Chapter 6

Phonetics and the I-Linguistics of Speech

A theory of language that focuses on the properties of external linguistic objects, the sort of thing we have been calling “E-language,” tends to limit itself to a pre-occupation with the nature of representations of those objects. For instance, in the period of structuralist phonemics, the theory of phonemic representations was the theory of sound structure. As we saw in chapter 4, it was on the basis of arguments about the nature of representation that generative phonology supplanted structuralism, even though the more basic issue distinguishing the two points of view hinged on the rather different issue of the place of rules in linguistic description.

Just because I-language theories are not limited to matters of representation, however, does not mean that they do not involve such questions. For instance, within the evolution of generative phonology, discussion of autosegmental structure and especially feature geometry (see Kenstowicz 1994 for a survey) marked a return to issues of representation in the study of sound structure, as other areas of inquiry seemed to lead to dead-ends in the 1970s. The focus in recent work on Optimality Theory as reviewed in chapter 5 has generally held issues of representation more or less constant, but as attention has turned to the phonetic motivation for constraints, the nature and content of phonetic representations has once again surfaced as an important issue.

Representations in general have increasingly been the focus of interest within contemporary cognitive science, as a consequence of its notion of the mind as an information-processing system. If this sort of picture is to become science, however, we must be careful to specify the form of the
information being manipulated, and the nature of the manipulations, with some precision. And that is exactly where we are required to provide theories of various sorts of representation. With luck, the practice that has been the focus of work by people working on language and on vision in this regard can become a model for other areas of cognition. In turn, however, as linguists take more seriously the cognitive context in which their work acquires much of its meaning, theories of linguistic representation have yet another set of concerns to respond to: they have to be appropriate for embedding in a theory of the mind, as well as being adequate to support the observed external facts of language as these are analyzed within particular theories of language per se.

Our knowledge of a language, as this is determined by the language organ we develop as a child on the basis of exposure to utterances in that language, includes what we know about contrasts, relations, and regularities within the set of linguistic objects. Obviously, though, it also includes what we know about the objects themselves. The structure of that knowledge is described by a theory of representations of the various sorts of object that form parts of our language. Seeing the foundation of these representations as an aspect of our knowledge (an I-language point of view) has somewhat different consequences from seeing them as based purely on externally determined properties, part of E-language. There may be much formal similarity between the actual representations that result from these two differing perspectives, but the conceptual content is still quite distinct.

In this chapter, we address the nature of the representations that seem to be most obviously and irreducibly based on observable, physically measurable properties: phonetic representations. We argue that when phonetics is seen as genuinely part of language, rather than a sub-part of physics or physiology, the resulting conception of phonetic representation (while still recognizable) differs in a number of important ways from what is often taught (or more accurately, assumed) in standard textbooks.

6.1 Representations and the study of sound structure

Most linguists assume, as we argued in chapter 4, that the principles of sound structure in a given language mediate between a phonological representation that indicates all and only the properties of an utterance in terms of which it contrasts with other utterances in that language, and a phonetic representation that provides a language-independent characterization of its
The nature of phonological representations has occasioned a great deal of discussion and a vast literature. Some of the most radical proposals within Optimality Theory have argued that representations of any sort other than surface phonetic ones are redundant, but the linguistic significance of some sort of phonological representation for linguistic items and utterances has generally not been in doubt. Many readers may be surprised to learn, though, that the status of phonetic representations themselves in linguistic theory has not always been quite so clear.

To be sure, there has often been some question about the extent to which phonetics is properly part of linguistics at all. If this kind of investigation of the articulatory, acoustic and perceptual properties of concrete acts of speaking is essentially a matter of more and more precise measurement of physiological, physical and neurological events, it seems to have little to do with linguistic structure per se, especially if we construe the latter as primarily cognitive in its basis. Phonetics would have the status of an auxiliary discipline — overlapping with, but properly included within, physics, physiology and the neurophysiology of the auditory system — that simply described the externally observable properties of the abstract objects with which linguistics is concerned. As Trubetzkoy (1939) put it, “phonetics is to phonology as numismatics is to economics.”

We argued in section 4.1 of chapter 4 that the kind of representation generally called ‘phonetic’ is a significant abstraction from the raw physical facts. Nonetheless, few would question the premise that acts of speaking do have some observable properties, and that the business of phonetics is to settle the facts of the matter as to what these are, as far as the language system is concerned. Relevant results of such observations can then be presented in some appropriate form, and who could question that such a ‘phonetic representation’ describes the things phonologists have to account for?

Leonard Bloomfield, in contrast, argued that there is no linguistic significance to phonetic representations (cf. Anderson 1985, 262ff.). His point was that insofar as these deviate in any way from a full physical record of the speech event (such as might be provided by a tape recording, supplemented with cineradiographic and other records of the articulatory details), they represent an arbitrary selection of some properties to the exclusion of others and cannot be said to be based on theoretically interesting principles. As such, they serve more as a record of biographical details about the phonetician (what properties he or she has learned to record and what to omit) than as a theoretically significant record of a linguistic event. Somewhat
similar objections have been revived (on different grounds) in an updated form by Pierrehumbert (1990).

Bloomfield’s objection is a very serious one, and one to which linguists have not always devoted enough attention — if only to be clear in their own minds about why they reject it. Why, after all, should we attribute to some particular subset of the physical properties of utterances the status of a fundamental characterization of language? The essential nature of language is that of a system of tacit knowledge as represented by the language organ, an aspect of the organization of the mind and the brain. In that light, the selection of some external properties of linguistic utterances as systematic to the potential exclusion of others requires at least some justification.

In this chapter, we will defend the claim that there is a significant notion of ‘phonetic representation’, one that is distinct both from a phonological representation and from a complete physical record of a speech event. This is part of our I-language system, and thus it merits the attention of linguists. The kind of representation to which we wish to attribute this status, however, is at least at first blush rather different from the sort of thing linguists typically teach their students to produce in a beginning phonetics course.1

Let us begin by asking about the factors that contribute to determining the physical properties of an act of speaking. A talker, let us assume, has something to say and initiates a sequence of gestures of the vocal organs which affect the surrounding air and are thus conveyed, perhaps, to the ears of potential listeners. Any particular speech event of this sort can be regarded as resulting from the interaction of a number of logically distinguishable aspects of the system that implements that intention, as in (6.1):

\[(6.1) \quad \begin{align*}
\text{a.} & \quad \text{The talker’s intention to produce a specific utterance (i.e., the properties that characterize the particular linguistic items — words, etc. — that compose it));} \\
\text{b.} & \quad \text{The fact that the utterance is produced by a speaker of a particular language (i.e., patterns of neuromuscular activity characteristic of the sound pattern of the particular language being spoken));} \\
\text{c.} & \quad \text{The fact that the utterance is a speech event (i.e., that its production invokes neurophysiological and motor control mechanisms}}
\end{align*}\]

1Another recent proposal involving a notable expansion of the notion of “phonetics” beyond the traditional is that of Kingston & Diehl 1994. The analyses and proposals of these authors are in some ways similar to the point of view presented below, though they differ in other technical and substantive respects that we do not go into here.
6.2. A LINGUISTIC BASIS FOR PHONETIC REPRESENTATION

that are brought into play in speech in general, as opposed, for instance, to the control regimes that are relevant to swallowing, breathing, etc.; and

d. The physical and physiological properties of the speech apparatus, the acoustics of such systems, etc.

Since we can decompose speech into its articulatory, acoustic, and perceptual aspects, we might envision (at least) three separate representations, one in each domain. Alternatively, we might seek a single representation that unifies all of these sorts of property in terms of one set of independent variables. Without going into the matter in more detail here, we should make it clear that we are quite persuaded by the arguments of advocates of a Motor Theory of Speech Perception (see Liberman & Mattingly 1985, Mattingly & Studdert-Kennedy 1991 for discussion) to the effect that the primes of phonetic specification lie in the articulatory domain, and not directly in the acoustics or in perception. This decision goes against much work in automated speech recognition, for example, which tends to be resolutely grounded in a “bottom up” approach to recovering linguistic structure on the basis of the structure of the acoustic signal alone.

Even granting the apparent difficulties that arise in the effort to specify an architecture for perception that implements a motor theory, we think that the problem for those who would model perception is to find a way to implement such an architecture. The tendency in much technologically oriented work on speech perception and recognition is rather to retreat into the notion that somehow the invariants must be out there in the acoustics if we will only keep looking, because perception would be more straightforward if they were. This is simply another example, we feel, of the drunk who persists in looking for his keys under the streetlight because the light is best there. We could refine our view of just what a Motor Theory is, but for present purposes that is not really necessary. Our reason for bringing the matter up at all is simply to be explicit about the assumption that it is articulatory activity that we wish to characterize.

6.2 A Linguistic Basis for Phonetic Representation

Returning to the problem of justifying the attribution of linguistic significance to a phonetic representation, suppose that we could arrive at a representation of the speech event that met the conditions in (6.2), one which
characterizes all and only those aspects of it that are under the control of the linguistic system (as opposed to those that are language-independent consequences of other factors). We submit that this could then be said to be genuinely linguistic, rather than accidental and external, and to furnish an appropriate notion of “phonetic representation” for inclusion in linguistic theory. Such a representation would clearly describe speech events from the point of view of the language organ.

(6.2) A phonetic representation characterizes all and only those aspects of a speech event that are under linguistic control, in the sense of being managed by the system of linguistic knowledge that the talker brings to bear in performance.

That means that we certainly want to include indications of properties that distinguish utterances from one another within a language. We also want to include indications of non-contrastive properties of an utterance that are under the control of the specific linguistic system within which it is produced — an indication, for example, of whether the occlusion of utterance-final stop consonants, or of stops in clusters, is released or not (recall the discussion of unreleased stops consonants in Korean in chapter 4). This property seems not to be used contrastively in any language. That is, there is apparently no language in which, for instance, [tap] and [tap'] might be distinct words. Despite this, it is still the case that languages can differ from one another in terms of it (cf. Anderson 1974), which means it is manipulated by the systems of individual languages.

On the other hand, we want to omit many things which we could, in principle, measure and record. By and large, for example, we can neglect the activity of the epiglottis in speech — not because it is completely inactive (it is not) or because its position has no auditory consequences (also untrue), but rather because, as far as we can tell, epiglottal activity is not something we manipulate per se, a dimension which is controlled differently in some languages than in others.

This is perhaps a controversial thing to assert, but let us be clear on why: some phoneticians have reported that in some languages epiglottal activity is manipulated for linguistic purposes. But in the absence of clear support for such claims, we would want to exclude the epiglottis from phonetic representation in spite of its physiological and acoustic importance unless and until it can be shown that epiglottal activity is independent of the control of other articulatory events, events which correspond to dimensions of irreducible linguistic importance.
What we seek, then, is a representation of all and only those aspects of a speech event that are under linguistic control, in the sense of being managed by the language organ: the system of linguistic knowledge that the talker brings to bear in performance. Another way to put this is to say that we want to characterize everything in the talker’s linguistic intention, as opposed to aspects of the event that follow from the physical, neurophysiological, and other extra-linguistic properties of the apparatus that is employed in talking. If we can provide such a representation, that would respond substantively to Bloomfield’s objection, by grounding the properties attributed to it in their role in the cognitive system that constitutes our linguistic knowledge. It would certainly be distinct from a full physical record of the speech event, since it would explicitly abandon the attempt to describe everything that is true of this event in favor of a description of everything that is linguistically determined about it.

One particular view of language in which the place of such a representation is clear is that of Chomsky’s recent Minimalist analyses (Chomsky 1995). In that approach, there are only three significant kinds of representation. Postponing one of these, the nature of lexical items, for a later chapter (7), the two key representations are the interfaces to sound, on the one hand (phonological form, or ‘PF’), and to meaning on the other (logical form, or ‘LF’). The first of these is intended to characterize all and only the aspects of an utterance’s form that are managed by the linguistic computational system, and it must be (exactly) adequate to serve as an interface to the language-independent systems of articulation and perception. PF in these terms is exactly the representation we seek.

6.2.1 Measurable but ‘unintended’ effects

It is, however, far from simple to construct such a representation. Disentangling the properties of the speech event that are due to linguistically formed intention from those that are due to the physics and physiology of the vocal apparatus, or to general motor control mechanisms that are operative in speech, is far from trivial. Particular properties certainly do not come labeled as to their source, and there are few generally applicable diagnostics in this area that can be applied directly to observations of physical events of speaking.
CHAPTER 6. PHONETICS AND THE I-LINGUISTICS OF SPEECH

Pitch micro-contours

We have said that the range of distinctions that separate linguistic elements from one another in a given language is a matter that should be resolved in the phonological representations characteristic of that language. However, the way in which particular phonological distinctions are realized phonetically is also a part of the phonology of the language. Phonetic representation must therefore provide us with a range of descriptive properties adequate to differentiate languages from one another in these terms. As a special case, we must also be able to discriminate on the one hand, between phonetic differences that correspond to different ways in which the vocal apparatus is controlled in distinct languages, and on the other, differences that correspond to the fact that some dimension which is specifically controlled in one language is left to the unguided control of the speech apparatus itself in another. In the first case, the grammars of the two languages in question differ in the way they specify the phonetic realization of the property in question. In the second case, the grammars differ in that one provides specific instructions for this realization, while the other leaves it unspecified, under the control of the speech apparatus.

Pre-systematically, here is the sort of distinction we wish to make: particular phonetic phenomena may be represented in the grammars of some languages but not others. Consider the fact that in general, vowels following syllable initial voiced obstruents begin at a somewhat lower pitch than their target value, and rise to that value over a relatively short period of time. Following voiceless obstruents, vowels begin on a slightly elevated pitch and fall to their target values.\(^2\) The basis of this effect might be the following: in order to overcome the fact that the aerodynamic conditions for vocal fold vibration are less than optimal when an obstruent constriction exists in the supra-laryngeal vocal tract, a compensatory internal adjustment of vocal fold elasticity or stiffness is a common accompaniment of voiced obstruents. Correspondingly, voiceless consonants are often accompanied by a stiffening that ensures the lack of vibration during their closure period. We do not need to understand the precise basis of the effect, however, to be sure that there is one: the testimony of the world’s languages is quite adequate on that score.

It is not hard to show that the vocal fold stiffness adjustments accompanying consonants are logically and phonologically independent of the dimensions of specification of vowel pitch. Nonetheless, such gestures associated with consonants have consequences for the rate of vibration, when, at some

\(^2\)See Lehiste 1970 for a review of the classic literature establishing these facts
6.2. A LINGUISTIC BASIS FOR PHONETIC REPRESENTATION

point shortly after the release of the occlusion, it occurs. This is a matter of the physics and physiology of the structures brought into play in speech production. On the other hand, we might well see the actual time course of the pitch contour between the consonant release and the following vowel as characteristic of the baseline control regime of the vocal organs as this applies in speech. The stiffness or slackness of the folds disappears with some particular latency and elasticity, and it seems reasonable to attribute the particular contour of pitch that results in a language like, say, English to general mechanisms of speech motor control, not to the grammar of English.

In general, we find essentially the same pitch contours in post-consonantal vowels across many languages. In some languages, however, such as Yoruba and Thai, where tone is linguistically significant, we find that the pitch contours of post-obstruent vowels are much sharper. We might suggest that in these cases, the language specifies an independent compensatory gesture that has the effect of bringing the vowel to its intended (and significant) pitch value more rapidly than in English. Compare the contours in Figure (6.1).

In English, as represented by the graphs in figure (6.1a), we see that there is still a difference in pitch in the vowel following a voiced consonant (e.g., [b]) as opposed to a voiceless one ([p]) more than 100 milliseconds after the stop release. In Yoruba, on the other hand, there are three contrastive tonal levels. We see in figure (6.1b) that a vowel following voiced vs. voiceless consonants displays the same (high, mid or low) pitch value within roughly 50 milliseconds after the stop release (slightly longer in the case of a high tone, but still sooner than in the non-tonal language English). Hombert (1976, p. 44) argues “that there is a tendency in tone languages (which does not exist in non-tonal languages) to actively minimize the intrinsic effect of prevocalic consonants.”

(Figure 6.1 about here)

We suggest that Yoruba phonetic representations thus should specify the presence of this property, which we can distinguish from matters of general physics and physiology (including the fact that the vowel pitch deviates at all from its target initially) as well as from matters of general speech motor control (such as the time course of pitch change that occurs, other things being equal). What matters is that in Yoruba, other things are not in fact equal. Phonetic representations in this language have to indicate a language-specific compensatory gesture that is part of its system, part of what the language organ of Yoruba speakers determines. Phonetic representations in English, on the other hand, ought not to say anything about the post-
consonantal pitch contours: the contours that we can observe physically are a matter simply of the vocal organs, not of the language organ.

Vowel length effects

In this instance, we can see how we want the answer to turn out, but in the general case there are many questions whose answers are not self-evident about what does and does not belong in a phonetic representation. A first problem, noted in Fowler 1990, results from the fact that not all of what is systematically carried out in speaking is due to linguistic intention. We might hope that any pattern which occurs in some languages, but not in others, must surely be under linguistic control, and thus suitable for inclusion in our phonetic representation. This is, in fact, a very standard form of argument in the linguistic phonetic literature.

Fowler (1990) points out, though, that this apparently obvious diagnostic may fail us. She notes that one of the most often cited phonetic regularities in the literature is the fact that, in most languages, vowels are shortened by ca. 20–30 msec. before voiceless obstruents as opposed to the duration they show in other environments. In a few languages, (cf. Anderson 1981 and
6.2. A LINGUISTIC BASIS FOR PHONETIC REPRESENTATION

references cited there), this effect is suppressed, and vowels display the same durations regardless of the nature of a following consonant. In English, on the other hand, the difference in length between vowels before voiced and voiceless consonants is considerably exaggerated, and (in contrast to the case in most others) well above the threshold of perceptibility.

Apparently, in languages like Saudi Arabic where the expected contextual vowel length difference is apparently suppressed, it is necessary to indicate this as a fact about the linguistically governed intentions of speakers. This is so despite the fact that the result is no difference at all in vowel duration regardless of context. And of course, it is also necessary to indicate in a representation of English utterances that vowels are considerably longer in voiced environments than would otherwise be expected. But what about the languages that simply show the expected small differences? Fowler argues that in fact these differences arise as a mechanical consequence of the different dynamics of closure gestures in voiceless as opposed to voiced obstruents: the talker ‘intends’ the same thing with respect to the vowels in both cases, but the implementation of that intention leads to an unintended consequence as a result of the coordination of the vowel and consonant gestures. In this case, then, the phonetic representations of vowels ought to be uniform, despite the fact that acoustically the vowel portions of utterances differ quite systematically by 20–30 msec. depending on the environment. Paradoxically, vowels of equal duration may thus come to be specified differently (the Saudi Arabic case) while ones of different duration will not differ in their phonetic representation where the differences are not intended per se. On the other hand, where the durational differences considerably exceed what can be attributed to the mechanical effects (as in the English case), we must again indicate the presence of some specific phonetic intention responsible for this.

**Laryngeal specifications in Korean**

This example leads us to the possibility that some of the observable properties of phonetic elements may result not from our active articulatory intentions, governed by the language organ as properties of a particular language, but rather as consequences of the necessities associated with other articulations. This notion has been discussed in the phonetic literature by Patricia Keating (1988) as a matter of underspecification.

Keating begins from the premise that parameters of a segment that are predictable should be unspecified in a phonological representation indicating contrasts, i.e. that phonological representations, at least, are underspecified
in the sense that not all segments have specified values for all possible properties. This is thus a version of the incompletely specified view of phonological representations discussed earlier in section 4.2.1. She goes further, however, in suggesting that under some circumstances such underspecification may persist into the phonetics. It may be the case, that is, that the values assumed along some dimension during some period of an utterance result not from any linguistically governed property of that part of the utterance, but simply by interpolation, as a result of the exigencies of getting from where you were to where you have to be. The spirit of such phonetic underspecification is entirely consistent with the view being developed here.

For example, Korean (cf. Kagaya 1971) contains three series of stops: “voiceless aspirated” stops, followed by a puff of breath as in for the [pʰ] in English pit, as opposed to “unaspirated” ones like the [p] of spit; and “voiceless glottalized” ones, associated with a simultaneous constriction of the vocal folds. Two of these, the voiceless aspirated and the voiceless glottalized, clearly have characteristic properties: in the one case, a spread glottis gesture (which is what produces the puff of breath) and in the other a constriction of the glottis. The remaining series, however, is what interests us.

These sounds are described as voiceless unaspirated initially, but voiced intervocally. Measurement of the time course of opening of the glottis (Kagaya 1971) during their production suggests that the voicelessness in initial position is actually a gradual transition from the open glottis position associated with the neutral position for speech in that environment to the approximated position required for the following vowel; and that intervocalic voicing represents merely the maintenance of the position appropriate to the consonant’s surroundings. The “unaspirated” stops, then, have no value of their own for the position of the larynx: they are phonetically underspecified for this property.

Vowel quality differences

Identifying such cases may in fact be quite difficult, since the mere fact of apparent predictability does not suffice to distinguish our linguistic intentions from the baseline behavior of the systems employed in speech. Sometimes properties that may seem thoroughly predictable nonetheless look as if they have linguistically enforced target values. Boyce et al. (1991) argue that for some speakers, at least, there is apparently an articulatory target for the property of lip rounding, under the control of the language organ even for segments like [t] where this would seem to be particularly irrelevant and
6.2. A LINGUISTIC BASIS FOR PHONETIC REPRESENTATION

thus most likely to follow from interpolation.

Fowler (1990) suggests that in some cases, what appears to be a gestural target may in fact follow from the assumption that the speech apparatus tends to return to the neutral position for speech insofar as it is otherwise unoccupied, and some apparent deviations from straightforward interpolation may result from this effect rather than from an active articulatory target. Sorting out these hypotheses is not easy. Nonetheless, even though many apparently predictable properties may turn out to be actively determined after all, it seems eminently plausible that a number of examples will remain in which some articulatory phenomenon results not from phonetically controlled gestural intentions but rather as consequences of interpolation between other, surrounding gestures.

An interesting case in which we can ask how much of what is articulated is actually a part of the (I-language) grammar concerns the vowel systems of languages with very few contrasting segments.

Anderson (1978) takes up a proposal originally due to Aert Kuipers (1960) concerning the Circassian languages, particularly Kabardian. This language has a full range of vowel types (including [i, e, a, o, u], etc.), and a listener with no knowledge of the language would notice nothing unusual about its vowels. On closer inspection, however, it turns out that in Kabardian, the actual quality of a vowel in any given position in a word is almost completely predictable from its surroundings. That is, both [i]’s and [u]’s occur in Kabardian words, for example, but [i] only occurs where adjacent to certain consonants, and [u] only adjacent to some others. The two are in complementary distribution in that there are no environments in which either can appear, as the basis of a contrast between words. Furthermore, the consonants surrounding [i] vs. [u] are not arbitrary: those favoring [u] are ones whose articulation is similar to that of an [u] in involving lip rounding and a relatively retracted tongue position, while those favoring [i] involve no rounding, but a relatively high, front tongue body position. Similar observations can be made for the environments in which the other phonetic vowels are found.

These facts suggest that in Kabardian, the wide phonetic variety of vowels corresponds to little or no difference in linguistic intention: essentially, the vowels consist of a vocalic transition from the position of a preceding consonant to that of a following one. In any given environment, there are in general only two possibly distinct qualities that can occur, which we can differentiate by saying that the transition is made with or without an accompanying deflection of the tongue body downward. The presence or absence of this deflection is thus the only way in which the articulation of one vowel
differs from that of another per se.

This kind of analysis was examined more deeply by John Choi (1992) in a UCLA dissertation. Choi explored a variety of possibilities quite closely, and examined articulatory and acoustic data on languages of this type. One such language which he analyzed in detail is Marshallese. Choi concluded that the front-back dimension in the realization of Marshallese vowels does not correspond to anything that is under independent linguistic control: no gesture of fronting or backing intervenes in the relatively smooth transition from preceding to following tongue body position as determined by the surrounding consonants, although tongue height is independently controlled. This state of affairs is thus intermediate between the almost completely underspecified vowels of Kabardian and a fully determined system.

When is a ‘schwa’ not a [a]?

Even in less ‘exotic’ languages, we may find facts that are suggestive of a similar conclusion. In particular, reduced vowels (e.g., those of unstressed syllables in English, where there are many fewer possible vowel qualities than in stressed syllables) may have some surprises in store for us as far as their phonetic content is concerned. Browman and Goldstein (1992c) examined the production of schwa (the vowel in the first syllable of tomato) in English, with the expectation that this would prove to have no distinct articulatory target associated with it. In fact, the acoustic character of the vowels we transcribe as [a] varies considerably as a function of the vowels in surrounding syllables (and to some extent, of adjacent consonants). English schwa is generally the result of a principle of reduction which neutralizes other differences in vowel quality under given conditions, especially absence of stress: compare the final syllables of relate, compare, precede, where there is a stress, with the corresponding syllables in relative, comparable, precedence, where that stress does not appear. Since the thing that seems to distinguish schwa from the other vowels of English is precisely its lack of potentially contrastive properties, it might be logical to expect it to have few or no positive properties of its own.

Nonetheless, Browman and Goldstein found evidence that schwas produced by English speakers in certain environments do indeed have a target associated with them: [ə] in /pVpVpVp/ (e.g., [pipəp]) has an articulatory target in the middle of the English vowel space (Browman & Goldstein 1992c), a sort of average of all of the other vowels, but they observe a distinct gesture deflecting the tongue body toward this position in the transition from a preceding vowel’s articulation to that of a following
one. The [a] target frequently has little or no direct acoustic consequence (beyond what we would expect of a smooth, targetless transition), but it is apparently present nonetheless in the speaker’s intention.

Interestingly, though, a later study (Browman & Goldstein 1992b) showed that while schwas in syllables within lexical items display this kind of articulatory target, schwas that appear epenthetically before the productive inflectional endings of the language may well be targetless after all. Thus, in “If Nita’d even KNOWN...” the underlined schwa shows a distinct lowering of the tongue dorsum, but in “It’s needed even NOW...” there is little or no such lowering. This suggests that the “schwas” in such endings, though often acoustically indistinguishable from other schwas, nonetheless correspond not to the speaker’s intention to produce a specific vowel, but rather to a dynamic effect in articulation whereby two similar gestures are sequenced in time so as to separate them, resulting in a brief interval whose only relevant property is the duration of the gestural offset. *Nita’d* thus contains two real, phonetic vowels, while *needed* contains only one, plus an additional vocalic interval that arises as a consequence of the need to separate the two d’s.

In all of these cases, we have portions of the speech event that bear properties which we could perfectly well measure, and record, and which may correspond closely to properties that are clearly linguistic in other cases; but for which a case can be made that in these instances, they correspond to nothing specific in the speaker’s intention. There is thus a clear difference in the extent to which the same E-language facts correspond to something present in the I-language of a given speaker, depending on the language. In a principled phonetic representation of the sort we are interested in here, then, we ought to omit them in the case where they are not in fact governed by the language organs of speakers of the relevant languages; but it does not appear that we can always determine that fact directly from examination of the physical facts alone.

### 6.2.2 Intended effects hidden from observation

We must, furthermore, cope with the fact that not all of what we intend in speech is actually realized physically (articulatorily and/or acoustically). Lindblom’s “H&H” theory (Lindblom 1990), for example, suggests that the phonetic characterization of phonetic elements might be in terms of a “hyper-articulated” form whose actual implementation within a larger context may be (indeed, generally is) incompletely achieved by virtue of other properties of the context.

This notion is further supported by Johnson, Flemming & Wright 1993,
who provide evidence that listeners identify vowels in terms of extreme, hyper-articulated targets rather than by their actual values. When asked to select the simple vowel that best corresponds to a vowel in a word in running speech, that is, they choose a ‘pure’ quality rather than one closer to the actual physical signal they have just heard. Here we have a variant of the traditional notion of “undershooting the articulatory target,” the point being that there is good reason to believe that the talker’s actual phonetic intention, as well as the listener’s interpretation of the event, is characterized by an ideal target and not by its undershot (or “hypo-articulated”) realization.

The most direct sort of argument for this general conclusion is provided by facts of casual speech cited by Browman & Goldstein 1990b. They observed that in relatively casual pronunciations of sequences such as must be, ground pressure, ‘phonetically’ [masbi[, [graumprəθ], there is apparent “loss” of the segments [t] and [d]. X-ray microbeam studies, however, reveal that an alveolar stop closure is actually made in each of these cases. It has no acoustic consequences, though, since it is (more or less) completely concealed by its overlap with the labial closure of the following [b] or [p]. The presence of the stop in the speaker’s I-language representation is confirmed by the articulatory data, even though this has no reflex in the actual signal, and thus would not be present in a traditional (E-language) phonetic representation.

In many forms of English, final stop consonants are lost in certain environments in fluent speech: thus, mist may be pronounced as [mis]. It is natural to ask whether the same sort of gestural ‘hiding’ is responsible for this. The phenomenon of variable final stop deletion in English is clearly very complex and related to a number of conditions. In some instances, such as the loss of final stops in morphologically simple (“monomorphemic”) forms before pause (e.g., the pronunciation of mist just cited), it is clear that gestural reduction rather than hiding must be at work, since there is no other gesture in the environment which could be said to overlap with the final [t] so as to obscure it. We must conclude that for such speakers, the reduction of the [t] in this case must be due to an I-language principle of phonetic realization.

Potentially, there is further empirical evidence to be found for Browman and Goldstein’s proposal: the gestural overlap account suggests that

---

4Guy (1991b, 1991a) provides a summary of what is known about this variation, and Kiparsky (1993) suggests an account in terms of Optimality Theory of the range of systems that have been found.
varying degrees of overlap ought to be found in a wide enough sample of speech, while the traditional (deletion) story is only up to describing variable probabilities for a categorical process deleting (or retaining) the stop. These considerations should not obscure the point of the example for our purposes, however: in at least some instances a gesture is actively intended, and indeed accomplished, despite the fact that other conditions interact with it so as to deprive it of any audible effect.

Such a gesture is obviously part of our intention, and surely appears in the instructions to the vocal apparatus (i.e., in PF, the \( I \)-language phonetic representation), despite the fact that it is quite undetectable in the acoustic event. Obviously, then, such a phonetic representation is not at all a neutral observation language for speech, simply reflecting physically measurable properties of the sounds produced. Instead it is a rather abstract and theoretical articulation of our understanding of what is orderly and coherent about speech events from the point of view of the dimensions controlled by the system of language, the language organ.

### 6.2.3 Discreteness of phonetic dimensions

Let us return to the issue of whether phonetic transcription is in fact a coherent notion that ought to play a role in linguistic theory. A recent argument that it is not is due to Pierrehumbert (1990), whose objections are based on the fact that traditional phonetic transcriptions are discrete and categorial (e.g. symbols of the International Phonetic Alphabet or ‘IPA’), but as noted in (6.3), there are several ways in which this fails to reflect linguistic reality.

(6.3) a. Effacement of segments in reduction (and epenthesis, such as the oral stop in nasal+fricative clusters) is a matter of “less and less” (or “more and more”), not “something vs. nothing.”

b. Language-specific variation may involve continuous detail, so there is no (finite) list of possible sounds.

c. Gradient (but systematic) phenomena such as intonation and laryngeal lenition (Pierrehumbert & Talkin 1992) require attention to hierarchical organization, not just fine transcription.

d. Phonetic details, as opposed to (categorical) phonological contrasts, are out of awareness, and thus not “cognitive”.

There is some cogency to these complaints about phonetic representations of the sort phoneticians teach beginning students to produce, but
their force is to show that this needs to be revised, not abandoned. We adopt here the gestural perspective of Articulatory Phonology (Browman & Goldstein 1990a, 1990b, 1992a, and elsewhere), on which the properties we specify correspond to dynamic *gestures* of the vocal organs, aspects of what a speaker intends to do in speaking (as opposed to frameworks based on describing static articulatory configurations or targets). Recall from chapter 4 that the representation of speech as a sequence of discrete, static segments is after all a theoretical idealization, warranted essentially by the insight it can bring to the study of sound in language. To the extent that that picture actually interferes with a coherent view of the relation between I-language structure and the implementation of speech, it needs to be replaced with a view that is more appropriate to its subject matter.

The gestural view provides us with a language for describing some of these phenomena, and for responding to the substance of the objections raised by Pierrehumbert. Browman and Goldstein’s model is not the only one that might serve for this purpose: the “C/D” model proposed by Fujimura (1994), for instance, has somewhat similar goals but interestingly different dynamic properties. This diversity suggests that there are quite genuine empirical issues involved in the determination of the properties of an appropriate phonetic representation, but that these issues are independent of the logical basis of phonetic representation as a characterization of the talker’s intentions. This is surely a desirable result (if not a very surprising one).

In responding to Pierrehumbert’s points, important advantages derive from the fact that the gestural representations of Browman and Goldstein do not have the exclusively categorial and segmentalized character of traditional transcriptions. Instead, gestures have magnitudes and temporal extent which are (at least in part) subject to specification and manipulation by the principles of individual linguistic systems. The representational vocabulary of I-language, on this account, thus includes some aspects of speech dynamics, as is apparently required, and not simply the description of a sequence of static configurations. Furthermore, the overall organizing principles of gestural synchronization and sequencing, and the description of the time course involved in the transition between distinct gestures, are also (at least in principle) parameterizable as a property of particular linguistic systems, thus determined as a property of individual language organs. This sort of specification of the overall system of gestural realization in a given language can be understood as something like a formalization (in part) of the traditional impressionistic notion of the “base of articulation” of the language (including statements like “French is spoken in the front of the
6.2.4 Timing effects

Let us assume, therefore, that the language organ deals with phonetic specifications in terms of intended gestures, rather than (as in much traditional terminology) semi-snapshots of ideal articulatory configurations or an acoustically based parameterization of speech events. An important component of gestural representations is the fact that they bring a non-discrete notion of time into consideration, in ways that traditional phonetic representations based on sequences of individual independent phonetic segments do not. Once non-discrete phenomena based on the detailed time course of speech are taken into account, the problem of isolating those dimensions of control that are under the control of linguistic intention (that is, that are part of I-language) and of identifying the range of specifiable variation along those dimensions becomes even more difficult. It must be stressed that phonologists — the people most concerned about language-particular characterization — have no serious descriptive language for low-level temporal dynamics, let alone an articulated theory of the independent variables in this domain.

In fact, phonologists have generally had little to say about the details of speech timing. Phoneticians know a lot about the time course of speech events, but phonologists have until quite recently worked hard to develop models in which there is literally no place for the details of time course: phonological elements are points in time that can succeed one another but that do not have duration, magnitude, elasticity, and so on. A widespread (but completely unsubstantiated) assumption among phonologists is that it is precisely in this domain that we can attribute everything we observe to the (language-independent) mechanics of speech articulation.

But of course the increased information content of non-discrete gestural representations potentially comes at a high price: it provides us with too much freedom in its raw form. In order to isolate the linguistically significant degrees of freedom in the domain of the time course of speech, we must separate the dimensions of rhythm and timing that are under the control of the language organ from the background dynamic properties of the articulatory system.

It seems clear that there are some important language-particular components to the determination of speech rhythm and timing: that is, there are properties of this sort that differentiate languages from one another, as anyone can testify who has vaguely heard a conversation across a crowded
room and recognized immediately that the speakers involved must be, say, Dutch, though without actually picking up a single word. See Beckman et al. (1992) for some discussion. And if these factors have a language-particular component, they must be reflected in phonetic representations of the sort we seek. But until we have phonological models that take time course seriously, we have quite literally no principled way to talk about what might be language-particular about speech dynamics. Such models are in fact beginning to emerge, but the problem has not, perhaps, received the attention it deserves from phonologists. And of course, until we know what aspects of speech organization are subject to language-particular determination, we do not know what we need to have a principled theory of in order to describe this aspect of the language organ.

Introduction of an appropriate specification of rhythmic and temporal effects into the description of the aspects of speech that fall under linguistic control is thus an important refinement of the range of problems to which phonologists should attend. We have assumed above that the form this might take is simply a specification of appropriate aspects of the time course of particular gestures in speech, but there is another aspect to the problem. Not only can the implementation of particular gestures over time fall under linguistic control, but independent of this, their synchronization and relative organization can provide yet another dimension of possible variability. Browman and Goldstein and their colleagues (Browman & Goldstein 1998, Browman, Goldstein, Honoroff, Jebbour & Selkirk 1998) develop this point in suggesting that otherwise comparable gestures in English and Tashlhiyt Berber, for example, differ in their ‘bonding’ or relative synchronization more than in their individual component timings. These ways of describing timing relations need further refinement and application to a variety of language types.

6.3 Speech Microprosody: A Research Program

A greater understanding of the ways in which languages can differ in their overall patterns of speech rhythm and timing requires much more study of speech micro-prosody: the low-level dynamics of pitch, intensity, and a variety of measures of the temporal control of articulation within and below the level of syllables. The problem in this investigation is to hold physical and phonetic factors (roughly) constant, while varying control of the

---

4See Byrd & Saltzman 1998 and Byrd, Kaun, Narayanan & Saltzman in press for one approach to a phonological model of speech dynamics
One possibility which can be exploited is that of comparing ordinary native speakers of (e.g.) English with speakers who have a good command of the language, but a “foreign accent.” A plausible hypothesis is that prosody is the hardest thing to learn, partly because it cannot be taught (since we do not know how to articulate our knowledge of it) and partly because in general we are not even aware of it. As we will note below in chapter 9, prosody is apparently controlled by the language organ at a very early stage of the language acquisition process — even before the appearance of the first words in the child’s production. A similar point is argued at length by Locke 1993. Although this is surely not all of the story, one large component of a “foreign accent” is likely to be prosodic anomaly. On the other hand, these speakers are still perfectly normal with respect to the basic phonetic mechanisms underlying articulation; and so we can assume that ways in which they deviate from the patterns evidenced by ordinary native speakers will reveal differences in the cognitive system controlling the dynamics of articulation, and thus aspects of the dimensionality of that system.

But there are several other populations of interest where we can hope to find anomalies vis à vis normal prosody. One of these is a class of clinical brain damaged patients who display what is called Acquired Foreign Accent Disorder. These are people who (sometimes only for a relatively short period) during recovery from cerebral insult (strokes and other sources of brain damage) are perceived to have acquired an accent — typically that of a language to which they have never been exposed. A notable example from the literature (Monrad-Krøhn 1947) is a Norwegian woman in WWII Norway who was perceived to have a German accent following head trauma (sustained, ironically, in a German bombing attack), and who was subsequently ostracized in her village on this account, when in fact she had never had any exposure to spoken German. Some rather similar symptoms can occasionally be induced by electrical stimulation of areas of the cortex in epileptic patients undergoing exploratory investigation of anatomical brain function prior to surgery intended to relieve intractable seizures.

\footnote{Some references on this topic include Graff-Radford, Cooper, Colsher & Damasio 1986, Blumstein, Alexander, Ryalls, Katz & Dworetzky 1987, Kurowski, Blumstein & Alexander 1996, and Carbary, Patterson & Snyder 2000. The interpretation of this disorder as one of a prosodic, rather than segmental, nature, as suggested below, has come to be fairly widely accepted, though details vary considerably among both patients and researchers. A particularly useful review with respect to this issue is provided by Berthier, Rutz, Massone, Starkstein & Leiguarda 1991.}
Examination of the phonetic details of the speech of such patients suggests that in general, whatever they are doing, it bears little or no specific resemblance to the speech of genuine foreign accented (normal) speakers. Furthermore, attempts to characterize what these patients are doing wrong in terms of segmental articulations seem to fail (or at least to be quite insufficient). Their similarity to foreign accented normal speakers lies in the fact that while they get (most of) the segmental phonology right, their prosody seems to be disordered.

Our hypothesis here is that the cognitive deficit underlying ‘acquired’ foreign accents is — or at least often involves — impaired control of rhythm and timing, the prosodic organization of the native language. When prosody sounds wrong, but the speech is still basically segmentally correct and intelligible, other speakers perceive this as a “foreign accent.” This is thus a circumstance in which the language organ’s control of speech prosody is disrupted, and comparison of these patients with normal speakers provides another potential basis for exploring the properties of the system itself.

Studying both natural and acquired foreign accents in comparison to the prosody of normal speech gives us two somewhat different windows on the question of what aspects of prosodic organization are part of the linguistic system, such that control of them could be lost (or fail to have been correctly acquired). There is also another kind of population that would be of interest: subjects in which a disorder exists that affects the system of articulation in general, rather than the cognitive systems that provide the instructions for control of articulation. The reason to look at this as well is to disentangle systematics that are, as it were, ‘post-PF’ (i.e., implementational matters external to the I-linguistic organization of speech) from those that represent dimensions of variation in a cognitively significant phonetic representation.

At least two such populations are potentially available for study, both involving sub-cortical damage that results in speech disorders. In cerebellar lesion patients with ataxic dysarthria, we find a perceived speech rhythm impairment, without (in general) corresponding impairment to higher level language functions. Somewhat similar symptoms are displayed by some early AIDS patients in whom subcortical degenerative damage has occurred. Our hypothesis is that the deficit in these patients is in the system responsible for general control and organization of rhythmic speech movements (cf. Keller 1987). Theirs is a problem in the general phonetics of prosody, not the language specific cognitive system.

These populations can be studied with respect to standard acoustic measures of timing, the time course of acoustic intensity and pitch, and also with articulatory measures (e.g. the time course of jaw movement, ultrasound...
6.4. CONCLUSION

studies of tongue movement, and other articulatory variables) as possible, depending on the conditions of access to individual subjects. The hope is that comparing these talkers along these dimensions will give us insight into what aspects of normal performance are attributable to what aspects of the overall speech and language system. Such investigations are still in early stages of development — there are major problems of principle and experimental design yet to be resolved, and few concrete results. The purpose of bringing up this work at all is simply to propose a methodology for gathering the sort of information we will need in order to arrive at an improved understanding of those dimensions of speech rhythm and timing that are part of the system of I-language, and which thus ought to be recognized as components of a linguistically significant phonetic representation.

6.4 Conclusion

If we take the lessons of the sections above to heart, the representation we arrive at for “PF” — the interface of the language organ with the mechanics of articulation — is quite distant from the sort of fine phonetic transcription using symbols of the International Phonetic Alphabet (or some other, similar system) which most current linguists learned as students (and generally continue to teach). These include at least the points in (6.4).

(6.4) Such a representation

a. fails to indicate some properties that are perfectly true, measurable, and systematic about utterances, insofar as these are the consequences, not specifically intended, of other aspects of speech;

b. indicates gestures that are intended even under circumstances where those gestures do not have their intended auditory (or perhaps even articulatory) effect;

c. specifies gestures with associated temporal dynamics, and not just timeless points, and which are related in hierarchical ways;

d. indicates gradient temporal relations rather than mere succession;

---

7The same representation is also appropriate as the interface between the language system and perception, if something like the Motor Theory referred to above is correct, as we presume.
A representation with these characteristics obviously cannot be taken to be the kind of physical observation language phonetics is often presumed to provide. It is rather a description of the cognitively real (though largely unconscious) representations that underlie speech motor control and that are in some sense recovered in perception. Despite its non-traditional character, it does seem to serve the larger goal of characterizing what is linguistically significant about the facts of speech. It is neither a full physical record of speech events, nor a restricted characterization of the minimal distinctive core that distinguishes higher-level linguistic elements from one another. As such, it serves as the output (or perhaps better, as the implicature) of the sound-structural regularities of a language. We claim that these are the things that ought to be demanded of a “phonetic representation” by those for whom such a notion finds its significance in the theory of language and the mind, rather than in physics or physiology. That is, it is an appropriate way of characterizing PF, the connection between the language organ and the organs of speech and hearing.