

Chapter 8: Contrast and Perceptual Distinctiveness*

Edward Flemming

From Bruce Hayes, Robert Kirchner, and Donca Steriade, eds., *Phonetically-Based Phonology*. 2004, Cambridge: Cambridge University Press. Section 4.3 deleted for this reading.

Note that this reading has a lot of IPA in it. For clarification, try downloading the official IPA chart from <http://www2.arts.gla.ac.uk/IPA/fullchart.html>

1 Introduction

Most ‘phonetically-driven’ or functionalist theories of phonology propose that two of the fundamental forces shaping phonology are the need to minimize effort on the part of the speaker and the need to minimize the likelihood of confusion on the part of the listener. The goal of this paper is to explore the perceptual side of this story, investigating the general character of the constraints imposed on phonology by the need to minimize confusion.

The need to avoid confusion is hypothesized to derive from the communicative function of language. Successful communication depends on listeners being able to recover what a speaker is saying. Therefore it is important to avoid perceptually confusable realisations of distinct categories; in particular distinct words should not be perceptually confusable. The phonology of a language regulates the differences that can minimally distinguish words, so one of the desiderata for a phonology is that it should not allow these minimal differences, or contrasts, to be too subtle perceptually. In Optimality Theoretic terms, this means that there are constraints favouring less confusable contrasts over more confusable contrasts.

There is nothing new about the broad outlines of this theory (cf. Lindblom 1986, 1990, Martinet 1955, Zipf 1949, among others), but it has important implications for the nature of phonology. First, it gives a central role to the auditory-perceptual properties of speech sounds since distinctiveness of contrasts is dependent on perceptual representation of speech sounds. This runs counter to the articulatory bias in phonological feature theory observed in Chomsky and Halle (1968) and its successors. Substantial evidence for the importance of perceptual considerations in phonology has already been accumulated (e.g. Boersma 1998, Flemming 1995, Jun this volume, Steriade 1995, 1997, Wright this volume; see also Hume and Johnson (2001) pp.1-2 and references cited there). This paper provides further evidence for this position, but the focus is on a second implication of the theory: the existence of constraints on contrasts. Constraints favouring distinct contrasts are constraints on the differences between forms rather than on the individual forms themselves. We will see that paradigmatic constraints of this kind have considerable implications for the architecture of phonology.

The next section discusses why we should expect perceptual markedness to be a property of contrasts rather than individual sounds and previews evidence that this is in

* I would like to thank the editors for detailed comments on this paper.

fact the case. Then constraints on contrast will be formalized within the context of a theory of phonological contrast. The remainder of the paper provides evidence for the key prediction of the theory: the markedness of a sound depends on the sounds that it contrasts with.

2 Perceptual markedness is a property of contrasts

The nature of the process of speech perception leads us to expect that any phonological constraints motivated by perceptual factors should be constraints on contrasts, such as the contrast between a back unrounded vowel and a back rounded vowel, not constraints on individual sounds, such as a back unrounded vowel. Speech perception involves segmenting a speech signal and categorizing the segments into a pre-determined set of categories such as phonetic segments and words. The cues for classification are necessarily cues that a stimulus belongs to one category as opposed to another. So we cannot talk about cues to a category, or how well a category is cued by a particular signal without knowing what the alternatives are. For example, it is not possible to say that a back unrounded vowel presents perceptual difficulties without knowing what it contrasts with. It is relatively difficult to distinguish a back unrounded vowel from a back rounded vowel so if a language allows this contrast the back unrounded vowel can be said to present perceptual difficulties, and the same can be said of the back rounded vowel. But if it is known that a back unrounded vowel is the only vowel which can appear in the relevant context, then all the listener needs to do is identify that a vowel is present as opposed to a consonant, which is likely to be unproblematic.

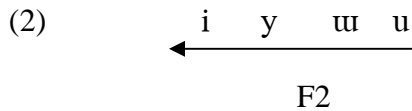
Perceptual difficulty is thus very different from articulatory difficulty. Articulatory difficulty can be regarded as a property of an individual sound in a particular context because it relates to the effort involved in producing that sound. There is no analogous notion of effort involved in perceiving a sound – perceptual difficulties don't arise because particular speech sounds tax the auditory system, the difficulty arises in correctly categorizing sounds. Thus it does not seem to be possible to provide a sound basis in perceptual phonetics for constraints on the markedness of sounds independent of the contrasts that they enter into. This point is assumed in Liljencrants and Lindblom's models of how perceptual factors shape vowel inventories (Liljencrants and Lindblom 1972, Lindblom 1986), and similar considerations are discussed in Steriade (1997).

The difference between regarding perceptual markedness as a property of contrasts rather than sounds can be clarified through consideration of alternative approaches to the analysis of correlations between backness and lip rounding in vowels. Cross-linguistically, front vowels are usually unrounded whereas non-low back vowels are usually rounded. This is true of the common five vowel inventory in (1), and in the UPSID database as a whole, 94.0% of front vowels are unrounded and 93.5% of back vowels are rounded (Maddieson 1984).

(1) i u
 e o
 a

The perceptual explanation for this pattern is that co-varying backness and rounding in this way maximizes the difference in second formant frequency (F2) between front and back vowels, thus making them more distinct. In general front and back vowels differ primarily in F2, with front vowels having a high F2 and back vowels having a low F2.

Lip-rounding lowers F2 so the maximally distinct F2 contrast is between front unrounded and back rounded vowels (Liljencrants and Lindblom 1972, Stevens, Keyser and Kawasaki 1986). This is illustrated in (2) which shows the approximate positions of front and back rounded and unrounded vowels on the F2 dimension. It can be seen that the distinctiveness of contrasts between front and back rounded vowels, e.g. [y-u], or between front and back unrounded vowels, e.g. [i-ʊ], is sub-optimal.



The standard phonological analysis of this pattern of covariation is to posit feature co-occurrence constraints against front rounded vowels and back unrounded vowels (3).

- (3) *[-back, +round]
 *[+back, -round]

This analysis does not correspond to the perceptual explanation outlined above. The constraints in (3) imply that front rounded vowels and back unrounded vowels are marked sounds, whereas the perceptual explanation implies that it is the contrasts involving front rounded vowels and back unrounded vowels that are dispreferred because they are less distinct than the contrast between a front unrounded vowel and a back rounded vowel. In Optimality Theoretic terms, there is a general principle that contrasts are more marked the less distinct they are, which implies a ranking of constraints as in (4), where *X-Y means that words should not be minimally differentiated by the contrast between sounds X and Y. (More general constraints which subsume these highly specific constraints will be formulated below).

- (4) *y÷ʊ >> *i÷ʊ, *y÷u >> *i÷u

These two accounts make very different predictions: Constraints on the distinctiveness of contrasts predict that a sound may be marked by virtue of the contrasts it enters into. If there are no constraints on contrasts, then the markedness of contrasts should depend simply on the markedness of the individual sounds, and should be insensitive to the system of contrasts. We will see a range of evidence that markedness of sounds is indeed dependent on the contrasts that they enter into – i.e. that there are markedness relations over contrasts as well as over sounds – and that the relative markedness of contrasts does correspond to their distinctiveness.

For example, the dispreference for front rounded vowels and back unrounded vowels extends to other vowels with intermediate F2 values, such as central vowels. Most languages contrast front and back vowels, and if they have central vowels, they are in addition to front and back vowels. The same explanation applies here also: since central vowels like [ɨ] fall in the middle of the F2 scale in (2), contrasts like [i-ɨ] and [ɨ-u] are less distinct than [i-u] and consequently dispreferred. But we will see in §4.1 that in the absence of front-back contrasts, vowels with intermediate F2 values, such as central vowels, are the unmarked case in many contexts. A number of languages, including Kabardian (Kuipers 1960, Choi 1991), and Marshallese (Bender 1968, Choi 1992), have

short vowel inventories which lack front-back contrasts. These so-called ‘vertical’ vowel systems consist of high and mid, or high, mid, and low vowels whose backness is conditioned by surrounding consonants, resulting in a variety of specific vowel qualities, many of which would be highly marked in a system with front-back contrasts, e.g. central vowels, back unrounded vowels, and short diphthongs. Crucially there are no vertical vowel inventories containing invariant [i] or [u], vowels which are ubiquitous in non-vertical inventories. That is, there are no vowel inventories such as [i, e, a] or [u, o, a].

This pattern makes perfect sense in terms of constraints on the distinctiveness on contrasts: as already discussed central vowels are not problematic in themselves, it is the contrast between front and central or back and central vowels which is marked ($*i-i$, $*i-u \gg *i-u$). In the absence of such F2-based contrasts, distinctiveness in F2 becomes irrelevant, and minimisation of effort becomes the key factor governing vowel backness. Effort minimisation dictates that vowels should accommodate to the articulatory requirements of neighboring consonants. This analysis is developed in §4.1.

These generalisations about vertical vowel systems show that the markedness of sounds depends on the contrasts that they enter into because sounds such as central vowels, which are marked when in contrast with front and back vowels, can be unmarked in the absence of such contrasts. The same pattern is observed in vowel reduction: when all vowel qualities are neutralized in unstressed syllables, as in English, the result is typically a ‘schwa’ vowel – a vowel type which is not permitted in stressed syllables in the same languages. This type of contrast-dependent markedness cannot be captured in terms of constraints on individual sounds. As Ní Chiosáin and Padgett (1997) point out, the cross-linguistic preference for front unrounded and back rounded vowels over central vowels suggests a universal ranking of segment markedness constraints as shown in (5), which implies any language with [i] will have [i, u] also. But this would imply that if only one of these vowels appears it should be [i] or [u], and certainly not a central vowel. More generally, this approach incorrectly predicts that if a sound type is unmarked, it should be unmarked regardless of the contrasts it enters into.

(5) $*i \gg *u, *i$

Constraints on the distinctiveness of contrasts, and their implications for phonology, are the focus of this paper. However it is also essential to consider general constraints, such as effort minimisation, that limit the distinctiveness of contrasts since actual contrasts are less than maximally distinct. So the first step is to situate constraints on the distinctiveness of contrasts within the context of a theory of phonological contrast. This is the topic of the next section. This model will then be applied to the analysis of particular phenomena, demonstrating the range of effects of distinctiveness constraints, and the difficulties that arise for models that do not include constraints on contrasts.

3 The dispersion theory of contrast

Constraints on the distinctiveness of contrasts are formalized here as part of a theory of contrast dubbed the ‘dispersion theory’ (Flemming 1995, 1996, 2001) after Lindblom’s (1986, 1990) ‘Theory of Adaptive Dispersion’, which it resembles in many respects. The core of the theory is the claim that the selection of phonological contrasts is subject to three functional goals:

- i. Maximize the distinctiveness of contrasts

- ii. Minimize articulatory effort
- iii. Maximize the number of contrasts

As noted above, a preference to maximize the distinctiveness of contrasts follows from language’s function as a means for the transmission of information. This tendency is hypothesized to be moderated by two conflicting goals. The first is a preference to minimize the expenditure of effort in speaking, which appears to be a general principle of human motor behaviour not specific to language. The second is a preference to maximize the number of phonological contrasts that are permitted in any given context in order to enable languages to differentiate a substantial vocabulary of words without words becoming excessively long.

The conflicts between these goals can be illustrated by considering the selection of contrasting sounds from a schematic two dimensional auditory space, shown in Figure 1. Figure 1a shows an inventory which includes only one contrast, but the contrast is maximally distinct, i.e. the two sound categories are well separated in the auditory space. If we try to fit more sounds into the same auditory space, the sounds will necessarily be closer together, i.e. the contrasts will be less distinct (Fig. 1b). Thus the goals of maximizing the number of contrasts and maximizing the distinctiveness of contrasts inherently conflict. Minimisation of effort also conflicts with maximizing distinctiveness. Assuming that not all sounds are equally easy to produce, attempting to minimize effort reduces the area of the auditory space available for selection of contrasts. For example, if we assume that sounds in the periphery of the space involve greater effort than those in the interior, then, to avoid effortful sounds it is necessary to restrict sounds to a reduced area of the space, thus the contrasts will be less distinct, as illustrated in fig. 1c. Note that while minimisation of effort and maximisation of the number of contrasts both conflict with maximisation of distinctiveness, they do not directly conflict with each other.

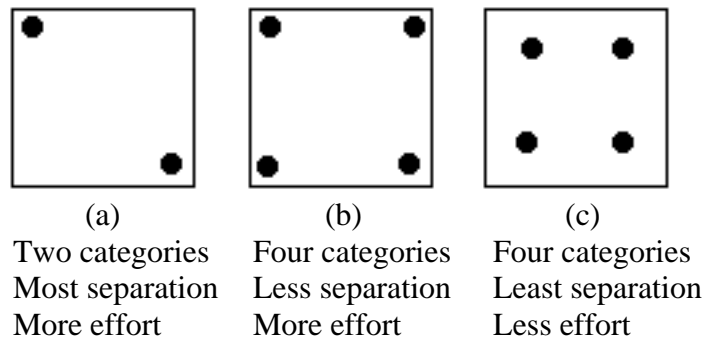


Fig 1. Selection of contrasts from a schematic auditory space.

3.1 Formulation of the constraints on contrast

Given that the three requirements on contrasts conflict, the selection of an inventory of contrasts involves achieving a balance between them. Optimality Theory (Prince and Smolensky 1993) provides a system for specifying the resolution of conflict between constraints, so this framework is adopted in formalizing dispersion theory. In this section

the functional goals for systems of contrasts posited above are formulated as Optimality-Theoretic constraints.

3.1.1 Maximize the distinctiveness of contrasts

Given the considerations outlined in §2, the measure of distinctiveness which is predicted to be relevant to the markedness of a contrast between two sounds is the probability of confusing the two sounds. Our understanding of the acoustic basis of confusability is limited, so any general model of distinctiveness is necessarily tentative. To allow the precise formulation of analyses, a fairly specific view of distinctiveness will be presented, but many of the details could be modified without affecting the central claims advanced here.

In psychological work on identification and categorisation it is common to conceive of stimuli (such as speech sounds) as being located in a multi-dimensional similarity space where the distance between stimuli is systematically related to the confusability of those stimuli – i.e. stimuli which are closer together in the space are more similar, and hence more confusable (e.g. Shepard 1957, Nosofsky 1992). This conception is adopted here. The domain in which we have the best understanding of perceptual space is vowel quality. There is good evidence that the main dimensions of the similarity space for vowels correspond well to the frequencies of the first two formants (Delattre, Liberman, Cooper, and Gerstman 1952, Plomp 1975, Shepard 1972), and less clear evidence for a dimension corresponding to the third formant (see Rosner and Pickering 1994:173ff. for a review).

A coarsely quantized three-dimensional vowel space, adequate for most of the analyses developed here, is shown in (6a-b) (cf. Liljencrants and Lindblom 1972). Sounds are specified by matrices of dimension values, e.g. [F1 1, F2 6, F3 3] for [i]. That is, dimensions are essentially scalar features so standard feature notation is used with the modification that dimensions take integer values rather than +/- . The locations of different vowel qualities are indicated as far as possible using IPA symbols. In some cases there is no IPA symbol for a particular vowel quality (e.g. the unrounded counterpart to [u] which might occupy [F1 2, F2 2]), while in many cases more than one vowel could occupy a given position in F1-F2 space due to the similar acoustic effects of lip rounding and tongue backing, e.g. central rounded [ø] occupies the same position as back unrounded [ɯ]. Also, the IPA low back unrounded vowel symbol [ɑ] is used for a wide range of vowel qualities in transcriptions of English dialects and could have been used to symbolize [F1 7, F2 2]. Similarly, [y] could also have been used for [F1 1, F2 5].

(6) a.

						F2						
						6	5	4	3	2	1	
						i	ɨ	y	ɨ	ʉ	u	1
							ɪ	ʏ			ʊ	2
							ɛ	ø		ɤ	ɔ	3
							e	ø	ə	ɣ	o	4 F1
								ɛ	ɐ	ʌ	ɔ	5
									æ	ɜ	ɑ	6
										a	ɑ	7

b.

			F3			
			3	2	1	
			i		y,ʉ,u	1
				ɪ	ʏ,ʊ	2
				ɛ	ø,ɤ,ɔ	3
				e	ø,ɣ,o	4 F1
				ɛ	ʌ,ɔ	5
					ɑ	6
					a	7

The distinctiveness of a pair of vowel qualities should then be calculated from the differences on each of these three dimensions. However, relative distinctiveness on a single dimension can be determined with much greater confidence than distinctiveness involving differences on multiple dimensions, so almost all of the analyses developed here are cases that can be formulated as the selection of a set of contrasting sounds along one perceptual dimension. Consequently we will concentrate on formalizing this restricted case. Contrasts on multiple dimensions are discussed in detail in Flemming (2001).

The requirement that the auditory distinctiveness of contrasts should be maximized can be decomposed into a ranked set of constraints requiring a specified minimal auditory distance between contrasting forms (7) (Flemming 1995). The required distance is indicated in the format *Dimension:distance*, e.g. 'MINDIST = F1:2' is satisfied by contrasting sounds that differ by at least 2 on the F1 dimension.

(7) MINDIST = F1:1 >> MINDIST = F1:2 >>... >> MINDIST = F1:4

To encode the fact that auditory distinctiveness should be maximized, MINDIST = *D:n* is ranked above MINDIST = *D:n+1*, i.e. the less distinct the contrast, the greater the violation.

3.1.2 Maximize the number of contrasts

The requirement that the number of contrasts should be maximized can be implemented in terms of a positive constraint, MAXIMIZE CONTRASTS, that counts the number of contrasts in the candidate inventory (Flemming 2001). The largest inventory or inventories are selected by this constraint, all others are eliminated. Of course the largest candidate inventories will usually have been eliminated by higher-ranked constraints, so this constraint actually selects the largest viable inventory.

3.1.3 Balancing the requirements on contrasts

The language-specific balance between these first two constraints on contrasts is modeled by specifying the language-specific ranking of the constraint MAXIMIZE CONTRASTS in the hierarchy of MINDIST constraints. Effectively, the first MINDIST constraint to outrank MAXIMIZE CONTRASTS sets a threshold distance, and the optimal inventory is the one that packs the most contrasting sounds onto the relevant dimension without any pair being closer than this threshold.

The conflict between the two constraints on contrasts is illustrated in the tableau in (8). This tableau shows inventories of contrasting vowel heights and their evaluation by MINDIST and MAXIMIZE contrasts constraints. We are considering constraints on contrasts so the candidates evaluated here are sets of contrasting forms rather than outputs for a given input. For simplicity, we are considering only a single perceptual dimension, so the individual vowels are representative of distinctive heights.

MINDIST constraints assign one mark for each pair of contrasting sounds which are not separated by at least the specified minimum distance. For example, candidate (b) violates MINDIST = F1:4 twice because the contrasting pairs [i-e] and [e-a] violate this constraint while [i-a] satisfies it, being separated by a distance of 6 on the F1 dimension. (Note that the number of violations will generally be irrelevant for MINDIST constraints ranked above MAXIMIZE CONTRASTS because it will always be possible to satisfy the MINDIST constraint by eliminating contrasts).


MAXIMIZE CONTRASTS is a positive scalar constraint, according to which more contrasts are better, so evaluation by this constraint is indicated using one check mark (✓) for each contrasting sound category – more check marks indicate a better candidate according to this constraint. The conflict between the two constraint types is readily apparent in (8): sets of vowel height contrasts which better satisfy MAXIMIZE F1 CONTRASTS incur worse violations of the MINDIST constraints.

(8)

	MINDIST = F1:1	MINDIST = F1:2	MINDIST = F1:3	MINDIST = F1:4	MINDIST = F1:5	MAXIMIZE CONTRASTS
a. i-a						✓✓
b. i-e-a				**	**	✓✓✓
c. i- <u>e</u> -e-a			***	***	*****	✓✓✓✓
d. i-I- <u>e</u> -e-a		**	*****	*****	*****	✓✓✓✓✓


The effect of ranking MAXIMIZE CONTRASTS at different points in the fixed hierarchy of MINDIST constraints is illustrated by the tableaux in (9) and (10). The ranking in (9) yields three distinct vowel heights – i.e. the winning candidate is (b). This candidate violates MINDIST = F1:3, but any attempt to satisfy this constraint by improving distinctiveness, as in candidate (a), violates higher-ranked MAXIMIZE CONTRASTS by selecting only two contrasting vowel heights. It is not possible to fit three contrasting vowels with a minimum separation of 3 features on the F1 dimension. Candidate (c) better satisfies MAXIMIZE CONTRASTS than (b), maintaining four contrasting vowel heights, but (c) violates higher-ranked MINDIST = F1:2 since [e-ε] and [ε-a] each differ by only 1 on the F1 dimension.

(9)

	MINDIST = F1:2	MINDIST = F1:3	MAXIMIZE CONTRASTS	MINDIST = F1:4	MINDIST = F1:5
a. i-a			✓✓!		
b.  i-e-a			✓✓✓	**	**
c. i-ε-ε-a		*!***	✓✓✓✓	***	*****

Thus the particular balance achieved here between maximizing the number of contrasts and maximizing the distinctiveness of the contrasts yields three contrasting heights. Altering the ranking of MAXIMIZE CONTRASTS results in a different balance. For example, if less weight is given to maximizing the number of contrasts, ranking MAXIMIZE CONTRASTS below MINDIST = F1:3, the winning candidate has just two contrasting vowel heights, differing maximally in F1. It is apparent that the maximally distinct F1 contrast [i-a] is preferred over any sub-maximal contrast, such as [i-æ] (which violates MINDIST = F1:6), although this comparison is not included in the tableau.

(10)

	MINDIST = F1:2	MINDIST = F1:3	MINDIST = F1:4	MAXIMIZE CONTRASTS	MINDIST = F1:5
a.  i-a				✓✓	
b. i-e-a			*!*	✓✓✓	**
c. i-ε-ε-a		*!***	***	✓✓✓✓	*****

Not all conceivable rankings of MAXIMIZE CONTRASTS correspond to possible languages. The balance between maximisation of the number of contrasts and maximisation of the distinctiveness of contrasts is determined by the ranking of MAXIMIZE CONTRASTS relative to the MINDIST constraints. Allowing all definable rankings predicts the existence of languages which value the number of contrasts very highly, resulting in a huge number of very fine contrasts, and languages which value distinctiveness very highly, resulting in a handful of maximally distinct contrasts. Neither of these extremes is attested. It seems that there is a lower bound on the distinctiveness required for a contrast to be functional, and that there is an upper bound beyond which additional distinctiveness provides a poor return on the effort expended. This could be implemented by specifying that certain MINDIST constraints, referring to the smallest

acceptable contrastive differences, are universally ranked above MAXIMIZE CONTRASTS, and that MAXIMIZE CONTRASTS is in turn universally ranked above another set of MINDIST constraints which make ‘excessive’ distinctiveness requirements. However it would be desirable to derive these bounds from general considerations of perceptibility and communicative efficiency rather than simply stipulating them.

Note that the need to place limits on possible constraint rankings is not novel to the dispersion theory. The same issue arises with respect to standard faithfulness constraints: If all faithfulness constraints are at the top of the ranking then all inputs will surface as well-formed outputs, i.e. this ranking would yield an unattested language with no restrictions on the form of words. Conversely, if all faithfulness constraints were at the bottom of the ranking then all inputs would be mapped to a single, maximally well-formed output (presumably the null output, i.e. silence).

3.1.4 Minimisation of effort

The analyses above do not include effort minimisation constraints. No general account of the effort involved in speech production will be proposed here, instead specific constraints such as ‘Don’t voice obstruents’ and ‘Don’t have short low vowels’ will be motivated as they become relevant. If a sound violates an effort constraint which outranks MAXIMIZE CONTRASTS, it will not be employed even if it would allow more contrasts or more distinct contrasts.

3.2 Some effects of MINDIST constraints

3.2.1 Dispersion

The most basic consequence of the distinctiveness constraints (MINDIST constraints) is a preference for distinct contrasts. This gives rise to dispersion effects whereby contrasting sounds tend to be evenly distributed over as much auditory space as effort constraints will allow (cf. Liljencrants and Lindblom 1972, Lindblom 1986). This effect has already been demonstrated above in relation to F1 contrasts, and the preference for front unrounded and back rounded vowels discussed in section 2 is another instance of this tendency, applied to contrasts on the F2 dimension. The acoustic effects of lip-rounding mean that the maximal F2 difference is between front unrounded vowels and back rounded vowels (11), so if maximisation of distinctiveness of F2 contrasts outranks maximizing the number of contrasts, these are the vowels that will be selected (12). F2 contrasts involving central vowels are necessarily sub-maximal, and thus are dispreferred. Of course, the appearance of non-peripheral vowels may be motivated by the desire to maximize contrasts – i.e. if MAXIMIZE CONTRASTS is ranked above MINDIST = F2:3.

(11) F2:

6	5	4	3	2	1
i	ɨ	y	ɨ	ʉ	u

(12)

		MINDIST = F2:3	MAXIMIZE CONTRASTS	MINDIST = F2:4	MINDIST = F2:5
a.	☞ i-u		✓✓		

b.	i-u		✓✓		*!
c.	y-u		✓✓		*!
d.	i-i		✓✓	*!	*
e.	i-i-u	*!	✓✓✓	**	**

This notion of dispersion of contrasting sounds is also closely related to the concept of ‘enhancement’, a term coined by Stevens, Keyser, and Kawasaki (1986). Stevens et al observe that ‘basic’ distinctive features are often accompanied by ‘redundant’ features which ‘strengthen the acoustic representation of distinctive features and contribute additional properties which help the listener to perceive the distinction’ (p.426). They regard the relationship between [back] and [round] in vowels as one of enhancement: [round] enhances distinctive [back]. In terms of the dispersion theory, this can be reformulated as the observation that independent articulations often combine to yield more distinct contrasts.

3.2.2 Neutralisation

A second basic effect of Mindist constraints, in interaction with the other dispersion theoretic constraints, is neutralisation of indistinct contrasts. In dispersion theory, neutralisation of a contrast results when constraints prevent it from achieving sufficient distinctiveness in some environment. That is, in a ranking of the form shown in (13) where *EFFORT is an effort minimisation constraint penalizing some articulation, a contrast will be neutralized in some context if it cannot be realized with a distinctiveness of d without violating *EFFORT.

$$(13) \quad \text{MINDIST} = d, *EFFORT \gg \text{MAXIMIZE CONTRASTS}$$

The distinctiveness that can be achieved for a given degree of effort varies across contexts. Some cues to contrasts are simply unavailable in certain contexts, for example release formant transitions are not available as a cue to consonant place if the consonant is not released into an approximant. In addition the articulatory effort involved in realizing a cue is generally highly context-dependent, for example voicing of an obstruent is more difficult following a voiceless sound than following a voiced sound because it is more difficult to initiate voicing than to sustain it (Westbury and Keating 1986). So the possibility of realizing a contrast that satisfies $\text{MINDIST} = d$ without violating *EFFORT depends on context, and consequently a given type of contrast may be selected as optimal in some contexts and not in others – i.e. the contrast is neutralized in those other contexts. For example, consonant place contrasts may be permitted before sonorants, but neutralized before obstruents, where stop bursts and release transitions are not available. Thus dispersion theory provides an account of Steriade’s (1995, 1997) generalisation that contrasts are neutralized first in environments where ‘the cues to the relevant contrast

would be diminished or obtainable only at the cost of additional articulatory maneuvers' (Steriade 1997:1).

It is important to note that the ranking of other constraints will typically be crucial in making the realisation of a distinct contrast more effortful in a particular context – e.g. stop bursts will only be absent before obstruents if some constraint requires the stop closure to overlap with the following consonant. In the example we will consider here metrical constraints on unstressed vowel duration make distinct vowel contrasts more difficult to realize in unstressed syllables.

The analysis of neutralisation will be exemplified with analyses of two common patterns of vowel reduction: reduction from a seven vowel inventory (14i) in primary stressed syllables to a five vowel inventory (14ii) in other syllables, as in Central Italian dialects (Maiden 1995), and reduction from a five vowel inventory (14ii) in primary stressed syllables to a three vowel inventory (14iii) elsewhere, as in Southern Italian dialects (Maiden 1995) and Russian (Halle 1959).

(14)	(i)	i	u	(ii)	i	u	(iii)	i	u
		e	o		e	o		a	
		ɛ	ɔ		a				
		a							

The Central Italian pattern is exemplified in (15) with data from standard Italian (as described in dictionaries). The pairs of words on each line are morphologically related so the parenthesized forms illustrate alternations that arise when stress is shifted off a vowel which cannot appear in an unstressed syllable.

(15)	<u>stressed vowels</u>		<u>unstressed vowels</u>	
	[i]	víno 'wine'	vinífero	'wine-producing'
	[e]	péska 'fishing'	peskáre	'to fish'
	[ɛ]	bél:o 'beautiful'	(bel:íno	'pretty')
	[a]	máno 'hand'	manuále	'manual'
	[ɔ]	mól:e 'soft'	(mol:eménte	'softly')
	[o]	nóme 'name'	nomináre	'to name, call'
	[u]	kúra 'care'	kuráre	'to treat'

The Southern Italian pattern is exemplified by the dialect of Mistretta, Sicily (Mazzola 1976) (16).

(16)	<u>stressed vowels</u>		<u>unstressed vowels</u>	
	[i]	vín:i 'he sells'	vin:ímu	'we sell'
	[e]	véni 'he comes'	(vin:ímu	'we come')
	[a]	ávi 'he has'	avíti	'he has'

[o]	móri	‘he dies’	(murímu	‘we die’)
[u]	úfi	‘he boils’	ufímu	‘we boil’

These patterns of reduction involve neutralisation of F1 contrasts only. According to the analysis of neutralisation outlined above, this implies that it is more difficult to produce distinct F1 contrasts in unstressed positions. The most likely source of that difference in difficulty is the difference in duration between primary stressed and other vowels in these languages. So the proposed analysis is that producing low vowels is increasingly difficult as vowel duration is reduced, and this motivates raising of short low vowels, leaving a smaller range of the F1 dimension for distinguishing F1 contrasts. This in turn can result in the selection of a smaller number of contrasts.¹

The most direct evidence for a relationship between vowel duration and the ability to achieve a high F1 comes from Lindblom’s (1963) finding that the F1 of Swedish non-high vowels decreases exponentially as vowel duration decreases. It is also well established that low vowels are longer than high vowels, other things being equal (Lehiste 1970). These effects are commonly attributed to the greater articulator movement involved in producing a low vowel between consonants: low vowels require an open upper vocal tract to produce a high F1, whereas all consonants (other than pharyngeals and laryngeals) require upper-vocal tract constrictions, so producing a low vowel between consonants requires substantial opening and closing movements. Westbury and Keating (1980, cited in Keating 1985) provide evidence that vowel duration differences are indeed related to distance moved: they found that vowels with lower jaw positions had longer durations in a study of English. Thus producing a low vowel with the same duration as a higher vowel will typically require faster, and consequently more effortful, movements. Reduction of low /a/ to [i] or [ə] in unstressed syllables is accordingly commonly reported both impressionistically, and in experimental studies such as Lindblom (1963).

This correlation between duration and raising of low vowels has been observed in Central Italian also: a study of vowels in the speech of five male Italian television news readers (Albano Leoni et al 1995) found that /a/ in a primary stressed syllable was twice as long as medial unstressed /a/, and the mean F1 of /a/ was 750 Hz in primary stressed syllables, but 553 Hz in medial unstressed syllables (close to the F1 of a stressed lower-mid vowel)². So the inventory in unstressed positions is more accurately transcribed as [i, e, ə, o, u], where [ə] is a lower-mid central vowel. High and higher-mid vowels had essentially the same F1 in stressed and unstressed positions.

While the direct effect of vowel shortening is to increase the difficulty of producing low vowels, this has obvious consequences for the selection of F1 contrasts: if the lowest vowel in an inventory is lower-mid ([F1 5] in the terms used here) this leaves less room for distinguishing F1 contrasts than in stressed syllables where the lowest vowel is truly

¹ Crosswhite (this volume) proposes a conceptually similar analysis of these patterns of vowel reduction, although the formalization is very different. She also suggests that vowel raising is desirable in unstressed syllables because it lowers the sonority of the vowel, resulting in a better correspondence between stress and vowel intensity, as well as being motivated by effort minimization.

² Word-final unstressed syllables were more variable in duration, probably because duration in this position is dependent on phrase-final lengthening effects. F1 of final /a/ was correspondingly more variable. The greater duration of phrase final vowels does not lead to a larger vowel inventory in this position – this is probably a ‘uniformity’ effect (Steriade 1997, 2000), i.e. it allows words to have a more consistent realization across phrasal positions.

low ([F1 7]), so it is not possible to maintain the same number of height contrasts with the same distinctiveness. Consequently three vowel heights are selected in unstressed syllables, and four in longer, stressed syllables.

This analysis can be formalized in terms of the constraint ranking in (17). The positions of relevant vowels on the F1 dimension are shown in (18).

- (17) UNSTRESSED VOWELS ARE SHORT, *SHORT LOW V, MINDIST = F1:2 >>
MAXIMIZE CONTRASTS >> MINDIST = F1:3

(18) F1: 7 6 5 4 3 2 1

a	ɐ	ɛ	e	ɛ̣	ɪ	i
		ɐ	ə			

UNSTRESSED VOWELS ARE SHORT is a place-holder for whatever constraints require unstressed vowels to be shorter than stressed vowels. *SHORT LOW V is a constraint against expending the effort to produce a short low vowel. This should properly be derived from a general model of articulatory effort (cf. Kirchner, this volume), but for present purposes we will formalize it as a constraint that penalizes short vowels with F1 of greater than 5 on the scale in (18).

In stressed syllables, the first two constraints are irrelevant, so the ranking yields four vowel heights, each separated by F1:2, as shown in (19). However, in unstressed syllables, high-ranking UNSTRESSED VOWELS ARE SHORT requires short vowels, so *SHORT LOW V is applicable also. This effort minimisation constraint penalizes low vowels, so the candidate [i-ɛ̣-ɛ-a] is now ruled out because [a] has [F2 7] (20a). Attempting to maintain four contrasts while avoiding low vowels, as in candidate (b), results in violations of MINDIST = F1:2 because [ɛ-ɐ] don't differ in F1. The winning candidate has three vowel heights, and so is evaluated as worse by MAXIMIZE CONTRASTS, but satisfies the higher-ranked minimum distance requirement.

- (19) Central Italian – Vowels in primary stressed syllables.

	*SHORT LOW V	MINDIST = F1:2	MAXIMIZE CONTRASTS	MINDIST = F1:3
a.	í-á		✓✓!	
b.	í-é-á		✓✓✓!	
c.	☞ í-ɛ̣-é-á		✓✓✓✓	***

- (20) Central Italian – Vowels in unstressed syllables.

	*SHORT LOW V	MINDIST = F1:2	MAXIMIZE CONTRASTS	MINDIST = F1:3
a.	i-ɛ̣-ɛ-a	*!	✓✓✓✓	***
b.	i-ɛ̣-ɛ-ɐ	*!	✓✓✓✓	***
c.	☞ i-ɛ̣-ɐ		✓✓✓	**

A similar ranking (21) derives the Southern Italian pattern in which three vowel heights are contrasted in primary stressed syllables, but only two elsewhere. The only difference is that both MINDIST constraints are ranked above MAXIMIZE CONTRASTS – i.e. distinctiveness requirements are more demanding.

- (21) UNSTRESSED VOWELS ARE SHORT, *SHORT LOW V, MINDIST = F1:2 >> MINDIST = F1:3 >> MAXIMIZE F1 CONTRASTS

This is the same ranking of MINDIST constraints used to derive three contrasting vowel heights in (9) above, and the same derivation applies in primary stressed syllables, where the top-ranked constraints are irrelevant. In unstressed syllables, *SHORT LOW V is applicable again, so the lowest vowel possible is [ɐ], and it is not possible to fit a vowel between [i] and [ɐ] while satisfying MINDIST = F1:3 (22b), so only two vowel heights are selected (22c).

- (22) Southern Italian – Vowels in unstressed syllables.

		*SHORT LOW V	MINDIST = F1:2	MINDIST = F1:3	MAXIMIZE CONTRASTS
a.	i-e-a	*!			✓✓✓
b.	i-ɛ-ɐ			*!*	✓✓✓
c.	☞ i-ɐ				✓✓

This analysis is based on the assumption that the difference between the two patterns of reduction lies in the ranking of MINDIST constraints, but there may also be differences in the characteristic durations of the unstressed vowels. The difficulty of producing a low vowel increases as vowel duration decreases, so if Southern Italian unstressed vowels are shorter than Central Italian unstressed vowels, then more raising of low vowels may occur, making reduction to a two-height system more desirable. Good evidence that different degrees of shortening can result in different degrees of reduction in this way is provided by Brazilian Portuguese. Brazilian Portuguese combines the two patterns of reduction: the seven vowel system (14i) is permitted in primary stressed syllables, the five vowel system (14ii) in syllables preceding the stress, and the three vowel system (14iii) in unstressed final syllables (stress is generally penultimate) (Mattoso Camara 1972). Since both patterns of reduction occur in the same language, they cannot be accounted for in terms of differences in the ranking of MINDIST constraints, but they can be accounted for in terms of differences in vowel duration. Major (1992) found that final unstressed syllables are substantially shorter than pre-stress syllables (which are in turn shorter than stressed syllables),³ so the same degree of effort should result in higher

³ Moraes (1998) found relatively small differences in duration between pre-tonic and final unstressed vowels, but he only measured high vowels which tend to be short in any case. Major measured low vowels, which are more relevant here. Moraes (1998) also shows that the duration difference can be eliminated by phrase-final lengthening of final unstressed syllables. As in Italian, it appears to be the phrase-medial

vowels in this position. The ‘low’ vowel is indeed higher in this position, as indicated by the standard impressionistic transcription of the final unstressed vowel system as [i, ə, u] (e.g. Mattoso Camara 1972). Acoustic data reported by Fails and Clegg (1992) shows a progressive decrease in F1 of the lowest vowel from primary stressed, to pre-stress, to final unstressed.

We will see that duration-based neutralisation is also central to some of the case studies that provide more direct evidence for distinctiveness constraints (4.1), but before turning to those cases, we will consider some additional issues in the formulation and application of dispersion theory.

3.2.3 Analysis of words and alternations


The analysis of vowel reduction raises two important issues concerning analyses using dispersion theory. First, we have only considered the selection of inventories of contrasting sounds but a phonology must characterize the set of well-formed possible words in a language. The implication of dispersion theory is that words must be evaluated with respect to paradigmatic constraints in addition to the familiar syntagmatic markedness constraints, such as effort minimisation and metrical constraints. That is, words must be sufficiently distinct from other minimally contrasting possible words (MINDIST), and there must be a sufficient number of such contrasting words (MAXIMIZE CONTRASTS). Deriving inventories of sounds in particular contexts is an important step towards the analysis of complete words because for a word to be well-formed, each sound in that word must be a member of the optimal inventory for its particular context. We will see in section 5 that developing this basic idea is not simple, but we will postpone that discussion until we have more thoroughly motivated the constraints on contrast.

The second issue raised by the analysis of vowel reduction is how morpheme alternations should be analysed. The analysis in §3.2.2 derives the distributional fact that [e] is not permitted in short, unstressed syllables in Sicilian Italian, but it says nothing about the fact that [e] alternates with [i] when stress shifts off it, e.g. [véni ~ vinímu] (16). In standard OT, the analysis of alternations centers on faithfulness to the underlying representation of a morpheme, but it is not possible to combine dispersion constraints with the faithfulness-based account of allomorphic similarity because the two are fundamentally incompatible. This is illustrated in (23). This tableau repeats the ranking used in (9) to derive three contrasting vowel heights [i-e-a], with the addition of a top-ranked faithfulness constraint IDENT [F1], which requires that output segments have the same [F1] value as the corresponding input segment – i.e. input values of [F1] must be preserved in the output. The problem arises where the input contains a vowel which is not part of the inventory derived by the dispersion constraints, as in (23). In the candidate inventories, the underlined form is the output corresponding to input /i/, whereas the other forms are the set of contrasting vowels required for the evaluation of constraints on contrast.

characteristics that are relevant to neutralizing vowel reduction. It is also interesting to note that Moraes found that final unstressed vowels have much lower intensity than vowels in other positions, and that this remains true even with final-lengthening (this should not be a consequence of vowel raising, since all vowels were high). Intensity should play some role in the perceptibility of vowel contrasts, but this factor is not analyzed here.

The inclusion of faithfulness constraints subverts the intended effect of the MINDIST and MAXIMIZE CONTRASTS constraints, because it makes the selected inventory of vowel height contrasts dependent on the input under consideration – the same constraints that are supposed to derive three vowel heights, as in (9), yield two [I-a] in (23) because faithfulness to the input F1 forces inclusion of [I] in an output form. The expected three vowel heights are derived if the input is /i/, for example.

(23)

/I/	IDENT [F1]	MINDIST = F1:3	MAXIMIZE CONTRASTS	MINDIST = F1:4
a. <u>i</u> -e-a	*!		✓✓✓	**
b. I-e-a		*!	✓✓✓	**
c.  I-a			✓✓	
d. <u>i</u> -a	*!		✓✓	

In Flemming (1995) it is proposed that allomorphy should be analysed in terms of a direct requirement of similarity between the surface forms of a morpheme, i.e. ‘output-output correspondence’ or ‘paradigm uniformity’ constraints. These constraints have been used to account for cyclicity and related effects (e.g. Benua 1997, Burzio 1998, Kenstowicz 1996, Steriade 1997, 2000), but, as Burzio (1998) observes, they can naturally be extended to account for all similarity relations between realisations of a morpheme including those observed in allomorphy, eliminating any role for an input. So [véni] alternates with [vínimu] because [i] is the most similar to [e] of the vowels that are permitted in unstressed syllables. However most of the analyses considered here concern distribution rather than alternations, so we will not pursue this line further here.

4 Evidence for constraints on contrasts

Now we have laid out the basics of a theory which incorporates constraints on the distinctiveness of contrasts, the theory will be applied in the analysis of phenomena which illustrate the effects of these constraints, and which are problematic for other theories. In general terms, the case studies provide evidence for the central prediction that the markedness of a sound depends on the sounds that it contrasts with. Without constraints on contrasts, markedness is predicted to be purely a property of sounds, so the markedness of a sound should be independent of the nature of the sounds that it contrasts with.

4.1 F2 contrasts and vowel dispersion

The first case study concerns the preference for front unrounded and back rounded vowels discussed in section 2. This pattern has already been analysed as a result of the preference for maximally distinct contrasts, i.e. the MINDIST constraints (§3.2.1): the maximal F2 difference is between front unrounded vowels and back rounded vowels, so if maximisation of distinctiveness of F2 contrasts outranks maximizing the number of contrasts, these are the vowels that will be selected. F2 contrasts involving non-peripheral vowels (central vowels, front rounded vowels, etc) are necessarily sub-maximal, and thus are dispreferred.

The novel prediction made by this analysis is that front unrounded and back rounded vowels should only be preferred where there are F2 contrasts. If there are no vowel contrasts that are primarily realized in terms of F2 differences, other constraints are predicted to govern backness and rounding of vowels, the most general of which are effort minimisation constraints. It is unusual for all vowel F2 contrasts to be neutralized, but there are two circumstances in which this happens: in ‘vertical’ vowel inventories, and in fully neutralizing vowel reduction in unstressed syllables, as in English reduction to ‘schwa’. In both cases the predictions of the dispersion-theoretic analysis are confirmed: we do not find front unrounded or back rounded vowels in most contexts, rather backness and rounding are governed by minimisation of effort. This means that they are realized as smooth transitions between preceding and following consonants, which frequently results in central or centralized vowel qualities.

4.1.1 Vowel qualities in the absence of F2 contrasts

Vowel inventories which lack front-back contrasts are found in Marshallese (Bender 1968, Choi 1992), Northwest Caucasian languages (Colarusso 1988) including Kabardian (Kuipers 1960, Colarusso 1992) and Shapsug (Smeets 1984), and some Ndu languages of Papua New Guinea including Iatmul (Laycock 1965, Staalsen 1966). These languages are typically described as having only central vowels, however this is a claim about the underlying vowel inventory posited as part of a derivational analysis, not an observation about the surface vowels. On the surface, all of these languages distinguish short vowels from longer vowels, with conventional F2 contrasts among the longer vowels, but no F2 contrasts among the short vowels. For example, the Northwest Caucasian languages Kabardian and Shapsug have a system of five normal length vowels [i, e, a, o, u] (Kuipers 1960:23f., Smeets 1984:123), and a ‘vertical’ system of two short vowels, which can be transcribed broadly as [i, ə].⁴ However the precise backness and rounding of these vowels depends on their consonantal context. Colarusso (1988) states that in NW Caucasian languages, ‘The sequence C₁iC₂ means “go from 1 to 2, letting your tongue follow the shortest path that permits an interval of sonorant voicing.” C₁əC₂ means “go from 1 to 2... but at the same time imposing on this trajectory an articulatory gesture which pulls the tongue body down and back.”’⁵ (p.307). Marshallese also has a long vowel inventory with F2 contrasts, but the medial short vowels /i, ə, a/ contrast in height only (Bender 1968). Again the backness and rounding of these vowels is dependent on consonant context: Choi (1992) shows that the F2 trajectory of these vowels is a nearly linear interpolation between F2 values determined by the preceding and following consonants.

These transitional vowel qualities are plausibly analysed as the result of effort minimisation. Although articulatory effort is not well understood, basic considerations imply that higher velocity movements should be more effortful (Kirchner this volume, Nelson 1983, Perkell 1997), and velocity of movement in a vowel is minimized by adopting a linear trajectory between preceding and following consonants. Some deviation from a linear vowel height trajectory is necessary to achieve a vocalic degree of stricture,

⁴ Kuipers actually transcribes the Kabardian high vowel as [ə], the mid-vowel as [a], and the ‘long’ low vowel as [ā], and Colarusso (1988) follows him in this, but their descriptions, Colarusso’s phonetic transcriptions, and acoustic data in Choi (1991) all indicate that the vowels are actually high and mid respectively.

⁵ The transcription of vowels has been altered in accordance with conventions adopted here.

and to realize F1 contrasts, but backness and rounding can be interpolated between preceding and following consonants, producing the near-linear F2 movements observed by Choi.

So vertical vowel systems are what we expect given the analysis of F2 dispersion above – where F2 contrasts are neutralized backness and rounding of vowels are determined by effort minimisation. The resulting vowel qualities are often central, back unrounded, front rounded, or short diphthongs involving these qualities. These are all vowel types which would be highly marked in the presence of F2 contrasts, so the markedness of vowel qualities depends on the contrasts that they enter into.

This conclusion holds even more clearly if we follow Choi (1992) in analyzing these vowels as being phonetically underspecified for [back] and [round] – i.e. these vowels lack specifications for these features in the output of the phonology, and the specific contextual allophones are generated through a process of phonetic interpolation. Such unspecified vowels only occur in the absence of F2 contrasts, so they are not just marked in the presence of F2 contrasts, they are unattested.

The other situation in which we find neutralisation of F2 contrasts is in vowel reduction. In languages such as English (Hayes 1995), Southern Italian dialects (Maiden 1995) and Dutch (Booij 1995) all vowel quality distinctions are neutralized in some unstressed syllables. The resulting vowel is usually referred to as ‘schwa’. Phonetic studies of schwa in Dutch (van Bergem 1994) and English (Kondo 1994) indicate that this vowel can also be analysed as the result of effort minimisation predominating where vowel contrasts are neutralized⁶. As in vertical vowels, F2 in schwa is an almost linear interpolation between values for adjacent consonants⁷. Since there are no height contrasts, F1 of schwa is expected to be transitional also. In most consonant contexts an opening movement is required to realize a vocalic stricture, but minimizing this opening movement results in a vowel with low F1, comparable to that of high vowels, as observed by Van Bergem (1994) and Kondo (1994).⁸ However these studies did not include schwas adjacent to non-high vowels, or separated from them by laryngeals (as in ‘saw another’). Examination of sequences of this kind in English suggests that F1 interpolates from the low vowel to the following consonant, which can result in a relatively high F1 during schwa.

Preliminary investigation of the Southern Italian dialect of Bari, based on recordings provided with Valente (1975), suggests that schwa is much the same as in English and Dutch. It is also predicted that the schwa vowels that break up consonant clusters in some Berber and Salishan languages (Dell and Elmedlaoui 1996, Flemming *et al* 1994) should be similar transitional vowels since there are no vowel quality contrasts in these positions. This appears to be correct for Montana Salish.

Schwa is not permitted where there are vowel quality contrasts (in stressed syllables), but is the unmarked vowel where quality contrasts are neutralized (e.g. in unstressed syllables). So reduction to schwa demonstrates a similar pattern of contrast-dependent markedness to that observed in vertical vowel languages. These patterns cannot be captured in terms of constraints on individual sounds. Ní Chiosáin and Padgett

⁶ Van Bergem (1994) also concludes that Dutch schwa is the minimum-effort vowel.

⁷ Consonant F2 is influenced by adjacent full vowels, so the vowel environment also influences schwa quality. There are no data on the influence of vowel environment on vertical vowel quality.

⁸ The use of IPA [ə] to transcribe this vowel is thus a source of confusion since [ə] is supposed to be a mid central vowel. The fact that schwa is typically a high vowel with transitional F2 helps to explain the use of the [ə] symbol to transcribe high vertical vowels in Caucasian (e.g. Kuipers 1960, Smeets 1984).

(1997) succinctly formulate the problem for theories without constraints on contrast as follows: the cross-linguistic preference for peripheral front unrounded and back rounded vowels over non-peripheral qualities vowels implies a universal ranking of segment markedness constraints as shown in (24).

(24) *i, *y, *ɯ >> *i, *u

This ranking implies that [i, u] should always be preferred to non-peripheral vowels, so [i] or [u] should be expected to appear in cases of neutralisation. For central vowels to appear as a result of neutralisation, *i would have to rank below *i, *u, but that would allow the derivation of unattested basic vowel inventories such as [i, a, u] or [i, i, a]. More generally, without constraints on contrast, inventories should always include the least marked sounds, no matter what the size of the inventory is. There is simply no way to directly capture contrast-dependent generalisations about markedness.

Socratic question #3: Assess the coherence of this proposal, assuming “classical” OT: “*i, if either /i/ or /u/ is present in the inventory of vowel phonemes.”

The same applies if it is assumed that transitional vowels are simply unspecified for [back] and [round], or [F2]. A constraint against such unspecified vowels would have to be ranked above constraints such as *i, *u to prevent transitional vowels from surfacing in F2 contrasts, but such a ranking implies that unspecified vowels should always be dispreferred, even in neutralisation.

It is often possible to propose a re-analysis of a pattern of contrast-dependent markedness as positional markedness. For example, vertical vowel inventories seem to be restricted to extra-short vowels, and the schwa found in neutralizing reduction is very short (see below), so it is possible to formulate a constraint against vowel qualities with non-transitional F2 among extra-short vowels and restrict the markedness constraints in (24) to apply only to longer vowels (25).

(25) *NORMAL DURATION[i] >> *NORMAL DURATION[i], *NORMAL DURATION[u]
 *EXTRA-SHORT[i], *EXTRA-SHORT[u] >> *EXTRA-SHORT[i]

This strategy runs into difficulties because the full typology of extra-short vowels is more complex. Neutralisation to a single vowel quality results in a schwa vowel in which both F1 and F2 are essentially transitional (although F1 must be above a certain minimum). This implies that schwa should be the least marked extra-short vowel. But if this is the case, schwa should be found in all inventories of extra-short vowels, which is not the case. Vertical vowels have specific F1 targets, and transitional schwa is also excluded from the extra-short vowel inventory [i, ə, u] (where [ə] is used in the IPA sense of a mid central rounded vowel), found in unstressed final vowels in Brazilian Portuguese (§3.2.2) and most unstressed syllables in Standard Russian (Crosswhite this volume).

So even among extra-short vowels, markedness depends on the system of contrasts, making it impossible to arrange vowel types in a single markedness hierarchy. Perhaps some basis could be found for differentiating the positions in which we find reduction to [i, ə, u], positions in which we find reduction to schwa, and positions in which we find vertical vowels. Then it would be possible to posit distinct hierarchies of vowel markedness for each type of position, but such a proliferation of increasingly specific constraints should prompt us to seek more general organizing principles, as we have done here. For example, positing position-specific hierarchies leaves it as a remarkable coincidence that vowels with transitional F2 are unmarked in precisely the hierarchies for positions where F2 contrasts are neutralized, and vowels with transitional F1 are unmarked in positions where F1 contrasts are neutralized.⁹

4.1.2 The motivation for neutralizing vowel F2 contrasts

In this section we will briefly address the motivation for neutralizing vowel F2 contrasts. We have seen that the dispersion-theoretic analysis correctly predicts the properties of vowel inventories without F2 contrasts, but we have not yet explained why a language would forgo F2 contrasts in the first place. The argument made above only depends on the outcome of F2 neutralisation, so the motivation for neutralisation is not directly relevant here, but it might be thought that vertical vowel inventories contradict MAXIMIZE CONTRASTS by failing to exploit F2 contrasts, so it is useful to show that this is not the case.

The analysis proposed here is that neutralisation of vowel F2 follows the standard pattern described in §3.2.2: F2 contrasts are neutralized in contexts where it is too difficult to realize them distinctly. A key factor that contributes to this difficulty is very short vowel duration. In §3.2.2. we saw evidence that short duration makes it difficult to produce high F1 in a vowel because there is little time for the necessary opening and closing movements. Similar considerations apply to the realisation of F2 contrasts in shorter vowels. Lindblom (1963) shows that F2 at the mid-point of a vowel in a CVC where both consonants are the same tends to move closer to an F2 value characteristic of the consonant as the duration of the vowel is reduced. As a vowel becomes shorter, it becomes more effortful to deviate from the least effort transition between preceding and following consonants by a significant amount, but deviation from the least effort transition is required to realize distinct F2 values for contrasting vowels. At short durations, the effort of realizing a distinct F2 contrast can be sufficient to make neutralisation optimal.

Schwa vowels are typically extremely short – Kondo (2000) reports a mean duration of 34 ms for English – so it is unsurprising that it is difficult to maintain either F1 or F2 contrasts in this context. Vertical vowels are also short, although generally longer than schwa. The Caucasian vowel length opposition is between short vowels and extra-short vowels, rather than between long and short vowels as in Japanese or Finnish (Choi 1991, Kuipers 1960:24, Smeets 1984:122, Colarusso 1988:349), and the Marshallese vertical vowels are comparable to these extra-short vowels (although the low vowel is longer –

⁹ Crosswhite (1999, this volume) proposes distinct markedness hierarchies for stressed and unstressed vowels where schwa is the most marked stressed vowel, but the least marked unstressed vowel. However, she uses [ə] to refer to both a mid-central vowel, as found in Brazilian Portuguese vowel reduction, and the transitional vowel found in complete neutralization, so this analysis fails to account for the distinct contexts in which these two types of vowel arise.

presumably this is necessary to reach a high F1) (Choi 1992). But in these languages the difficulty presented by short duration is exacerbated by rich inventories of consonant place contrasts. F2 transitions play an important role in realizing these contrasts, so it is less possible to facilitate vowel contrasts by co-producing vowels with consonants. Marshallese has an extensive system of palatalisation, velarisation, and labio-velarisation contrasts (e.g. [p^j-p^v], [k-k^w]), and sequences such as [p^jup^j] and [p^vip^v] obviously require substantial tongue body movement. The Caucasian languages contrast large sets of places of articulation, together with some secondary articulations. So to some extent it appears that vertical vowel inventories are trading vowel F2 contrasts for consonant-centered F2 contrasts. Indeed, analysts have varied between characterizing Arrernte as a vertical vowel language with extensive labio-velarisation contrasts, or as a language with vowel F2 contrasts, and a smaller consonant inventory (Ladefoged and Maddieson 1996:357). However it is apparent that rich consonant contrasts alone do not give rise to neutralisation, because F2 contrasts are maintained among longer vowels in the same consonant contexts.

The constraint ranking in (26) is a partial formalisation of this analysis of vertical vowels. The constraint *HIGHEFFORT is intended to penalize particularly rapid movements – specifically, with very short vowels, it rules out anything more than small deviations from a smooth transition between tongue body and lip positions for preceding and following consonants. The MINDIST constraint imposes a substantial minimum distance for vowel contrasts in F2, and for contrasts based primarily on F2 during consonant release transitions. This constraint is satisfied by contrasts between fully front and back vowels (e.g. i-u, e-o) or between palatalized and velarized consonants (see sample F2 specifications in 27).

(26) *HIGHEFFORT, MINDIST = F2:4 >> MAXIMIZE CONTRASTS

(27) F2:

6	5	4	3	2	1
i	i̠	ị	i	u	u
	e	ə̣	ə	ɤ	o
C ^j				C ^v	

The operation of these constraints is illustrated by the tableau in (28) which shows the selection of an inventory of CVCs with extra-short vowels, considering only secondary articulation contrasts as representatives of consonant contrasts and only F2 contrasts among vowels.

(28)

		*HIGH EFFORT	MINDIST = F2:4	MAXIMIZE CONTRASTS
a.	C ^j ₁ C ^j C ^j _u C ^j C ^j ₁ C ^v C ^j _u C ^v C ^v ₁ C ^j C ^v _u C ^j C ^v ₁ C ^v C ^v _u C ^v	*!*****		8
b.	C ^j ₁ C ^j C ^j ₁ C ^j C ^j ₁ C ^v C ^j ₁ C ^v C ^v ₁ C ^j C ^v ₁ C ^j C ^v ₁ C ^v C ^v _u C ^v		*!***	8

c.	C ^h iC ^j	C ^h iC ^y			4
	C ^y iC ^j	C ^y iC ^y			
d.	CiC	CüC			2!

Candidate (a) best satisfies MAXIMIZE CONTRASTS since it allows palatalisation-velarisation contrasts on consonants in all positions, and front-back contrasts in vowels, however CVCs such as [p^hup^j, p^yip^y] involve substantial violations of *HIGHEFFORT since they involve movement from two full front-back movements in a short duration. Candidate (b) is intended to include CVCs which barely satisfy *HIGHEFFORT, i.e. they represent the maximum allowable effort, while maintaining vowel and consonant contrasts in all positions (indicated by somewhat arbitrary transcriptions). However, with such short vowels, the maximum allowable effort results in only small deviations from a smooth transition between the secondary articulations of the consonants (indicated by somewhat arbitrary transcriptions), and consequently indistinct F2 contrasts, so candidate (b) violates the MINDIST constraint. The winning candidate, (c), satisfies *HIGHEFFORT since it involves only transitional vowels, transcribed here with central vowel symbols. There are no vowel F2 contrasts, and the palatalisation-velarisation contrasts satisfy the MINDIST constraint.

Candidate (d) satisfies the *HIGHEFFORT and MINDIST constraints by neutralizing consonant contrasts rather than vowel contrasts. This candidate loses out to (c) because it realizes fewer contrasts. Probably other considerations contribute to this outcome – e.g. consonant place contrasts will typically be cued by a release burst or during the consonant constriction itself, as well as by formant transitions, so they may be more distinct than extra-short vowel F2 contrasts – but it seems likely that one advantage of adopting a vertical vowel system is that many consonant contrasts can be differentiated in a relatively short duration (consonant constriction plus transitions) whereas distinct vowel contrasts take longer to realize. So abandoning vowel F2 contrasts may actually be motivated by MAXIMIZE CONTRASTS rather than being in conflict with this constraint.

This analysis suggests that vertical vowels are similar to the schwa vowels of Berber and Salish in that they serve primarily present to allow the realisation of consonant contrasts, but F1 is not generally implicated in consonant contrasts, so it is possible to simultaneously realize vowel F1 contrasts if vowels are permitted to be somewhat longer than the Berber or Salish schwa.

4.1.3 Related phenomena

Dispersion theory predicts that where no contrasts are primarily realized on a given dimension, then realisation on that dimension will be governed by minimisation of effort, or other contextual markedness constraints. Neutralisation of F2 contrasts in vertical vowel inventories and in fully neutralizing vowel reduction are examples of this phenomenon. There are probably many other examples of this pattern, but in some cases they can be difficult to detect because the least effort realisation of a sound type is similar to a sound found in contrast. For example, in many contexts, the least-effort laryngeal state for an obstruent will be voicelessness, due to aerodynamic factors discussed in the next section. However, voiceless stops also provide a distinct contrast with voiced stops, so least effort stops may be similar to contrastively voiceless stops in many contexts. Dispersion theory leads us to expect that non-contrastive voiceless stops should be more

prone to partial voicing following a preceding sonorant because effort minimisation disfavours active measures to promote voicelessness, but the differences involved can only be identified by instrumental analysis, so we do not have relevant data for many languages (but see Hsu 1998 for evidence of this pattern in Taiwanese). However, there is good evidence for the related prediction that effortful enhancements of stop voicing should only apply where there are voicing contrasts, as shown in the next section.

Contextual nasalisation of vowels provides another possible example of this type of pattern. It is a slightly more complex case because nasalisation does effect the distinctiveness of vowel quality contrasts, particularly those involving F1 (Wright 1986, Beddor 1993), but it obviously has a much greater effect on vowel nasalisation contrasts. So although we expect some general resistance to nasalisation of vowels, it is to be expected that oral vowels should be more tolerant of contextual nasalisation in the absence of nasalisation contrasts.

Again, differences in the magnitude and extent of partial nasalisation can only be determined by instrumental methods. Cohn (1990) shows that contrastive oral vowels in French undergo much less contextual nasalisation than English vowels preceding a nasal, but there is no obvious difference following a nasal. In any case, French may not be the most relevant example since the vowel nasalisation contrasts are generally accompanied by differences in vowel quality. There is evidence that the extreme measure of denasalisation of nasals to avoid contextual vowel nasalisation is only adopted where there are vowel nasalisation contrasts.

The only way to ensure that a vowel adjacent to a nasal is completely oral is to execute the velum movement during the stop closure, resulting in a brief oral stop. This pattern is observed in a wide variety of languages (Anderson 1976, Herbert 1986), the most striking instance being Kaingang, where nasals are prenasalised preceding an oral vowel (42b), post-nasalised following an oral vowel (42c), and “medio-nasalised” between oral vowels (42d).

- (29) a. $\tilde{V}m\tilde{V}$
 b. $\tilde{V}m^bV$
 c. $V^bm\tilde{V}$
 d. V^bm^bV

Herbert (1986) claims that this pattern of realisation is only observed in languages with contrastive nasalisation, as one would expect if partial denasalisation is motivated by the pressure to maximize the distinctiveness of vowel nasalisation contrasts. That is, replacing a nasal by a more marked partially-nasalised stop is only justified where it serves to maximize the distinctiveness of a contrast with nasalised vowels, because allophonic partial nasalisation does little damage to the distinctiveness of vowel quality contrasts.

The schematic ranking in (30) shows the outlines of a dispersion-theoretic formulation of this analysis. The dispreference for partially nasalised stops universally outranks constraints against contrasts between partially-nasalised vowel qualities (e.g. $*\tilde{i}-\tilde{e}$, where a single tilde indicates partial nasalisation), so allophonic partial nasalisation of

vowels is always preferred to denasalisation of nasal consonants¹⁰. But the distinctiveness constraint against contrasts between partially and fully nasalised vowels * \tilde{V} - \tilde{V} (where a double tilde [$\tilde{\tilde{V}}$] marks a fully nasalised vowel) can outrank *PARTIALLY NASALISED STOP, so denasalisation can be conditioned by a contrastively oral vowel.

(30) * \tilde{V} - $\tilde{\tilde{V}}$, *PARTIALLY NASALISED STOP >> * \tilde{i} - \tilde{e} , * \tilde{e} - \tilde{a} , etc.

Without constraints on contrast, it is not possible to account for the fact that denasalisation requires vowel nasalisation contrasts since any constraint that favoured denasalisation adjacent to oral vowels would necessarily apply to all oral vowels, whether or not they contrast with nasalised vowels.

There is one exception to Herbert's generalisation: in some Australian languages, including Gupapuyŋu (Butcher 1999), nasals are optionally pre-stopped post-vocally although there are no vowel nasalisation contrasts. Butcher suggests that this partial denasalisation serves to ensure that the closure transitions are oral, avoiding the destructive effect of nasalisation on the distinctiveness of formant patterns (Repp and Svastikula 1988, Wright 1986). The distinctiveness of formant transitions is particularly important because the relevant languages distinguish 4-6 places of articulation among nasals. So this exceptional case also appears to be motivated by distinctiveness constraints.¹¹

4.2 Enhancement of stop voicing contrasts

Another example of contrast-dependent markedness is provided by the typology of laryngeal contrasts among stops. A number of languages contrast prenasalised or implosive stops with voiceless unaspirated stops, but do not have plain voiced stops. The preference for prenasalised or implosive stops over plain voiced stops is explained on the grounds that prenasalised and implosive stops are more distinct from voiceless stops (cf. Iverson and Salmons 1996). However these sounds are also more effortful than plain voiced stops, so most languages forgo these enhancements. Crucially enhancement of stop voicing does not occur in the absence of contrast – we do not find prenasalisation or implosivisation of intervocally voiced stops for example. This is expected if the only reason for exerting the additional effort involved in producing these sounds is to satisfy a

¹⁰ It is not clear whether *PARTIALLY NASALIZED STOP is properly a constraint against the effort involved in moving the velum with sufficient rapidity and precision to produce a nasalization contour, or whether partially nasalized stops are dispreferred relative to full nasals because they yield inferior contrasts with some ubiquitous sound category such as voiceless stops. However it does seem that languages do not have partially nasalized stops unless they also have nasals, either in contrast or in alternation with the partially-nasalized stops (Herbert 1986:16ff.).

¹¹ A form of post-nasalization can also arise without vowel nasalization contrasts through a process of 'pre-obstruentisation' discussed by Steriade (1993) (e.g. Diyari, Icelandic). Steriade argues that this process is not denasalisation per se because it is accompanied by pre-stopping of laterals in the same environments (l → dl). Further reason for doubting that it is the orality of vowels which conditions this process of post-nasalization comes from the fact that they are not conditioned by all oral vowels – it only applies to post-stress or geminate nasals.

constraint on the distinctiveness of contrasts, but, like other contrast-dependent patterns of distribution, it is difficult to account for without constraints on contrast.

Prenasalised and implosive stops are often thought of as more marked than plain voiced stops. While it is true that they are less frequent than plain voiced stops, there is no implicational relationship between these sound types: a substantial number of languages have prenasalised or implosive stops without having plain voiced stops, e.g. San Juan Colorado Mixtec has prenasalised stops but no plain voiced stops (31) (Campbell, Peterson & Lorenzo Cruz 1986). This pattern is discussed by Iverson and Salmons (1996) in relation to Mixtec, and by Herbert (1986:16ff.) who cites a number of other examples, including Fijian, Lobaha, Reef Islands-Santa Cruz languages, and South Gomen. Other examples include Southern Barasano (Smith & Smith 1971) and Guaraní (Gregores & Suárez 1967).

(31) San Juan Colorado Mixtec stops

p	t	tʲ	k
ᵐb	ⁿd	ⁿdʲ	

Languages which contrast voiceless and implosive stops but lack plain voiced stops seem to be less common (Maddieson 1984:28), but the UPSID database of phonological inventories (Maddieson 1984) includes two examples: Nyangi and Maasai (both Eastern Sudanic). The stops of Nyangi are shown in (32). In addition, Vietnamese voiced stops are often implosive (Nguyen 1970), and Ladefoged and Maddieson (1996) report that ‘fully voiced stops in many diverse languages (e.g. Maidu, Thai and Zulu) are often accompanied by downward movements of the larynx that make them slightly implosive’ (p.78).

(32) Nyangi stops:

p	t	c	k
b	d	f	g

Voiced stops are distinguished from voiceless stops by a variety of cues. One of the most important is Voice Onset Time (Lisker and Abramson 1964, Lisker 1975), but the presence of voicing during closure (indicated by periodicity and low-frequency energy) is also significant (Stevens and Blumstein 1981). Implosive and prenasalised stops are more strongly voiced than plain voiced stops, and so are better distinguished from voiceless stops in this respect. It is difficult to sustain high intensity of voicing during a stop closure because pressure builds up behind the closure until there is no longer a pressure drop across the glottis. Without a sufficient pressure drop there is no airflow through the glottis, and voicing ceases (Ohala 1983, Westbury and Keating 1986). So voicing tends to decline in intensity through a voiced stop closure. Lowering the velum allows air to be vented from the vocal tract, mitigating the pressure build-up, and thus facilitating the maintenance of high intensity of voicing. In addition, radiation from the nose results in higher intensity of the speech signal than radiation through the neck, which is the only source of sound in an oral stop (Stevens et al 1986:439).

Similarly, lowering the larynx during the stop closure, as in implosive stops, expands the oral cavity, reducing the build-up of pressure. Consequently implosives are

characteristically strongly voiced. Lindau (1984) found that the amplitude of voicing actually increases through the course of an implosive closure. Implosives also have very low-intensity release bursts because the intensity of the burst depends on oral pressure at release (Ladefoged and Maddieson 1996:82). Intensity of the release burst has been shown to be a significant cue to stop voicing contrasts in English (Repp 1979), so this is also likely to make implosives more distinct from voiceless stops than plain voiced stops.

Given these considerations, it seems likely that languages like Mixtec and Nyangi prefer prenasalised-voiceless and implosive-voiceless stop contrasts over the more common voiced-voiceless contrast because the former are more distinct contrasts (Henton, Ladefoged, and Maddieson 1992, Iverson and Salmons 1996). The conflicting constraint that leads many languages to forgo maximizing distinctiveness is probably effort minimisation. Implosives involve more effort than plain voiced stops because they involve an additional larynx-lowering gesture. Prenasalised stops require rapid raising of the velum to produce oral and nasal phases within the same stop.

This analysis can be formalized as follows. We will assume that at least two dimensions distinguish voiced and voiceless stops: VOT and strength of voicing ([voice]) (33), which could be quantified in terms of the intensity of the periodic part of the speech signal, for example.

(33)	VOT:	0	d, ⁿ d, d̥		Voice:	0	t
		1	t			1	d
		2	t ^h			2	ⁿ d, d̥

A larger difference on either dimension results in a more distinct contrast, as reflected in the universal ranking of MINDIST constraints in (34). Distances on multiple dimensions are indicated by joining the distances on individual dimensions with a ‘+’ sign. So MINDIST = VOT:1+VOICE:2 is satisfied by contrasts such as [ⁿd-t] and [d-t] which differ by 1 on the VOT dimension and 2 on the Voice dimension. The contrast [d-t] violates this constraint, but satisfies the higher-ranked MINDIST constraint.

$$(34) \quad \text{MINDIST} = \text{VOT:1+VOICE:1} \gg \text{MINDIST} = \text{VOT:1+VOICE:2}$$

For present purposes the fact that prenasalised stops and implosives involve greater effort than plain voiced stops will be implemented as a fixed ranking of constraints against these sound types (35).

$$(35) \quad *IMPLOSIVE, *PRENASALISED \text{ STOP} \gg *VOICED \text{ STOP}$$

Then a language like Nyangi, with implosives in place of voiced stops, is derived by the following ranking, as shown in (37).

$$(36) \quad \text{MINDIST} = \text{VOT:1+VOICE:1} \gg \text{MINDIST} = \text{VOT:1+VOICE:2}, \text{ MAXIMIZE} \\ \text{CONTRASTS}, *PRENASALISED \text{ STOP} \gg *IMPLOSIVE \gg *VOICED \text{ STOP}$$

(37)		MINDIST = VOT:1+ VOICE:1	MINDIST = VOT:1+ VOICE:2	MAXIMIZE CONTRASTS	*PRENASALISED STOP	*IMPLOSIVE	*VOICED STOP
a.	t-d		*!	✓✓			*
b.	t-d			✓✓		*	
c.	t ⁿ d			✓✓	*!		
d.	t			✓!			

We will assume for now that the preference for implosives over prenasalised stops depends purely on the relative ranking of the effort-minimisation constraints against these sounds types, so the ranking in (36) derives implosives where *PRENASALISED STOP >> *IMPLOSIVE (cf. 37c), while prenasalised stops are derived if this ranking is reversed. The more common voiced-voiceless contrast is derived if MINDIST = VOT:1 + VOICE 2 is ranked below both of these effort minimisation constraints (38).

(38)		MINDIST = VOT:1+ VOICE:1	MAXIMIZE CONTRASTS	*PRENASALISED STOP	*IMPLOSIVE	MINDIST = VOT:1+ VOICE:2	*VOICED STOP
a.	t-d		✓✓			*	*
b.	t-d		✓✓		*!		
c.	t ⁿ d		✓✓	*!			
d.	t		✓!				

If a voicing contrast is not maintained, the distinctiveness of voicing contrasts is irrelevant, so voicing of stops is determined primarily by effort minimisation. In many contexts effort minimisation prefers devoicing of stops due to aerodynamic factors reviewed above, but in some contexts, e.g. following a nasal or in short stops between vowels, voicing appears to be easier to produce and many languages follow effort minimisation, resulting in allophonically voiced stops in these contexts (Westbury and Keating 1986, Kirchner 1998). For example, stops are voiced intervocalically and following nasals in Tümpisa Shoshone (Dayley 1989, Kirchner 1998). Implosives and prenasalised stops, on the other hand, are never preferred by effort minimisation constraints, so these sounds are only expected in contrast with voiceless stops.

The patterns of distribution analysed here involve a contrast-dependent generalisation: implosives and prenasalised stops can be preferred to voiced stops where they contrast with voiceless stops, but they are never preferred to voiced stops where there is no voicing contrast. That is, there is no post-nasal implosivisation or intervocalic prenasalisation. This situation is difficult to account for without constraints on contrasts because any simple way of deriving implosives/prenasalised stops in place of voiced stops without these constraints is liable to predict that these sounds could also be preferred in the absence of contrast.

In a theory without constraints on contrasts, a preference for implosives over voiced stops implies a ranking of constraints with the effect of that shown in (39). The exact

formulation of *VOICED STOP and *IMPLOSIVE is not important, it is only necessary that one favours implosive stops over voiced stops, and the other effectively imposes the reverse preference. We must also assume that faithfulness to the feature that differentiates implosives from plain voiced stops (e.g. [lowered larynx]) is low-ranked throughout to explain the absence of contrasts between plain voiced and implosive stops in the relevant languages.

(39) IDENT[VOICE], *VOICED STOP >> *IMPLOSIVE

The reverse ranking of *VOICED STOP and *IMPLOSIVE would also have to be allowed to derive the usual voiced-voiceless contrast:

(40) IDENT[VOICE], *IMPLOSIVE >> *VOICED STOP


The problem arises when these ranking possibilities are combined with rankings required to analyze allophonic variation in languages without voicing contrasts. The basic ranking for a language without stop voicing contrasts has to place IDENT[VOICE] below the effort minimisation constraints:

(41) *IMPLOSIVE >> *VOICED STOP >> IDENT[VOICE]

To derive intervocalic voicing, it is necessary to differentiate the markedness of voiced stops between vowels from their markedness in other contexts. A simple approach is to posit a constraint against intervocalic voiceless stops, *VOICELESS STOP/V_V, ranked above the general constraint against voiced stops (42). But we have already seen that *VOICED STOP must be able to out-rank *IMPLOSIVE to account for languages with implosives but no plain voiced stops. So nothing prevents reversing the ranking of these constraints, as in (55-56), which derives the unattested phenomenon of intervocalic voicing implosivisation, i.e. stops are implosive between vowels (43), but voiceless elsewhere (44), because this ranking makes it preferable to replace any voiced stop by an implosive.

(42) *VOICELESS STOP/V_V, *IMPLOSIVE >> *VOICED STOP >> IDENT[VOICE]

(43)

	/ata/	*VOICELESS STOP/V_V	*VOICED STOP	*IMPLOSIVE	IDENT [VOICE]
a.	ata	*!			
b.	ada		*!		*
c.	 ada			*	*

(44)	/ad/	*VOICELESS STOP/V_V	*VOICED STOP	*IMPLOSIVE	IDENT [VOICE]
a.	☞ at				*
b.	ad		*!		
c.	ad̥			*!	

The problem with this approach is that it is not possible to express the fact that implosives and prenasalised stops are only favored because they yield more distinct contrasts with voiceless stops (or an additional contrast), so nothing favours implosives in the absence of stop voicing contrasts. Without constraints on contrasts, it is necessary to posit constraints favouring implosives and prenasalised stops independent of contrast, which then predicts that these sounds could be preferred over plain voiced stops in the absence of contrast. The preference for implosives and prenasalised stops must be strictly dependent on the presence of a contrast, which implies constraints on contrasts.

4.3 Allophonic and contrastive nasalisation

[Section deleted for this reading assignment; available from BH.]

5 Conclusion: working with constraints on contrast

We have now seen substantial evidence that phonology includes constraints on contrasts, specifically constraints that favor maximizing the distinctiveness of contrasts (MINDIST), and a constraint that favours maximizing the number of contrasts (MAXIMIZE CONTRASTS). We have also seen that these constraints do not operate independently from more familiar syntagmatic markedness constraints, e.g. as a theory of inventories, somehow operating outside of conventional phonological analyses. The interaction between syntagmatic and paradigmatic constraints is central to the derivation of basic phenomena such as neutralisation (3.2.2) and blocking in harmony processes (4.3). According to the dispersion theory, the set of well-formed words in a language represents an optimal balance between the number and distinctiveness of the contrasts between words, and constraints that define preferred sound sequences, such as effort minimisation and metrical constraints. However combining paradigmatic and syntagmatic constraints in this way does result in a system with very different properties from an OT grammar based on conventional constraints because constraints on the distinctiveness of contrasts evaluate relationships between forms. So if we want to determine whether a putative word is well-formed, we must consider whether it is sufficiently distinct from neighbouring words. But these words must also be well-formed, which implies assessing their distinctiveness from neighbouring words, and so on. Thus it seems that we cannot evaluate the well-formedness of a single word without determining the set of all possible words.

The analyses above avoid this problem by considering only the evaluation of inventories of contrasting sounds (or short strings of sounds) in a particular context rather than evaluating complete words. For example, evaluating vowel inventories effectively involves determining the set of contrasting sounds that are permitted in a syllable nucleus. This makes the evaluation of MINDIST and MAXIMIZE CONTRASTS

straightforward since only a small number of contrasting sounds are possible in a given context. This simplification is valid given certain assumptions. First, the context must be well-formed. For example, if we are evaluating the set of vowels that can appear before a nasal stop, it must be true that nasal stops are part of an inventory of consonant contrasts that can occur in postvocalic position. Second, nothing outside of the specified context should be relevant - that is, no constraint that is ranked high enough to affect the well-formedness of the inventory should refer to material outside of the specified context.

More generally, the strategy for avoiding the problem of mass comparisons is to derive generalisations about the set of possible words in a language – e.g. stressed vowels are all drawn from a certain set – rather than deriving particular words. But this strategy is not actually novel, it is the usual approach to phonological analysis. Even if it is possible to determine whether an individual word is well-formed with respect to a constraint ranking, the result of such an exercise is usually not very significant. Showing that a grammar can derive an individual word is not usually the goal of phonological analysis of a language, the goal is to devise a grammar that derives all and only the possible words of that language. The usual intermediate goal is to derive generalisations about all the possible words of the language, exactly as in the analyses here.

For example, in analysing a language it is usual to restrict attention to a single process, e.g. place assimilation between nasals and stops, ignoring stress assignment, distribution of vowels, etc. Such an analysis may be illustrated by deriving complete words, e.g. /kanpa/ → [kampa], but in itself this is uninteresting. The real goal is to derive the generalisation that nasals are always homorganic to following stops. Properly, establishing such a generalisation requires showing that no contrary output is derived if all possible inputs are passed through the grammar (Prince and Smolensky 1993:91). So with or without paradigmatic constraints, there is an important distinction between deriving individual words using a grammar and reasoning about the properties of the set of words derived by that grammar. Constraints on contrast make complete derivation of individual words difficult, but that does not preclude deriving generalisations about possible words.

To approach the derivation of complete words, it is necessary to derive increasingly comprehensive descriptions of the set of possible words. Such a description need not be a list of possible words, it could be a grammar that generates the possible words. That is, one way to deal with the need to evaluate all words simultaneously could be to evaluate candidate grammars which provide compact characterisations of candidate sets of possible words. Any such solution involves substantial additions to the analytical machinery of phonology, but we have seen that these steps are well-motivated.

References

- Albano Leoni, F., M.R. Caputo, L. Cerrato, F. Cutugno, P. Maturi, and R. Savy (1995). Il vocalismo dell'Italiano. Analisi di un campione televisivo. *Studi Italiani di Linguistica Teorica e Applicata* 24, 405-411.
- Anderson, Stephen R. (1976). Nasality and the internal structure of segments. *Language* 52, 326-344.
- Beddor, Patrice S. (1993). The perception of nasal vowels. Marie K. Huffman and Rena A. Krakow (eds.) *Nasals, Nasalization, and the Velum, Phonetics and Phonology vol. 5*. Academic Press, San Diego, 171-196.
- Bender, Byron W. (1968). Marshallese phonology. *Oceanic Linguistics* 7, 16-35.
- Benua, Laura (1997). *Transderivational Identity*. Ph.D. dissertation, University of Massachusetts, Amherst.
- Boersma, Paul (1998). *Functional Phonology*. PhD dissertation, University of Amsterdam.
- Booij, Geert (1995). *The Phonology of Dutch*. Oxford University Press, Oxford.
- Burgess, Eunice, and Patricia Ham (1968). Multilevel conditioning of phoneme variants in Apinayé. *Linguistics* 41, 5-18.

- Burzio, Luigi (1998). Multiple correspondence. *Lingua* 104, 79-109.
- Butcher, Andrew (1999). What speakers of Australian aboriginal languages do with their velums and why: The phonetics of the nasal/oral contrast. *Proceedings of the International Congress of Phonetic Sciences 1999*, 479-482.
- Byrd, Dani (1992). Marshallese suffixal reduplication. Jonathon Mead (ed.) *Proceedings of WCCFL XI*, 61-77.
- Campbell, Sara Stark, Andrea Johnson Peterson, and Filiberto Lorenzo Cruz (1986). *Diccionario Mixteco de San Juan Colorado*. Instituto Lingüístico de Verano, México, D.F.
- Choi, John D. (1991). An acoustic study of Kabardian vowels. *Journal of the International Phonetic Association* 21, 4-12.
- Choi, John D. (1992). *Phonetic Underspecification and Target Interpolation: An Acoustic Study of Marshallese Vowel Allophony (UCLA Working Papers in Phonetics 82)*. Ph.D. dissertation, University of California, Los Angeles.
- Chomsky, Noam, and Morris Halle (1968). *The Sound Pattern of English*. New York: Harper and Row.
- Cohn, Abigail C. (1990). *Phonetic and Phonological Rules of Nasalization (UCLA Working Papers in Phonetics 76)*. Ph.D. dissertation, University of California, Los Angeles.
- Cohn, Abigail C. (1993a). A survey of the phonology of the feature [nasal]. *Working Papers of the Cornell Phonetics Laboratory* 8, 141-203.
- Colarusso, John (1988). *The Northwest Caucasian Languages: A Phonological Survey*. Garland Publishing, New York.
- Colarusso, John (1992). *A Grammar of the Kabardian Language*. University of Calgary Press, Calgary.
- Dayley, Jon P. (1989). *Tümpisa (Panamint) Shoshone grammar*. University of California Press, Berkeley.
- Delattre, Pierre C., Alvin M. Liberman, Franklin S. Cooper, and Louis J. Gerstman (1952). An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word* 8, 195-210.
- Dell, François, and Mohammed Elmedlaoui (1996). On consonant releases in Imdlawn Tashlhiyt Berber. *Linguistics* 34, 357-395.
- Fails, Willis C., and J. Halvor Clegg (1992). A spectrographic analysis of Portuguese stressed and unstressed vowels. Donald P. Macedo and Dale A. Koike (eds.) *Romance Linguistics: The Portuguese Context*. Bergin and Garvey, Westport, 31-42.
- Ferguson, Charles A. (1963). Assumptions about nasals: A sample study in phonological universals. Joseph Greenberg (ed.) *Universals of Language*. MIT Press, Cambridge, 53-60.
- Flemming, Edward (1995). *Auditory Representations in Phonology*. PhD dissertation, UCLA.
- Flemming, Edward (1996). Evidence for constraints on contrast: The dispersion theory of contrast. Chai-Shune K. Hsu (ed.) *UCLA Working Papers in Phonology* 1, 86-106.
- Flemming, Edward (2001). *Auditory Representations in Phonology*. Garland Press, New York [Revised version of Flemming (1995)].
- Flemming, Edward, Peter Ladefoged and Sarah G. Thomason (1994). The phonetic structures of Montana Salish. *UCLA Working Papers in Phonetics* 87, 1-34.
- Fujimura, Osamu (1963). Analysis of nasal consonants. *Journal of the Acoustical Society of America* 32, 1865-1875.
- Gregores, Emma, and Jorge A. Suárez (1967). *A Description of Colloquial Guaraní*. Mouton, The Hague.
- Halle, Morris (1959). *The Sound Pattern of Russian*. Mouton, The Hague.
- Hayes, Bruce (1995). *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press, Chicago.
- Henton, Caroline, Peter Ladefoged, and Ian Maddieson (1992). Stops in the world's languages. *Phonetica* 49, 65-101.
- Herbert, Robert K. (1986). *Language Universals, Markedness Theory and Natural Phonetic Processes*. Mouton de Gruyter, Berlin.
- Hsu, Chai-Shune K. (1996). Voicing underspecification in Taiwanese word-final consonants. *UCLA Working Papers in Phonetics* 96, 90-105.
- Hume, Elizabeth, and Keith Johnson (2001). A model of the interplay of speech perception and phonology. Elizabeth Hume and Keith Johnson (eds.) *The Role of Speech Perception in Phonology*. Academic Press, New York.
- Iverson, Gregory K., and Joseph C. Salmons (1996). Mixtec prenasalization as hypervoicing. *International Journal of American Linguistics* 62, 165-175.
- Keating, Patricia A. (1985). Universal phonetics and the organization of grammars. Victoria A. Fromkin (ed.) *Phonetic Linguistics*. New York: Academic Press, 115-32.
- Kenstowicz, Michael (1996). Uniform exponence and base identity. J. Durand and B. Laks (eds.) *Current Trends in Phonology*. CNRS and University of Salford Publications, 363-393.
- Kirchner, Robert (1997). Contrastiveness and faithfulness. *Phonology* 14, 83-111.
- Kirchner, Robert (1998). *An Effort-Based Approach to Consonant Lenition*. PhD dissertation, UCLA.
- Kondo, Yuko (1994). Targetless schwa: is that how we get the impression of stress timing in English? *Proceedings of the Edinburgh Linguistics Department Conference '94*, 63-76.
- Kondo, Yuko (2000). Production of schwa by Japanese speakers of English: an acoustic study of shifts in coarticulatory strategies from L1 to L2. Michael Broe and Janet Pierrehumbert (eds.) *Papers in Laboratory Phonology 5: Acquisition and the Lexicon*. Cambridge University Press, Cambridge, 29-39.
- Kuipers, Aert H. (1960). *Phoneme and Morpheme in Kabardian (Eastern Adyghe)*. *Janua Linguarum*, series minor, no. 8. Mouton, The Hague.

- Ladefoged, Peter, and Ian Maddieson (1996). *The Sounds of the World's Languages*. Blackwell, Oxford.
- Laycock, D.C. (1965). *The Ndu Language Family*. Linguistics Circle of Canberra, Series C, no.1. Australian National University, Canberra.
- Lehiste, Ilse (1970). *Suprasegmentals*. MIT Press, Cambridge, MA.
- Liljencrants, Johan, and Björn Lindblom (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48, 839-62.
- Lindau, Mona (1984). Phonetic differences in glottalic consonants. *Journal of Phonetics* 12, 147-155.
- Lindblom, Björn (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 1773-1781.
- Lindblom, Björn (1986). Phonetic universals in vowel systems. J.J. Ohala and J.J. Jaeger (eds.) *Experimental Phonology*. Academic Press.
- Lindblom, Björn (1990). Phonetic content in phonology. *PERILUS* 11, 101-118.
- Lisker, Leigh (1975). Is it VOT or a first formant detector? *Journal of the Acoustical Society of America* 57, 1547-1551.
- Lisker, Leigh, and Arthur S. Abramson (1964). A cross-language study of voicing in initial stops. *Word* 20, 384-422.
- Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge University Press, Cambridge.
- Maeda, Shinji (1993). Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. Marie K. Huffman and Rena A. Krakow (eds.) *Nasals, Nasalization, and the Velum, Phonetics and Phonology vol. 5*. Academic Press, San Diego, 147-170.
- Maiden, Martin (1995). Vowel systems. Martin Maiden and Mair Parry (eds.) *The Dialects of Italy*. Routledge, London, 7-14.
- Major, Roy C. (1992). Stress and rhythm in Brazilian Portuguese. Donald P. Macedo and Dale A. Koike (eds.) *Romance Linguistics: The Portuguese Context*. Bergin and Garvey, Westport, 3-30.
- Martinet, Andre (1952). Function, structure, and sound change. *Word* 8, 1-32.
- Martinet, Andre (1955). *Economie des Changements Phonétiques*. Berne: Francke.
- Mattoso Camara, Joaquim (1972). *The Portuguese Language*. University of Chicago Press, Chicago.
- Mazzola, Michael L. (1976). *Proto-Romance and Sicilian*. Peter de Ridder, Lisse.
- Moraes, João A. de (1998). Brazilian Portuguese. Daniel Hirst and Albert Di Cristo (eds.) *Intonation Systems*. Cambridge University Press, Cambridge.
- Nelson, W.L. (1983). Physical principles for economies of skilled movement. *Biological Cybernetics* 46, 135-147.
- Nguyen, Dang Liem (1970). *A contrastive phonological analysis of English and Vietnamese*, vol. 4. Pacific Linguistics Series C, No. 8. Australian National University, Canberra.
- Ní Chiosáin, Máire, and Jaye Padgett (1997). Markedness, segment realization, and locality in spreading. Report no. LRC-97-01, Linguistics Research Center, University of California, Santa Cruz.
- Nosofsky, Robert M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology* 43, 25-53.
- Ohala, John J. (1975). Phonetic explanations for nasal sound patterns. Charles A. Ferguson, Larry M. Hyman, and John J. Ohala (eds.) *Nasalfest: Papers from a Symposium on Nasals and Nasalization*. Stanford University, 289-316.
- Ohala, John J. (1983). The origins of sound patterns in vocal tract constraints. Peter F. MacNeilage (ed.) *The Production of Speech*. Springer-Verlag, New York, 189-216.
- Ohala, John J., and Manjari Ohala (1993). The phonetics of nasal phonology: Theorems and data. Marie K. Huffman and Rena A. Krakow (eds.) *Nasals, Nasalization, and the Velum, Phonetics and Phonology vol. 5*. Academic Press, San Diego, 225-250.
- Onn, Farid M. (1980). *Aspects of Malay Phonology and Morphology*. Ph.D. dissertation, Universiti Kebangsaan Malaysia, Bangi.
- Perkell, Joseph S. (1997). Articulatory processes. William J. Hardcastle and John Laver (eds.) *The Handbook of Phonetic Sciences*. Blackwell, Oxford, 333-370.
- Piggott, Glynne (1992). Variability in feature dependency: The case of nasality. *Natural Language and Linguistic Theory* 10, 33-78.
- Plomp, Reinier (1975). Auditory analysis and timbre perception. Gunnar Fant and Michel Tatham (eds.) *Auditory Analysis and Perception of Speech*. Academic Press, New York, 7-22.
- Prince, Alan, and Paul Smolensky. (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*. To appear: Cambridge, MA: MIT Press.
- Repp, Bruno H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech* 27, 173-189.
- Repp, Bruno H., and K. Svastikula (1988). Perception of the [m]-[n] distinction in VC syllables. *Journal of the Acoustical Society of America* 83, 238-247.
- Robins, R.H. (1957). Vowel nasality in Sundanese. *Studies in Linguistic Analysis*. Blackwell, London, 87-103.
- Rosner, B.S., and J.B. Pickering (1994). *Vowel Perception and Production*. Oxford University Press, Oxford.
- Schourup, Lawrence (1972). Characteristics of vowel nasalization. *Papers in Linguistics* 5, 530-548.
- Schwartz, Jean-Luc, Denis Beutemps, Christian Abry, and Pierre Escudier (1993). Inter-individual and cross-linguistic strategies for the production of the [i] vs. [y] contrast. *Journal of Phonetics* 21, 411-425.
- Shepard, Roger N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika* 22, 325-345.

- Shepard, Roger N. (1972). Psychological representation of speech sounds. Edward David and Peter Denes (eds.) *Human Communication: A Unified View*. McGraw-Hill, New York, 67-113.
- Smeets, Riëks (1984). *Studies in West Circassian Phonology and Morphology*. Hakuchi Press, Leiden.
- Smith, Richard, and Connie Smith (1971). Southern Barasano phonemics. *Linguistics* 75, 80-85.
- Staalsen, Philip (1966). The phonemes of Iatmul. *Papers in New Guinea Linguistics* 5 (Linguistics Circle of Canberra, Series A, No.7), 69-76.
- Steriade, Donca (1993). Orality and markedness. *Berkeley Linguistics Society* 19, 334-347.
- Steriade, Donca (1995). Neutralization and the expression of contrast. Ms, UCLA.
- Steriade, Donca (1997). Phonetics in phonology: the case of laryngeal neutralization. Ms, UCLA.
- Steriade, Donca (2000). Paradigm uniformity and the phonetics-phonology boundary. Michael Broe and Janet Pierrehumbert (eds.) *Papers in Laboratory Phonology 5: Acquisition and the Lexicon*. Cambridge University Press, Cambridge, 313-334.
- Stevens, Kenneth N. (1999). *Acoustic Phonetics*. MIT Press, Cambridge.
- Stevens, Kenneth N., and Sheila E. Blumstein (1981). The search for invariant acoustic correlates of phonetic features. Peter D. Eimas and Joanne L. Miller (eds.) *Perspectives on the study of speech*. Lawrence Erlbaum, Hillsdale, 1-38.
- Stevens, Kenneth N., S. Jay Keyser, and Haruko Kawasaki (1986). Toward a phonetic and phonological theory of redundant features. Joseph S. Perkell and Dennis H. Klatt (eds) *Invariance and Variability in Speech Processes*. Lawrence Erlbaum, Hillsdale, 426-449.
- Valente, V. (1975). Puglia. Manlio Cortelazzo (ed.) *Profilo dei dialetti italiani vol. 15*. Pacini, Pisa.
- Van Bergem, Dick R. (1994). A model of coarticulatory effects on the schwa. *Speech Communication* 14, 143-162.
- Walker, Rachel (1998). *Nasalization, Neutral Segments, and Opacity Effects*. Ph.D. dissertation, University of California, Santa Cruz.
- Walker, Rachel, and Geoffrey K. Pullum (1999). Possible and impossible segments. *Language* 75, 764-780.
- Westbury, John, and Patricia A. Keating (1980). Central representation of vowel duration. *Journal of the Acoustical Society of America* 67, S37A.
- Westbury, John, and Patricia A. Keating (1986). On the naturalness of stop consonant voicing. *Journal of Linguistics* 22, 145-66.
- Wright, James T. (1986). The behavior of nasalised vowels in the perceptual vowel space. John J. Ohala and Jeri J. Jaeger (eds.) *Experimental Phonology*. Academic Press, Orlando, 45-68.
- Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge.