Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT)

Jean-Luc Schwartz, Louis-Jean Boë, Christian Abry

8

Linking Dispersion–Focalization Theory and the Maximum Utilization of the Available Distinctive Features Principle in a Perception-for-Action-Control Theory

Jean-Luc Schwartz, Louis-Jean Boë, and Christian Abry

8.1 INTRODUCTION

The ''substance-based'' approach to phonology was born some 35 years ago with two seminal contributions, one by Liljencrants and Lindblom (1972)—the first of Lindblom's many variations on dispersion and perceptual contrast (*Dispersion Theories*: Lindblom 1986, 1990b)—the other, Stevens's *Quantal Theory* (1972, 1989). These contributions constituted the starting point for a rich tradition of descriptive and theoretical phonetics, in which the aim is not to refute the existence of a formal phonological level with its intrinsic formal principles and rules, but, instead, to determine and, possibly, model how the emergence of such formal systems could be shaped by the perceptuo-motor substance of speech communication.

The link between substance and form, however, is still not completely clear. In 1979, John Ohala questioned the role of maximization of perceptual dispersion, suggesting that with this principle ''we should undoubtedly reach the patently false prediction that a seven-consonant system should include something like the following set: ɗ, kˇ, ts, ɬ, m, r, l'' (Ohala 1979: 185), that is, a mixed combination of seven manner and three place contrasts, supposed to enhance perceptual distinctiveness.

This work owes a great deal to Ian Maddieson, whose UPSID database offers researchers a fundamental and invaluable tool. We thank the editor and the two anonymous reviewers for their helpful criticisms on the first versions of this text. We are grateful to Pauline Welby for her helpful comments and for her help with the English translation

He suggested that systems instead tend to limit their use of phonetic features, and that if a new feature is introduced in a system, it tends to combine systematically with the existing features in the system. This is what he called the *Maximum Utilization of the Available Features* (MUAF) principle (see also Clements 2003*a*, *b*).

In this chapter, we first review some of the major facts about vowel and consonant systems in human languages and show that both dispersion and MUAF principles seem to be at work in the shaping of these systems. We then present the Dispersion–Focalization Theory (DFT) that we proposed several years ago for predicting vowel systems, adding a new concept, focalization, to the classical dispersion concept inspired by Lindblom (Schwartz *et al*. 1997*b*). We propose that the integration of dispersion and MUAF forces us to consider speech units as resulting from a perceptuo-motor process. We present the Perception for Action Control Theory (PACT, Schwartz *et al*. 2002), which provides an integrated perceptuo-motor framework in which perception is conceived of as a set of mechanisms that allow the listener to recover speech gestures shaped by multisensory processing. Finally, we discuss how dispersion, focalization, and MUAF can be integrated in PACT in order to predict some of the main components of vowel and consonant systems in human languages.

8.2 DISPERSION AND MUAF IN VOWEL AND CONSONANT INVENTORIES

8.2.1 Systems

Based on databases of phoneme inventories, we estimate that there are about a thousand phonemes in the world's languages. The UPSID Database (Maddieson 1986*a*; Maddieson and Precoda 1990), which groups 451 languages (hereafter $UPSID_{451}$) drawn from the families and sub-families of languages defined in the Stanford classification and selected with a criterion of genetic distance, includes 920 phonemes with 179 vowels, 89 diphthongs, and 652 consonants. (The most recent UPSID version with 566 languages (Maddieson 2001) does not modify this picture.)

As early as 1928, the first typologies of phonological systems of the world's languages proposed by Trubetzkoy (1939) revealed that languages make use of relatively limited choices among all the phoneme possibilities determined by a simple combinatory rule. Thus, there is a strong bias in favor of systems with five vowels and 22 consonants. Within vowel systems, considering the 28 IPA vowel qualities, the theoretical number of five-vowel systems is about 10^5. There are, however, no more than 25 different five-vowel systems attested in the world's languages (Vallée 1994). Moreover, it is important to note that major typological and structural trends bear no clear relationship to linguistic families, either considering typologies proposed long ago by historical linguists (Meillet and Cohen 1924), or the framework of multilateral comparisons proposed by Greenberg (1963). For example, the Indo-European family

includes vowel systems ranging from five to 28 vowels considering the sample used by Maddieson in the $UPSID_{451}$ database comprising Greek, Irish, Breton, German, Norwegian, Lithuanian, Russian, Bulgarian, French, Spanish, Farsi, Pashto, Kurdish, Hindi-Urdu, Bengali, Kashmiri, Shinalese, Albanian, Armenian, Nepali, Konkani, and Ormuri. The fact that certain features (e.g. nasality, length, rounding, missing /p/ or missing /g/, three-vowel systems) are associated with certain geographical areas can only confirm that typological classifications of sound structures and genetic classifications according to linguistic families are far from identical, and that languages exhibit geographical (*Sprachbund*) tendencies rather than genetic ones.

8.2.2 Vowels

The 179 vowels in $UPSID_{451}$ are based on 38 plain vowel qualities, combined with secondary features such as nasalization, length, and pharyngealization. The number of units in the vowel systems varies from three for certain North American (e.g. Alabama), South American (e.g. Amuesha) and Australian languages (e.g. Arrente), to 24 (!Xu, Khoisan family) and even 28 (Kasmmiri, Indo-European). However, there is a strong preference for five vowels (comprising 20% of the $UPSID_{451}$ languages), or five vowel qualities (28% of the $UPSID_{451}$ languages, considering /a/, /ã/ and /aː/, for example, as having the same vowel quality).

In general terms, vowel systems seem to combine dispersion and MUAF principles. In fact, if a language has only three vowels, these are the three extreme plain vowels /i a u/, rather than /ə θ ɜ/, which are perceptually too close to each other, or a combination of quality and secondary articulation contrasts (e.g. nasality, length, pharyngealisation) as in /i ã uː/, as argued by Ohala for consonants (1979: 185). In a similar vein, if a language has five vowels, they are mainly the well-formed plain series /i e a o u/, (Crothers 1978; Maddieson 1984; Vallée 1994; Boë 1997; Schwartz *et al*. 1997*a*), accounting for 20 percent of languages in $UPSID_{451}$, rather than / ə θ ɕ ɜ ɐ/, with no clear perceptual distinctiveness, or /i/ /e/ /a/ /o/ /u/ combined with nasal, breathy, laryngeal, or pharyngeal secondary features. If a language has nine vowels, the preponderant system is /i ɪ e ɛ a ɔ o ʊ u/ (e.g. Tampulma, Niger Congo). Two-thirds of $UPSID_{451}$ languages have only plain vowels, with no secondary features. In systems with more than nine vowels, there generally appears a secondary series, one plain (e.g., /i e a o u/) and the other typically nasal (e.g., /ĩ ẽ ã õ ũ/ in Beembe or Kpan, Niger Congo), or long (e.g. /iː eː aː oː uː/ in Tonkawa, Amerindian). With 24 vowels, as in !Xu, the basic /i e a o u/ series is combined with subsets of secondary articulations combining nasality, length, and pharyngealization. This value of nine provides a kind of threshold above which a sub-space of plain vowels, obeying dispersion criteria, is combined with other features with some trend for MUAF (corroborating different observations; e.g. Crothers 1978: 113; Maddieson 1984: 128, 131; Lindblom and Engstrand 1989: 113; Engstrand and Krull 1991: 13–15; Vallée 1994: 95–6).

Previous analyses allowed us to show that the schwa vowel /ə/ seems to play a specific role in this pattern (Schwartz *et al*. 1997*a*), escaping in some sense from the traditional vowel space. Our assumption was that schwa, when it does exist in a given system, might be produced by a kind of systematic relaxation procedure based on vowel reduction (see van Bergem 1994), making it a sort of parallel system. To test this idea, we introduced a ''transparency rule'' which specifies whether or not a vowel interferes with the overall structure of the vowel system. The principle is that if a unit in a given system is ''transparent'', its presence or absence in the system should not modify the structure of the system. Other units, on the other hand, should do so because of relational interactions patterning the sound systems. For example, /i/ is never a transparent vowel since removing it from preferred systems like /i a u/ or /i e a o u/ leads to /a u/ or /e a o u/, which are unattested in human languages. We showed that schwa is the only vowel which respects this transparency rule since the relative distribution of systems with or without /ə/ is exactly the same. This reinforces the assumption that schwa is a parallel vowel, which exists because of intrinsic principles (probably based on vowel reduction) different from those of other vowels.

The vowels /i/, /a/, and /u/ are systematically used in vowel inventories. Potential counter-examples of systems without these three vowels have been proposed, such as short vowels in Arabic (Kennedy 1960; Mitchell 1962; Tomiche 1964), some Australian languages, or Indo-European reconstructions */e a o/ (about 4000 BC) proposed by Saussure (1879). Note the following points, however.

1. The acoustical analysis of dialectal Arabic vowels (e.g. Allatif and Abry 2004, for Syrian) shows that utterances of short vowels actually have [e] and [o] as their most frequent realizations, but also display clear cases of [iː] and [uː] for their longer counterparts.
2. Though a reduced system typically appears in the 3-vowel Australian indigenous languages (Butcher 1994), extreme [i] configurations are displayed, e.g., for accented phrase final vowels in female speech (Fletcher and Butcher 2003).
3. The oldest attested Anatolian Indo-European languages do not display the abnormal */e a o/ system (no matter how algebraic a speculation it might be, cf. the ± ATR proposals in Greenberg's Eurasiatic), since we regularly find /i a u/ in Luvian, and /i e a u/ in Hittite, Palaic, and Lycian (Melchert 1994).

Therefore it seems that predictions should not only exploit perceptual differentiation, but also incorporate perceptual representation spaces based on a hierarchy of features: first plain (100%) and then, after saturation of this space (generally after nine vowels), various combinations of features such as nasality (22%), length (11%), nasality and length (2%), reproducing to a certain extent the basic /i e a o u/ schema. Perceptual distances should be computed separately in each of these spaces, though not independently, as shown by the MUAF principle.

8.2.3 Consonants

Consonants constitute about three quarters of the available phonemes in the world's languages. Therefore, most languages (97% of the $UPSID_{451}$ languages) have more consonants than vowels. Consonant systems have mainly between 18 and 25 units (minimum 6 for Rotokas and maximum 95 for !Xufi, with a large series of clicks), with a peak at 22. The most frequent consonant systems contain at least six plosives /p t k b d g/, two to four nasals (/m n/, or /m n ŋ/ or / m n ɳ ŋ/), four fricatives including /f s h/, two approximants /j l/ and two affricates /ts tʃ/. It is only with a very large number of consonants that ejectives or clicks appear. $UPSID_{451}$ displays 12 places of articulation (see Fig. 8.1). The distribution of consonant place and manner of articulation is now quite well-known (Laver 1994; Boë 1997) (Table 8.1). In the following sections we focus on plosive systems, to provide some elements for modeling that will be incorporated in Section 8.4.

Like vowel systems, consonant systems seem to combine dispersion and MUAF principles. Indeed, considering plosives, if a language has three plosives (as in 3% of $UPSID_{451}$ languages; Ainu, Maasai, Nicobarese, and Yagua are examples), it has /p t k/, rather than other combinations of place (e.g. a coronal, a palatal, and a velar) or a combination of place and secondary articulation contrasts (e.g. /p t^h k^w/). If a language has six plosives, which is the most frequent number (found in 24% of the languages of $UPSID_{451}$), it has /p t k b d g/, rather than /p t c k q ʔ/, /b d g G ʔ/, or /p ^{n}t c^h $k^ʕ$ q^w ʔ/. With nine plosives, the basic /p t k b d g/ series combines with

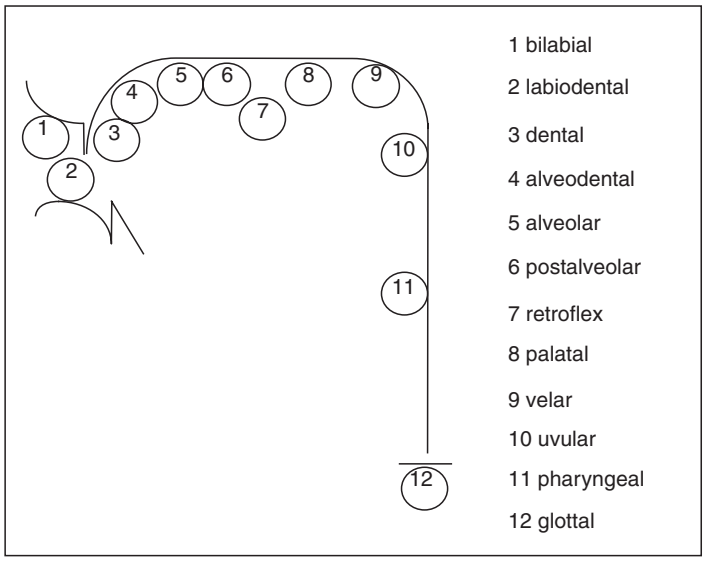

FIGURE 8.1. The twelve consonantal articulation places in UPSID. In the analyses in Section 2.3, the "coronal" articulation groups positions 3, 4, 5, and 6 (dental, alveodental, alveolar, and postalveolar)

TABLE 8.1. Percentage of place-of-articulation occurrences (in columns) for the five consonant categories (in rows) in $UPSID_{451}$. The most frequent places for each category are shown in bold type (from Boë 1997)

	Bilabial	**Alveodental**	Postalveolar	Palatal	**Velar**	Uvular	Pharyngeal	Glottal
Plosives	**99**	**100**	6	16	**99**	13	1	48
Nasals	**95**	**96**	10	31	**53**	0	0	0
Fricatives	**58**	**85**	43	8	29	11	4	**62**
Affricates	0	**85**	**49**	4	1	1	0	0
Approximants	**79**	**78**	3	**85**	**75**	0	0	0

secondary feature sets such as aspiration, prenasalization, palatalization, or laryngealization (see Maddieson 1986: 116–20 for details). Altogether, in the UPSID extension to 556 languages (Maddieson 2001), 45 percent of the systems include the six plosives /p t k b d g/. Therefore, it seems that there is both a best-place set including a labial, a coronal, and a velar (possibly for dispersion reasons, as we shall show in section 8.4) and a combination à la MUAF of this place set with other features, first voicing and then secondary features.

As with schwa in vowel systems, it may be suggested that the glottal articulation could play the role of a ''transparent'' unit, considering that it often emerges from a ''complete reduction'' of the consonantal supraglottal gesture, just as schwa emerges from a reduction of the vocalic supraglottal articulation. Actually, it appears that the ''transparency'' criterion works quite well with /ʔ/. In fact, systems with the three places of articulation /labial, coronal, velar/ constitute 33 percent of the languages in $UPSID_{451}$, while systems with the four places of articulation [labial, coronal, palatal, velar] constitute 7.5% of the $UPSID_{451}$ systems. Strikingly, the values for systems with the same distribution plus the glottal articulation are almost the same: 31% for [labial, coronal, velar, glottal] and 5% for [labial, coronal, palatal, velar, glottal]. Thus the glottal articulation does not seem to intervene in the structural relationship among plosive places. Basically, a given place system (not considering the glottal articulation) has about a 50 percent chance of containing a glottal articulation in addition to its basic structure; this leads to similar frequencies of systems with and without this consonant.

Consonant systems contain many fewer nasals than plosives. Indeed, about 3.5 percent of $UPSID_{451}$ languages have no nasals at all, 6 percent have only one (/m/ or /n/), 28 percent have two (mainly /m n/), 27 percent have three (generally /m n ŋ/), and 27 percent have four, adding /ɳ/. The nasality feature can be combined with other features, as shown by $UPSID_{451}$ languages containing nasals that are retroflex (Khanty, Ural-Altaic), long (Wolof, Niger-Kordofanian), voiceless (Sui, Austro-Tai), laryngealized (Nez Percé, Amerindian), or breathy (Hindi-Urdu, Indic). As with vowels and oral stops, these secondary features appear only if the system contains over a certain number of units, four for nasals.

The overall picture for plosives does not seem so different from that for vowels, with a preferred set of places /p t k/ (corresponding to the preferred /i a u/ or /i e a o u/ sets for vowels), an addition of supplementary features (such as voicing, nasality, and secondary features) more or less in line with the MUAF principle, and the existence of a transparent unit escaping from the structure by the *transparency rule*. An important point to be addressed in the case of nasals is the potential role of the *visual* modality in language system patterns. Of the languages in $UPSID_{451}$, 94 percent contain a contrast between a bilabial /m/ and a coronal /n/. The contrast between these two phonemes is quite easy to lipread, but is acoustically quite weak, as shown by the fact that blind children have some difficulty in learning to distinguish the two (Mills 1987). Hence it is not unreasonable to assume that the high visibility of this contrast plays a part in the fact that it is almost universal. Of course, visual perception is likely to play a role for all other sounds, but it is particularly noticeable in the case of the /m/–/n/ pair.

8.3 THE DISPERSION–FOCALIZATION THEORY OF SOUND SYSTEMS

The first quantitative simulations of vowel inventories are, of course, due to Liljencrants and Lindblom's (1972) Dispersion Theory (DT), based on maximization of perceptual distances in the (F1, F2) or (F1, F′2) plane (F′2 being the so-called ''perceptual second formant'' integrating F2, F3, and F4 if the higher formants are in the vicinity of F2: see Carlson *et al.* 1970). The basic principle underlying *Dispersion–Focalization Theory* (DFT, Schwartz *et al.* 1997*b*) is to associate a structural dispersion cost based on inter-vowel perceptual distances (dispersion) and a local cost based on intra-vowel perceptual salience (focalization). The DFT assumes that for a given number of vowels, the preferred system (i.e. the most frequently observed in the world's languages) is obtained by minimizing the sum of these two components, applied to acoustic parameters (formants expressed in Barks) characterizing each vowel located in the Maximal Vowel Space (MVS). This space (Boë *et al.* 1989) groups all possible productions of the human vocal tract, and it is determined from simulations on simplified (e.g. Fant's 4-tube model, 1960) or anthropomorphic (Maeda 1990; Beautemps *et al.* 2001) vocal tract models. The three cardinal vowels [i], [a], and [u] are the three corners of the maximum vowel space in the (F1, F2) plane, [y] being the fourth corner in the (F2, F3) space, allowing a better representation of the rounding dimension (Fig. 8.2).

The energy function of a given system with n vowels is given by: $E_{DF} = E_D + \alpha E_F$ where E_D is the dispersion cost (related to vowel structure) and E_F the focalization cost (related to the nature of each vowel) weighted by a factor α. E_D is defined, as in DT, by the sum of the squared inverse of the perceptual distances (using the

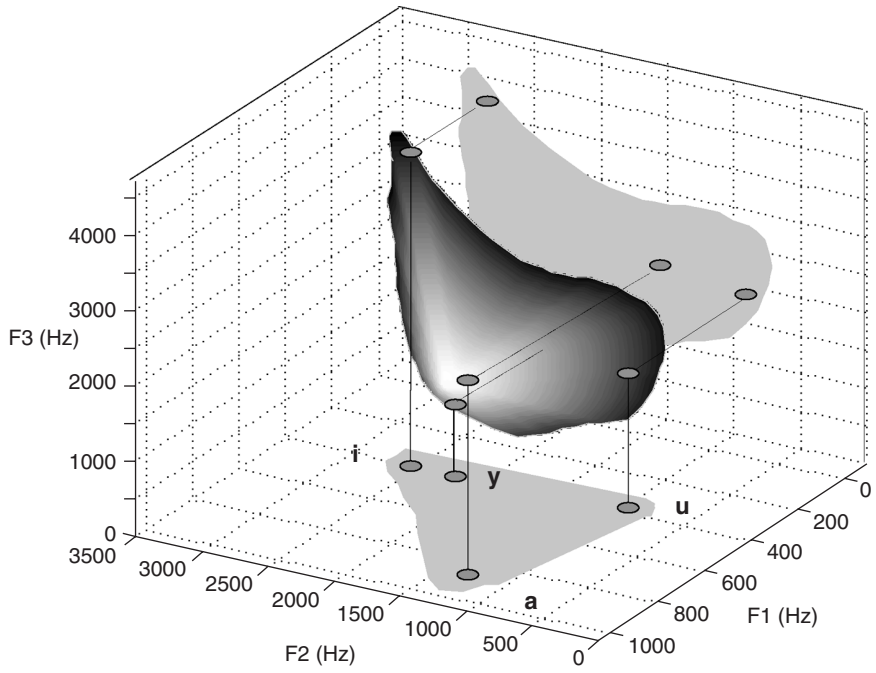

FIGURE 8.2. The F1–F2–F3 Maximal Vowel Space defined with an anthropological model (see text). The four corner-point vowels in this space are [i y a u]

perceptual second formant F2′) between each pair of vowels. In order to deal with the excessive number of high non-peripheral vowels in the DT predictions, we introduce a stretching of the acoustic space along the F1 dimension, assuming that higher formants play a minor role in the phonetic quality of the vowels. The λ parameter sets the weight between F2′ and F1. Simulations suggest that a λ value around 0.3 is necessary adequately to reproduce vowel systems. This raises the question of the possible explanation for a lesser role of higher formants in perceptual dispersion. We suggested three possible reasons for this. The first is due to Lindblom (1986), who showed that if a formant-based distance is replaced by an auditory spectral distance, this results in decreasing the ratio between the [i]–[u] and the [i]–[a] distance by a factor of around 2, which means applying a horizontal shrinking factor λ of 0.5. The same line of reasoning has recently been exploited and refined by Diehl *et al.* (2003) using more realistic auditory representations incorporating temporal neural coding. Secondly, various types of perceptual data suggest that lower-frequency formants are better perceived than higher-frequency ones (e.g. Delattre *et al.* 1952). This may be related to psycho-acoustic facts about the auditory representation of vowels and complex sounds, in which the representation of F1 is shown to be systematically enhanced relative to the representation of higher formants, because of remote suppression of higher-frequency by low-frequency components (see e.g. Moore and Glasberg 1983, Stelmachowicz *et al.* 1982, and Tyler and Lindblom 1982). The third

possible cause of F1 stretching relative to F2 and higher formants is based on a non-auditory argument from proprioception. Indeed, it has been proposed that the close–open dimension, mainly related to F1, could be better represented in proprioception than the front–back dimension, mainly related to F2 (Lindblom and Lubker 1985).

The additional focalization term, specific to DFT (and controlled by the second parameter α), diminishes the energy of configurations with vowels with F1 close to F2, F2 close to F3, or F3, close to F4. The focal vowels are produced by articulatory maneuvers changing the formant-to-cavity affiliations (Badin *et al*. 1990), hence making such configurations more stable (see Abry *et al*. 1989 for the relation between focalization and Quantal Theory). This is a very specific way of producing a spectral concentration of energy, which favors integration by the auditory system. Recent production (Ménard *et al*. 2006) and perception (Polka and Bohn 2003; Schwartz *et al*. 2005) data suggest that focalization plays a role in vowel perception just as focal colors play a role in color perception (Brown and Lenneberg 1954; Rosch-Heider 1972).

We thus obtain an energy function combining two terms depending on the λ and α parameters, respectively. DFT allows us to predict not only preferred systems, but also a number of possible variants in the so-called (λ, α ''phase space''; Schwartz *et al*. 1997*b*), which allows us to simulate about 85 percent of systems with three to seven plain vowel qualities, which is about 65 percent of the $UPSID_{451}$ systems (Boë *et al*. 2002). We illustrate such a ''phase space'' in Figure 8.3, showing that for an

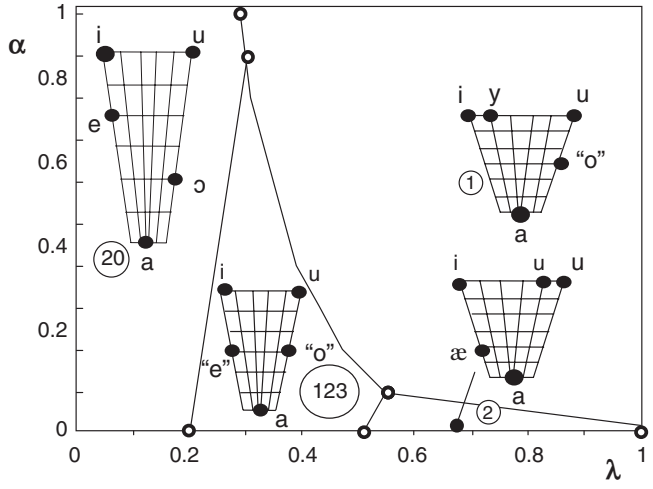

FIGURE 8.3. The DFT five-vowel-systems phase space. The λ-parameter controls the F′2 vs. F1 stretching. Notice how increasing λ leads to adding one vowel between [i] and [u]. The α parameter controls focalization. Notice how increasing α enables incorporation of [y] into the system. The phase space is divided into (α, λ) regions for which a specific system is preferred (i.e. has a lower energy in the DFT framework). All the predicted systems do occur in $UPSID_{451}$ (their number of occurrences is inserted near each system, inside a circle). The preferred system in $UPSID_{451}$ is /i e a o u/ (123 occurrences in $UPSID_{451}$), which is predicted for a λ value around 0.3

appropriate value of λ around 0.3, the best /i e a o u/ system, present in 123 of the 451 languages of the $UPSID_{451}$ database, is in fact selected. Other simulations allowed us to determine that α should be set to around 0.3 to account for the existence of /y/ in about 8 percent of vowel systems. Unfortunately, though the existence of focalization now seems well established by various experiments, there is no precise evaluation of the α value, apart from the present estimation derived from data fitting. This will be a challenge for future experiments.

8.4 LINKING DFT AND MUAF IN THE PERCEPTION-FOR-ACTION-CONTROL THEORY

8.4.1 Percepts and controls

The description of universals in vowel and consonant systems clearly points to a theoretical difficulty. Sound systems in language inventories seem to exploit both structural and local costs in the perceptual domain (i.e. dispersion and focalization) and a combination principle, which in some sense escapes dispersion and replaces it by some kind of phonological feature economy (Clements 2003*a*, *b*). How can we account for this?

It is tempting to assume that MUAF is based on maximal use of available controls rather than on features, in the sense that once a new articulatory control of the vocal tract has been discovered (e.g. control of the velum for nasals vs. oral stops, or control of the glottis for voiced vs. unvoiced consonants), systems are driven systematically to combine it with other available controls. This is in fact what Lindblom (1998) proposed in his ''lexical recalibration'' model, according to which speech units are considered both as sounds that should be easy to distinguish (i.e. dispersion) and as gestures that should be easy to learn in the course of development. Learnability would induce a maximum use of available controls, since a system with fewer controls would be preferred over a system with more controls, which should be more difficult to learn. This would lead to a preference for /i a u/ rather than, /e œ o/, for instance (because of dispersion) or /i aː ũ/ (because of learnability). The key question here is to understand better how to combine the concepts of dispersion and available control.

8.4.2 The Perception-for-Action-Control Theory

The Perception-for-Action-Control Theory (PACT) goes one step further in the sensori-motor route. There is a long history of debate between auditory theories (e.g. Nearey 1997; Massaro 1987) and motor theories (A. M. Liberman and Mattingly 1985; Fowler and Rosenblum 1991) of speech perception. Simplifying somewhat, auditory theories consider that speech perception works without action, that is, without incorporating any knowledge about the way sounds are produced by the

articulatory system. On the other hand, motor theories consider that the objects of speech perception are gestures and not sounds. Hence in some sense motor theories consider speech perception without audition (this is most clearly expressed in the ''speech is special'' view, according to which audition does not intervene *per se* in the .processing of speech gestures: see, for example e.g., Whalen and Liberman 1987 and Whalen *et al*. 2006).

Our view is that speech perception is shaped both by auditory processing and motor knowledge (Schwartz *et al*. 2002). PACT assumes that speech perception not only allows listeners to follow the vocalizations of their conversation partner in order to understand then, but also to imitate and learn. In other words, perception allows listeners to specify the control of their future actions as a speaker. There is in this view an integrated process, combining perceptual shaping together with an inversion mechanism, allowing the listener to recover articulatory control in relation to his or her understanding of the perceptual goal. This process is different from both a pure ''auditory'' and a pure ''motor'' theory of speech perception. It integrates perceptual processing and articulatory knowledge (possibly in a computational ''speech robotics'' framework; see Abry and Badin 1996; Serkhane *et al*. 2005). To illustrate this better, let us examine two examples from the patterning of vowel systems.

First, assuming that the Motor Theory of Speech Perception is correct, what would be the possible predictions in terms of oral vowel systems in human languages? There are basically three degrees of freedom for producing oral vowels: height, front-back position, and rounding. This results in a 3-D articulatory space, illustrated in Fig. 8.4a (with a shrinking of the space for open configurations, for which the front–back and rounding dimensions play a less important role). What would be the best three-vowel system in this space? The system /i a u/ is a very good choice, in terms of articulatory dispersion, and it is compatible with the $UPSID_{451}$ data. However, [y a ɯ] should be as

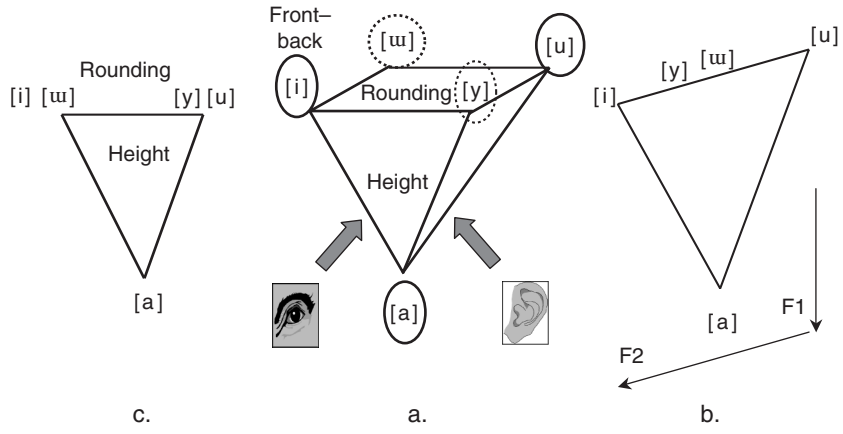

FIGURE 8.4. The articulatory three-dimension space of oral vowels (a), together with its auditory (b) and visual (c) projections

good a choice. It combines articulatory features differently, but the difference cannot be assessed in articulatory terms. However, this second system never appears in human languages. The reason for this is clearly auditory. Auditory perception is a kind of lateral projection of this 3-D space, in a 2-D (F1, F2) space (Fig. 8.4b) in which [i u] is of course much better (in terms of dispersion) than [y ɯ]. Note that vision (lipreading) can also be incorporated in this schema (Fig. 8.4c). It provides a view from the front, where only height and rounding emerge while the front–back dimension is lost (Robert-Ribes *et al.* 1998), hence [i u] and [y ɯ] are equivalent in this projection. Altogether, the prevalence of /i a u/ and the absence of /y a ɯ/ clearly shows that gestures are shaped by perception.

On the other hand, there are many examples showing that articulatory knowledge seems to intervene in speech perception (e.g. Fowler 1986; A.M. Liberman and Whalen 2000). Let us take the example of vowel reduction. It has long been claimed that listeners are able to recover targets from coarticulated speech and particularly from reduced speech (e.g. Lindblom 1963). Some work from our lab has shown that a stable articulatory target [a] can be recovered by acoustic-to-articulatory inversion, in spite of acoustic variability due to reduction in an [iai] sequence (Lœvenbruck and Perrier 1997). In this vein, there is a very clear case, provided by the coarticulated /u/ in a fronting context, for example, in French *doute* ''doubt'' /dut/, which becomes acoustically close to a front [y] because of coarticulation. However, it is striking that though most languages contain sequences such as /C_1uC_2/ where C_1 and C_2 are two coronal consonants, /y/ exists in fewer than 10 percent of the languages (see section 8.2). This shows that listeners are able to recover speaker's intentions, hence the need for introducing ''procedural knowledge'' (Viviani and Stucchi 1992) about speech production in the course of speech perception.

To summarize, the objects of speech perception for PACT are multi-sensory percepts regularized by knowledge of speech gestures, or speech gestures shaped by perceptual processes. This view has gained some support from recent transcranial magnetic stimulation (TMS) data showing that listening to coronals specifically modulates the excitability of neurons driving the tongue (Fadiga *et al.* 2002), while listening to or looking at labials specifically modulates the excitability of neurons driving the lips (Watkins *et al.* 2003). More generally, the recent literature on mirror neurons and on a ''perceptual action understanding system'' (e.g. Rizzolati and Arbib 1998) adds a strong neuro-anatomical background to PACT. A plausible cortical circuit for PACT is provided by the ''dorsal route'' (Hickok and Poeppel 2000), connecting perceptual processes in the temporal region (Superior Temporal Sulcus) with action understanding in the frontal lobe (including Broca's area and motor and premotor areas) passing by parietal regions matching sensory and motor representations (see, e.g., Sato *et al.* 2004). In the framework of the current paper, the advantage of PACT is that it intrinsically combines perceptual distinctiveness (modeled in the DFT) with theories of the control of speech gestures, including a number of cognitive aspects about their development and neuro-anatomy.

8.4.3 Vowel and consonant systems resulting from unfolding actions quantally shaped by perception

Since sound systems in language inventories are conceived by PACT as speech gestures shaped by perceptual processing (auditory and visual), we need to know how speech gestures are produced, step-by-step, in a phylogenetically and developmentally plausible scenario.

A central piece here is provided by the Frame/Content theory (hereafter FC theory) developed by MacNeilage and Davis (MacNeilage 1998; MacNeilage and Davis 1990*a*, *b*, 1993), which includes ontogenetic, phylogenetic, and neuroanatomical components. FC theory claims that speech production begins with babbling with phonation associated with jaw cycles and no other articulator being controlled online apart from vocal tract pre-settings, which are kept stable throughout the jaw cycles. Besides their alimentary function, jaw cycles would induce alternation of closing and opening patterns (*closants* and *vocants*) and constitute the *frame* of speech. The segmental speech *content* (independent control of consonants and vowels inside the frame) would then progressively emerge from the development of the central and peripheral motor control of the other articulators. Hence, the FC theory provides a natural unfolding sequence of speech gestures. Furthermore, the perceptual shaping of jaw movements is strongly nonlinear. Indeed, jaw closing results in switching from a laminar to a turbulent airflow, and finally to a complete obstruction with no sound. This is a typical case to which Quantal Theory can be applied extremely efficiently (Fig. 8.5a).

Considering that the division between vowels and consonants continues to be generally accepted in phonetics and phonology (Boë and Durand 2001), the FC theory thus provides a natural developmental and evolutionary pathway towards the birth of consonants and vowels, with the emergence of a consonantal /obstruent/ closed pole provided by syllable onset (and maximal jaw closing) and a vocalic /sonorant/ open pole provided by syllable rhyme (and maximal jaw opening). Consonants and vowels do in fact emerge in the FC theory as different objects which occupy different ''slots'' in the jaw cycle and hence also in the mental representations of language, as displayed in a number of psycholinguistic behaviors (MacNeilage 1998; Levelt 1989). Perceptually, they also correspond to different representation spaces, for instance, formant onset values or burst spectral characteristics after the vocal tract closure for plosives; friction noise for fricatives; harmonic spectrum or formants for vowels (Fig. 8.5). In addition, experimental data suggest that they are indeed processed as separate perceptual streams (Fowler 1986; Fowler and Smith 1986).

8.4.4 [b]–[d]–[g] as a universal triangle as acoustically optimal as [i]–[a]–[u]

From the basic action generator provided by jaw cycles producing frames in the FC theory, the next step is the progressive mastery of vocal-tract shaping. This mastery is

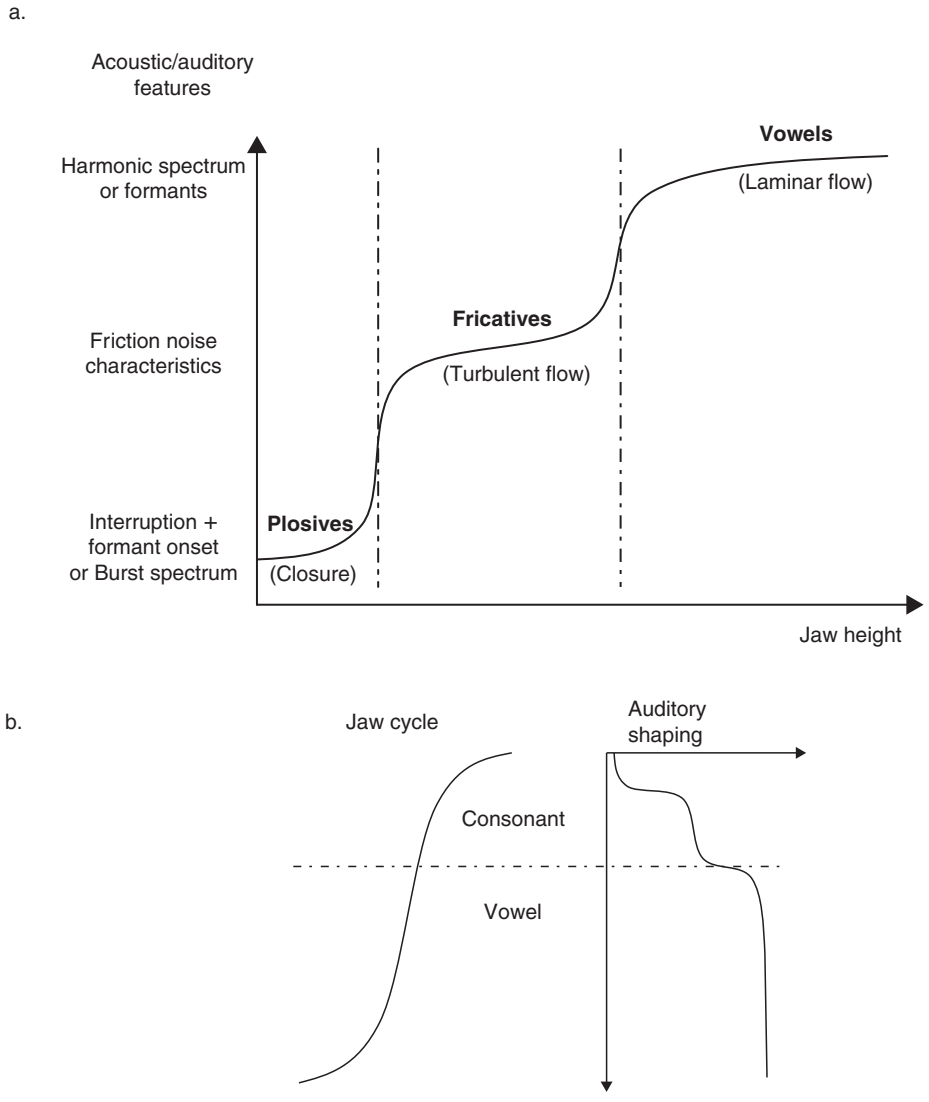

FIGURE 8.5. The articulatory-to-acoustic/auditory quantal transformation along jaw height (a) When jaw position varies from low to high, the acoustic/auditory features change suddenly around two boundary values, separating plosives from fricatives and fricatives from vowels. This is a typically quantal relationship in the sense of Stevens (1972, 1989) (see text). (b) The consequence is that inside jaw cycles, there is a natural division between a consonantal pole for high jaw positions and a vocalic pole for low jaw positions

developed independently for the control of contacts in plosives and the control of the global tract shape for vowels (discarding fricatives in the discussion to follow). This is the probable sequence of the acquisition of control in speech development (Abry *et al.* forthcoming), and where perceptual shaping comes into play; it should allow us to

understand how languages select and combine individual articulators for controlling the vocal tract shape in a perceptually efficient way. In mastering the vocal-tract shaping, the DFT is applied separately to vowels and consonants. For vowels, we showed in Section 8.3 how the height, front–back, and lip rounding dimensions were structured by dispersion and focalization. For plosives, it is also possible to propose the same kind of analysis (Abry 2003).

Indeed, it is noticeable that available Haskins's patterns for the Playback synthesizer allowed one to represent—long before locus equations—a triangle of CV transitions (Cooper *et al*. 1952). In a syllable such as /ga/, when moving towards [a]—in an F2–F3 plane—we can take as a starting point the ''hub'' locus (velar pinch) at about the [g] release, where F2 equals F3 (Fig. 8.6a). From that point on F3 rises and F2 falls. The pattern for [b] shows F2–F3 rising together towards [a], while they are both falling for [d]. Plotting these starting points (plosive release values) on the F2–F3 plane in Figure 8.6b, we obtain a plosive acoustic triangle [b d g] mirroring the famous acoustic F1–F2 vowel triangle [i a u]. In this right triangle of consonantal formant movements, [g] is at the 90 degrees angle, moving to [a], close to the hypotenuse, while [d] and [b] are also converging to [a] along this hypotenuse. Voiced consonants have their F1 motion in common, which is basically of no use in parsing their place, but only their consonantal nature—their rising formant movement—from wall vibrations (180 Hz) to F1 vowel target value.

Of course, the diagrams in Figure 8.6 are schematic plots rather than articulatory simulations, though they can be interpreted clearly in articulatory terms according to

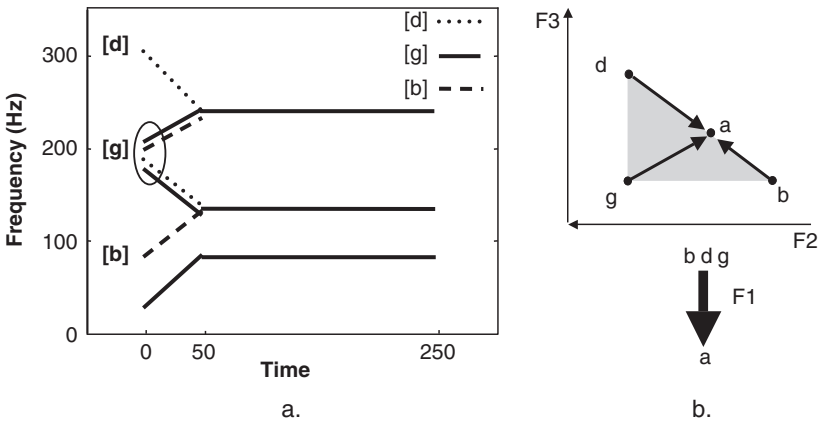

FIGURE 8.6. (a) Schema of VCV transitions for [aba], [aga], and [ada]. All the F1 transitions are the same, and they characterize the transition from vowel to plosive, back to vowel, rather than the plosive's place of articulation. From the consonantal release to the [a] vowel, F2 and F3 change together with similar slopes: both rising for [b], both falling for [d]. [g] is characterized by a convergence of F2–F3 (''hub'' locus, circled in the figure) from which F2 decreases and F3 increases towards [a]. (b) This provides a possible ''plosive triangle'' in the F2–F3 space, while the F1 trajectory is independent of plosive articulation place

the developmental sequence described earlier. Indeed, plosives are generated in the FC theory by a progressive mastering of contacts inside the vocal tract, produced by upward jaw movements progressively upgraded by the carried articulators (lips and tongue), applied to a basically neutral vocal tract shape. It has long been known that [b], [d], and [g] appear as extrema of F2–F3 configurations in nomograms varying the place of a constriction applied to a neutral tube (see e.g. Fant 1960; Stevens 1998). We provide in Figure 8.7 simplified possible two- and three-tube configurations that correspond rather well to the [b], [d], and [g] configurations that could be generated in the FC theory. Simply closing a neutral tube for labials (Fig. 8.7a); raising and hence flattening the tongue, carried by the jaw, all along the front of the vocal tract towards an [i]-like configuration for [d], typically producing a [djə] as commonly produced by infants in babbling (Fig. 8.7b); or for [g] making a constriction connecting a back closed-closed 8 cm long cavity with a half-wavelength resonance at 2180 Hz, and a front closed–open cavity with a fourth-wavelength resonance at the same value, thus

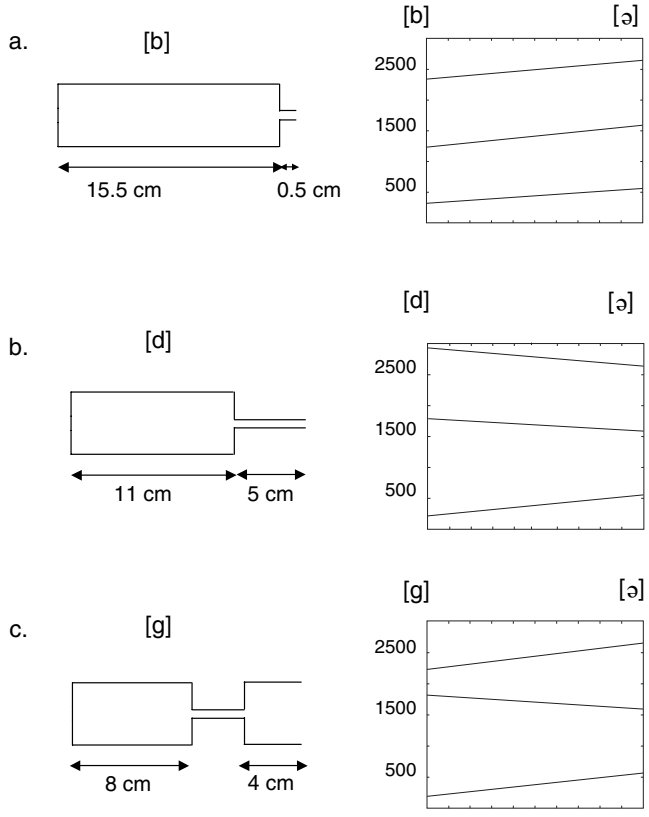

FIGURE 8.7. Simplified geometric configurations for a labial (a), a coronal (b) and a velar (c) constriction applied to a neutral tube (see text). For each configuration, the corresponding formant trajectory towards a 16-cm-long neutral tube is provided on the right

producing a perfect hub locus with the convergence of F2 and F3 (Fig. 8.7c). Simulations with an articulatory model give more realistic CV patterns (Berrah *et al*. 1995). Therefore, [b], [d], and [g] appear as a possible solution to the perceptual dispersion maximization problem, provided that this problem is raised in a plausible developmental framework in which these three places emerge; they should provide a strong perceptual contrast (in terms of maximal formant transitions towards a neutral or [a] configuration) from basic upward movements of the jaw to which movements of the lips and tongue have been added.

In speech development, F1 is the acoustic/audible movement corresponding to the carrier of speech, the lower jaw, with typical labial [bababa] or coronal [dadada] ''frames'' (less often velars [gVgV]), following MacNeilage and Davis's (1993) proposal. When carried articulators (lip and tongue) progressively become independent from the jaw, as early as the beginning of the second year, the F2–F3 stream carries information on contact place. Thus the F2–F3 plane is developmentally orthogonal to F1. When coarticulation emerges—typically, after the first year—the [a] vowel in a CV syllable can be produced during the closure phase of the consonant (as evidenced by a case study; see Sussman *et al*. 1999). This means that the vocalic F1 value is coproduced with the intrinsic consonantal F1 closing–opening trajectory. Until four years of age, the mastering of the control of the whole vocal tract for [i] and [u] will be in progress, if they are present in the mother tongue, as is the case for most languages (for the development of timing control for the rounding contrast in high front vowels [i] vs. [y] in French, see Noiray *et al*. 2004). The differentiation process is thus comparable for consonants, in the F2–F3 plane, and for vowels, basically in the F1–F2 plane, and the Dispersion Theory proves efficient for structuring both acoustic planes. Of course, these two orthogonal triangles in the F1–F2–F3 space adjoin each other since they are generated by the acoustic theory of the vocal-tract tube with one or two constrictions (Fig. 8.8). An additional dimension may be used, such as F3 for vowels (e.g. to contrast [i] vs. [y], as in French), or F1 for consonants (contrasting pharyngealized vs. plain segments, as in Arabic).

8.4.5 MUAF emerging from PACT: a computational proposal

In PACT, actions provide the degrees of freedom to combine, and perceptual dispersion and focalization provide the mechanisms driving combination. How could this result in MUAF? The answer is related to the fact that some articulatory degrees of freedom might operate on acoustic dimensions different from those related to another subgroup of articulatory commands. For example, in the case of vowels, lip rounding, and tongue front–back placing basically operate on the same parameters (F2, and to a lesser extent F1 and F3). Adding height, these three articulatory controls interact in the acoustic/auditory (F1–F2–F3 or F1–F′2) space, hence dispersion structures their combination. However, control of the glottis operates on dimensions (e.g. vowel duration or voice quality) other than formant frequencies (though not

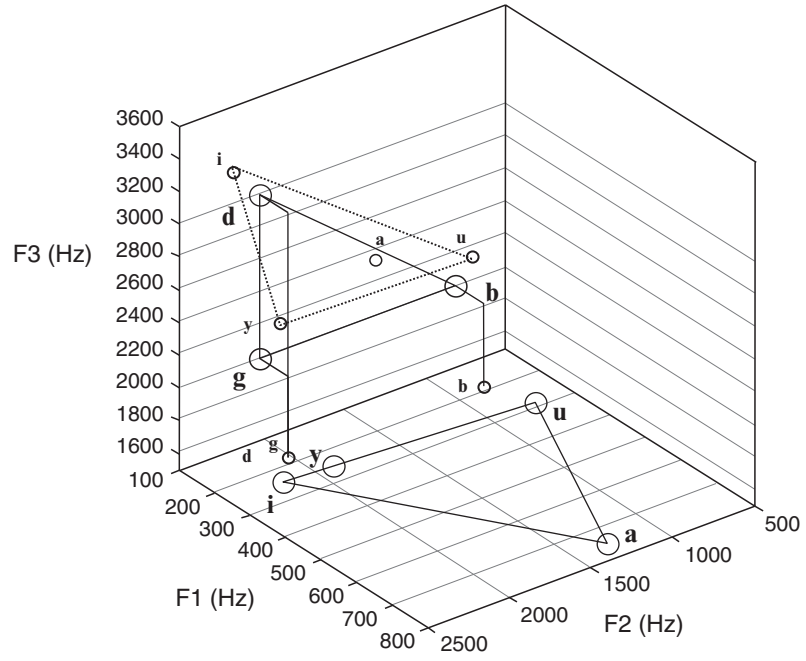

FIGURE 8.8. The three dimensional vowel/plosive place-space. In gray, the F1–F2 vowel triangle and the F2–F3 plosive triangle are shown. Circles display focal vowels and maximally distant plosives

completely unrelated), therefore the combination of place and voice controls may escape dispersion. How then can MUAF (or ''feature economy'') appear?

A tentative computational answer was proposed a few years ago by Berrah and Laboissière (1997) and Berrah (1998) using a series of simulations of the emergence of a common phonetic code in a society of robots using evolutionary techniques. Robots in this study were simple agents able to produce sustained vowels represented in a perceptual space (including formants, and, possibly, an additional dimension intended to be separable from formants). Each robot had a lexicon composed of a fixed number of randomly initialized items. Pairs of robots, randomly selected, communicated through transactions in which one robot, the speaker, emitted one of its items. The other robot, the listener, related the perceived item to its own repertoire by selecting the nearest item in terms of perceptual distance, moving this item slightly towards the perceived item, and moving all the other items away, in order to avoid further confusions and to maximize perceptual distinction (dispersion). A first series of simulations in a formant space allowed Berrah and Laboissière to reproduce some of the major trends of vowel systems (see also de Boer 2000). In a second series of simulations, Berrah and Laboissière added a ''secondary'' acoustic dimension (for example, vowel duration). They explored a modified version of their model in which repulsion was not applied simultaneously to all acoustic/auditory parameters, but only to the subset (either formants or secondary dimensions) for which repulsion was

the most efficient, that is, most increased the distance between the target item and the competitor items. This means that items were considered as sounds generated by subcomponents of an action system selected as perceptually efficient. Typical results are illustrated in Figure 8.9. The right-hand panels represent positions of the items for the whole set of robots on the 3D acoustic space (F1–F2 plus the additional dimension), for a five-vowel (top row), an eight-vowel (middle row) and a ten-vowel system. The left-hand panels display the item projections in the F1–F2 space. It appears that when the number of vowels in the lexicon is small enough (top row), dispersion applies in the F1–F2 space, but the secondary dimension stays unexploited. However, when the number of vowels increases (middle and bottom row), there is an expansion towards the secondary dimension, with a progressive trend for MUAF (partial in the eight-vowel case, and complete in the ten-vowel case). This rather adequately replicates the trends presented in Section 8.2.2, gathered from UPSID vowel inventories. Though

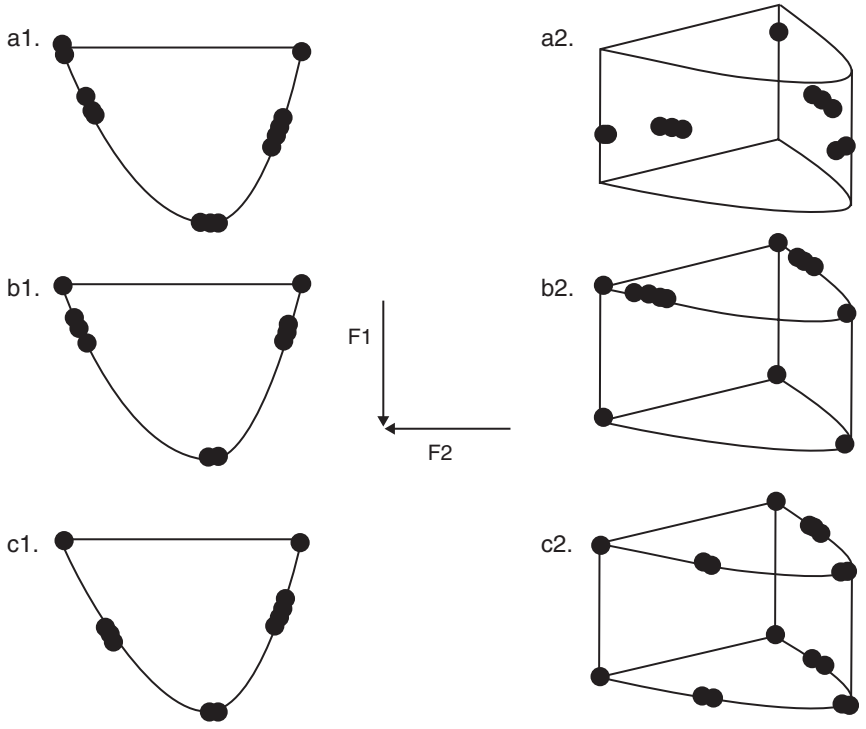

FIGURE 8.9. Vowel-system simulations in a computational evolutionary approach involving a ''society of robots'' (from Berrah and Laboissière 1997). Robots exchange sounds and adapt their repertoire to their partners, by increasing perceptual distances between items in a 3-D space made of F1–F2 plus an additional secondary parameter (e.g. vowel duration). On the left: projection of the items after convergence in the F1–F2 space. On the right: position of the items after convergence in the 3-D space. From top to bottom: five-, eight-, and ten-vowel systems. Notice how F1–F2 projections are similar in the three rows (due to dispersion), while the use of the third dimension appears only in the second and third row, that is, for a sufficient number of vowels in the system (this simulates MUAF)

preliminary, these simulations show that if a given set of articulatory degrees of freedom is provided, with the possibility of dissociating their acoustic consequences into different sub-spaces leading us to define sub-groups in the set of articulatory controls, PACT may result in both dispersion principles in the specific sub-groups and feature economy. This can lead to either discarding the ''secondary'' articulatory dimension as in Figure 8.9a or to a partial or systematic combination of the secondary dimension with the primary group as in Figures 8.9b and c.

8.5 ''NO INTERFACE BETWEEN PHONOLOGY AND PHONETICS'' (J. OHALA 1990*C*): CONCLUDING REMARKS ABOUT AN INTEGRATED VIEW

Substance-based linguistics is an old dream for phoneticians, and more generally for speech scientists. This was nicely expressed by Lindblom (1984: 78) with his formula ''derive language from non-language.'' We have tried to show that it is now possible to use a theory relying on perceptuo-motor (non-linguistic or pre-linguistic) interactions, a computational framework, and a quantitative methodology inspired by proposals originally formulated by Lindblom and Liljencrants, Stevens, and enriched by J. Ohala, MacNeilage, and Davis for ontogenetical aspects. This provides a number of proposals for predictions of vowel and consonant systems in the PACT framework. This approach has led to further successful predictions concerning, for example, fricative systems (Boë *et al.* 2000) or consonant sequences and the ''labial–coronal'' effect (Rochet-Capellan and Schwartz 2005*a*, *b*). Of course, we do not claim that phonological tendencies in the world's languages can be reduced to explanations formulated solely in terms of linguistic substance. Moreover, a large number of questions remain unsolved or insufficiently developed, and many new simulations will be necessary to provide a completely implemented version of PACT (for example, for extensively simulating vowel and consonant systems). Probably, however, much can be done to anchor a number of seemingly formal regularities, such as MUAF, within perceptuo-motor substance.

Speech is by nature an interdisciplinary area of research, lying at the crossroads of several sensori-motor systems involved in the production and perception of biological communication signals, and of a major human competence, language. The expanding interests and capabilities of phonetics have triggered a reorganization of the scientific connections between phonetic sciences and adjacent disciplines. New interactions are clearly underway between linguistics, cognitive science, and certain sectors of the physical and engineering sciences (Boë 1997). Particularly interesting in this respect is the trend towards ''laboratory phonology'' which combines experimental phonetics, experimental psychology, and phonological theory (J. Ohala and Jaeger 1986). This

approach aims at subjecting hypotheses of phonological organization to the kinds of validation used in the experimental sciences, which has been lacking to date in generative phonology. Phonetic knowledge can thus explore and specify the natural constraints that all phonological theories must respect in order to satisfy concerns of (neuro)physiological plausibility. The integration of phonology into the natural order of things no longer needs to involve a subordinate relationship between the two disciplines.