inventories with the same worst pair.

2.4. Inventory stabilization and evaluation

In the simulation, inventories were optimized by time-constrained simulated annealing: For a given size \( n \) and known jaw/dorsum displacement constraints, an inventory was created with \( n \) vowels with
random acoustic qualities, yielding a particular score determined mostly by the least dispersed vowel pair. These two vowels were then changed by small random perturbations of their variables, thus changing their dispersion and the inventory score. If the new score was higher, the resulting inventory served as the source for the next perturbation (perhaps with a different worst pair). This improvement-by-perturbation process was repeated cyclically, and was terminated after 40+5x\(n\) cycles without significant improvement or after 3000+500x\(n\) cycles in total. In order to prevent this cyclic process from getting stuck at a minor local extremum, the source inventory was also replaced by a new one with a lower score, if a low (and gradually decreasing) probability game was successful.

This process approximates inventory learning: inventory size can be inferred from same/different relationships between vowel exemplars stored in the learner’s lexicon (e.g. minimal pairs), and the dorsum and jaw constraints can be inferred from observing other speakers. Since running speech exemplars cover the whole (constrained) acoustic continuum, the search for prototypes must cover this whole space, which is approximated by a random starting point and random perturbations. The time constraint mimics the size dependency and finiteness of the acquisition period.

An inventory attained by this process is predicted to be perceptually learnable. The set of such inventories may be narrowed down by additional constraints such as an articulatory ban on low rounded vowels or an organizational requirement for symmetry in non-low vowels [4]. Such non-perceptual (and, in the case of symmetry, non-phoneic) constraints are ignored here. It is also acknowledged that vowels might undergo quality shifts motivated by external factors (which may be phonetic, socio-linguistic etc.), yielding inventories that cannot be predicted. It is assumed, however, that such inventories are not perceptually learnable, and undergo reorganization within the next generation of speakers.

In order to explore the set of learnable inventories predicted and how they mimic unity and diversity of strategies in real language inventories, \(3x\(n\)\) inventories were optimized for each combination of size and articulatory constraints. Using the 65 prototypes, the vowels of the optimized inventories were categorized by a simple distance metric, and inventories of identical categorical composition were counted and grouped under the best-scored inventory among them.

3. RESULTS AND DISCUSSION

Fig. 4 summarizes representative results by plotting, for selected configurations, two predictions including the top-scored one. Most other inventories predicted for these configurations differ from those shown here in one adjacent category (e.g. /\(\lambda\)/ instead of /a/). The full set of predictions (with formant frequencies) may be obtained from the author.

While the results demonstrate some over-generation, they clearly mimic universal traits. Compatibility with all previous models for \(3\leq n\leq 7\) is maintained in the sense that all their accurate predic-
tions are learnable by the current model, and usually emerge more than once and for various configurations. Genuine improvements for these inventory sizes include (a) /iəʊ/ reminiscent of a vertical 3-vowel inventory, (b) consistent emergence of /iaou/ as a learnable 4-vowel inventory, (c) /iæʊəʊ/ as a learnable 5-vowel inventory (d) non-point /u/ (FCF≈850Hz) and /ʊ/ (FCF≈1050Hz) in 5-vowel inventories even for extreme dorsum displacement, (e) a tendency of back rounded vowels to be slightly lower than corresponding front unrounded vowels, (f) identical scores for intermediate displacement 5-vowel and extreme displacement 7-vowel inventories [14], and (g) series of front-rounded vowels /yəʊ/ in 7-vowel inventories (with the /ɔ/~/æ/ height asymmetry similar to e.g. Hungarian short vowels).

Predictions for 8- and 9-vowel inventories are also encouraging: Peripheral-only strategies lose prestige but remain learnable (for intermediate dorsum and extreme jaw displacement), while a competition between strategies with one and two internal vowels emerges. For the latter strategy, both front-rounded (e.g. Finnish, Norwegian) and central-unrounded (e.g. Korean) series are constantly predicted, as well as the (attested) /yəʊ/ and (over-generated) /iʊəʊ/ strategy, which would have been filtered out (like most other over-generated predictions) by a passive phonological symmetry constraint. For extreme dorsum and slight jaw displacement as in inventories with rather high low vowels (F₁<650Hz) but near-cardinal vowels otherwise, e.g. Turkish, Swedish and Estonian, a strategy of three internal vowels emerges, including the /ɔ/, o~/æ/ series and a high vowel with FCF≈1300Hz (/u~ɪu/). This prediction is rather accurate for the languages mentioned, and the extra-front /y/ required to accommodate /u~ɪu/ is true at least for Swedish [13].

Figure 4: Simulation results (count;score) for selected configurations (boldfaced: n, dorsum-displacement, jaw-displacement). ‘Ex.’, ‘In.’ and ‘Sl.’ respectively stand for extreme, intermediate and slight displacement conditions.
In summary, this revised DFT-model overcomes the theoretical drawbacks of its predecessor and yet predicts more accurately inventories with up to 9-vowels, and more importantly, shows that the last word about the role of perception in shaping vowel inventories has not yet been said.

4. REFERENCES

perception: matching and discrimination studies. *UCLA-WPP* 83, 47-54


