PERCEPTUAL CONTRAST AND STABILITY IN VOWEL SYSTEMS: 
A 3-D SIMULATION STUDY

J.L. Schwartz(1) L.J. Boë(2) P. Perrier(1) B. Guérin(1) P. Escudier(1)

(1) INPG/ENSERG - 46 Av. Felix-Viallet 38031 Grenoble Cedex
(2) Université Stendhal - Domaine Universitaire 38040 Grenoble Cedex BP 25X

ABSTRACT

Since the beginning of the 70s, several attempts have been made 

to explain the phonetic structure of vowel systems by introducing 

extralinguistic principles, listener-orientated (perceptual contrast 

and stability) or speaker-orientated (articulatory contrast and 

economy). The best predictions have been realized with the 

perceptual contrast theory (PCT), but two main problems remain: 

the too great number of high non-peripheral vowels and the 

impossibility to predict the [i,y,u] series within the high vowels 

set. We try to get rid of these difficulties while staying within the 

field of listener-orientated principles. First, we study the PCT in 

the F1-F2-F3 space, in order to better account for the role of 

higher formants in the perception of front vowels. In this space, 

we show that the problem of high non-peripheral vowels can be 

solved with an increased weight of F1, but the case of [y] can 

only be understood by reinforcing the stability of the [i]-[y] 

pair. This is done by means of a "focalization" principle, 

according to which vowels with strong formant convergence - 

[i] characterized by a strong F3-F4 convergence and [y] 

characterized by a strong F2-F3 convergence - would be 

perceptually preferred.

INTRODUCTION

Since the beginning of the 70s, a number of proposals have 

been made to explain the general structure of vowel systems on 

the basis of extra-linguistic principles. The first - 

listener-orientated - idea is that sounds can be used as a support 

for communication only if they can provide reliable linguistic 

oppositions. This led to two more or less complementary theories: 

(i) distinct vowels have to be perceived as different, hence 

the corresponding spectra have to be as distant as possible - the 

so-called "perceptual contrast theory", PCT in the following, due 

to Liljencrants and Lindblom [8,9]: 

(ii) it must be possible to produce a single vowel with 

slightly variable articulatory positions, hence a vowel must 

correspond to "stable" articulatory configurations, where 

articulatory variations have poor acoustic consequences - the 

so-called "quantal theory" introduced by Stevens [16,18].

One can also think of speaker-orientated proposals. Various 

suggestions have been made from this point of view. 

(i) distinctivity of articulatory configurations [9]; 

(ii) cost of articulatory configurations [5]; 

(iii) precision of the articulatory control [1].

While listener-orientated arguments can be quantitatively 
evaluated on firm grounds, speaker-orientated ones can be, at 
this time, only qualitatively estimated. That is why we decided to 
try to go as far as possible in the perceptual domain, in order to 
be able to describe precisely what can and cannot be explained in 
the structure of vocalic systems with listener-orientated theories.

1. SOME PROBLEMS IN 
PREDICTION OF VOWEL SYSTEMS

The first simulation based upon the perceptual contrast theory 
(PCT) introduced by Liljencrants and Lindblom (L&L in the 

following) gave very promising results [8]. Further 

developments by Crothers [6], Lindblom [9] and the very 

thoughtful study by S.F. Disner [7] on the UCLA Phonological 

Segment Inventory Database UPSID [10] clearly show that the 

main trends of vocalic systems are well captured by the PCT. 

However, a few problems remain. We shall consider here only 
those that are linked with high vowels, assuming that most other 
difficulties would be ruled out by introducing regularization 

principles.

1.1. Importance of peripheral vowels

One of the strongest drawbacks of the first simulation by L&L 

concerns the respective number of "peripheral" (front unrounded 

or back rounded) versus "non peripheral" (front rounded, central 

or back unrounded) vowels: the 6-vowel prediction contained 3 

high vowels [i, ,u] and prediction with 7 or more contained 4 

high vowels [i ,u] while phonological inventories show that 

systems with only 2 high vowels are common for 6- or 

7-vowel systems (namely [i ,e ,o ,a] and [i ,e ,u ,o ,a]) 

and that [i ,u] appear together only with a total of at least 9 

vowels.

This severe drawback was at the basis of Lindblom's further 

modification of the model [9], but the introduction of new 
distances based on psychoacoustic concepts were not really 

conclusive, and not only in this respect.

1.2. The case of [u]

A fact that generally received little consideration in studies on 

vocalic systems concerns the difference in status within the set of 

non-peripheral high vowels. A first statistical study on the 

UPSID database shows that 24 of the 312 languages in the base 

(7.7%) contain [u], 23 (7.2%) contain [u] and 44 (14.1%) contain 

[u] or [u]. This seems to fit well with PCT: vowels 

which are the most at the center are the most probable, while 
vowels displaced towards the left [i] or the right [u] have 
amost equally lowered probabilities of appearance.

However, closer inspection of the base reveal strong 

asymmetries between [i] and [u]. On 23 occurrences, [u] 

appears 10 times with [i] only, 8 times with [i,u] and 5 times 

with [i,y,u]. The [i,y,u] structure is best predicted by the 

PCT, the corresponding F2 values of these vowels being 

respectively around 16, 13, 10 and 7 Barks. As was noticed by 

Crothers [6], the [i,u,u] structure is also acceptable,
considering that $F_2$ for [u] can reach about 1400 Hz, which gives an $F_2$ series of 16, 11 and 7 Barks. Finally, the [i,u] structure is of course quite unpredictable, and can perhaps be interpreted by the low structural articulatory cost of an [i,u]-based system with no lip action. On the contrary, on 24 occurrences in the base, [y] never appears with [u] only, but appears 14 times with [i,u] and 10 times with [i,u] and one vowel within the [u , i, u] set. The rather great number of [i , y, u] structures is this time in very strong contradiction with PCT, with a corresponding $F'_2$ series of 16, 13 and 7 Barks. The existence of this structure thus constitutes a major problem that we tried to deal with in our simulations.

II. 3-D PCT-BASED SIMULATIONS

A model similar to the one proposed by L&L is defined by (1) an acoustic domain where vowels must be confined, due to articulatory constraints, (2) a distance defined on the considered acoustic parameters - formants or effective formants - (3) a criterion to select the "best" configurations for a fixed number of vowels. Let us consider our choice for each of these three points.

II.1. Definition of an acoustic F1-F2-F3 domain

We chose to use a simulation in the F1-F2-F3 domain, in order to better account for the role of higher formants in the perception of high vowels. The extensive study in our laboratory [13] of the articulatory model elaborated by Maeda [11] allowed us define a "maximal space" in which all of the 200000 points of our articulatory-acoustic dictionary are located (Fig.1). This maximum space was parametrized by $F_3$ in a rather efficient manner (Fig.2). Hence we dispose of a set of analytic equations that allow us decide whether a given (F1,F2,F3) configuration is acceptable or not.

II.2. Distances in the F1-F2-F3 space

These distances are necessary for computing an energy function:

$$E_0 = \sum_{ij} (1/d_{ij})^2$$

(1)

where $d_{ij}$ describes the spectral distance between two vowels prototypes in the simulation. $E_0$ must then be minimized for optimal contrast. Since the global spectral distances used by Lindblom [9] were not really convincing, we limited ourselves to direct computation on the formant values. Considering two points $p_a(F_{1a},F_{2a},F_{3a})$ and $p_b(F_{1b},F_{2b},F_{3b})$ three distances were computed:

(i) an (F1,F2,F3) Euclidian distance $d_1$

$$d_1 = [(F_{1a}-F_{1b})^2 + (F_{2a}-F_{2b})^2 + (F_{3a}-F_{3b})^2]^{1/2}$$

(2)

(ii) an (F1,F2) Euclidian distance $d_II$

$$d_{II} = [(F_{1a}-F_{1b})^2 + (F_{2a}-F_{2b})^2]^{1/2}$$

(3)

where $F_2$ was computed from $F_2$, $F_3$ and $F_4$ thanks to a model we proposed in a previous work [12], and $F_4$ was supposed fixed at 3350 Hz.

(iii) and (F1,F2) weighted Euclidian distance $d_{III}$

$$d_{III} = [(F_{1a}-F_{1b})^2 + \lambda (F_{2a}-F_{2b})^2]^{1/2}$$

(4)

where $\lambda$ could be chosen at any value lower than 1, assuming that higher formants play less part in vowel phonetic quality than lower ones. This last distance was introduced to implement in an easier and more general way - though less related to experimental knowledge in auditory perception - Lindblom's suggestion that $F_1$ weight was underestimated in the original simulations by L&L.

![Fig.1 - 3D acoustic domain defined by Maeda's articulatory model](image1.png)

A series of F1-F2 cuts for fixed $F_3$ are presented, with $F_3$ varying from 2000 Hz to 2900 Hz

![Fig.2 - Parametrization by $F_3$ of the 3D space presented in Fig.1](image2.png)

A series of F1-F2 cuts for fixed $F_3$ are presented, with $F_3$ varying from 2000 Hz to 2900 Hz

II.3. Criterion for the choice of vowel configurations

Various criteria have been proposed in the literature, namely systems with the highest contrast [8], sets of systems with the highest contrasts [9] or systems with sufficiently high contrast [6]. We also use here a new criterion: select only those systems that are associated to locally highest contrasts - local minima in the energy function - and hence which are stable when small acoustic modifications are applied.
II.4. Results

As it could be predicted, results with \( d_1 \) are very bad, since the unweighted F3 dimension allowed two vowels to be distinkted only on this dimension, which produced completely unacceptable configurations.

Results with \( d_2 \) are of course much similar to the ones obtained by L&L. Stable systems with the highest contrast are listed in Table I for 3 to 9 vowels, and presented in Fig.3.

<table>
<thead>
<tr>
<th>number of vowels</th>
<th>predicted system</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>[i,a,u]</td>
</tr>
<tr>
<td>4</td>
<td>[i,e,a,u]</td>
</tr>
<tr>
<td>5</td>
<td>[i,e,a,u,i]</td>
</tr>
<tr>
<td>6</td>
<td>[i,e,a,o,u,i]</td>
</tr>
<tr>
<td>7</td>
<td>[i,e,a,o,u,i,y,u]</td>
</tr>
<tr>
<td>8</td>
<td>[i,e,a,o,u,i,y,u,w]</td>
</tr>
<tr>
<td>9</td>
<td>[i,e,a,o,u,i,y,u,w]</td>
</tr>
</tbody>
</table>

Table I - Best systems in the 3-D PCT

Weighting of F2 in \( d_3 \) by the \( \lambda \)-factor could in theory help us predict the low number of high vowels observed in the UPSID base. In fact, we observed that the [i,e,u,o,ɔ] and [i,e,u,o,ɔ,a] systems are stable only for very low \( \lambda \)-values (\( \lambda \approx 0.2 \)). It shows that F1 should have a weight about 5 times greater than F2 to correctly predict the data. We shall come back later on this seemingly too great ratio. Finally, the [i,y,u] structure is never predicted, which shows once more that the PCT is not enough to account for the case of [y].

III. FOCALIZATION

III.1. Introduction of the focalization principle

The existence of the [i,y,u] structure shows that the [i]-[y] pair must be associated with a lower level of energy than one would predict on the basis of (F1,F2) distances, and hence that it can introduce a new term in the energy function. This term could be found in the “focalization” concept developed in our laboratory [3,4]. We proposed in this concept that “focal points” - articulatory configurations with maximum formant convergence due to an exchange of affiliation of acoustic resonances - were indeed reference or preferred points in the vowel space. We then showed that F2-F3 convergence for [y] and F3-F4 convergence for [i] allowed definition of two stable domains separated by a very strong instability in the articulatory-perceptual passage, which led to the existence of a “natural” frontier and two quantal classes - in the meaning of Stevens - around [i] and [y] [1]. Finally, we showed in a perceptual experiment [15] that formant convergence was associated with a better perceptual stability, while patterns with equal F2-F3 and F3-F4 distances were more difficult to memorize.
III.2. Simulations based on contrast and focalization

We decided to add to the classical terms in the energy function (namely $E_0$) a new set of terms for focalization, diminishing the energy of configurations with vowels close F1 and F2, F2 and F3 or F3 and F4. Our new energy function was now:

$$E = E_0 + E_{12} + E_{23} + E_{34}$$

with

$$E_{12} = \sum_i 1/(F_{2i} - F_{1i})^2$$

$$E_{23} = \sum_i 1/(F_{3i} - F_{2i})^2$$

$$E_{34} = \sum_i 1/(F_{4i} - F_{3i})^2$$

We used only distance $d_{ij}$ with various $\lambda$ values in the $E_0$ term.

III.3. Results

We obtained three main results.

(i) Introduction of terms $E_{12}$ and $E_{23}$ reinforcing stability for close F1 and F2 or F2 and F3 results in a greater number of peripheral vowels. Indeed, the $[i, o, E, U, O, 3]$ and $[i, O, E, U, O, 3, B]$ systems are obtained for a more reasonable F2 weight ($\lambda = 0.3$).

(ii) Terms $E_{23}$ and $E_{34}$ increase the stability of configurations including pair [i,y], [i] with a maximal F3, [y] with a minimal distance F3-F2). We give in Fig. 4 a case of stable system with [i,y,u] that could provide by regularization the basis for the French vowel system.

(iii) Focalization terms $E_{12}$ and $E_{23}$ diminish the relative stability of vowels with F2 far from both F1 and F3, which fits well a prediction that vowels must "choose" between an F1-F2 and an F2-F3 mass, and escape a "hole" around (F1=250Hz, F2=1000Hz), with no or few vowels in the phonological inventories [2,14,17].

Fig. 4 - An 8-vowel stable system with the [i,y,u] series
Vowels are this time presented both in the F1-F2 and F2-F3 planes. Notice the position of [y] (vowel 8 in the simulation) with minimal F3-F2 distance.

CONCLUSION

We have now obtained a reasonably good level of prediction of vocalic systems by means of listener-orientated principles, namely perceptual contrast and focalization. The main trends observed in the UPSID inventory are well captured. Regularization principles should then suffice to allow a good systematic prediction. Speaker-orientated principles could of course modify our predictions. They could also be used to a better evaluation of the selection between all stable systems. However, quantitative models at this level still seem to represent a big challenge for further research.

REFERENCES