Phonetics and Phonology  

Abigail C. Cohn  

1

Privativity, underspecification, and consequences for an adequate theory of phonetic implementation  

Abigail C. Cohn  

39

The prosodic structure of Burmese: a constraint-based approach  

Antony Green  

67

Speaker normalization in the perception of Mandarin Chinese tones  

Corinne B. Moore  

97

English loans in Japanese: constraints in loanword phonology  

Ayako Tsuchida  

145

Non-native productions of Thai: acoustic measurements and accentedness ratings  

Ratreed Wayland  

165
The Cornell Phonetics and Phonology Group:

Niken Adisasmito, Graduate Student in Linguistics
Jon Alcantara, Graduate Student in Linguistics
David Brandt, Graduate Student in Linguistics
Ioana Chitoran, Graduate Student in Linguistics
Abigail Cohn, Assistant Professor of Linguistics
Tobey Doelman, Graduate Student in Linguistics
David Garrett, Undergraduate Researcher
Beverley Goodman, Graduate Student in Linguistics
Bill Ham, Graduate Student in Linguistics
Jeong-Im Han, Graduate Student in Linguistics
Susan Hertz, Senior Research Associate
Michael Jessen, Graduate Student in Linguistics
Allard Jongman, Assistant Professor of Linguistics
Hyunsoo Kim, Graduate Student in Linguistics
Lisa Lavoie, Graduate Student in Linguistics
Becky Letterman, Graduate Student in Linguistics
Corinne Moore, Graduate Student in Linguistics
David Parkinson, Graduate Student in Linguistics
Joan Sereno, Acting Assistant Professor of Linguistics and Psychology
Elizabeth Slack, Pew Undergraduate Researcher
Ayako Tsuchida, Graduate Student in Linguistics
Ratree Wayland, Graduate Student in Linguistics
Munther Younes, Senior Lecturer in Near Eastern Studies
Draga Zec, Assistant Professor of Linguistics

Technical and Administrative Staff:
Scott Gargash, Systems Manager
Neal Jacobs, Audio-Video Technician
Marge Leonard, Administrative Aide
Steve Masiclat, Director, Language Learning Center
Previous Issues

WPCPL No. 2 (April 1988, 176 pp.): Research in Laboratory Phonology ..............................................................$7.00

WPCPL No. 3 (June 1988, 109 pp.): Stress, Tone and Intonation .................................................................$6.00

WPCPL No. 4 (December 1990, 262 pp.) ...........................................................................................................$8.00

WPCPL No. 5 (September 1991, 123 pp.): Phonetic and Phonological Studies on Vowel Features .................$8.00

WPCPL No. 6 (October 1991, 356 pp.) ..............................................................................................................$8.00

WPCPL No. 7 (March 1992, 179 pp.) ...............................................................................................................$8.00

WPCPL No. 8 (December 1993, 258 pp.) .........................................................................................................$10.00
  M. Jessen: Stress Conditions on Vowel Quality and Quantity in German
  A. Bradlow: A Cross-language Comparison of vowel production and perception:
    Language-specific and universal Aspects
  C. Moore: Phonetic Observations on Stress and Tones in Mandarin Chinese
  J. A. Sereno & A. Jongman: Acoustic Correlates of Form Class
  A.C. Cohn: A Survey of the Phonology of the Feature [± nasal]
  S. Inkelas & D. Zec: Auxiliary Reduction Without Empty Categories: A Prosodic Account

WPCPL No. 9 (December 1994, 283 pp.) .........................................................................................................$11.00
  J. Alcantara: Rethinking English Vowel Alternations
  I. Chitoran: Acoustic Investigation of Georgian Harmonic Clusters
  A. C. Cohn & K. Lockwood: A Phonetic Description of Madurese and its Phonological Implications
  K. De Jong: On the Status of Redundant Features: the Case of Backing and Rounding in American English
  J. Han: Perceptual Cues for Korean Tense and Lax Consonants
  H. Kim & A. Jongman: Complete Neutralization of Manner of Articulation in Korean
  R. Letterman: A Phonetic Study of Sinhala Syllable Rhymes
  A. Tsuchida: Fricative-Vowel Coarticulation in Japanese Devoiced Syllables: Acoustic and Perceptual Evidence
  D. Zec: Footed Tones and Tonal Feet: Rhythmic Constituency in a Pitch Accent Language

WPCPL No. 10 (December 1995, 189 pp.) ......................................................................................................$11.00

This issue of the Working Papers was edited by Ioana Chitoran and Ratree Wayland. Each paper was reviewed by one faculty member and one graduate student. We wish to thank all the reviewers.

Copies of these reports are distributed free of charge to phonetics, phonology and speech laboratories interested in exchanging their own reports with us, and are otherwise available from us for the prices indicated above, plus a shipping charge of $2.00 for the first book, $1.00 for each additional. Add $1.00 for orders to be shipped outside North America. Checks must be in U.S. funds, drawn on a U.S. bank, payable to "DMLL". Orders and inquiries should be sent to: DMLL Publications, Morrill Hall, Cornell University, Ithaca, NY 14853, USA. Our e-mail address is books@plab.dmll.cornell.edu.
Phonetics and Phonology*

Abigail C. Cohn

In this chapter, I investigate a number of issues about phonology, phonetics, and their relationship. First, I discuss the nature of both phonology and phonetics, by considering the nature of rules, representation, and underspecification in each domain. I then focus on the nature of the relationship between phonetics and phonology and the question of distinguishing between the two. It is concluded that phonology and phonetics are indeed formally distinct in that what they manipulate is different: in the phonology, abstract qualitative representations are manipulated and in the phonetics such representations are realized quantitatively in both time and space. It is argued that the mapping from phonology to phonetics can be insightfully represented by a target-interpolation model. Such a model is taken as a starting point to account for phonetic implementation of the feature Nasal, the goal of the subsequent chapters of the study.

Phonetics and phonology are usually defined as distinct areas of study. Phonetics is the study of the physical properties of sounds used in human speech: their production, their acoustics, and their perception. Phonology is the study of how speech sounds pattern together. Yet there is also an implicit derivational relationship between the two: a phonological representation indicates the abstract, linguistic characteristics of sounds; the phonetic representation is the physical output or realization of the phonological representation, what the speaker actually produces or the hearer perceives. This relationship can be investigated from the point of view of production or perception.

There is evidence from psycholinguistics that abstract phonological representations exist for speakers and hearers and many researchers conclude that this is part of what speakers/hearers know about their language. Following this view, the nature of the relationship between abstract linguistic representations and the actual physical output is important for our understanding of language as a human cognitive process, since it offers us insight into one aspect of the linguistic behavior of both speakers and hearers in using human language. Thus not only the abstract patterns of sounds, but also the mapping processes between distinctive contrasts and physical events are central issues in our understanding of phonology, as one facet of linguistic knowledge.

---

* This work is the introduction of a book in progress, entitled From Phonology to Phonetics. It takes chapter 1 of my dissertation, Cohn (1990), as a starting point. Some of these issues are also taken up in Cohn (1993b) §2. Thanks to Ioana Chitoran, Allard Jongmin, Pat Keating, and Draga Zec for comments. Thanks also to Bruce Hayes, Pat Keating, Peter Ladefoged, Donka Minkova, and Carlos Otero for general discussion of the issues presented here.
It is often assumed that mapping from an abstract representation to the physical output is a fairly mechanical process; the specifics of this process are usually only hinted at. However, recent work on this topic shows that the mapping from the phonology to the phonetics is in no sense trivial. (See notably Pierrehumbert 1980, Beckman and Pierrehumbert 1986, Pierrehumbert and Beckman 1988, and Keating 1984; 1985a; and 1988a; see also Browman and Goldstein 1992 and work cited therein for a somewhat different view.) In this study, I investigate the mapping from an abstract phonological representation, viewed in terms of discrete and timeless segments, to a quantitative, phonetic representation, realized in time and space, through a detailed study of the feature Nasal, including its phonological behavior and its phonetic implementation. In this first chapter, I discuss the nature of phonology and phonetics and their relationship.

The widely held view of phonology and phonetics as two distinct areas of investigation is in fitting with a modular view of the linguistic grammar, a tenet of modern linguistic theory (see Chomsky 1988). Phonology and phonetics, while both concerned with sound structure, have distinct goals and are governed by different principles. Within a generative perspective, phonology provides a set of representations and rules for any given language. The set of representations allows all of the meaningful contrasts in a particular language to be made; while the rules flesh out these representations to provide a surface representation interpretable by the phonetics. The phonology serves the dual role of representing the linguistically relevant aspects of sound structure and providing the input to the phonetic realization.

Within the model of generative phonology, the phonology must encode the non-predictable aspects of pronunciation of all of the lexical items within any human language. It is generally agreed that only non-predictable information is represented in the mental lexicon and that predictable information is introduced by rule. A range of evidence strongly supports the view that phonological representations involve discrete units with a segmental component. (See Halle and Stevens 1990 for discussion of this point.) A form such as will /w1l/ is assumed to consist of three discrete sounds or phonemes: /w/, /l/, /l/. This view is based on distributional evidence, i.e. the fact that will /w1l/ is distinct from bill /btl/, from well /wel/, and from win /w1n/, and psycholinguistic evidence, e.g. slips of the tongue.

Within most current phonological theories, phonological representations consist of bundles of distinctive features. Although a priori, such features could be arbitrary, strong evidence exists for the view that features are characterized in phonetic terms. This set also serves to define systematic properties of linguistic sound systems, such as phonotactic
patterns and systematic alternations, which are affected by both articulatory and perceptual considerations. It is an important result that the same set of features serves to characterize these different sound properties and that this set of features is best defined in phonetic (articulatory and/or acoustic) terms. Although some debate exists, the set of features generally assumed for phonological representations is roughly the set proposed by Chomsky and Halle (1968, hereafter SPE), with some slight modifications. (See Keating 1988c for an overview of post-SPE features.)

A lexical item such as will /wɪl/ is represented abstractly as three distinct feature bundles, as schematized in Figure 1A. (The specific features used in this example are not central to the point being made.) The phonetic realization of such a form is not so clearly segmentable into three units, as shown by the spectrogram presented in Figure 1B.

A.  
\[ \begin{array}{ccc}
\text{w} & \text{l} \\
-\text{cons} & +\text{cons} \\
+\text{son} & +\text{son} \\
+\text{back} & -\text{back} \\
+\text{round} & +\text{high} \\
\ldots & -\text{tense} & \ldots \\
\end{array} \]

B.  
\[ \begin{array}{c}
\text{4000} \\
\text{3000} \\
\text{2000} \\
\text{1000} \\
\text{Hz} \\
\text{time} \\
\end{array} \]

Figure 1. A. A possible phonological representation of American English will /wɪl/; B. A wideband spectrogram of the same form, spoken by a female speaker of American English. F1 and F2 have been highlighted with a black line.

---

1 While Chomsky and Halle (1968) focus on the articulatory aspects of their feature system, they explicitly acknowledge the importance of acoustic and perceptual correlates: "We shall speak of the acoustical and perceptual correlates of a feature only occasionally, not because we regard these aspects as either less interesting or less important, but rather because such discussions would make this section... much too long." (p. 299) It is important to bear in mind the relevance of both articulation and perception for phonology.
This spectrogram is an acoustic representation, one phonetic representation of the form will. In the spectrogram, we observe that the formants (the characteristic overtones that acoustically identify the quality of vowels and sonorants) do not show clear breaking points between the /w/ and /u/ or the /u/ and /i/, rather they are continuously varying. It is difficult to identify a precise point at which one segment ends and another begins. Hockett's (1955) analogy of an assembly line of mashed Easter eggs graphically captures this point.

"Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other." (p. 210)

He then goes on to discuss an inspector who he likens to the hearer, whose job is not to put the eggs back together, but to identify what they were.

In a sense, this is the crux of the problem in relating phonological representations to phonetic ones. How can we characterize the relationship between the abstract discrete representation in Figure 1A, and the continuous one in Figure 1B? Note that we can ask this same question, going either from an abstract linguistic representation to a physical one or from a physical to an abstract linguistic one. I frame the question here in terms of the mapping from an abstract representation to a physical one.

In the remainder of this chapter, I turn first to a discussion of the nature of phonological rules and representations (§1), and phonetic ones (§2) before delving into the issue of distinguishing between phonetics and phonology (§3). I conclude with an outline of the remaining chapters (§4).

1. Phonological background and assumptions

Although strong evidence exists that segments are a psychologically relevant level of representation, it is also clear that segments consist of smaller units or features. Distinctive feature theory as developed by the Prague School of Linguistics is a central tenet of modern phonology (see notably Trubetzkoy 1939 and Jakobson, Fant, and Halle 1963 and also Anderson 1985 for discussion of these historical developments). In more recent work, it has also been argued that particular features may function independently, as separate tiers (Goldsmith 1976, Clements 1976) and that there is some kind of logical hierarchical structure among features internal to the segment, termed feature geometry (Clements 1985, Sagey 1986, McCarthy 1988). A well-formed underlying phonological representation involves feature specifications assumed to bear a universally determined structural relationship to one another.
Rules of the phonology consist of elemental processes of feature association: spreading (assimilation), schematized in (1a); delinking (deletion, or loss), shown in (1b), and deletion followed by feature fill-in (dissimilation), as in (1c).

(1)  

\[
\begin{array}{cccc}
\text{a.} & \cdot & \cdot & \cdot \\
\text{b.} & \cdot & \cdot & \cdot \\
\text{c.} & \cdot & \cdot & \cdot & \cdot \\
+\text{F} & +\text{F} & +\text{F} & +\text{F} & -\text{F} & +\text{F}
\end{array}
\]

Within this approach, explicit predictions are made about the simplicity of rules. Those involving only these elemental processes are the simplest, and are thus predicted to be very common cross-linguistically. Further, only individual features or those dominated by a single abstract node in the feature geometry are predicted to pattern together in these processes. The feature geometry is one part of a well-formed phonological structure, which includes both a segmental representation and a prosodic one (that is, syllabic structure, metrical structure, and so forth).

1.1 Phonological rules

The recent phonological literature includes much discussion of patterns of phonological rule behavior and how they interact.\(^2\) As summarized in (2), the following clustering of properties has been observed.

(2) sensitive to morphological structure \hspace{1em} not sensitive to morphological structure \\
    cyclic \hspace{1em} non-cyclic \\
    derived environments \hspace{1em} across-the-board \\
    exceptions \hspace{1em} no exceptions \\
    structure preserving \hspace{1em} introduce new sounds

A large class of phonological rules has been identified that are sensitive to morphological structure. Most commonly such rules might apply if a form is morphologically complex, but not if it is monomorphic. Such rules typically apply only in derived environments (unless they are structure building), apply cyclically, allow exceptions, and usually manipulate only those sounds which are part of the underlying inventory of the language. On the other hand, rules which are not sensitive to morphology typically apply across-the-board, are non-cyclic and apply in an exceptionless fashion. This clustering of properties is accounted for within the framework of Lexical Phonology (Kiparsky 1982, 1983 and

---

\(^2\) Some recent work leads to a rethinking of the status of phonological rules in general (see notably Goldsmith 1993, McCarthy and Prince 1993, and Prince and Smolensky 1993); in particular the derivational metaphors that have dominated the generative literature are questioned. However for ease of exposition, I maintain the canonical generative phonology view of rules in the discussion here.
Mohanan 1982), by positing two distinct sorts of rules. The rules that interact with the morphology are said to be *lexical* rules; they are interleaved with the morphology and have the chance to reapply after each morphological operation. These rules are inherently cyclic.³ Rules which apply in an across-the-board fashion, those which are not affected by morphological structure, are called *postlexical* and are argued to apply at the end of the grammar, after all of the lexical rules. This division gives a view of the grammar such as the one schematized in Figure 2.

![Diagram](image)

**Figure 2.** View of the grammar following Lexical Phonology.

The lexical rules correspond with what has been called the *deep phonology*. The postlexical rules are nearer the surface and include the class of allophonic rules. Mohanan (1982) argues that the last level of the lexical phonology has a special status. In particular, speakers appear to judge sameness and distinctiveness at this level, not at the underlying level or a more surface level. Speakers' awareness of a rule's effect may be taken as support for something being part of the lexical phonology; e.g. speakers of English are typically sensitive to alternations due to Trisyllabic Laxing (*opaque ~ opacity*) a lexical rule. In contrast, naive native speakers are usually insensitive to allophonic variation in their language; e.g. speakers of English typically judge the /p/ of *pit* [pʰɪt] and the /p/ of *spit* [spɪt] to be the same (Mohanan 1982, Kenstowicz 1994). These distinctions will be important below in §3 in the discussion of rule types.

1.2 Phonological representations

It is worth considering in some detail the nature of phonological representation, in both time and space. Within the phonology, representations are discrete and categorical. As

---

³ Some debate exists about the nature of cyclicity and the status of level ordering, another independent tenet of Lexical Phonology, see Kaisse and Shaw (1985) for an overview and Hargus and Kaisse (1993) for a range of recent perspectives on Lexical Phonology and also Halle and Vergnaud (1987) for a somewhat different point of view.
mentioned above, it is widely assumed that *segments* consist of a hierarchically structured set of distinctive features, organized into a tree. While most researchers agree on the basic outline of this hierarchical relationship, debate still exists over the specific set of features and their precise relationships. Additionally there are two somewhat different views of the nature of, or motivation for, this hierarchical structure. The view put forth by Clements (1985) and McCarthy (1988), among others, is that this structure is a logical structure, motivated by phonological evidence. The view espoused by Halle (1983) and Sagey (1986), among others, is that this phonological structure is articulatorily motivated, with the structure mirroring the vocal tract. The position taken here is that the feature geometry must necessarily be a logical phonological structure and the degree to which this structure mirrors the vocal tract is an empirical question. It is important to bear in mind that such abstract phonological representations are necessarily both articulatorily and perceptually motivated. I take the geometry proposed by McCarthy (1988), schematized in Figure 3, as a starting point. One of the many issues, as yet not fully resolved is the location of the feature Nasal in the geometry, a point to which we return in Chapter 5.

![Feature Geometry Diagram](image)

**Figure 3.** The feature geometry, following McCarthy (1988).

The space dimension is represented as the logical structure of the feature geometry with feature values for each feature. A binary feature may be positively specified (+), negatively specified (−), or unspecified (Ø). Thus, possible representations for the feature Nasal include only [+nasal], [-nasal], and [Ønasal] (assuming it is a binary feature), as exemplified in the English word *can* /kæn/, shown in (3).
Segments also include an abstract duration or timing unit, represented independently from the feature matrix itself. It is the timing or abstract duration of the segment that allows the segment to be related to higher level prosodic structure, which organizes segments into syllables, feet, and so forth. The possible timing contrasts available within the phonology are very limited. The inherent durations of different segment types do not play a role in the phonology. Only a very gross characterization of timing appears to be relevant. The relevant categories within a segment appear to be two units, one unit, and part of a unit, unspecified for specific duration or proportion. (Steriade 1990 makes a similar point.) More complex structures might arise through phonological derivation, e.g. Inouye’s (1989) analysis of flaps in English, as well as her reanalysis of medio-nasals in Kaingang (as discussed by Anderson 1976), both of which are argued to involve structures with three branches. But there is little evidence for structures with more than two branches in underlying representation.

These categories are captured within current phonological theory as timing units (either C’s and V’s (e.g. Clements and Keyser 1983) or x’s (e.g. Levin 1985), or as moras or weight units (e.g. Hyman 1985, McCarthy and Prince 1986). The mapping between segment quality, that is the features, and timing units represents phonologically relevant relationships, as exemplified in (4) (using x’s).

(4)  

<table>
<thead>
<tr>
<th>a. x</th>
<th>b. x x</th>
<th>c. x</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>/</td>
<td>α β</td>
</tr>
</tbody>
</table>

The representation in (4a) is that of most segments, one set of features for one timing unit. (4b) represents one set of features taking up two timing units, such as a geminate consonant or a long vowel. (4c) represents a contour segment, such as an affricate, a prenasalized stop, or a short diphthong. These can be viewed as two feature matrices attached to one timing unit, or a feature matrix with one branching feature (see Sagey 1986). It has been observed that, although geminates are consistently longer in duration than single segments, the relative timing of single segments and long segments varies from language to language. This does not invalidate the notion of timing units, but rather shows that they are abstract entities. On an abstract level only, geminates are twice as long. But the actual physical timing is a matter of phonetic realization which is, to some degree, language specific.
The representation in (4c) implies that two feature specifications share one timing unit. Each specification gets part of the whole. It does not say what the specific timing relationships are; this again would be language specific. It has been shown by Chan and Ren (1987) that prenasalized stops vary in their relative timing, yet I assume that this would not affect the phonological representations. A prediction that follows from this is that no language will distinguish between single unit contour segments of different proportions (e.g. [mb] vs. [m^b], where superscript denotes shorter duration). To my knowledge, no language makes a phonological contrast of this sort. In recent work, Steriade (1993, 1994) has proposed a rather different type of structure to account for contour segments, arguing that only [-continuant] consonants may show these contour effects, as they consist of both a closure and release phase. Thus a prenasalized stop consists of a closure phase which is nasal and a release which is oral. This provides a much more constrained theory of contour segments. The implications of such representations will be explored in Chapter 4.

1.3 Degree of phonological feature specification

An area of interest in the recent phonological literature has been the degree of specification in phonological representations. In other words, is it the case that for every segment each feature is specified for a plus or minus value? If so, the representation is a fully specified one. If not, if no specification is possible, there is said to be underspecification. In SPE, even though lexical representations were underspecified, with values provided through markedness and predictability of certain feature values from other feature specifications, it was explicitly stated that all features must be fully specified in the phonology. Thus all values were filled in at the level of underlying representation of the phonology. This led to a constrained view of feature representation, but a rather unsightful approach to certain phonological processes, such as harmony rules, which were represented as feature changing rules. Recently this constraint on full specification has been rejected (Kiparsky 1982, Pulleyblank 1983, and Archangeli 1984); it is now generally assumed that some degree of underspecification is used by the phonology. Yet there is much debate in the literature as to the degree of underspecification allowed, the way in which underspecification should be constrained or restricted, and how rules interact with representations which are not fully specified (cf. Kiparsky 1985, Archangeli & Pulleyblank 1986, Steriade 1987, Archangeli 1988, Christdas 1988, Clements 1988, 1993, Mester and Itô 1989, and Steriade 1995 for discussion of these issues).

Two types of approaches to underspecification have played a central role in these discussions: (a) the approach taken by Kiparsky (1985) and Archangeli and Pulleyblank
(1986), often termed Radical Underspecification, in which no reference is made underlyingly to either redundant or unmarked values; following this view, the unmarked value of a feature cannot be referred to underlyingly or in the phonology, until filled in by rule; (b) the type of approach taken by Steriade (1987) and Clements (1988), referred to here as Contrastive Underspecification, where in case of contrast, both feature values are specified in the underlying representation, but in the case of redundant specification, no value is present. See Archangeli (1988) and Mester and Itô (1989) for insightful discussion of the development of underspecification theory and comparison of these two types of theories, and Clements 1993 and Steriade 1995 for a recent perspective on some of these issues.

Relevant to the present discussion is the widely agreed upon assumption that redundant feature values are underlyingly unspecified; I assume that there is some degree of underspecification and that certain phonological patterns may be due, not to individual feature specification, but to more general principles. For the sake of explicitness, I follow a Contrastive Underspecification approach. While in some of the cases investigated here Contrastive Underspecification offers a more transparent interpretation of the observed patterns than Radical Underspecification does, nothing crucial hinges on this choice and the facts in the present study are compatible with both types of approaches.

As an example, consider the specification of the feature Nasal in both French and English, languages differing in their phonological use of the feature, as presented in (5).

(5) Phonological specification for Nasal

<table>
<thead>
<tr>
<th>Feature</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Stops</td>
<td>[-nasal]</td>
<td>[-nasal]</td>
</tr>
<tr>
<td>Nasal Consonants</td>
<td>[+nasal]</td>
<td>[+nasal]</td>
</tr>
<tr>
<td>Continuant Consonants</td>
<td>[Ønasal]</td>
<td>[Ønasal]</td>
</tr>
<tr>
<td>Vowels</td>
<td>oral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nasal</td>
<td>[Ønasal]</td>
</tr>
</tbody>
</table>

Both French and English make a contrast between nasal and oral stops, e.g. French net /net/ 'neat' ~ dette /det/ 'debt', English net /net/ ~ debt /dêt/. I therefore assume that, in both languages, the nasal consonants are specified as [+nasal] and the oral stops as [-nasal]. In both languages, the [+continuant] consonants show no contrast and I assume remain unspecified for the feature Nasal. French, however, has a contrast among the vowels as well, e.g. beau /bø/ 'beautiful' ~ bon /bɔn/ 'good'; I conclude that the nasal
vowels are specified as [+nasal] and the oral vowels as [-nasal]. English shows no such contrast among the vowels; thus, in the absence of evidence to the contrary, I assume that vowels in English remain unspecified for Nasal throughout the phonology.

An alternative view of feature specification is that some or all features are privative, that is, they are present or absent (see Lombardi 1991 and Steriade 1995). In particular, some researchers (e.g. Itô and Mester 1989, Steriade 1993, 1995, and Trigo 1993) have argued that the feature Nasal is privative. This claim has important implications for phonological representation, as well as for phonetic implementation; the phonological and phonetic consequences of this view are discussed in Chapter 3. For the moment, I maintain the view that Nasal is a binary feature.

In sum, within the phonology, phonological features are hierarchically arranged; feature specifications are represented as only plus, minus, and unspecified; and timing is represented in abstract units: x, two x's, or part of x. At the end of a phonological derivation, the output is still a discrete, static one, with logical structure, relative ordering relationships between segments (or feature specifications), but no actual (concrete) durations, or quantities along the physical dimensions corresponding to features. Rather it is the job of the phonetics to assign real durations and actual quantities.

2. Phonetic rules and representations

In this section, I consider the nature of phonetic representation (§2.1), phonetic implementation (§2.2), and the role of underspecification in the phonetics (§2.3).

2.1 Phonetic representations

In SPE, it is assumed that there is a phonetic level of feature representation, the phonetic transcription, at which point binary values are translated into a small number of discrete categories. The phonetic transcription is described as follows:

The phonetic transcription can therefore be taken to be a two-dimensional matrix in which the columns stand for consecutive units and the rows stand for different features. At this level of representation each feature is to be thought of as a scale. A particular entry in the matrix, then, indicates the position of the unit in question on the given scale. The total set of features is identical with the set of phonetic properties that can in principle be controlled in speech; they represent the phonetic capabilities of man and, we would assume, are therefore the same for all languages. (pp. 294-5.)

---

4 It has been widely debated in the literature whether the nasal vowels of French are underlying or derived (cf. Schane 1968, Dell 1970, Tranel 1981, Prunet 1986, among others). What is important here is that, in either case, a contrast exists at the level of representation relevant to the present discussion; since, if the nasal vowels are derived, it is in the deep phonology.
This, then, is a linguistically relevant, discrete, phonetic representation. Everything beyond the phonetic transcription is assumed to be universal. Anderson (1974) offers a concrete instantiation of such a phonetic representation. As an example, he discusses three possible phonetic values for the feature Nasal in the Breton dialect of Plougastel – [0nasal], [.3nasal] and [.7nasal], on a scale of zero to one – to describe oral vowels in an oral context, oral vowels next to a nasal, and distinctively nasal vowels, respectively. Such an approach, in effect, codifies the phonetic values that exist at the level of the phonetic representation (which Anderson notes are only significant in their relative relationship to each other).

This SPE-type view gives a formal status to the phonetic representation – a still discrete, but detailed, representation. It is not obvious that such a level of representation is necessarily warranted. (See Pierrehumbert 1990 for discussion of this point.) It might also be the case that the status of the phonetic representation is different for different features. For any particular feature, we can ask whether there are a certain number of points along the scale that are linguistically relevant. This is an empirical question. Keating (1984) argues that, for the feature Voice, there is evidence of an intermediate formal level of phonetic representation in terms of categories distinct from the phonological categories. She offers a detailed and formal proposal for such a level of representation for Voice. She argues that the phonological feature Voice is represented at an intermediate stage as three phonetic categories: [voiced], [voiceless unaspirated], [voiceless aspirated]. Different languages will map the two binary values [+voice] and [-voice] differently into the phonetic mapping. This accounts for the kind of similar phonological processes observed across languages with respect to Voice, even though what is voiced in one language may count as voiceless in another. (However, in more recent work, Keating 1990b no longer ascribes the same formal status to an intermediate level of representation.) But even if this is a correct view for Voice, it does not follow that such a level of representation necessarily exists for other features.

Much of the discussion of the relationship between phonology and phonetics has focused on the nature of the phonetic representation (Chomsky and Halle 1968, Anderson 1974, Ladefoged 1977, van Reenen 1982, Mohanan 1986). This emphasis on representation rather than process results, I believe, from the formal status attributed to the phonetic transcription as defined by Chomsky and Halle. But such a level of phonetic representation may or may not be relevant to implementation. Following the view of implementation proposed by Pierrehumbert (1980) and Pierrehumbert and Beckman
(1988), the phonetic representation is an extension of the phonological representation. The phonologically discrete units get increasingly blurred through the implementation of different parameters. What is taken as the phonetic representation is a result of the phonetic implementation; it may or may not be the case that such representations should be formally codified. Thus for example, the [.3nasal] level observed on vowels next to a nasal consonant in Breton might arise from the transition between an oral segment and a nasal segment. The [.3nasal], although descriptively accurate, may not be directly relevant to our understanding of the mapping. It might be the case that such a value is the result not of a phonetic target, but of a transition between two more extreme values. Values such as [.3nasal] may arise as a result of change over time. If this is the case, we can still describe the phonetic event in these terms, but do not necessarily want to ascribe a formal status to this description. In the analysis in this study, although phonetic representations are presented, I focus primarily on the process of implementation, in the belief that the phonetic representation of the feature Nasal follows from its implementation.

2.2 Phonetic implementation

The formal properties of phonological rules have been a topic of ongoing research in generative phonology; the formal properties of phonetic implementation rules are much less studied and much less clearly understood. The nature of phonetic implementation is a complicated issue for a number of reasons.

First, phonetics, representing the physical properties of speech, includes a number of possible representations. Denes & Pinson's (1993) characterization of the speech chain, presented in Figure 4, illustrates this multifaceted nature of phonetics.

![Figure 4. The speech chain (adapted from Denes & Pinson 1993, p. 5).](image)

At a minimum, we can talk about the articulatory event, the acoustic event, and the auditory event, as being part of the phonetics. We might further distinguish between the higher
level articulatory gesture and lower level motor events and between auditory and perceptual events. The implicit question that shapes researchers' assumptions about the nature of phonetics is which facets of phonetic activity are under speaker or hearer control. When a speaker produces a labial gesture, as in the sound /b/, are the individual muscles commanded to fire, or are the individual muscle firings a result of a more global command to make a labial gesture? Research shows that higher level aspects of such events are under linguistic control, as are some of the lower level ones. I focus here on higher level events; while in a sense this is an arbitrary decision, it seems reasonable to assume that higher level events are more likely to be relevant to linguistic structure.

Second, even if we could fully resolve the first issue, we would still be faced with the issue of how should we best describe or model phonetic information. Consider the schematic example of an articulatory gesture, a movement of the velum, presented in Figure 5A. Here the velum begins in a high position at the beginning of the vowel (i), starts to lower part-way through the vowel, reaching a low position by the beginning of the nasal consonant (ii) which is maintained throughout the duration of the nasal consonant (during the oral closure of the consonant), then it starts to rise again during the following vowel (iii).

![Diagram](image)

**Figure 5.** A schematic velic gesture in /ana/, A. movement of the velum, B. static model of this movement, C. dynamic model of this movement.

We might argue that this phonetic event is best characterized by the high and low positions of the velum, with transitions connecting the targets. This view suggests that the
fundamental information is provided by the maxima and minima, while the transitions are, in effect, a by-product. I characterize this as a static view of the phonetic representation, since it is the endpoints, not the movement, which are taken to be primary. Such an approach is basically a target-interpolation model: maxima and minima are targets, connected through interpolation. Target-interpolation models of phonetic implementation have been discussed in the literature in two different contexts. First there are the rule-based systems of speech synthesis (e.g. Holmes, Mattingly, and Shearne 1964; Mermelstein 1973; Hertz 1982; Allen, Hunnicutt, and Klatt 1987, among others; see Klatt 1987 for a review). More recently target-interpolation models have been applied to the problem of the mapping of phonology to phonetics (Pierrehumbert 1980, Pierrehumbert and Beckman 1988, Keating 1985a). (These two domains of investigation are not necessarily distinct as shown by the linguistically based speech synthesis work of Hertz 1991 within her Delta system.) Within a target-interpolation model, we might model the event in Figure 5A as shown in Figure 5B.

Alternatively we might take the movement to be the primary event and model the velic gesture as a gesture or an oscillation, as schematized in Figure 5C. I characterize this sort of view as a dynamic view of the phonetic representation: the movement is primary and the endpoints are a byproduct of the timing and amplitude of the oscillation. This is basically the view espoused by Browman and Goldstein in their recent work (e.g. Browman and Goldstein 1992).\(^5\)

A third issue that influences the notion of phonetic representation is the theoretical position we take vis-à-vis the relationship between phonetics and phonology. As argued below in §3.1, evidence supports the conclusion that the phonetics is not just the automatic, universal realization of phonological patterns. Rather the phonetics is, at least in part, language specific and therefore under linguistic control. Phonetic patterns are quite systematic and therefore assumed to be rule governed, the basic tenet of the growing field of generative phonetics. Furthermore, our notion of the phonology and the representation of the output of the phonology will influence what we take to be a simpler or more straightforward model of the phonetics.

Models of phonetic implementation can be characterized in terms of the form of the phonological output, discrete or continuous, categorical or gradient, static or dynamic; the nature of the phonetics, static or dynamic; and the role of timing within the phonetics, as intrinsic or extrinsic to phonetic events. Most models involve an explicit point of

\(^5\) For Browman and Goldstein the phonology is also characterized in gestural terms.
interpretation or translation, though some provide a more seamless view of the connection between phonology and phonetics. (See Pierrehumbert 1990 for discussion of different types of models.)

As discussed above, the view taken here is that phonological representations are discrete, categorical, and static. One possible view of phonetic implementation would be to add concrete durations to autosegmental associations. A model consistent with this view of phonology is the Look Ahead model of Henke (see Keating 1985a for a discussion of this model and its history). As Keating discusses, in this model a feature value changes as soon as it is able to. One can think of this as on-off switches, with the principle being to change to the next value as soon as possible, predicting extensive anticipatory, to no carry over coarticulation. As has been argued elsewhere in the literature, such a model is not adequate to account for the facts of nasalization (see Kent, Carney, and Severeid 1974 and Benguerel, Hirose, Sawashima, and Ushijima 1977, among others). However, more generally, such an approach alone would not be sufficient, as autosegmental representations are inherently qualitative; rather a quantitative evaluation of such abstract representations is required.

A richer phonetic interpretation of phonological information is needed. The job of phonetic rules is nicely characterized by Pierrehumbert and Beckman (1988, pp. 4-5):

The phonetic rules are like phonological rules in that they seek to describe complex regularities in sound structure through the interaction of a few general principles. They differ from phonological rules in the representations that they manipulate. They take as input phonological representations, but their output consists of quantitative functions, representing facts about articulations or sounds.

The class of target-interpolation models allows quantitative evaluation of such representations with few assumptions. Target-interpolation models are typically assumed to assign static targets to phonological feature values. Recent work suggests that targets may have inherent duration (Pierrehumbert and Beckman 1988, Huffman 1989) or more than one target per phonological segment may be required (Huffman 1989, Laniran 1992). Additionally, nothing excludes the possibility that discrete phonological representations could be mapped to targets which are basically gestural in nature (a point to which we return in Chapter 6); such a view is argued for by Zsiga (1993).

An outstanding example of phonetic implementation within a target-interpolation model is Pierrehumbert's (1980) dissertation, in which she models intonation in English going from an abstract (sparse) phonological representation to observed physical output. This approach has been developed in further work (notably Beckman and Pierrehumbert 1986
and Pierrehumbert and Beckman 1988). Such an analysis relies heavily on phonetic implementation rules. Pierrehumbert develops the traditional distinction in implementation between evaluation and interpolation rules. The former rules evaluate feature values, translating them into phonetic targets located in both time and space. The latter rules connect up targets. This approach offers a possible model for examining segmental processes, as proposed by Keating (1985a).

Within a target-interpolation model, feature values leaving the phonology are translated into phonetic targets, such that a plus value translates to relatively more of the physical value that implements a particular feature than a minus value; these targets are then hooked up through interpolation, as illustrated in Figure 6:

<table>
<thead>
<tr>
<th>Phonological Output:</th>
<th>+F</th>
<th>-F</th>
<th>+F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic Targets:</td>
<td>high</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td></td>
<td>low</td>
<td></td>
<td>•</td>
</tr>
<tr>
<td>Phonetic Interpolation:</td>
<td>high</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.** The target-interpolation model of phonetic implementation.

In this example, we see a sequence of segments with values for some feature F. The vertical scale is the physical dimension. The target evaluation rules assign phonetic targets based on phonological specification, with the +F specifications receiving a high target, and -F a low target, for the particular phonetic parameter (or parameters) that corresponds to the phonological feature. I assume for the moment that targets consist of single points located in the middle of the duration of a segment, with linear interpolation between the targets.

This model raises issues about the nature of targets, their assignment, and interpolation. We return to the nature of targets in Chapter 3. Many patterns of interpolation appear to follow from the linear connection of targets. This is the case for some of the intonational patterns described by Pierrehumbert or the case of jaw lowering (see Keating 1987). But consideration of a broader range of data makes it clear that a restrictive theory of interpolation which allows only for straight lines is not empirically adequate. Some of the patterns presented by Pierrehumbert, as well as the case of tongue backing in Arabic discussed by Keating (1987) show greater complexity. Pierrehumbert (1980) proposes a model of interpolation, including three types of curves to account for the patterns that she
presents. Choi (1992) also finds both linear and non-linear patterns of interpolation in his phonetic study of frontness vs. backness in Marshallese. See Choi (1992) for a comprehensive discussion of the issues relating to the nature of interpolation. In the course of the present study, I assume linear interpolation, although this is something of an oversimplification.

This view of phonological and phonetic structure and the nature of the mapping between them can be contrasted with the kind of approach taken by Browman and Goldstein (1992 and work cited therein), within the model of Articulatory Phonology. Browman and Goldstein argue that both phonology and phonetics can be represented in terms of gestures, articulatory events with inherent timing, indicating local constrictions in the vocal tract. Following this view, the relationship between phonology and phonetics is essentially seamless, as the same structures are manipulated by both the phonological and phonetic components. The categorical nature of lexical representations is accounted for by imposing constraints on gestural overlap and timing. This model could be characterized as one which is inherently quantitative, with the more qualitative aspects of the phonology accounted for by the imposition of constraints. (Such constraints have yet to be explicitly laid out. Until such constraints are made more explicit, the categorical representations of the phonology appear ad hoc.) This view takes articulation to be primary at both the phonetic and phonological levels.

Both approaches have been used quite successfully to model phonetic events and in each case a number of developments and amplifications have been made. While Articulatory Phonology is attractive in a number of respects, I follow Steriade (1990), Clements (1992), and others, in arguing that an autosegmental, feature geometric approach offers a more predictive, constrained view of phonology. I take such a view as a starting point for the mapping from phonology to phonetics accounted for by a target-interpolation model. My primary goal here is to investigate how explicit phonological representations can be interpreted by the phonetics. While I will argue that phonetic patterns of nasalization are insightfully captured within a target-interpolation model, this does not rule out the possibility that features are not all uniform in this respect. Features may differ as to whether they are best modeled in static or dynamic terms. It should be noted that the phonetic representations that I develop are still quite abstract in a number of respects. In particular, I assume the existence of an extrinsic timing model, not developed here.
2.3 Phonetic underspecification

An issue basic to the discussion of phonetic implementation is whether there is necessarily full specification of features at the output of the phonology. Full specification leaving the phonology is of course assumed in any theory which does not allow phonological underspecification. But even under certain views of phonological underspecification, it is often assumed that there is full specification at the end of the phonological derivation, due to default fill-in rules (e.g. Archangeli and Pulleyblank 1986). In contrast, in order to account for patterns of pitch, as measured from F0 traces, Pierrehumbert (1980) assumes a sparse phonological output. Not every syllable which is potentially the bearer of a pitch accent receives such an accent. In effect then, the phonological output is not fully specified along the dimension that is implemented as intonation. Other examples include the realization of the vowel-like quality of intervocalic [h], which Keating (1988b) argues is phonologically unspecified for vowel quality, the vowels of Marshallese, argued by Choi (1992) to be unspecified for backness, and the nasalization of continuant consonants in Sundanese, as discussed below in Chapter 5 and by Cohn (1993a). In order to account for these observations, it is necessary to abandon the assumption that there is full specification leaving the phonology.

A target-interpolation model taken together with the possibility that there is not full feature specification leaving the phonology leads to the following schematic view of implementation: a plus value of a feature should translate to a relatively high level; a minus value to a low level, and an unspecified value would be expected to be determined by phonetic context. The rules of the phonetic implementation evaluate the targets and connect them through interpolation. Three examples are given in Figure 7.

<table>
<thead>
<tr>
<th></th>
<th>a.</th>
<th>b.</th>
<th>c.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonological Output:</td>
<td>+F -F -F</td>
<td>+F ØF -F</td>
<td>+F ØF +F</td>
</tr>
<tr>
<td>Phonetic Targets:</td>
<td>high ⋅</td>
<td>high ⋅</td>
<td>high ⋅</td>
</tr>
<tr>
<td></td>
<td>low ⋅</td>
<td>low ⋅</td>
<td>⋅</td>
</tr>
<tr>
<td>Phonetic Interpolation:</td>
<td>high ⋅</td>
<td>high ⋅</td>
<td>high ⋅- - - - -</td>
</tr>
<tr>
<td></td>
<td>low - - - - -</td>
<td>low - - - - -</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7.** Schematic examples of three patterns of phonetic interpolation allowing for phonetic underspecification.

In Figure 7A the form is fully specified for the feature F leaving the phonology. Targets are assigned, along some scale for the particular feature. These targets are then hooked up
through interpolation, showing a fairly rapid transition between the neighboring high and low targets. In Figure 7B, the first segment is [+F], the second segment is unspecified leaving the phonology and the final segment is [−F]. Only the first and last segments receive phonetic targets. There is interpolation between the targets and the middle segment receives a transitional amount of the scale for the feature F from the phonetic context. In Figure 7C, both the first and last segments are specified as [+F] and the middle segment is again unspecified. Targets are assigned and in this case, there is interpolation straight through the middle segment. The unspecified segment receives a high value of the relevant phonetic parameter due to the phonetic context, giving the (erroneous) impression that it had a relatively high target for Feature F, when in fact it has no target of its own.

Let us look more closely at the vowel-like qualities of intervocalic [h], as discussed by Keating (1988b), as an example. It has commonly been assumed that [h] gets its vowel-like qualities from a phonological assimilation rule, where [h] assimilates to the vowel features of the following vowel (e.g. Ladefoged 1993). Were this the case, we would expect that [h] would have similar formant structure (the characteristic acoustic pattern of a vowel or vowel-like segment, indirectly related to the phonological vowel quality features) throughout most of its duration as that of the following vowel. Keating studied formant patterns of [h] and concluded that the formant structure was not due to a phonological rule, but due rather to phonetic context. Consider relevant spectrographic examples, [iha] and [aha] presented in Figure 8A & B, respectively.

A. [iha]  

B. [aha]

---

**Figure 8.** Wideband spectrograms of a female speaker of American English saying A. [iha] and B. [aha].
As shown in Figure 8A, [i] is characterized by having a low F1 (centered here at about 300 Hz) and a high F2 (centered at about 2700 Hz); [a] is characterized by a high F1 (centered at about 1000 Hz) and a low F2 (centered at about 1500 Hz). The formant structure is constant throughout the duration of the vowel. Were it the case that [h] gets its formant structure from a phonological rule of assimilation from the following vowel, we would expect the [h] to have formant values like the following [a] for most of its duration. This is not the case; rather, the F1 and F2 values are transitional throughout the full duration of the [h]. This can be explained in a straightforward fashion, if it is assumed that [h] is unspecified for vowel quality leaving the phonology; it therefore gets no phonetic targets of its own and its vowel-like qualities are from the phonetic context, due to interpolation. Note that this conclusion is partially dependent on the segmentation of the [h] and the adjacent vowels. It is difficult to determine precise segmentation of this sequence, since the [h] is voiced throughout. The realization of F1 and F2 for [h], then, looks very much like the pattern of phonetic implementation schematized above in Figure 7B.

Consider now the spectrogram in Figure 8B. Here again the [a]'s both preceding and following the [h] are characterized by high F1's and low F2's. The formant structure of [h] is also characterized by a high F1 and low F2. Had we not considered the evidence in Figure 8A, we might (erroneously) assume that [h] has its own targets, high F1 and low F2, but based on that evidence, it is much simpler to assume that here too [h] has no vowel feature specifications of its own, therefore no phonetic targets, and that it too gets its vowel-like quality from the phonetic context, through interpolation between the neighboring targets. This then is like the pattern of phonetic implementation schematized in Figure 7C. It is important to note that such cases can easily be misinterpreted as all three segments, including the intermediate segment, having the same value specified. Comparison with forms where a transition might be expected throughout are crucial for determining such cases of underspecification leaving the phonology.

Keating’s account of intervocalic [h] offers strong support for the view that there is not necessarily full feature specification leaving the phonology. The facts are amenable to an analysis in terms of the target-interpolation model (but could clearly be accounted for with other models of phonetic implementation as well). The assumption that there is not full specification leaving the phonology has important consequences. (1) It is not necessarily the case that there are rules which assign phonological feature default values at the end of the phonology. (2) Since a well-formed phonological output might not be fully specified along all dimensions, it is an empirical issue to determine whether a particular
representation is fully specified and whether or not default phonological values have been assigned.

A particular view of phonetic underspecification is Keating's (1988a, 1990a) Window Model. Keating proposes an explicit model of interpolation and phonetic spreading which assumes that rather than having a single target in the vertical dimension (space) for each specified value, there are ranges within which interpolation must stay. Thus a segment with a very narrow window would appear to be one with a specific target and one with a wide window would appear to be underspecified. This approach offers an account for certain kinds of variability, as well as a means of constraining possible types of interpolation.

In the present study, I take a target-interpolation model as the starting point to account for the phonetic implementation of the feature Nasal. I argue that such a model needs to be modified and extended to account for the observed facts. The characterization of targets in terms of both time and space is extended from the concept of a single point located in the horizontal and vertical dimensions. I argue that targets have inherent duration and a specific account of such duration will be proposed. Further, I argue that windows, or ranges of values, play a role in implementation.

3. Phonology vs. phonetics

With this brief sketch of rules and representations, both within the phonetics and phonology, we can now turn to the question of the relationship between phonetics and phonology.

3.1 Background on the relationship between phonetics and phonology

In order to study the relationship between phonology and phonetics and the nature of phonetic implementation, we have to be able to distinguish between them. Within the framework of generative phonology, phonological representations and derivations have been assumed to be a central part of the grammar, but until recently phonetics was often assumed to fall outside of the domain of the linguistic grammar. A widely held view, following SPE, was that phonetic implementation was universal. This was discussed explicitly in terms of coarticulation. Coarticulation, the overlapping or blending of neighboring segments, was assumed to follow from general non-linguistic principles, such as the physiology of the vocal tract. Stated more generally, phonetic implementation or the physical realization of the abstract patterns represented by the phonology was assumed to be mechanical. As a consequence, a phonological output was assumed to have a unique physical realization. For example, in English, vowels are longer preceding a voiced stop
than a voiceless one: the vowel in *bad* /bæd/ is longer than in *bat* /bæt/. This was not taken as evidence of a phonological length difference in English, but, rather, was assumed to follow from mechanical implementation of the voicing contrast. It was also assumed that the same differences occurred cross-linguistically. Under this view, the distinction between phonetics and phonology appeared clear-cut. Phonology involved language specific rules, whereas phonetics was the universal, mechanical realization of the phonology. This view, where only the phonological rules were part of the linguistic grammar, is schematized in Figure 9.

\[
\begin{array}{c}
\text{phonological rules} \\
\downarrow \\
\text{universal phonetic implementation rules} \\
\downarrow \\
\text{physical output}
\end{array}
\]

*Figure 9.* Traditional view of the linguistic grammar (SPE).

Since the phonetics was *automatic*, there was no notion of there being a phonetic rule system as such. As the mapping between phonological rules and phonetic implementation was thought to be universal, little attention was paid to the phonetic implementation of phonological representations from a linguistic point of view. Such questions were the purview of phonetics alone and not thought to be of interest to the phonologist.

More recently, this view has changed. Extensive evidence has been provided of the language specific nature of phonetics. Many phonetic processes that were assumed to be mechanical, on closer inspection, turn out to show significant differences between languages. In such cases, a purely mechanical explanation is no longer tenable. One such example, widely discussed in the literature, is the observation about vowel length mentioned above (see Chen 1970, Fromkin 1976, Anderson 1981 and Keating 1985b, among others). Physiological explanations have been offered, but these do not account for the observed cross-language differences, in which some languages suppress these differences and others enhance it.

Another such example is velum movement. It has been argued that contextual nasalization of vowels in English is due to the *sluggishness* of the velum. But studies have shown that the velum can move much more quickly than the degree of coarticulation
observed in English would suggest (see Ohala 1975 for discussion). Furthermore, cross-linguistic work has shown significant language differences with respect to the timing of contextual nasalization (Clumeck 1976). At best, we can say that these effects are physiologically motivated tendencies (see Anderson 1981).

Phonetic differences can even occur between closely related dialects. Such is the case with the realization of the time course of nasal vowels in European French and Canadian French, as shown in Figure 10.

![Time course of nasal vowels](Image)

**Figure 10.** Time course of nasal vowels of European French & Canadian French, after van Reenen 1982 p. 74, showing a graphic representation of averages of 12 European French and 31 Canadian French nasal vowels in nonnasal environments, based on measurements from tracings in Brichler-Labaeye (1970) and Charbonneau (1971). Time in centiseconds on the x-axis and the value for N% on the y-axis, 

$$N\% = \frac{\text{nose coupling}}{\text{mouth coupling} + \text{nose coupling}} \times 100,$$

where N = Nose Coupling, the opening in mm$^2$ of the nasal port in a cross-section perpendicular to the airstream at the point of greatest constriction between the velum and the pharyngeal wall.

In Figure 10, we see the time course of velum movement during a nasal vowel in European French compared with Canadian French (e.g. [n] as in *bon* [bɔ̃] 'good' (m.)). This figure (adapted from van Reenen 1982) shows averages of nasal opening vs. time for nasal vowels in non-nasal environments, measured from x-ray tracings for European French and Canadian French from Brichler-Labaeye (1970) and Charbonneau (1971) respectively. We see that there is a different time course, with the onset of velum lowering (onset of nasalization) occurring much sooner in European French than Canadian French.
The difference is not a mechanical one; there is no reason to assume that the physiology of Canadian speakers of French is different from that of European speakers. Yet it cannot be accounted for in an obvious way by differences in the phonological system, since the value of nasal vowels in the two systems is the same. Finally, these language specific differences are systematic ones, and thus appear to be rule governed.

Based on such evidence, it has been convincingly argued that at least some phonetic rules are indeed part of the linguistic grammar and are not to be relegated to the *universal phonetic component*, leading to a view such as the one schematized in Figure 11, where both the phonological and the language specific phonetic rules are part of the grammar.

![Diagram](image)

Figure 11. Current view of the linguistic grammar.

This more complex view of phonetics leads us to seek a principled characterization of the difference between phonological and phonetic rules which follows from the mechanisms involved. Language-specific differences in phonetic realization identify aspects of the phonetics which are not universal and therefore not mechanical. Such differences may serve as a device for identifying linguistically relevant phonetics. Yet how can these cases be systematically distinguished from the phonology?

One possible view of the relationship is that language specific phonetic rules are a subset of phonological ones and should be accounted for with phonological rule mechanisms and representations. Another view is that phonological rules and language specific phonetic rules differ crucially in that what they manipulate is different: phonological rules manipulate discrete categorical representations, whereas phonetic ones manipulate quantitative ones. This fundamental difference necessitates that different formal mechanisms are involved. It is this latter position that will be taken here. We should
therefore seek a principled characterization of the difference between phonological and phonetic rules which follows from these different properties.

3.2 Distinguishing between phonetics and phonology

The conclusion that there is indeed a class of language specific phonetic rules leads us to seek a principled way of characterizing such rules and distinguishing them from phonological rules. As discussed in the preceding subsection, language specific differences in phonetic realization identify aspects of the phonetics that are not universal and therefore not mechanical. Such differences may serve as a device for identifying linguistically relevant phonetics. On the other hand, if it is claimed that such rules are distinct from the phonology, we need a systematic way of making this distinction.

Keating (1990a, p. 452) argues that phonological rules are expected to affect most of a segment in a significant way, whereas phonetic rules can have more gradient effects. "Phonetic rules can thus, for example, assign a segment only a slight amount of some property, or assign an amount that changes over time during the segment". Keating (1987, 1988a) proposes the following clustering of properties:

(6) \begin{align*}
\text{Phonological rules} & \quad \text{Phonetic rules} \\
categorical & \quad \text{gradient/quantitative} \\
discrete & \quad \text{continuous in time and space} \\
& \quad \text{segments may vary in quality continuously} \\
static & \quad \text{part of a segment affected} \\
effects & \\
full & \end{align*}

By static, we mean output that does not change noticeably during the duration of time assumed to be associated with a particular segment or sequence of segments; such effects might be characterized as a plateau (along the space dimension). Gradient effects might be of different types; of particular interest to us here are cases of change in space over time, resulting in a cline-like effect. That static, categorical outputs are the result of phonology and gradient outputs from transitions between segments are the result of phonetics is not surprising. More interesting is the fact that phonetic rules may affect a full segment or more in a continuously varying way. This result follows directly from certain assumptions about the nature of phonetic implementation, most importantly the observation that there is not necessarily full specification of features leaving the phonology. The distinction is not always an obvious one. Phonetic output may at times appear to be categorical. Recall our discussion of the vowel-like qualities of [h], where we saw cline-like effects throughout the [h] when the preceding and following vowel were different, but static effects when the
adjacent vowels were the same. We also cannot exclude the possibility that a phonological
distinction may appear gradient; as when, for example, a target has only a brief point-like
duration. Such a clustering of properties can be, at best, taken as guidelines.

Of particular interest are cases which appear to involve gradience in the phonology or
where other aspects of what has been described as phonological rule application look phonetic. Keating (1988a, 1990a) discusses a number of cases of assimilation involving
gradient rule application that she argues are phonetic and not, as previously assumed,
phonological. As discussed above, she suggests that the vowel-like qualities of [h] are a
result of phonetic implementation and not phonological assimilation. Additional cases
where Keating argues a phonetic rule (applying gradiently) has been misanalyzed as being
phonological (applying categorically) include vowel allophony in Russian and the
spreading of emphasis from the emphatic consonants in some dialects of Arabic.
Pierrehumbert and Beckman (1988) argue that some aspects of pitch accent in Japanese
result from phonetic implementation, not phonological tone spreading. It is important to
note that such observations depend crucially on instrumental phonetic data as the difference
between gradient and categorical rule application cannot necessarily be determined through
impressionistic observation. Careful listening alone is not sufficient to determine the status
of a rule, since we as hearers impose categories (see Repp 1984 and work cited therein),
even on gradient phenomena.

There are several interesting cases discussed by Kiparsky (1985), in which gradience is
involved. He discusses cases where the same rule seems to apply at different points in the
grammar, once categorically, once gradiently. He argues that this is a result of lexical and
postlexical application of the same rule. I believe that these cases may be amenable to
analysis in terms of phonetic interpolation through an unspecified span. A case discussed
by Kiparsky, of particular interest here, is the case of Nasal Harmony in Guaraní. It has
been widely observed that there are long distance spreading effects of nasalization in
Guaraní. Noteworthy is the fact that some of these effects appear to be categorical,
wheras some are gradient. Kiparsky proposes that these effects are the result of the lexical
and postlexical application of the same rule. It is nasalization of the class of unstressed
continuants which results in gradient effects. This is a class of segments which
characteristically does not contrast for the feature Nasal cross-linguistically. There does not
appear to be evidence of the nasal specification of these segments playing a role in the
phonology. It seems plausible that this class of segments remains unspecified for the
feature Nasal and that these segments receive quantitative levels of nasalization through
phonetic interpolation. An analysis along these lines is proposed for the case of Sundanese nasalization, as discussed in Chapter 5.

Identification of gradient behavior is taken as evidence of phonetic, not phonological, rule application precisely because the mechanisms of phonology and phonetics are distinct: Phonological rules manipulate discrete, timeless segments, whereas phonetic rules manipulate variables which are continuous in time and space. Gradient or cline-like behavior is a result of phonetic implementation, while categorical output or plateaus result from phonological rule application. (See Pierrehumbert 1990 for an interesting discussion of the mechanisms involved in phonology and phonetics respectively.)

Note, I distinguish here between gradient and variable. I take gradient to refer to change in space over time, while variable refers to applicability of a process independent of phonological context. A process may be variable in two senses, first it may or may not apply (this is often accounted for by saying that a rule is optional), second it may apply to varying degrees, e.g. deletion can be seen as the endpoint along a continuum of shortening. Some researchers assume that gradience and variability go hand in hand. Although an interesting hypothesis (not yet explicitly tested to my knowledge), it does not follow from either definition. Optional rule application seems equally consistent with phonological or phonetic rule application, while variability in the second sense suggests phonetic implementation. I return to the issue of variability in Chapter 4.

Taking a somewhat different approach to the question of gradient vs. categorical behavior, Browman and Goldstein (1992, p. 171) observe that "processes occurring during the act of talking will cause gradient changes that can ultimately be perceived as a categorically different gestural structure". Clements (1992, p. 12) notes that Browman and Goldstein's view suggests that "speech is produced in a gradient fashion, but perceived (and thus represented) categorically". This interpretation offers another explanation of why phonetics is inherently gradient, while mental representations (the phonology) are categorical.

The distinction between gradient and categorical behavior is intuitively attractive, yet what sorts of independent evidence might be adduced to support the claim that such behavior correlates with phonetic vs. phonological rule application, respectively?

As discussed above, within the framework of Lexical Phonology (Kiparsky 1982), a clustering of properties is used to determine the status of rules, lexical vs. postlexical. Although the distinction between lexical and postlexical does not correlate directly with phonological vs. phonetic, use of similar criteria should be helpful in this case as well. If a rule is shown to have characteristic behavior of a lexical rule (applying cyclically, applying
only in derived environments, applying within certain grammatical domains, respecting Structure Preservation, or ordered before such a rule), we can conclude that it is indeed phonological. Such criterion are used to corroborate evidence from nasal airflow data in the following chapters.

Some researchers have equated phonetic rules with postlexical rules. For example, Liberman and Pierrehumbert (1984) suggest that, except for phrasal rules, postlexical rules are phonetic implementation rules. Kiparsky (1985, p. 94) argues that this position is too strong. He observes that there are two types of postlexical rules, those with a categorical output (as a result of feature changing or default fill-in) and those which are essentially phonetic in nature, with gradient, variable output. (Kaisse's 1990 distinction between two types of postlexical rules, P1 and P2 can be seen to parallel Kiparsky's characterization.)

Although the distinction is mainly terminological, I take the position that postlexical phonological rules and phonetic implementation rules are distinct (rather than there being two types of postlexical rules), since the rule mechanisms involved are distinct. A logical argument in support of this position can be made from the process of historical change. One common type of historical change is that a phonetic characteristic (typically based on some sort of universal tendency) is exaggerated until it is perceived as distinct from the conditioning factor. This is what Hyman (1976) has termed phonologization. Consider, for example, the case of the shortening and eventual deletion of a segment. A segment might have a tendency to shorten in certain prosodic positions or at faster rates of speech. The most extreme version of shortening is deletion. As a phonetic process we would expect the results of this to be gradient and perhaps also variable; even if the segment is deleted in some cases, there is evidence through the gradient and variable nature of the process that the segment is indeed present underlyingly. Yet once the deletion becomes increasingly systematic, it could easily be reinterpreted by the hearer or language learner as not being present at all. For such members of the speaker community, the process would be realized as a categorical one of deletion. There is no reason to suppose that, at the precise point that the effects of a rule are reinterpreted as categorical, it becomes a lexical rule. It follows that there are either two distinct sorts of postlexical rules or as assumed here that phonetic implementation rules are distinct from postlexical ones.

While logically postlexical phonological rule and phonetic implementation rules are distinct, differentiating between the two may nevertheless be difficult. Insight in such cases can be gained by considering more clearcut cases first. If we can identify cases which are clearly phonological and others which are phonetic and show that realization of these patterns correlates with the expected plateau vs. cline distinction, this enables us to
use this distinction to evaluate less obvious cases. In the subsequent chapters, we will proceed in this way, first studying patterns of nasalization where the phonological status is independently ascertainable.

We will find that the patterns of nasalization investigated in this study are indeed amenable to characterization in these terms; but even in light of these results, it does not follow that all features and types of data will necessarily be interpretable in such a transparent fashion. A genuine counterexample to this characterization of phonetic vs. phonological would be a case where there is a rule, argued to be phonological, due to cyclic application or rule ordering, which is shown to apply in a gradient fashion or where an underlying contrast is realized gradiently. One such possible exception is the case of realization of Hausa tone (Inkelas, Leben, and Cobler 1987) where each syllable must be specified for a tone and a sequence of high tones appears to result in a gradient output.

At this point it might be useful to consider a schematic representation of the grammar, presented in Figure 12.

```
lexical phonology

postlexical phonology

phonetic implementation:
  target evaluation
  interpolation
  phonetic constraints
```

**Figure 12.** Schematic view of the grammar.

The implication of this model is that there should be interaction between lexical phonological rules and morphological ones, but not between postlexical rules and lexical ones, nor between the phonetic implementation and the phonological rules. This means that rules manipulating discrete and timeless units are not interleaved with those that manipulate quantitative values. Although following the model of Pierrehumbert and Beckman (1988), where phonetic implementation adds information, rather than replacing it, it is not impossible that phonological rules follow phonetic ones. Some cases of what appear to be such ordering paradoxes have been cited in the literature, e.g. Anderson
(1975). The ordering of phonological and phonetic rules is an area that warrants further investigation and will be able to be addressed more systematically as detailed work on the phonetics-phonology interface continues.

4. Organization of the study

In this chapter, the background has been laid out for the remainder of the study. I have investigated a number of issues about phonological and phonetic rules and representations and their relationship. A target-interpolation model has been proposed as a starting point to account for the implementation of the feature Nasal. In the subsequent chapters, this model will be extended and modified to account for patterns of nasalization in French, English, and Sundanese.

The structure of the remaining chapters is as follows. In Chapter 2, more specific background for the study is presented and methodology is discussed, including description of the apparatus used for data collection and methods of analysis. In Chapter 3, we start with the facts of French, where a wide range of possible contexts of nasal-oral sequences is studied. In order to account for the observed patterns, certain modifications are proposed to the target-interpolation model, including assigning inherent duration to targets and developing the role of phonetic constraints. In Chapter 4, we turn to the data of English, where again a range of contexts of nasal-oral sequences are studied. It is observed that the facts are basically compatible with the model proposed in Chapter 3, although variability must be taken into account. In Chapter 5, we extend our study to Sundanese, a language which exhibits long-distance effects of nasalization. In Chapter 6, the concluding chapter, some comparison between the three languages is presented, the proposed model is discussed, and implications of the analysis are considered.

5. References


Privativity, Underspecification, and Consequences for an Adequate Theory of Phonetic Implementation*

Abigail C. Cohn

The consequences of privative features for an adequate model of phonetic implementation are explored. First some background on recent theories of underspecification and privativity is presented. Then these issues are explored by examining the observed patterns of nasalization in French. First an account of the phonetic implementation of these patterns is sketched out assuming Nasal as a binary feature and then considering what would be required for Nasal to be interpreted as a privative feature. The results require a "smart phonetics" which has access to a wide range of information, not necessarily generally assumed to be part of a phonetic representation.

1. Introduction

The view of phonological representations as consisting of sets of feature specifications is one of the cornerstones of modern phonology. Recent proposals include the view that all phonological features are fully specified, as well as views of partial specification (falling as a class under the term "underspecification"). Closely related to the degree of specification is the issue of whether features are binary, as assumed in Chomsky and Halle (1968, hereafter SPE) or privative (building on Trubetzkoy’s 1939 concept), as assumed more recently by a number of researchers. The nature of phonological representation, particularly the surface representation, has direct implications for our understanding of the relationship between phonetics and phonology and the question of phonetic implementation. In this paper, I would like to explore the nature of feature specification and its implications for the phonology-phonetics interface.

First in the remainder of this section, I review some issues about phonological specification, starting with the view espoused in SPE (§1.1) and then I briefly discuss the relationship between phonology and phonetics (§1.2). In §2, I explore the implications of privative features for an adequate phonetic implementation by looking at a specific case, the phonetic patterns of nasalization in French.

* This is a revised version of a paper entitled “Privative Features: Consequences for an Adequate Theory of Phonetic Implementation” presented at the Phonetics-Phonology Workshop at the LSA Linguistics Institute, Columbus Ohio, August 1993. Versions of this paper have also been presented at Cornell University and University of Maryland, College Park. Thanks to each of these audiences and Ioana Chitoran for useful discussion and thanks especially to Donca Steriade for a provocative and interesting response at the Phonetics-Phonology Workshop. This research has been supported in part by NSF grant # BNS-9111155.
1.1 Underspecification and Privativity

In SPE, an explicit distinctive feature theory is proposed, building on the seminal work of Jakobson, Fant, and Halle (1963). The SPE system has been enormously influential and serves as the foundation of many current assumptions, explicit or implicit, about the nature of phonological representations. Following Jakobson, Fant, and Halle (1963), in SPE phonological representation consisted of bundles of a limited number of binary features (with phonetic correlates). Full specification of all features was assumed within the phonology. Generalizations about markedness were captured in the lexicon, with a set of markedness conventions, but these were implemented prior to the input to the phonology. The input to the phonology, or the underlying representation, consisted of fully specified feature matrices. Following this assumption, all phonological rules were feature changing.

The advantages of the SPE system were that it offered a constrained and explicit theory of phonological structure. More recently, many arguments have been put forward in support of other views of feature structure and specification, claimed to offer a more insightful view of the phonology. I consider here two closely related proposals: underspecification and privativity.

Most recent work in phonology rejects the notion that there is full specification at the input into the phonology; rather it is widely assumed that there is at least some degree of underspecification of phonological representation (see Steriade 1987, Archangeli 1988, Mester and Itô 1989, and more recently Clements 1993, Steriade 1995). Two main views of underspecification have been dominant:

(1)  • Contrastive Underspecification = no redundant values, binary values in case of contrast (e.g. Steriade 1987, Clements 1988)
       • Radical Underspecification = only a single value present underlingly, unmarked values filled in (e.g. Kiparsky 1982, Archangeli 1984)

Contrastive and Radical Underspecification share the assumption that redundant information is excluded from underlying representations and may be filled in at some point in the phonological derivation. They differ in their treatment of contrastive "unmarked values", assumed to be present underlingly in the case of Contrastive Underspecification, but absent in the case of Radical Underspecification, filled in by redundancy rules during the course of the derivation. Under a strong view of Radical Underspecification (e.g. Kiparsky 1982, Archangeli 1984), it is assumed that there is full specification at the output of the phonology, while Contrastive Underspecification is agnostic on this point. Evidence
for these views of underspecification comes from behavior of morpheme structure constraints and phonological processes.

Another important source of evidence for the underspecification of phonological representations comes from what has been termed "phonetic underspecification" (Keating 1988), cases where patterns of phonetic realization are affected solely by the phonetic context and not the presence of a feature value. As argued by Keating (1988) and Cohn (1990), such cases provide independent support for the nature of phonological specification and also argue against the strong view of Radical Underspecification.

Several recent proposals suggest that many, if not all, distinctive features should be viewed as privative (e.g. Lombardi 1991, Steriade 1995). What does it mean for a feature to be privative? In contrast to binary features, where reference is made to both plus and minus values, phonological contrast is captured by the presence or absence of a particular feature. The strongest hypothesis would be that no need ever arises to refer to more than one value of a particular feature in the phonology.

The term "privative" comes from Trubetzkoy's (1939) use of the term in his system of oppositions. In particular, within his system distinctions are made between privative, equipollent, and gradual oppositions. Of interest to us here are privative vs. equipollent oppositions:

\[ a. \textit{Privative} \text{ oppositions are oppositions in which one member is characterized by the presence, the other by the absence, of a mark. For example: "voiced"/"voiceless," "nasalized"/"nonnasalized," "rounded"/"unrounded." The opposition member that is characterized by the presence of the mark is called "marked," the member characterized by its absence "unmarked." . . .} \]

[b. \textit{Gradual}]

\[ c. \textit{Equipollent} \text{ oppositions are oppositions in which both members are logically equivalent, that is, they are neither considered as two degrees of one property nor as the absence or presence of a property. For example: German } p-t \text{ and } f-k. \text{ (Trubetzkoy 1939, translated by Baltaxe 1969, p. 75) } \]

First, we note the importance of markedness in Trubetzkoy's definition. The presence of a mark—markedness—is defined based on language specific evidence, determined by neutralization. These terms are relative for Trubetzkoy and may be interpreted in a language specific manner; it is the structure and the function of the system that allows this determination. The same set of sounds may have a different status in two different languages, thus this notion of privative opposition is different from the current notion, where privative features are generally assumed to be universally defined as such. It is also
important to remember that for Trubetzkoys these were oppositions within a phonological system and not distinctive features as such.

A number of different sorts of arguments have recently been put forth in favor of privative features. First, a privative feature theory can be argued to be more restrictive, as there are fewer possible representations. Unlike theories of underspecification where only one value is allowed at some point in the derivation, under a privative theory only one feature value is allowed ever. Second, a privative feature theory has been argued to be more empirically adequate. For example, very few convincing cases of [-coronal] have been identified, and this observation is part of the motivation for arguing that Coronal and the other basic places of articulation should be represented with articulator nodes, rather than with binary features (see Sagey 1986 and McCarthy 1988). Finally a privative theory has been argued to offer a more coherent theory of neutralization (Lombardi 1991). Such an approach avoids the problem where a minus value and an unspecified value represent the same thing, e.g. [-voice] and [Øvoice] for obstruents. It is important to note that it might be that some features are privative and others are binary.

To compare these differing views of feature specification, I take as an example English voicing specifications, as shown in (2).

(2) English voicing specifications

<table>
<thead>
<tr>
<th>input</th>
<th>e. g.</th>
<th>full spec</th>
<th>contrastive</th>
<th>radical</th>
<th>privative</th>
</tr>
</thead>
<tbody>
<tr>
<td>obstruents</td>
<td>t, s, ċ</td>
<td>-V</td>
<td>-V</td>
<td>ØV</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>d, z, ȳ</td>
<td>+V</td>
<td>+V</td>
<td>+V</td>
<td>V</td>
</tr>
<tr>
<td>nasals</td>
<td>n</td>
<td>+V</td>
<td>ØV</td>
<td>ØV</td>
<td></td>
</tr>
<tr>
<td>liquids/glides</td>
<td>l, y</td>
<td>+V</td>
<td>ØV</td>
<td>ØV</td>
<td></td>
</tr>
<tr>
<td>vowels</td>
<td>i, a, u</td>
<td>+V</td>
<td>ØV</td>
<td>ØV</td>
<td></td>
</tr>
<tr>
<td>output</td>
<td>same</td>
<td>specs may be added</td>
<td>full spec by default</td>
<td>same</td>
<td></td>
</tr>
</tbody>
</table>

Under a view of full specification, all segments are underlyingly specified as [+voice] or [-voice] and remain specified throughout the phonological derivation. Any rules therefore involve features changing, e.g. voicing assimilation in clusters would involve a [-αvoice] specification becoming [αvoice] in a particular context. Under Contrastive Underspecification, both feature values would be underlyingly specified within the class of obstruents, but no specifications would be present underlyingly within the class of sonorants. These redundant values might or might not get filled in during the course of the
derivation. Under Radical Underspecification, only the contrastive, marked values would be present underlingly, that is, [+voice] in obstruents; both the unmarked [-voice] values in the obstruents and the redundant [+voice] values in the sonorants would be filled in by rule, required before any other rule could apply making references to those values. At the output of the phonology, the specifications under these three views might potentially be the same. A much sparser representation results if Voice is privative, underlingly only the voiced obstruents would be marked as Voice (just like the case of Radical Underspecification), and under the strictest interpretation, this would be the case throughout the derivation, with no [-voice] values ever filled in.

One goal of an adequate phonological theory is to incorporate markedness. Such a theory should account not only for occurring vs. non-occurring structures, but should also inform us about common vs. uncommon structures. Both Privativity and Radical Underspecification claim to incorporate markedness. Privativity incorporates markedness in the most direct way, by saying that unmarked values do not exist. Within Radical Underspecification, markedness is achieved by relegating unmarked values to a secondary position. They are not present underlingly, being assigned by rule in the course of the derivation. Additionally, unmarked values have been argued to be language specific; for example, Archangeli and Pulleyblank (1989) argue that the unmarked value in an ATR system is a language specific property. Privativity's position vis-à-vis markedness is clearly the stronger one.

An important question to ask is whether Privativity is substantively different from Radical Underspecification. Until the point at which unmarked values are filled in for a radically underspecified feature, Privativity and Radical Underspecification do not appear to be distinct. The real issue is whether the unmarked values are filled in at some point in the phonology. Radical Underspecification does not identify a specific point at which such values are filled in, while the strong version of Privativity asserts that such values play no role at any point in the phonology. Other versions of Privativity allow values to be filled in post-lexically or phonetically, making these theories more constrained than Radical Underspecification only in that an explicit point for feature fill-in is identified (see Lombardi, ms. for discussion of this point).

Recent research by Prince and Smolensky (1993) and McCarthy and Prince (1993) results in a rather dramatic rethinking of many basic assumptions of phonology. Following this view, phonology is no longer seen in derivational terms. One area that clearly requires rethinking is the nature of phonological specification. Theories of underspecification have largely been derivational theories, asking how information is filled in over the course of the
derivation. Yet the goals of these theories—an account of empirical adequacy, markedness, contrast vs. redundancy—still need to be achieved. These issues need to be reconsidered in non-derivational terms. (See Clements 1993 and Steriade 1995 for discussion of related issues.)

Smolensky (1993) argues that Radical Underspecification is inadequate as a treatment of markedness. Optimality Theory offers a different view of markedness, literally, unmarked things have fewer marks or violations. (However, Lombardi to appear argues that certain features may indeed still be privative.) But what about underspecification more generally, should we return to the SPE view of full specification? The basic Optimality Theory position seems to be to assume full specification, unless there is evidence to the contrary; while the default derivational view has been to assume as little information as necessary. What should we take for evidence? Inkelas (1994) argues that underspecification is still required in certain cases of alternations in the shape of a morpheme and cases where three values of a feature appear to play a role. Similarly, evidence from phonetic underspecification also supports this conclusion. We might refer to these sorts of evidence as "empirically motivated" underspecification. There is also an important issue of a non-derivational view of contrastive vs. redundant information. Optimality Theory potentially offers other ways of capturing redundant vs. contrastive information besides, or in addition to, underspecification. Such differences might also be captured through underparsing and the relative ranking of parse constraints. Systematic discussion of these very important issues is beyond the scope of the present paper. Here, I will continue to use some derivational metaphors, but from time to time will step back and explicitly consider the consequences for Optimality Theory.

If some or all phonological features are indeed privative, this has important implications not only for phonological representations, but also for the relationship between phonetics and phonology.

1.2 The relationship between phonology and phonetics

A theory of privative phonological features has direct consequences for the phonetics-phonology interface. Before laying out this issue more explicitly, I start by characterizing the relationship between phonetics and phonology that I am assuming here.

The phonology consists of abstract sound patterns, while the phonetics is the physical realization of these abstract patterns. In considering the relationship between phonology and phonetics, I make two important assumptions: (1) Both phonology and phonetics may be language specific and therefore constitute part of the linguistic grammar. (2) Phonology
and phonetics are distinct, as the mechanisms are distinct—phonology manipulates discrete abstract units, while phonetics manipulates gradient quantitative structures. (See Cohn 1993b for discussion of these assumptions.)

I assume that the phonology-phonetics interface consists of translation of a static representation into a quantitative one, realized in both time and space. Following Pierrehumbert (1980), Keating (1985), and others, I assume a target-interpolation model of this relationship, whereby feature specifications leaving the phonology are translated into phonetic targets, such that a plus value translates to relatively more of the physical value that implements a particular feature than a minus value. The targets are then hooked up through phonetic interpolation. Based on empirical evidence, phonetic targets are argued to have inherent duration (Pierrehumbert and Beckman 1988, Huffman 1989). In the case of the feature Nasal, targets last for most of the duration of a segment (Cohn 1990). Furthermore, as discussed above, underspecification may persist into the phonetics (Keating 1988). In the case of Nasal, assuming for the moment that it is a binary feature, three possible values are present at the output of the phonology: +, −, Ø.

A schematic example of the mapping from the phonology to the phonetics for the feature Nasal following this view is given in (3).

(3) Phonological Output: a. -N +N +N b. -N ØN +N

Phonetic Targets: & interpolation

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td>low</td>
</tr>
</tbody>
</table>

In (3a), we see an example where each segment is fully specified at the output of the phonology. Each segment has a target lasting for most of the duration of the segment and rapid transitions between the segments are expected. In (3b), we see a case where the first and third segments are specified, but the middle segment is unspecified at the output of the phonology. We expect the first segment to be oral for most of its duration and the third segment nasal, with the middle segment’s pattern of nasalization being defined by the phonetic context, showing a clinalike pattern throughout the duration of the segment.

At this juncture it is worth asking whether Optimality Theory requires a rethinking of the phonology-phonetics interface. In one sense, the answer to this question is no: since it is the surface phonological representation that serves as input to the phonetics, it does not matter whether this representation has arisen as a final step in a traditional derivation or as the most harmonic member of a candidate set, generated by Gen and evaluated by a set of ranked constraints. However, recently generative phonetic analyses have been cast in rule
based terms; this suggests a rethinking of the phonetics in non-derivational terms. We might define a non-derivational phonetics as a set of quantitative constraints, resulting from articulatory, aerodynamic, and perceptual requirements, evaluated in light of phonological structure. A non-derivational phonetics is compatible with many of the assumptions already made within generative phonetics. First it is generally assumed that there should not be any intrinsic ordering of rules within the phonetics. Implementation is necessarily constrained, both by the representation it is implementing and certain independent requirements of phonetic well-formedness. And in this sense, it seems straightforward to recast the notion of phonetic implementation in terms of constraints. (For fuller discussion of these issues, see Cohn 1995.)

With this brief review of issues of underspecification and privativity and the nature of the relationship between phonology and phonetics, I would like to explore the implications of privative features for the phonology-phonetics interface: the question to ask is how would the phonetics implement a privative feature (as opposed to a binary one). If we assume "weak privativity", whereby a privative feature becomes binary later in the phonology, there are really no implications for the phonetics distinct from the implementation of binary features, since by the output of the phonology, there will be both plus and minus values referred to. The substantive question to ask, then, is what are the consequences of "strong privativity": if the output of the phonology, that is, the input to the phonetics, consists of only privative feature specifications, what would the phonetics have to look like?

As noted by Keating (1991), common phoneticians' practice interprets [ØF] as distinct from [-F]:

\[
\begin{align*}
(+F) & \quad \text{do some specific thing} \\
(-F) & \quad \text{do the opposite specific thing} \\
(ØF) & \quad \text{do nothing with passive dependence on context}
\end{align*}
\]

Keating notes that this results in a paradox: "If [-F] is collapsed with [ØF] because it seems phonologically invisible, then it cannot be visible later on to block phonetic interactions." Can we fully maintain a privative phonetics, with no distinction between [-F] and [ØF] and at what cost?

2. Implications of privative features

In this section, I explore the implications of privative features for the phonology-phonetics interface by looking at the phonetic realization of patterns of nasalization in French. I start in §2.1, by considering the phonological evidence for the feature Nasal as a
privative feature, as well as the feature Voice, since this will be relevant to the phonetic implementation of Nasal. I then turn to the study itself in §2.2 & §2.3 and the implications for privativity §2.4.

2.1 The features Nasal and Voice

A number of recent proposals argue that some if not all features may be privative; we consider here the proposals and evidence for Nasal and Voice.

Several researchers have recently argued that the feature Nasal is privative (see, e.g. Itô and Mester 1989, Steriade 1993a, b, Trigo 1993). Steriade (1993b) discusses the observed patterns of markedness and notes that Nasal is a case of what she terms "context free markedness": "If a language possesses a [+nasal] series of sounds then it will possess the corresponding [-nasal] series." (Steriade 1993b, p. 334). Conceptually Nasal fits the criterion laid out by Steriade for a privative feature.

One set of proposals assumes a representation, where the feature Nasal is dominated by a soft palate node (see e.g. Itô and Mester 1989, building on the structure proposed by Sagey 1986), as shown in (5).

(5) A representation of Nasal as a privative feature

```
SoftPalate
| Nasal
```

As Yip (1989) argues, there are a number of reasons to find this representation undesirable; basically it just encodes a binary distinction through an extra dummy feature, which has no phonetic content of its own. A more substantive proposal for Nasal as a privative feature comes from Steriade's work within her aperture node theory (Steriade 1993a, b). Steriade argues that [-continuant] segments are phonologically characterized by both a closure and release phase. Her proposed representations for plain, prenasalized, and postnasalized nasals are sketched out in (6).

(6) nasal stop  prenasalized stop  postnasalized stop

```
[nasal]  / \  A0 Amax
[m]  [mb]  [b\text{m}]
```

This leaves us with the empirical issue of accounting for putative cases of reference to [-nasal] within the phonology of any language. As argued by Steriade, there are few
convincing cases spreading [-nasal] and a clear asymmetry exists in that no long distance cases of [-nasal] spreading exist. The main use of [-nasal] is in the representation of pre- and post-nasalized stops. Steriade's representations obviate the need for reference to [-nasal] for these segments, as shown in (6) and also capture an important asymmetry, not predicted in other representations, that only noncontinuants can be contour. The case for Nasal as a phonologically privative feature is quite strong. There are, however, a few problematic cases which require attention. These include a local [-nasal] spreading rule in Sundanese, shown by Cohn (1993a) to be phonological and cases of references to three values of Nasal, as seen, for example, in Tucano (see Noske 1993).

A number of researchers have also argued that the feature Voice is privative (see, for example, Mester and Itô 1989, Cho 1990, Lombardi 1991). In discussing the status of Voice, it is important to note that what is meant is a widely assumed post-SPE version of the feature (following e.g. Clements 1985), where it is the presence or absence of voicing which is the phonetic correlate, not the position of the vocal cords (as argued in SPE). Mester and Itô and Cho argue for Voice as privative, while Lombardi and Steriade (1994) argue that the laryngeal features as a class are privative. (The evidence for both Spread Glottis and Constricted Glottis as privative is quite convincing, but will not concern us here.) Here too the observed patterns of markedness are consistent with Steriade's criterion: "the presence of a voiced obstruent series implies the presence of the corresponding voiceless series, but not vice versa." (Mester and Itô 1989, p. 263, following Maddieson 1984a).

There is a large set of putative cases of phonological reference to [-voice] (as discussed by Mester and Itô 1989, Cho 1990, Lombardi 1991, Kenstowicz 1994). One such class of cases is symmetrical spreading rules: [voice] —> [\alpha voice] / _____ [\alpha voice]. The solution proposed by a number of researchers is that these cases really consist of two distinct rules, one of delinking and the other of spreading. Such analyses make a clear prediction that the voiced sequences, but not the voiceless ones, constitute linked structures. We therefore might expect to see integrity effects in the former, but not the latter structures, but I do not know of any such cases that have been discussed in the literature. Another class involves cases where what was assumed to be [-voice] is argued to be something else, e.g. "voiceless nasals" as being [+Spread Glottis] (see e.g. Cho 1991). Still other cases can be argued to be the result of phonetic implementation and not the phonology. Sonorant Devoicing in English might be such a case. As has been widely observed, a sonorant following a voiceless stop is devoiced, e.g. play [pley]. Under the view that Voice is privative, the explanation cannot be that [\j] is [-voice]. It has been suggested that this is
due to a phonological assimilation rule, whereby [+SG], rather than [-voice] is spread to the liquid (see Iverson and Salmon 1994 for such an account), though the timing facts may be more complex than would be predicted following this view. On the other hand, it is quite likely that this is really a question of phonetic implementation and interpolation of voicing through an unspecified span (as argued by Cohn 1990).

Yet other cases of phonological [-voice] specifications and redundant [+voice] in sonorants have led some researchers to propose what I will term "weak privativity", whereby at a later point in the grammar a privative feature becomes binary. Cho (1991) suggests that phonologically Voice is privative, but that phonetically distinctions are made between [+voice] and [-voice]. Lombardi (ms) has suggested that there may be post-lexical reference to [-voice]. These proposals are in a sense variants of Radical Underspecification, but more constrained in that such values only become available at a recognized point in the grammar. Central to this approach is the claim that such rules always apply late in the grammar. Another type of solution to difficult cases of characterizing sonorant voicing vs. obstruent voicing is to propose that sonorant and obstruent voicing are distinct features, see Rice (1993) and Steriade (1995) for proposals along these lines.1 While many researchers hold to the view that Voice is phonologically privative, this is clearly a complex question and one that I believe has not been fully resolved.

2.2 The phonetic realization of patterns of nasalization in French

In this study, nasal airflow was taken as a measure of the phonetic output for the feature Nasal. Data from a wide range of phonological contexts was examined for two speakers. There were five repetitions of each utterance, said in a frame /dit _____ dø fwa/dites_____ deux fois 'say _____ twice'. The results were quite consistent across repetitions for both speakers. (See Cohn 1990 for a fuller discussion of the methodology used in the study.)

It is the feature specifications at the output of the phonology, which serve as input to the phonetics. Assuming for the moment Contrastive Underspecification, both [+nasal] and [-nasal] specifications are present in the case of contrast, that is, for both [-continuant] consonants and vowels. Where Nasal is redundant, in the case of [+continuant]

1 Note that the two features in these sorts of solutions are distinct from the two features Stiff and Slack proposed by Halle and Stevens (1971), which between them define a continuum.
consonants, I assume that no values are specified. This results in three possible values, [+nasal], [-nasal] and [Ønasal] at the output of the phonology. This is summarized in (7).

(7) French phonological specifications

<table>
<thead>
<tr>
<th>specification</th>
<th>class</th>
<th>abbreviation</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>[+nasal]</td>
<td>nasal consonants</td>
<td>N</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>nasal vowels</td>
<td>∨</td>
<td>ε, ɛ</td>
</tr>
<tr>
<td>[-nasal]</td>
<td>voiceless stops</td>
<td>T</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>voiced stops</td>
<td>D</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>oral vowels</td>
<td>V</td>
<td>e, o</td>
</tr>
<tr>
<td>[Ønasal]</td>
<td>continuants</td>
<td>L</td>
<td>1</td>
</tr>
</tbody>
</table>

The question we need to address is how are patterns of nasal airflow realized for each of these classes of sounds — oral, nasal, and unspecified — in a range of contexts. In §2.3, I first consider a range of patterns of nasal airflow in French and present an analysis assuming Nasal as a binary feature, with [+nasal], [-nasal] and [Ønasal] values at the output of the phonology. Then in §2.4, I turn to the issue of how we might account for the observed patterns with Nasal as a privative feature, excluding [-nasal] values.

2.3 Patterns of nasal airflow, Nasal as a binary feature

The patterns of nasal airflow presented here are part of a larger study, reported in Cohn (1990). In our discussion, I focus on a subset of the patterns and abstract away from many of the finer details of the airflow patterns. I consider here a range of sequences of oral, nasal, and unspecified sounds, by looking at representative examples of nasal airflow traces.

I start by considering the realization of oral segments followed by nasal segments. Assuming phonological feature specifications are mapped to targets lasting most of the duration of the segment, we would expect to see the pattern of phonetic implementation schematized in (8), where the oral segment is fully oral, the nasal segment nasal, with a rapid transition in between.

---

2 Optimality Theory may lead to a rethinking of some of these assumptions, particularly with respect to the treatment of redundant specifications. The different status of contrastive vs. redundant information might be treated by relative ranking of Parse constraints, with ParseContrastive outranking ParseRedundant. These are issues for further research, but here I will continue to assume that redundant values are unspecified, unless there is empirical evidence to the contrary.
Consider now the representative examples of this pattern involving both [-continuant] consonants and vowels, shown in Figure 1.

Figure 1. Representative examples of nasal airflow in [-nasal] [+nasal] sequences, a. DVT *botte* /bɒt/ 'boot'; b. VN *bonnet* /ˈbɒnɛt/ 'bonnet'; c. TN *(di)tes nez* /tʃɛn/ 'say nose'; d. TV *thon (deux)* /tʃd/ 'tuna'.

In Figure 1a, a sequence of oral segments is shown. Here virtually no nasal airflow is observed. This example sets an effective baseline for the other cases. In Figure 1b, we see an example of an oral vowel followed by a nasal consonant. The vowel is oral for most of its duration, followed by a rapid transition into the nasal consonant, which is fully nasal for its duration. In Figure 1c, we see an oral stop followed by a nasal consonant, again with a rapid transition between the fully oral stop and fully nasal consonant. In Figure 1d, we see a similar pattern of an oral stop followed by a nasal vowel. Here the transition, again rapid, occurs after the release of the stop. (In Cohn 1990, an explicit analysis of the patterns of transitions is given, where it is shown that different segments control the transition to various degrees. I abstract away from these issues here.)
To generalize across these patterns, we see that the oral segments are characterized by a lack of nasal airflow for most of their duration, while during the nasal segments, both consonants and vowels, significant nasal airflow is observed for most or all of their duration, with rapid transitions in between, like the pattern schematized in (3a). These patterns of nasal airflow in French allow a direct interpretation of the phonological feature specification of Nasal. Those segments phonologically characterized as oral, which I assume to be specified as [-nasal] at the output of the phonology, show no significant nasal airflow, while those which are phonologically nasal ([+nasal]) show significant nasal airflow both temporally and spatially. Such segments are nasalized for most or all of their duration and the patterns of nasalization are seen to be quite plateau-like, not gradient or transitional.

Let us now compare these patterns with a sequence of an unspecified segment followed by a nasal consonant. The predicted pattern of phonetic implementation is schematized in (9).

\[(9) \ (-N) \ ØN \ +N\]

\[\text{Phonological output: } \quad \begin{array}{ccc}
  x & x & x \\
  \mid & \mid & \\
  -N & +N & \\
\end{array}\]

\[\text{Phonetic implementation: }\]

In this case we expect to see the preceding oral segment be fully oral, the following nasal segment fully nasal and the unspecified segment partially nasalized due to the context, with a cline-like pattern of increasing nasal airflow throughout the duration of the segment. Let us consider the representative examples of this pattern presented in Figure 2.

![Figure 2](image)

**Figure 2.** Representative examples of nasal airflow in [Ønasal] [+nasal] sequences, a. VLN *belle Ne(l) /bel#ne/ 'pretty Nell'; b. TLV *(di)tes long /#l#l/ 'long'.

\[100\text{ms}\]
In Figure 2a, we see that the preceding oral vowel is fully oral, the following nasal consonant fully nasal and intervening [l] shows increasing partial nasalization throughout its duration. This is quite different from the pattern seen for an oral vowel or stop before a nasal consonant, illustrated in Figure 1. Here the [l] appears not to contribute an oral target of its own, consistent with the assumption that it is phonologically unspecified. This pattern is accounted for in my analysis as interpolation through an unspecified span, resulting in a pattern that looks like the schematic example, shown above in (3b).

But not all [l]'s (or [+continuant] consonants more generally) are the same. In Figure 2b, the [l] is fully oral. This difference appears to be due to syllable position. An [l] in the onset of a syllable behaves as if it were [-nasal]. In Cohn (1990), I propose a phonological rule which assigns [-nasal] to an onset otherwise unspecified for Nasal in order to account for these sorts of cases. This could also be treated non-derivationally. Intuitively, what is at issue here is that fuller specification is required in onset position, resulting in the unmarked [-nasal] specification in the absence of any other specification.

In these cases then, we see evidence for three classes of segments—[+nasal], [-nasal], [Ønasal]—at the output of the phonology, as well as a rule or constraint requiring specification of the feature Nasal in onset position. Now let us turn to carryover cases of +N-N and +N ØN -N. The predicted patterns for these cases—the mirror images of -N+N and -NØN+N—are schematized in (10) and representative examples are presented in Figure 3.

\[
(10) \quad \begin{array}{c}
+\text{N-N and } +\text{N } \emptyset \text{N } \text{-N} \\
\text{Phonological output: } & x \quad x \quad & x \quad x \quad x \\
& \text{| } \text{|} \quad \text{| } \text{|} \quad \text{| } \text{|} \\
& +\text{N} \quad -\text{N} \quad & +\text{N} \quad -\text{N} \\
\text{Phonetic implementation: } & \underline{x} & \underline{x} \quad & \underline{x} \quad \underline{x} \quad \underline{x} \\
\end{array}
\]
100ms

**Figure 3.** Representative examples of nasal airflow in [+nasal][-nasal] and [+nasal][Ønasal][-nasal] sequences, a. VT *bonté* /boNtle/ 'goodness'; b. NLV *(b)onne lèt(re)l* /on#lel/ 'good letter'; c. ND *(b)onne dele(tte)l* /on#del/ 'good debt'; d. NV *nez d(eux)l* /ne#d/ 'nose twice'.

In Figure 3a, the pattern of a nasal consonant followed by an oral (voiceless) stop is what we would predict, with the former being fully oral and the latter fully nasal, with a rapid transition in between. In Figure 3b, with an [l] following a nasal consonant, we find a clinelike pattern throughout the [l], again what we would expect for an unspecified segment. (Note, however that the [l] is in onset position, but nevertheless seems to be unspecified.) But the patterns in Figure 3c & d, which we would a priori expect to look like the pattern in Figure 3a, are very different. In both of these cases, the segment following the nasal consonant, an oral voiced stop or vowel, appears to be unspecified for Nasal in this context, exhibiting a clinelike pattern of nasalization, just like [l]. In Cohn (1990), I propose a rule of [-Nasal] Delinking, whereby the [-nasal] specification of a voiced segment is deleted following a [+nasal] specification. Intuitively, what is happening here is that the nasal specification is spreading to the right, resulting in a phonetic effect of nasalization, and this spreading overrides the presence of a [-nasal] specification. The result of this is distinct behavior of voiceless and voiced oral segments, yielding an asymmetrical pattern for voiced stops and vowels preceding and following nasals.

In (11) I summarize the observed patterns of nasal airflow.
(11) Patterns of nasal airflow in French

N, \( \checkmark \) always nasal, independent of context
T always oral, independent of context
L oral, or a cline, dependent on context and syllable structure
D, V asymmetrical behavior, oral before +N, partially nasal after +N

Nasals and voiceless stops exhibit consistent patterns of airflow independent of context, while voiced stops, continuant consonants, and oral vowels all exhibit patterns dependent on context. In addition, the pattern for oral stops and vowels is different from that seen for [+continuant] consonants. Thus we see several distinct patterns of behavior among those segments which are not [+nasal].

In Cohn (1990), a systematic comparison of nasalization in French and English is presented. Here, I focus on the contexts most relevant to our discussion; the VN and NV contexts in English are exemplified in Figure 4.

\[ \begin{array}{ccc}
\text{a.} & \text{d} & \text{e} & \text{n} \\
-\text{N} & \text{ØN} & +\text{N}
\end{array} \]

\[ \begin{array}{ccc}
\text{b.} & \text{n} & \text{e} & \text{d} \\
+\text{N} & \text{ØN} & -\text{N}
\end{array} \]

100ms

**Figure 4.** Representative examples of nasal airflow in VN and NV sequences in English, a. DVN *den* /den/ b. NVD *ned* /ned/.

The VN pattern in English looks like the French LN pattern, seen above in Figure 2a. In Cohn (1990, 1993b), I argue that this is because vowels in English are unspecified for Nasal, due to lack of contrast; thus the clinelike pattern results from interpolation through an unspecified span. The NV pattern in English is very similar to that seen above for French. This is again due to the lack of specification in English, resulting in a symmetrical pattern in the VN and NV cases in English, in contrast to the asymmetrical pattern observed in French. Within a binary system of feature specification of Nasal, the language specific differences between French and English follow from the difference in feature specification, directly reflecting the difference in contrast between the two languages.

While D, V, and L appear to contribute nothing in terms of orality following a nasal consonant, when we look at the most restrictive context, an oral segment between two [+nasal] segments, we see that subtle, but significant, differences emerge. The predicted
realization of +N-N+N and +NØN+N patterns is schematized in (12) and representative examples are presented in Figure 5.

(12) +N -N +N and +N ØN +N patterns

Phonological output:

```
+X +X +X
+X +X +X
```

Phonetic implementation:

```
```

Figure 5. Representative examples of nasal airflow in [+nasal][-nasal][+nasal] and [+nasal][Ønasal][+nasal] sequences, a. DTV (b)on than /'G#h/ 'good tuna'; b. DLV (b)on lin /'Gh/ 'good flax'; c. DGV (d)indon /'Gd/ 'turkey'; d. NVN nonne /'nun/ 'nun'.

In Figure 5a, we see that [t] is still fully oral and in Figure 5b, we see that [l] is nasalized throughout, imposing no orality of its own. These are the expected patterns for a [-nasal] and [Ønasal] segment respectively in this context. However the patterns with D and V are different from what we saw above in Figure 3. In Figure 5c, we see that D imposes some orality; it is not fully nasalized and therefore not fully unspecified. Similarly in Figure 5d, while the "oral" vowel is nasalized throughout, it exhibits a low level of nasalization, not the degree of nasalization expected of a fully nasalized vowel with the same quality (such as

---

3 The difference in level of nasal airflow between [ŋ] and [ɛ] seen in these examples is the result of an impedance effect, due to the differing degree of oral openings in these two vowels. I abstract away from these differences in this discussion.
seen in Figure 5a). Thus D and V impose certain weak phonetic requirements. Different classes of segments appear to impose distinct phonetic requirements or conditions on phonetic well-formedness of a segment. These are precisely the sorts of quantitative constraints that we would expect in the phonetics.

As we have seen, the phonetic requirements of voiced and voiceless stops are different. Phonetically, voiced stops minimally have oral releases, whereas voiceless stops must be oral for most of their full duration. We might account for this difference through reference to closure and release, that is, voiced stops are amenable to having their closure portion nasalized (in effect becoming a prenasalized stop) while voiceless stops are not. This difference is schematized in (13).

(13) Phonetic requirements of voiced and voiceless stops

[d] [t]

We can account for the behavior of oral vowels, following Keating's (1990) Window Model, assuming that the target for an oral vowel is not necessarily complete orality, but rather constitutes a limited range in space between fully oral and partially nasal, while a nasal vowel has a range of values falling within a range of full nasalization. As schematized in (14), this would account for the apparent lack of target of an oral vowel in the context following a [+nasal] segment, but preceding a [-nasal] one (14a), compared with the case between two [+nasal] segments (14b).

(14) Phonetic requirements of vowels

a. NVT

b. NVN

In (15), I summarize the phonological specifications and phonetic requirements of each class of segments.
(15) Phonological specifications and phonetic requirements

<table>
<thead>
<tr>
<th>phon spec</th>
<th>phonetic realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Nasal consonants [+nasal]</td>
<td></td>
</tr>
<tr>
<td>Nasal vowels [+nasal]</td>
<td></td>
</tr>
<tr>
<td>Voiceless stops [-nasal]</td>
<td></td>
</tr>
<tr>
<td>b. Voiced stops [-nasal]</td>
<td></td>
</tr>
<tr>
<td>Vowels [-nasal]</td>
<td></td>
</tr>
<tr>
<td>/l/ [Ønasal]</td>
<td></td>
</tr>
</tbody>
</table>

In (15a), the patterns of phonetic realization reflect direct interpretation of what I have assumed to be the phonological specifications of these classes of segments. As noted above, for these classes of sounds, the pattern of phonetic implementation is independent from phonetic context. The situation in (15b) is more complex. For these sounds, the pattern of realization is dependent on context and shows a stronger or weaker requirement depending on the phonetic context. Voiced stops and oral vowels are fully oral preceding a [+nasal] segment, but impose only weak constraints after a [+nasal] specification. [+continuant] consonants, exemplified by [l], impose no requirements, except in onset position before a [+nasal] specification. I believe that these well-formedness requirements follow from articulatory (aerodynamic) or perceptual requirements, but I state them here as descriptive requirements based on empirical observations of the patterns, since clearly more work is needed to fully understand the nature of these requirements.

Under an analysis assuming Nasal is a binary feature, an account of these facts is fairly straightforward. The independent patterns seen in (15a) follow directly from the target-interpolation model assumed here. In the cases in (15b), it is assumed that segments may have phonetic well-formedness conditions, weaker than the effect of a phonological specification, that emerge in certain contexts. In a binary account, the difference between the strong vs. weak requirements is due to the presence or absence of a [-nasal] specification. In Cohn (1990), I assume that this is due to the deletion of a [-nasal] specification in certain contexts. In non-derivative terms this could be accounted for as the relative ranking of Parse[-nasal] and phonetic permeation of nasalization.
2.4 The phonetic consequences of Nasal as privative

The importance of phonetic requirements in accounting for the observed patterns leads directly to the question of whether we can account for these patterns WITHOUT reference to [-nasal]. In other words, can the observed patterns be accounted for without making a distinction between [-nasal] and [Ønasal], thereby avoiding Keating's paradox. In (16), I summarize the work that the [-nasal] vs. [Ønasal] distinction is doing in the analysis.

(16) Uses of [-nasal] vs. [Ønasal]
  a. Contrast between T, D, V vs. L
  b. Context dependence of D and V
  c. Orality in syllable onsets for L
  d. Differences between French and English

First a three-way contrast was assumed, between nasal segments, oral ones (T, D, V), and unspecified ones (L), to account for the distinct behavior of each of these classes of sounds. Additionally direct reference was made to [-nasal] both in accounting for the context dependent pattern of realization in D and V and in accounting for the sensitivity to syllable structure in characterizing the realization of L. Further, certain language specific patterns, such as the difference between VN in French (seen above in Figure 1b) and in English (seen in Figure 4a), where a clinelike pattern is observed, is straightforwardly accounted for by making a distinction between [-nasal] in the French case and [Ønasal] in the English case.

In order to account for the observed patterns without reference to [-nasal], the phonetics has to be enriched in a number of ways: (1) the assignment of phonetic targets would have to be based on reference to other feature values and context; (2) reference to syllable structure would have to be made; (3) reference to a rest or neutral position would be necessary; (4) language specific patterns of implementation would be required.

The assignment of phonetic targets to segments specified as [+nasal] would be straightforward, but in order to know how to interpret the pattern of nasalization for a segment unspecified for Nasal, reference would have to be made to other feature specifications besides Nasal. It would be necessary to distinguish between vowels and consonants, between continuants and noncontinuants, and between voiced and voiceless stops. This strategy for phonetic implementation of privative features is suggested by Keating (1991) and would work in cases where the required information would also be present under a privative feature theory. But the situation for the difference between voiced stops and voiceless ones becomes very tricky. Whether we characterize this difference
phonologically (with a rule spreading nasalization to the closure of voiced stops) or phonetically (through phonetic requirements), these segments differ only in their voice specification, thus it appears that reference would need to be made to [+voice] vs. [-voice]. And it is the class of [-voice] segments which impose the more stringent requirements. If we maintain the view that Nasal is privative, this suggests that Voice is not privative, at least not by the input to the phonetics (consistent with "weak privativity", as proposed by both Cho and Lombardi). Other evidence, such as the phonetic evidence from tone splitting and tonogenesis (as discussed by Hombert et al. 1979 and others), also supports this conclusion. What is found is that voiced consonants lower the pitch of the beginning of the following vowel, while voiceless consonants raise the pitch (rather than being neutral). These effects cut across obstruent and sonorant voicing, as discussed by Maddieson (1984b) in the case of Burmese.

The assignment of phonetic targets to oral segments would also have to be sensitive to context, in order to account the pattern of VN vs. NV. This difference is not just due to the window effect, since in the VN case, basically the strong requirement is required. Under a binary analysis, this is due to the presence of a [-nasal] specification in the VN case. Such a solution is not available under the privative view. Thus in order to account for the assignment of phonetic targets, not only is reference to the Nasal feature specifications needed, but access to both paradigmatic and syntagmatic information is required.

Direct reference to syllable structure in the phonetics would also be required. Under a binary view, the transparent behavior of [l], except in syllable initial position, was accounted for with a phonological assignment of [-nasal] syllable initially. If Nasal is privative, the phonetic requirements of [l] would have to be sensitive to syllable structure. Phonetic requirements would be assigned based both on inherent featural content of segments as well as syllable structure. In addition, here too syntagmatic information would be required, since the presence of a preceding [+nasal] specification overrides the effects of syllable structure.

There would also need to be a notion of rest or neutral position, to account for the fact that spans of segments unspecified for Nasal are not just nasalized throughout. Here different classes of segments show distinct behavior. An [l] in a N_\_N environment is indeed nasalized throughout, but other classes of segments are not. In the case of D, the contextual effects of nasalization are quite local, limited to the closure of the segment, with the release being oral. In the case of NVD, the D is fully oral, even though we might expect the closure portion to be amenable to nasalization. To account for the locality of these effects, reference to rest position is needed, that is, after a certain duration (perhaps
defined quite abstractly) the velum reverts to rest position. Articulatorily the velum shows a speech ready position (raised), distinct from breathing position (lowered), so reversion to the speech ready position is not implausible, but how is this mechanism integrated to the phonetics and constrained in its behavior?

Finally, phonetic implementation of feature specifications would have to be language specific to account for the observed differences in French and English. Under a privative view, this difference could not be due to a phonological difference in specification, between [-nasal] and [Ønasal] and it could not be due to differences in window sizes, since the NV cases are quite similar in the two languages. This would require not only language specific, but again context dependent, phonetic interpretation. Good evidence of the language specific nature of phonetic implementation exists, but in this case the phonetics would be required to implement a difference that stems directly from the nature of phonological contrast in the two systems.

In summary then, the consequences of a privative phonological theory are the following. While the basic patterns in French can be accounted for without phonological reference to [-nasal], this can only be done by making the phonetics quite powerful. It requires what we might term "smart phonetics". The absence of Nasal is interpreted through other features, including Voice as a binary feature, and through context, thus requiring both paradigmatic and syntagmatic information. Additionally reference to syllable structure and rest position would be necessary and all of this would have to be language specific to account for differing patterns in e.g. French and English. Whether we interpret this as a desirable result depends in large part on the degree to which there is independent evidence that the phonetics needs to be able to refer to these sorts of information.

3. Conclusions

In the above discussion I have explored the consequences of privative features in the phonology for an adequate model of phonetic implementation. I considered these issues by examining the observed patterns of nasalization in French, sketching out first an account of the phonetic implementation assuming Nasal as a binary feature and then considering what would be required for Nasal to be interpreted as a privative feature. The results require a "smart phonetics" which has access to a wide range of information, not necessarily generally assumed to be part of a phonetic representation. The result is to turn the traditional (SPE) view of language specific phonology and universal phonetics on its head, with much of the language specific work being done in the phonetics. The viability of this
approach depends on independent evidence that the phonetics should have access to this range of information.

While it seems possible to implement Nasal as a privative feature, one of the consequences seems to be to keep Voice a binary feature. This leaves us with the question of how different Voice and Nasal are as features. This in turn has implications for our understanding of markedness. Can we distinguish in a principled manner between features like Nasal and Voice? Steriade (1995) suggests that there are two different kinds of features, those which are privative and characterized by having a neutral position and those which are equipollent, not showing the same asymmetries. Yet as seen above, those arguing for Voice as privative make arguments about markedness similar to those for Nasal, so is there a real difference?

Finally, I have mentioned a number of issues concerning feature specification and the nature of underspecification within a non-derivational view of phonology. Much work remains to be done on issues of feature specification and its relationship to phonetics.

4. References


Iverson, G. and J. Salmons (1994) *Spread Glottis implementation in Germanic*. University of Wisconsin-Milwaukee and Purdue University ms.


Lombardi, L. (ms.) *Allophonic rules and distinctive features*. University of Maryland ms.


The Prosodic Structure of Burmese: A Constraint-Based Approach

Antony Green*

In this paper I shall explain the structure of the syllable, foot, and prosodic word in Burmese, using a constraint-based framework following Optimality Theory. In particular, I shall propose a family of constraints called UNARITY, which states that a prosodic category (PrWd, Ft, σ) is permitted to contain no more than one of the next lower prosodic category (Ft, σ, μ, respectively). This constraint is responsible for several types of idiosyncratic prosodic behavior seen in Burmese: the difference between and distribution of major (heavy) and minor (light) syllables; the fact that a foot (obligatorily bimoraic in Burmese as elsewhere) can be only a single heavy syllable; the resulting lack of [L L] feet, and the difference between so-called “reducing” compounds (in which a member is reduced from a major syllable to a minor syllable) and “nonreducing” compounds (in which no such reduction takes place).

1. Introduction

The Burmese language has been discussed descriptively by a variety of authors, including Bernot (1963, 1980), Okell (1969), and Wheatley (1987), but little work on the theoretical aspects of Burmese phonology has been done.1 In this paper I propose to explain the structure of the syllable, foot, and prosodic word in Burmese, using a constraint-based framework. In particular, I shall discuss several issues in the prosodic structure of Burmese, including the distinction between, and distribution of major and minor syllables and the nature of the foot in Burmese. To this end I shall propose that in Burmese, a prosodic category is permitted to contain no more than one of the next lower prosodic categories; this constraint is responsible for the idiosyncratic prosodic behavior seen in Burmese.

In the rest of section 1 I give the inventory of Burmese surface phones—vowels, tones, and consonants. In section 2 I discuss major and minor syllables. In section 3 I argue that the foot in Burmese is a single heavy syllable, not the iamb proposed by Bennett (1994) for Thai and by Griffith (1991) for Cambodian. In section 4 I discuss the prosodic word (PrWd) with especial attention to compounds and superlong words. In section 5 I discuss the maximum size of prosodic categories and propose a family of constraints called UNARITY to limit the size of any prosodic category to exactly one of the next lower prosodic category. Finally, in section 6 I present other constraints necessary to account for the prosodic behavior of Burmese.

---

*I should like to thank the following people for their help and advice: Fraser Bennett, John Bowers, Abby Cohn, Terri Griffith, Allard Jongman, Martha Larson, Asun Martínez-Arbelaitz, Denise Meyer, David Parkinson, San San Hnin Tun, Ratreer Wayland, Julian Wheatley, Moira Yip, Draga Zec, and an anonymous reviewer. Any mistakes are mine alone.

1Detailed phonetic studies include Mehnert and Richter (1972–77) and Thein Tun (1982).
1.1 Vowels and tones

The surface vowels and tones of Burmese are as shown in (1).2

(1) The surface vowels and tones of Burmese

<table>
<thead>
<tr>
<th>Monophthongs</th>
<th>Diphthongs</th>
<th>Tones</th>
</tr>
</thead>
<tbody>
<tr>
<td>i(:)</td>
<td>ei</td>
<td>high: ã etc.</td>
</tr>
<tr>
<td>u(;)</td>
<td>ou</td>
<td>low: a etc.</td>
</tr>
<tr>
<td>eː</td>
<td>ai</td>
<td>creaky: ã etc.</td>
</tr>
<tr>
<td>oː</td>
<td>au</td>
<td>checked: aʔ etc.</td>
</tr>
<tr>
<td>ø</td>
<td></td>
<td></td>
</tr>
<tr>
<td>æ(ː)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The syllable structure of Burmese is basically C(G)V((V)C), which is to say the onset may consist of a consonant optionally followed by a glide,3 and the rhyme may consist of a vowel alone, a vowel with a consonant, or a diphthong with a consonant. Vowel length in Burmese is not phonemic, but predictable: vowels (except æ) are long in open syllables, and short in closed syllables. Some representative words are shown in (2).4

(2)

a. CV       meː       'girl'
b. CVC      meʔ       'crave'
c. CCV       myeː       'earth'
d. CCVC      myeʔ       'eye'
e. CVVC      fʔauʔ      'address (a superior)'
f. CCVVC     pyeʔN      'be stupid'

2There is very little agreement from one author to another on the designation of the tones. Some authors use terms such as low, high, creaky, checked, falling, heavy, glottalized, etc.; others number the tones 1–4. Among the authors who use numbers, there is even variation as to which tone is given which number. There is also wide variation as to the transliteration of Burmese. Throughout this paper, I shall be using the same names of tones as used by Wheatley (1987), and the same marks to indicate them, except that while Wheatley leaves the low tone unmarked I mark it by a, etc., to disambiguate low tone from absence of tone. Other differences in this paper compared with Wheatley's usage are the transliteration of the placeless nasal as N rather than n and spelling aspirated sounds as pʰ, sʰ, hₘ etc., rather than hp hs hm etc. As Bennett and Lehman (1994, 19) point out, the aspirated nasals are phonetically preaspirated, whence the transcription hₘ.

3But some C + glide clusters are prohibited, e.g. *ky, *ty, etc.

4All examples in this paper are from Bernot (1963), Okell (1969), or Wheatley (1987), unless otherwise noted.
Burmese, like many languages of Southeast Asia, has a distinction between major and minor syllables. The exact definitions of major and minor syllables vary from language to language, but in general a major syllable is a surface-heavy (mü) syllable, and a minor syllable a surface-light (μ) syllable. In Burmese, the characteristics of a major syllable are: (i) it is always heavy (thus open syllables are long); (ii) it may contain any vowel or diphthong except ø; and (iii) it bears tone. All the words in (2) above are examples of major syllables. The characteristics of a minor syllable in Burmese are: (i) it must contain the vowel ø and no other vowel; (ii) it must be a light (therefore open) syllable; (iii) it must not bear tone; and (iv) it must not be the final syllable of the word. A result of this last restriction is that a word may not contain only minor syllables. Examples of words containing minor syllables are shown in (3).

(3)

a. kʰəlou?
  ‘knob’
b. pəlwɛ:
  ‘flute’
c. θəyɛ:
  ‘mock’
d. kəle?
  ‘be wanton’
e. tʰəməyɛ:
  ‘rice-water’
f. məiNməwu?
  ‘women’s clothing’

Examples of illicit minor syllables are shown in (4). (4a) and (b) violate the restriction that minor syllables must be light; (4c) violates the restriction that minor syllables must not bear tone; (4d–f) violate the restriction that minor syllables must not be word-final; and (4e–f) also violate the restriction that a word must have at least one major syllable.

(4)

a. *məNməwu?
b. *kʰəlou?
c. *θəyɛ:
d. *bəwɛzə
e. *zəbə
f. *zə

The terms “major syllable” and “minor syllable” seem to have been used first by Henderson (1952) for Cambodian and Shorto (1960) for Palaung.
Syllables will be discussed more fully in part 2 below.

1.2 Consonants

The consonants of Burmese are shown in (5).

(5) The consonants of Burmese

<table>
<thead>
<tr>
<th>Stops (the palatals are affricates)</th>
<th>Nasals (N is placeless)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p t c k?</td>
<td>m n n η N</td>
</tr>
<tr>
<td>pʰ tʰ cʰ kʰ</td>
<td>hₘ hₙ hₚ hₚ</td>
</tr>
<tr>
<td>b d j g</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fricatives</th>
<th>Liquids and glides</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ s f h</td>
<td>l y w (r)</td>
</tr>
<tr>
<td>sʰ</td>
<td>lʰ (wʰ)</td>
</tr>
<tr>
<td>δ z</td>
<td></td>
</tr>
</tbody>
</table>

r and wʰ are rare.

Only placeless consonants are allowed in the coda position, namely ? (which can assimilate to the onset consonant of the following syllable) and a placeless nasal N (which also can assimilate to the place of articulation of the following consonant; otherwise it is realized as nasalization on the vowel with an approximate coronal articulation after monophthongs and an approximate velar articulation after diphthongs (Bennett and Lehman 1994). Before the aspirated sonorants, ? tends to drop rather than assimilate (Okell 1969, 6–7). Examples of the assimilation of ? and N, which tends to happen only in rapid speech, are seen in (6).

---

6See Trigo (1988) for a full discussion of the behavior of placeless nasals (she calls them nasal glides) across languages.

7In this paper I transcribe words with internal ? or N in their unassimilated forms (yaʔkweʔ, pyàuNquN, etc.), chiefly to disambiguate such forms as ?eimmeʔ: ‘sleep-IRREALIS’, which is from ?eiʔmeʔ, and ?eimmeʔ: ‘at home’, which is from ?eiNmeʔ.
(6)
a. yakkwe? < yaʔkwe?
   seittŋIN < seiʔtŋIN
   lousza? < louʔza?
   myeʔnaʔ < myeʔhnaʔ
   ‘area, quarter’
   ‘opinion’
   ‘fictitious story’
   ‘face’
b. pyàʊŋkouN < pyàʊNkouN
   ṭəpyjntʰweʔ? < ṭəpyjNtʰweʔ?
   ṭəŋjeʔ? < ṭəNjeʔ?
   ‘alter completely’
   ‘go outside’
   ‘iron hook’

2. Syllables
2.1 Major syllables

As alluded to above, a word in Burmese must contain at least one major syllable, which may be defined as a syllable whose nucleus is a full vowel—i.e. any monophthong or diphthong except a. Major syllables are always bimoraic, as vowels are long in open syllables, and short in closed syllables (7).

(7)
a. kʰa:
   ‘shake’
b. kʰaN
   ‘undergo’
c. kʰa?
   ‘draw off’

Since there is no underlying vowel-length distinction in Burmese, [kʰa:] is undoubtedly the surface representation of /kʰa/; the vowel is lengthened in an open syllable in compliance with the principle of the minimum word (McCarthy and Prince 1986, 1990, 1993b, 1995), which says that a lexical word must correspond to a PrWd. The optimality-theoretic constraint for this is LX=PR (8).

(8) \( \text{LX}=\text{PR} \) (Prince and Smolensky 1993, 101)

A lexical word must correspond to a prosodic word.

A PrWd must contain at least one foot, and FOOTBINARITY (9) says that feet are binary at some level of analysis—either the syllable or the mora.
(9) **FOOTBINARITY** (Prince and Smolensky 1993, 47)

Feet are binary at some level of analysis (σ, μ).

Since Burmese is a quantity-sensitive language, it is appropriate to propose that feet in Burmese are binary at the moraic level of analysis. Therefore the word in Burmese must be at least bimoraic. Lengthening a short vowel violates the constraint **FILL** (10).

(10) **FILL** (Prince and Smolensky 1993, 25)

Syllable positions are filled with segmental material.

The ranking \( \text{LX=PR, FTBIN} \rightarrow \text{FILL} \) is shown in the tableau in (11). A PrWd is demarcated by [ ]₀, and a foot is demarcated by parentheses ( ). The first candidate in (11) is a candidate with no prosodic structure at all, which could not be realized phonetically.⁸

<table>
<thead>
<tr>
<th>(11)</th>
<th>/kʰə/</th>
<th>LX=PR</th>
<th>FTBIN</th>
<th>FILL</th>
</tr>
</thead>
<tbody>
<tr>
<td>kʰə</td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[kʰə]₀</td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ñ [kʰə]₀</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

Major syllables are also always associated with one of the four tones (12).

(12)

a. high:  
   kʰə:  ‘be bitter’
   kʰəN  ‘dry up’

b. low:   
   kʰə:  ‘shake’
   kʰəN  ‘undergo’

c. creaky: 
   kʰə:  ‘fee’
   kʰəN  ‘appoint’

d. checked:  
   kʰə?  ‘draw off’

---

⁸In optimality tableau, an asterisk (*) indicates violation of a constraint, an exclamation point (!) indicates that a violation is fatal, and a pointing finger (♀) indicates the optimal candidate. Angled brackets indicate the failure to parse an underlying segment (e.g. (x) means x is not parsed), and outlining indicates the addition of a segment that is not underlying (e.g. x means x is added).
The phonotactic restrictions on major syllables are the following: The diphthongs *ei, ai, ou, au* MUST be closed by one of the coda consonants *?* or *N* (13); the mid monophthongs *e, o, u* MUST occur in open syllables (14); *e* may occur in an open syllable or a syllable closed by *?*, but no syllable may end in *εN* (15).

(13)

a.  *?ei?*  ‘sleep’  *?eiN*  ‘house’
b.  *sʰai?*  ‘arrive’  *tʰaiN*  ‘sit’
c.  *cʰou?*  ‘sew’  *cʰouN*  ‘overspread’
d.  *cau?*  ‘stone’  *cauN*  ‘cat’

(14)

a.  *?ɛ:*  ‘be cold’
b.  *cʰo:*  ‘be sweet’
c.  *cɔ:*  ‘fry’

(15)

a.  *θwɛ:*  ‘connect by thread etc.’
b.  *θwe?*  ‘be fluent’
c.  **θwɛN*

This complementary distribution of *e, o, u* with *ei, ou, au* can be explained by prohibiting all place features from the coda of a syllable. The Coda Filter (Steriade 1982, Itô 1986, 1989, Yip 1991) was devised as a way of restricting the occurrence of features in the coda; for example, by prohibiting place features. A Coda Filter doing just this was formalized by Itô (1989) as in (16).

(16)  Coda Filter

* C ^1σ

[PLACE]
Burmese patently obeys this constraint, as \( ? \) and \( N \) are both placeless. The Coda Filter has traditionally applied only to consonants. But if the Coda Filter is extended to coda VOWELS as well, we can explain the complementary distribution of diphthongs and monophthongs seen above. Specifically, if coda vowels are barred from having place specifications, diphthongs can be excluded from open syllables, because to allow diphthongs in open syllables is to allow the place features of the second element of the diphthong to occur in coda position. An optimality-theoretic constraint called CODA-COND, excluding all place features from coda position, can be formulated as in (17).

(17) **CODA-COND**

A coda segment can have no place specification of its own at all.

Other relevant constraints here are PARSE (18) and \(*COMPLEX^{\text{Nuc}}\) (19).

(18) **PARSE** (Prince and Smolensky 1993, 85)

Underlying segments must be parsed into syllable structure.

(19) **\(*COMPLEX^{\text{Nuc}}\)** (Prince and Smolensky 1993, 87)

No more than one segment may associate to the nucleus.

Let us consider a form such as \( c^{h\varphi} \): ‘be sweet’ (14b), and assume the diphthong \( ou \) to be underlying; in this case, the structure shown in (20a) violates CODA-COND by placing the place features of \( u \) in the syllable coda. In (20b) the coda is filled by the \( o \) of the nucleus (which licenses place features), and the \( u \) simply fails to be parsed. In (20c) both elements of the diphthong are governed by the first mora, and the second mora is left empty, in violation of FILL.9

---

9According to Prince and Smolensky (1993, 50), we can assume “that unfilled moras are interpreted in the output as continuations of a tautosyllabic vowel.” Although in their example the vowel is a monophthong, we can extend this interpretation to the second vowel in the diphthong of \( l^{\text{c\varphi}}ou \); thus (b) below is an interpretation of (a). According to Prince and Smolensky’s analysis, both (a) and (b) are violations of FILL.

\[
\begin{align*}
\text{(a)} & \quad \begin{array}{c}
\sigma \\
\mu \\
\mu \\
c^{\varphi} u
\end{array} \\
\text{(b)} & \quad \begin{array}{c}
\sigma \\
\mu \\
\mu \\
c^{\varphi} u
\end{array}
\end{align*}
\]
(20)

(a) \* σ
(b) σ
(c) \* σ

As shown in the tableau in (21), the constraint ranking CODA-COND, FILL » PARSE » *COMPLEX_Nuc derives the correct results.

<table>
<thead>
<tr>
<th>(21)</th>
<th>/cʰou/</th>
<th>CODA-COND</th>
<th>FILL</th>
<th>PARSE</th>
<th>*COMPLEX_Nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20a) = cʰ[ʊ]_u[?]_μ</td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(20b) = cʰ[ʊ]_u(u)</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(20c) = cʰ[ʊ]_u[?]_μ</td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the other hand, when the diphthong is followed by a (placeless) coda consonant, as in cʰou? ‘sew’ (10c), the coda is filled with the consonant, and the diphthong is placed under a branching nucleus (22a). This violates *COMPLEX_Nuc, but other candidates (22b, c) violate the more highly ranked constraints CODA-COND and PARSE. The tableau illustrating this is seen in (23).

(22)

(a) σ
(b) σ
(c) σ

<table>
<thead>
<tr>
<th>(23)</th>
<th>/cʰou?/</th>
<th>CODA-COND</th>
<th>PARSE</th>
<th>*COMPLEX_Nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(22a) = cʰ[ʊ]_u[?]_μ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(22b) = cʰ[ʊ]_u[?]_μ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(22c) = cʰ[ʊ]_u(u)[?]_μ</td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This explanation will hold for the other monophthong/diphthong pairs, seen above in (13–14): ʔeː ~ ʔeiN, ʔeiʔ; cɔː ~ cuN, cauʔ. Thus we may conclude that the monophthongs e o ə are surface allophones of underlying /ei ou au/.

Since e and ai both occur before ʔ (e.g. ʔkɛʔ ‘offer respectfully’ ~ ʔkaiʔ ‘arrive’; ʔteʔ ‘go up’ ~ ʔtaiʔ ‘attack’), they are separate phonemes /e/ ~ /ai/, but the distinction between them is neutralized as ɛː in open syllables. It should be emphasized, however, that since no phonological or morphological process in Burmese will convert an open syllable into a closed syllable, it is impossible to determine whether words like ʔwɛɛ: ‘connect by thread’, bɛː ‘duck’, ʔzwɛː ‘table’, ʔɛː ‘district’, ʔwɛː ‘buy’, pɛː ‘peas’, etc., have underlying /e/ or /ai/. From the point of view of learnability, therefore, it may be preferable to assume that /ai/ simply does not occur in open syllables.10

We may regard the absence of /ɛN/ as an accidental gap, since the other underlying monophthongs permit a nasal coda: /iN/, /uN/, /aN/ surface straightforwardly as [iN], [uN], [aN].

The underlying full vowel system of Burmese is shown in (24). The diphthongs /ei/, /ai/, /ou/, /au/ have diphthongal allophones only in closed syllables: in open syllables they surface as [ɛː], [ɛː], [ɔː], [ɔː] respectively.

\begin{tabular}{|c|c|}
\hline
Monophthongs & Diphthongs \\
\hline
i & ei \\
u & ou \\
ɛ & ai \\
a & au \\
\hline
\end{tabular}

The complete inventory of rhymes of major syllables, including tones, is shown in (25). The high vowels i u are pronounced lax in closed syllables (iN, iʔ; uN, uʔ) (Wheatley 1987, 840), but as this is not relevant to syllable structure I shall transcribe such rhymes simply as iN, iʔ; uN, uʔ.

---

10 Historically, Burmese -ɛː has three Proto-Tibeto-Burman sources, namely *-e, *-ay, and *-a₂y, as shown, for example, by pɛː ‘peas’ < TB *be, lɛː ‘change, exchange’ < TB *lay, and rɛː ‘knead’ < TB *nay (Benedict 1972, 59 ff.)
(25) The fifty rhymes of major syllables

<table>
<thead>
<tr>
<th></th>
<th>Nonnasal rhyme</th>
<th>Nasal rhyme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>/i/</td>
<td>1. ิ</td>
<td>2. ิ</td>
</tr>
<tr>
<td>/e/</td>
<td>22. ำ</td>
<td>23. ำ</td>
</tr>
<tr>
<td>/ai/</td>
<td>(ェ)</td>
<td>(ェ)</td>
</tr>
<tr>
<td>/ei/</td>
<td>30. อ</td>
<td>31. อ</td>
</tr>
<tr>
<td>/au/</td>
<td>37. อ</td>
<td>38. อ</td>
</tr>
<tr>
<td>/ou/</td>
<td>44. อ</td>
<td>45. อ</td>
</tr>
</tbody>
</table>

2.2 Minor syllables

Minor syllables contain only the vowel ə in Burmese. They contain neither tone nor coda consonants, and may never appear in the final position of a word; a minor syllable must always be followed by a major syllable. Examples of licit and illicit minor syllables were seen above in (3)–(4); they are repeated here for reference in (26–27).

(26)

a. ကောလ် ˈkəlou?: ‘knob’
b. ပြေးဗ: ˈpəlweː: ‘flute’
c. သုဗ: ˈθəyːɡ: ‘mock’
d. ကြည် ˈkələ?: ‘be wanton’
e. သေမေူးဗ: ˈtəməyeː: ‘rice-water’
f. မောင်မော်ဗ ˈmeiNməwə?: ‘women’s clothing’

(27)

a. *မောင်မော်ဗ?
b. *ကောလ်?
c. *သုဗ?
d. *ဗော်ဗ

e. *ဇော်
f.  *zə

As we have seen, major syllables are bimoraic; it is reasonable to suppose that minor syllables are monomoraic. This supposition is borne out by the phonetic evidence of Mehner and Richter (1972–77), in which the duration of minor syllables has a range of 25–50 ms, while the duration of major syllables has a range of 150–600 ms. Following standard definitions of feet (Hayes 1995, Prince 1990, Allen 1973, inter alios), we may assume that heavy (µµ) syllables are in fact feet; they may in principle be either iambs or moraic trochees. The difference between major and minor syllables is thus to be explained in terms of foot structure.

3. Feet
3.1 The case for monosyllabic feet

Although most words in Burmese are either [L H] or [H], there is good reason to suppose that the foot in Burmese is not an iamb—either (L ¹H) or (¹H)—but rather a single heavy syllable (¹H) only. In this case, light syllables in [L H], [L L H], and [H L H] words would remain unfooted, as shown in (28).

(28)  Monosyllabic feet in Burmese

a.  (bèː)  ‘duck’  d.  zə(bwɛː)  ‘table’
b.  (pàN)  ‘flower’  e.  təmə yatɛː)  ‘rice-water’
c.  (niʔ)  ‘year’  f.  (mèN)mə(wuʔ)  ‘women’s clothing’

As mentioned above, the nonoccurrence in Burmese of [L] words can be explained by invoking the principle of the minimum word.

The descriptive fact that ə is unfooted in Burmese is reminiscent of Cohn and McCarthy’s (1994) analysis of Indonesian, where ə resists footing, where possible, by virtue of a constraint NON-Foot(ə), which states that a syllable headed by ə has no metrical projection, in effect prohibiting the inclusion of ə in a foot (29).

(29)  NON-Foot(ə) (Cohn and McCarthy 1994, 21)

Schwa-headed syllables have no metrical projection.
This constraint is also relevant in Burmese, since all light syllables contain ə in their nuclei, and all light syllables are unfooted. Thus, leaving aside the issue of tone, the full prosodic structures of the words in (28) above are as shown in (30).

(30) a. Ft b. Ft c. Ft

\[ \sigma \]

\[ \mu \mu \]

\[ b \hat{e} \]

\[ \mu \mu \]

\[ p \hat{a} N \]

\[ 'n i ? \]

d. Ft e. Ft f. Ft Ft

\[ \sigma \]

\[ \mu \mu \]

\[ z \hat{e} b \hat{e} \]

\[ \mu \mu \]

\[ t' \hat{e} m \hat{e} y \hat{e} \]

\[ m \hat{e} i N m \hat{a} w u ? \]

Now, according to Grouping Harmony (Prince 1990), for both iambs and moraic trochees, (L L) is as harmonious a foot as (H); the two should be equivalent. But although a Burmese word may consist of a single heavy syllable, as shown above, no Burmese word consists of two light syllables; *zəbə is not a possible word, and therefore presumably not a possible foot. One result of this phenomenon is that it is impossible to determine whether the (H) foot of Burmese is a trochee or an iamb. If (L L) feet were allowed, the issue would be resolved by the placement of stress (or tone): (l'L L) would mean a trochaic system, (L 'L) an iambic system. But (l'H) alone is ambiguous. Before addressing the issue of why (L L) feet are prohibited in Burmese, I shall first argue against an iambic analysis, and then discuss compound words and the PrWd.

3.2 The case against iambs

At first glance, it may appear that Burmese has iambic feet. The Grouping Harmony Principle of Prince (1990) shows that iambs are preferably of the shape (L 'H), otherwise

---

11 In a future version of this paper, I hope to show that tone in Burmese is licensed by the foot.
(L 'L) or (H). These foot shapes correspond to his Weight-to-Stress Principle, which says, 
"if heavy, then stressed," and to its contraposition, "if unstressed, then light." (In 
Burmese and Thai—discussed below—tone-bearing syllables fill the role of stressed sylla-
bles.) As mentioned in the discussion of major and minor syllables above, Burmese al-
 lows words of the shape [L 'H] and ['H], but not [L 'L]. If Burmese is to be analyzed as 
iambic, the nonoccurrence of [L L] words (*zaaba) must be explained.

At least two other Southeast Asian languages have been described as iambic: Thai by 
Bennett (1994), and Cambodian by Griffith (1991) and Hayes (1995). Both languages, 
like Burmese, allow [L 'H] and ['H] words, and disallow [L 'L] words. Thai at least also 
allows [L L 'H] words, just as Burmese does; if Cambodian allows [L L 'H] words, Grif-
fith does not mention them. It seems desirable that a theory to explain the behavior of feet 
in one of these languages should be able to explain all three languages as uniformly as 
possible.

Griffith’s (1991) discussion of Cambodian shows how iambic feet can be built on 
words, as shown in (31) (= Griffith’s (6)). In these Cambodian words, vowel length is in-
dicated by doubling the vowel letter; primary stress is marked by ‘; secondary stress by ’;
aa, etc., indicates a long (bimoraic) diphthong; o indicates a short (monomoraic) diph-
thong. Note that while Burmese and Thai use tone to indicate the head syllables of words, 
Cambodian (which does not have tones per se) uses stress. (The stress marks in (31) are 
as given by Griffith. I do not know why some syllables with a single grid mark are 
marked as bearing primary stress, while others are marked as bearing secondary stress, 
and others are unmarked for stress. But the point is that every syllable with two grid 
marks bears primary stress.)

(31) (X) (X) (X) (X) (X)
     (X) (X) (X) (X) (X)
     — — — — —
bat chii baŋkāet siəwphiw
‘to close’ ‘to be ill’ ‘to originate’ ‘book’

(X) (X) (X) (X)
     (X) (X) (X)
     — — — — —
būŋ-sūŋ thōmmedáa pūttaśāhnaa
‘to pray’ ‘ordinary’ ‘Buddhism’
While Griffith’s analysis accounts for the attested forms, she does not address why (L ¹L) iambs are not encountered in Cambodian.

Bennett's (1994) optimality-theoretic analysis of Thai reveals that Thai has not only [¹H] and [L ¹H] words, as shown in (32), but underlying [¹L] and [L ¹L] words as well. These latter type are augmented by a coda ʔ to become surface [¹H] and [L ¹H], as shown in (33). In the Thai examples, ́, ʰ, and ʔ mark high, falling, and low tones, respectively; mid tone is unmarked.

(32)

a. /tôn/  [tôn]  ‘stem’
b. /taa/  [taa]  ‘eye’
c. /tɔɔ/  [tɔɔ]  ‘to reply’
d. /tapuu/  [tapuu]  ‘nail’
e. /sanùk/  [sanùk]  ‘fun’
f. /lamùt/  [lamùt]  ‘sp. fruit’
g. /pratʰêet/  [pratʰêet]  ‘country’

(33)

a. /tÔ/  [tÔ?]  ‘table’
b. /dù/  [dù?]  ‘fierce’
c. /kɔɔ/  [kɔɔ?]  ‘island’
d. /katʰi/  [katʰi?]  ‘coconut milk’
e. /pracù/  [pracù?]  ‘to fill up’
f. /rajá/  [rajá?]  ‘space, period’

Bennett proposes that this augmentation takes place in order to comply with the Stress-to-Weight principle, which says, “if stressed, then heavy.” Bennett attributes this principle to Prince (1990), but actually, Prince is arguing for the converse: “if heavy, then
stressed" (the Weight-to-Stress Principle, or WSP). He says (358), “we specifically deny” that “if stressed, then heavy” is a principle. On the other hand, Hayes (1995) argues that (L 1′L) iambs violate his “Iambic/Trochaic Law”—actually more of a functionally based tendency—which says, in essence, that the strong and weak syllables of iambs should contrast in duration, while the strong and weak syllables of trochees should contrast in intensity. In other words, (L 1′H) is a better iamb than (L 1′L), but (1′L L) is a better trochee than (′H L). Prince’s (1990) Grouping Harmony principle makes the same prediction, but ranks (H) equal to (L L) for both iambs and trochees.

Hayes (1995, 83 and 205 ff.) lists some two dozen languages with iambic lengthening rules that change (L 1′L) feet into (L 1′H) feet. Of course, it is not obligatory in iambic languages that (L 1′L) feet get lengthened to (L 1′H); Hayes (1995, 64 ff. and 211 ff.) lists Seminole/Creek and Unami and Munsee Delaware as iambic languages with iambs of all three possible shapes: (L 1′H), (1′H), and (L 1′L).

The case of [L L L] words in Thai is interesting: one might think that the natural footing (with iambic ?-augmentation) would be either [(L H)(H)] or [(H)(L H)]. But neither of these is the case; the actual result (following Bennett’s iambic analysis) is [L (L H)], with the first syllable left unfooted, as shown in (34).

(34) /kʰarəhá/ ‘fortune’
* [kʰaraʔhá?]
* [kʰaʔrahá?]
 [kʰarəhá?]

Bennett proposes that [L (L H)] is preferable to either [(L H)(H)] or [(H)(L H)] because of FtBIN. Bennett proposes that the foot in Thai is binary at the syllabic level, so that the monosyllabic (H) feet in [(L H)(H)] and [(H)(L H)] violate FtBIN.¹² But this analysis is theoretically worrisome: Bennett is proposing that although Thai is a weight-sensitive language (as proved by his Stress-to-Weight Principle), Foot Binarity operates on the syllabic level. Prince and Smolensky allowed Foot Binarity at the syllabic level to account for syllabic trochees, which are of the shape (′σ σ) and tend to occur in quantity-insensitive languages (though Hayes 1995, 102 lists ten exceptions). Neither Hayes

¹² Monosyllabic words like tό ‘table’ and taa ‘eye’ would thus also violate FtBIN, but satisfaction of FtBIN in these cases, for example by *ʔatό? and *ʔataa, would violate a higher-ranked constraint against epenthetic syllables (FILL-σ).
(1995) nor Prince (1990) permits “syllabic iambs”, as such would seem to entail quantity-insensitive iambic systems, which Hayes argues do not exist. Thus Bennett would seem to be taking unfair advantage of optimality theory’s chief flaw—that it is too powerful. There is nothing intrinsic in the OT framework that prevents FtBIN from operating at the syllabic level in languages with iambic systems, yet Hayes’ results indicate that such a restriction is desirable.

The facts in Cambodian and Thai can be accounted for with a theory of monosyllabic (H) feet, as I proposed for Burmese above, without the problems of Bennett’s iambic analysis. The -augmentation of underlying [L] and [L L] words in Thai can be seen as compliance with FtBIN at the moraic level—theoretically desirable in quantity-sensitive languages. Examples (35–36) show monosyllabic feet in Cambodian and Thai.

(35) Monosyllabic feet in Cambodian
   a. (bat) ‘to close’
   b. kα(káay) ‘to scratch about’
   c. pra(kán) ‘to object’

(36) Monosyllabic feet in Thai
   a. (tó?) ‘table’
   b. ka(tʰi?) ‘coconut milk’
   c. (taa) ‘eye’
   d. kʰara(há?) ‘fortune’
   e. ta(puu) ‘nail’

Although iambic lengthening in Thai is unnecessary under this analysis, and although I am skeptical of Bennett’s “Stress-to-Weight Principle,” it should not be thought that I am arguing against iambic lengthening as a valid process. Hayes (1995, 205 ff.) gives many examples of languages with iambic lengthening, e.g. Hixkaryana, Choctaw/Chickasaw, Cayuga, and several of the Yupik languages. Hung (1994, 57 ff., 117–9) discusses Hixkaryana and Choctaw in optimality-theoretic terms. She proposes a constraint IAMBIC QUANTITY that says, “In a rhythmic unit (W S), S is heavy” (63), to achieve iambic leng-

---

13Hayes (1995, 266–8) shows that apparent cases of syllabic iambs, as in Southern Paiute, Araucanian, Onondaga, and Dakota, are actually cases of (L L) iambs in languages without heavy syllables; this absence of heavy syllables precludes the possibility of (L H) or (H) iambs.
thening. This is less general than Bennett’s Stress-to-Weight Principle, but in the case of a (W S) iamb, the two have the same result. Following Hung’s account of Hixkaryana, one could argue for iambicity in Thai thus: FtbIn applies at the moraic level, augmenting /tô/ ‘table’ to [(tô)]; IAMBIC QUANTITY will then augment /katʰI/ ‘coconut milk’ to [(katʰI)]. Bennett’s theoretically tenuous Stress-to-Weight Principle is thus unnecessary. Nevertheless, I still believe that the only well-formed foot in Thai may well be a single heavy syllable, as it is in Burmese.

4. Compounds, superlong words, and the prosodic word

Most Burmese words either are monosyllabic, or have an initial minor syllable followed by a major syllable. In either case, the word can be seen as consisting of one PrWd and one foot (37).

(37)

a. [(θwätə)]ω  ‘go’
b. [(lɛtə)]ω  ‘be heavy’
c. [(sɑʔtə)]ω  ‘writing’
d. [(ʔɛiN)]ω  ‘house’
e. [θə(yɛʔ)]ω  ‘mock, satirize’
f. [kʰθ(louʔ)]ω  ‘knob’

Words of more than one foot (henceforth, polypodic words) are chiefly compounds (38a) or loanwords (38b) in Burmese. Very few polypodic words are apparently neither compounds nor loanwords (38c). I shall refer to noncompound polypodic words, such as (38b–c), as “superlong” words. I shall discuss the status of the PrWd in polypodic words later.

(38)

a. (nɛt)(tʰaiN) = (nɛt) + (tʰaiN)  ‘reside’ = ‘stay’ + ‘sit’
b. (cʰɛʔkə(lɛʔ))  ‘chocolate’ < English
c. (mʊtN)(daiN)  ‘storm’
4.1 Nonreducing compounds

There are two types of compounds in Burmese; I call them NONREDUcing and REDUCing compounds. Nonreducing compounds are straightforward: two or more words are strung together to form a single word. The individual members of these compounds probably retain their original PrWds, which are then parsed recursively into a single, larger PrWd, as shown in (39).

(39) Nonreducing compounds: \( \omega + \omega > [\omega \omega]_\omega \)

a. \([\text{məs}]_\omega + [(\text{tʰa} \text{ίN})]_\omega > [\text{məs}](\text{tʰa} \text{ίN})]_\omega\]
   ‘stay’ + ‘sit’ > ‘reside’

b. \([\text{yàN}]_\omega + [(\text{wɛ} \text{ɛ})]_\omega > [\text{yàN}](\text{wɛ} \text{ɛ})]_\omega\]
   ‘sell’ + ‘buy’ > ‘trade’

c. \([\text{cɛʔ}]_\omega + [(\text{sʰ} \text{ɛN})]_\omega > [\text{cɛʔ}](\text{sʰ} \text{ɛN})]_\omega\]
   ‘fowl’ + ‘elephant’ > ‘turkey’

d. \([\text{ʔə}(\text{yɛ} \text{ɛ})]_\omega + [(\text{ʔə} \text{cʰ} \text{ɛN})]_\omega > [\text{ʔə}(\text{yɛ} \text{ɛ})](\text{ʔə} \text{cʰ} \text{ɛN})]_\omega\]
   ‘qualification’ + ‘quality’ > ‘standard’

e. \([\text{pʰɛz}]_\omega + [(\text{pʰ} \text{ɛz})]_\omega + [(\text{kʰ} \text{a} \text{ίN})]_\omega > [\text{pʰɛz}](\text{pʰ} \text{ɛz})(\text{kʰ} \text{a} \text{ίN})]_\omega\]
   ‘send’ + ‘to’ + ‘tell’ > ‘tell (him) to send (it)’

f. \([\text{cʰi} \text{ʃ}]_\omega + [(\text{lɔ} \text{ʰi})]_\omega + [(\text{kàN})]_\omega > [\text{cʰi} \text{ʃ}](\text{lɔ} \text{ʰi})(\text{kàN})]_\omega\]
   ‘look’ + ‘-ing’ + ‘be good’ > ‘be good to look at’

g. \([\text{ka} \text{ʔ}]_\omega + [(\text{pɛz})]_\omega + [\text{ʔə}(\text{θi} \text{ʔi})]_\omega + [\text{ʔə}(\text{həN})]_\omega > [\text{ka} \text{ʔ}](\text{pɛz})(\text{θi} \text{ʔi})(\text{həN})]_\omega\]
   ‘paddy’ + ‘peas’ + ‘fruit’ + ‘grain’ > ‘crops’

h. \([\text{ʔə}(\text{ɛz})]_\omega + [(\text{ʔɛN})]_\omega + [(\text{kʰwe} \text{ʔ}]_\omega + [(\text{yəN})]_\omega > [\text{ʔə}(\text{ɛz})](\text{ʔɛN})(\text{kʰwe} \text{ʔ})(\text{yəN})]_\omega\]
   ‘pot’ + ‘bowl’ + ‘cup’ + ‘ladle’ > ‘household goods’

(I am not concerned here with certain effects of compounding, such as voicing, as seen in (39e), or the loss of the prefix ʔə in some forms (39g), but not in others (39d).)

---

14 It is not unknown for a language to have more than one type of compound. Mohanan (1982, 1986), Aronoff and Sridhar (1983), Sproat (1986), and Inkelas (1989) discuss compounds in Kannada and Malayalam, where “sub-compounds” and “co-compounds” have different phonological effects from each other. Unlike Kannada and Malayalam, Burmese does not seem to have an obvious semantic distinction between the two types of compounds.

Examination of the glosses in the examples will show that many “compounds” in Burmese are not compounds in the traditional sense at all, but rather “concatenation[s] of lexical words and grammatical formates, presumably under a single X-bar category (X0)” (Bennett and Lehman 1994, where the reader is referred also to Lehman 1986 and 1993).

15 See Inkelas (1989) and McCarthy and Prince (1993a) on the recursiveness of the PrWd.
Compound words consisting of more than one PrWd are well attested: in Igbo (Zsiga 1992), Malayalam (Sproat 1986, Inkelas 1989), Sanskrit (Selkirk 1980), and Turkish and Hungarian (Nespor and Vogel 1986), for example, certain compounds must contain more than one PrWd. Also in the history of Welsh, as described by Jackson (1953), the boundary between members of a compound word functions like any other word boundary.¹⁶ Further evidence that the members of nonreducing compounds in Burmese keep their original PrWds will be presented in section 5 below.

4.2 Reducing compounds

In reducing compounds, the penultimate member of the compound becomes a minor syllable, losing its tone and its second mora (whether vocalic or consonantal), and reducing the vowel quality to a. Prosodically speaking, in reducing compounds, two PrWds are combined into one monopodic PrWd (40a-e). Three PrWds are combined into two monopodic PrWds, which may themselves be part of a single, larger PrWd (40f-i).

(40) Reducing compounds: ω + ω > ω

a. [(càN)]ω + [(pòr)]ω > [cà(bòr)]ω
   ‘floor’ + ‘insect’ > ‘bug’

b. [(ηàz)]ω + [(?]ú:]ω > [ηà(?]ú:)]ω
   ‘fish’ + ‘egg’ > ‘fish-spawn’

c. [(θwà:]ω + [(yfz)]ω > [θà(yfz)]ω
   ‘tooth’ + ‘juice’ > ‘saliva’

d. [(tɔ(mÌN)]ω + [(yfz)]ω > [tɔmà(yfz)]ω
   ‘rice’ + ‘water’ > ‘rice-water’

e. [kÀ(là:)]ω + [(pyfz)]ω > [kàlà(byfz)]ω
   ‘Indian’ + ‘country’ > ‘India’

f. [(lu?)]ω + [(piz)]ω + [(là:)]ω > [(lu?)]ω[pà(là:)]ω
   ‘free’ + ‘is’ + Q > ‘is (he) free?’

g. [(nàfz)]ω + [(mÀfz)]ω + [(lÀ:)]ω > [(nàfz)]ω[mà(lÀ:)]ω
   ‘stay’ + IRREALIS + Q > ‘will (he) stay?’

¹⁶Specifically: the final vowel of the first member drops just like the final vowel of a word, but there is no general rule of internal syncope; also certain consonants (g, s, y, w) pattern at the beginning of the second member of a compound as they pattern at the beginnings of words, rather than as they regularly pattern in word-internal position. So, for example, the Proto-Brittonic sequence *-ogn- developed into Welsh -oent- internally (*ognos > oen ‘lamb’), but *-o+gn- became -ə+n- (*oino+gne > un+ne ‘having one color’, like *eso gne > ei ne ‘his color’).
h. \([(\text{mei}N)]_{\omega} + [(\text{má}z)]_{\omega} + [(\text{wu}?)_{\omega}] > [((\text{mei}N))_{\omega}[(\text{má}z)(\text{wu}?)]]_{\omega}\)

'woman' + fem. + 'clothing' > 'women's clothing'

i. \([(\text{lánN}])_{\omega} + [(\text{má}z)]_{\omega} + [(\text{tǎ:p})_{\omega}] > [((\text{lánN}))_{\omega}[(\text{má}z)(\text{tǎ:p})]]_{\omega}\)

'road' + 'main' + honorific > 'main road'

Whether a compound will be of the reducing or nonreducing type cannot be determined phonologically. There is no phonological reason why (34a) 'reside' must be \(nɛ\text{-}\text{thq}i\text{n}\), not *\(nɛ\text{thqi}N\), nor why (35a) 'bug' must be \(cə\text{bdo}\); not *\(cə\text{nbd}\). For our purposes, we may assume that \(nɛ\text{thqi}N\) is generated with two PrWds, and \(cə\text{bdo}\) with one.

4.3 Superlong words

There are in Burmese very few superlong words, by which I mean morphologically simplex (i.e. not compound) polypodic words. Most but not all of them are loan words (41).

(41)

a. \(cə\text{unca}\) 'be anxious'

b. \(mʊ\text{ndàiN}\) 'storm'

c. \(t\text{ʰa:pañ}\) 'enshrine' < Pāli \(p\text{ha:pa}\)

d. \(bʊ\text{dá}\) 'Buddha' < Pāli

e. \(?ə\text{pari}s\text{ey}i?\) 'appreciate' < English

f. \(c\text{ʰs:ke}\text{le}\) 'chocolate' < English

It is unclear at first whether these superlong words have a single PrWd, or whether they are like the polypodic compounds in (39) and (40f–i), each foot being parsed into a PrWd, and then the PrWds being parsed into a larger PrWd. If this latter suggestion is true, the words in (41) have the prosodic structures in (42).

(42)

a. \([(\text{c\text{ʰ}uN})_{\omega}[(\text{c\text{ʰ}a})_{\omega}]]_{\omega}\)

b. \([(\text{m\text{ʰ}uN})_{\omega}[(\text{d\text{ʰ}a})_{\omega}]]_{\omega}\)

c. \([(t\text{ʰ}a:)_{\omega}[(p\text{ʰ}a(n\text{ʰ}a:)_{\omega}]]_{\omega}\)

d. \([(bʊ?)_{\omega}[(\text{d\text{ʰ}a})_{\omega}]]_{\omega}\)

e. \([(?ə\text{p\text{ʰ}a}(r\text{ʰ}i:)_{\omega}[(ə(y\text{e}?)_{\omega}]]_{\omega}\)

f. \([(c\text{ʰ}s:)_{\omega}[(r\text{ʰ}a(\text{le}?)_{\omega}]]_{\omega}\)
Although there is no direct evidence for the internal, monopodic PrWds, we shall see below that this analysis is theoretically advantageous.

5. Maximum size of prosodic categories

Let us next consider the maximum size of the prosodic categories word, foot, and syllable in Burmese. We have seen that the PrWd, in most cases at least, takes just one foot: The vast majority of words in Burmese consist either of a single foot (43a–c), or an unfooted minor syllable followed by a foot (43d–f). We may therefore make the generalization stated in (44).

(43)

a. \[ [(mɛt)]_o \]  ‘girl’

b. \[ [(fəuʔ)]_o \]  ‘address (a superior)’

c. \[ [(pəɛɨn)]_o \]  ‘be stupid’

d. \[ [(kʰə(ʔəuʔ))]_o \]  ‘knob’

e. \[ [(pə(ʔəuʔ))]_o \]  ‘flute’

f. \[ [(θə(ʔəuʔ))]_o \]  ‘mock’

(44) A PrWd preferably contains at most one foot.

We have also seen that the foot in Burmese consists of just one syllable. Thus, while in theory bimoraic feet may be either of the form (H) or of the form (L L), in Burmese (L L) feet are not allowed: there are no words of the shape *zəʔə. We can therefore make the generalization stated in (45).

(45) A foot contains at most one syllable.

Let us now examine the possibility of collapsing the generalizations stated in (44) and (45) into a single generalization that a prosodic category contains no more than one of the next lower prosodic category. This generalization may be stated as an optimality-theoretic constraint named UNARITY, as in (46).
(46) **UNARITY**

A prosodic category $\pi$ contains no more than one of the next lower prosodic category $\pi - 1$.

This constraint depends, of course, on the prosodic hierarchy (47), as discussed in Selkirk (1980), McCarthy and Prince (1986, 1993a), Nespor and Vogel (1986), Zec (1988), Inkelas (1989), and elsewhere.

(47) **Prosodic Hierarchy**

```
PrWd
   | Ft
   |   σ
   |   μ
```

The most striking prediction UNARITY makes is that syllables should be monomoraic; as we saw in (43), most syllables are in fact major syllables, and therefore bimoraic. But we can propose that FTBIN outranks UNARITY in Burmese, and that a syllable may be bimoraic only in order that FTBIN not be violated. UNARITY predicts that unfooted syllables must be monomoraic, since FTBIN is not an issue, and we have seen that unfooted syllables in Burmese are indeed all minor and therefore monomoraic. This is the reason why the first syllable in a reducing compound such as $\text{cəbə}: \text{'bug' (40a)}$ is reduced to $\text{ə}$: a bimoraic syllable would induce an extra violation of UNARITY at the syllabic level, and since this word has a single PrWd, footing the first syllable violates UNARITY at the PrWd level. Leaving syllables unfooted violates PARSE-$\sigma$ (48), but this constraint is not ranked high.

(48) **PARSE-$\sigma$**

Syllables are parsed into feet.

UNARITY is not a single constraint so much as a family of constraints: UNARITY$^{\text{PrWd}}$, UNARITY$^{\text{Ft}}$, and UNARITY$^{\sigma}$ can—in this instance in fact, must—have separate rankings.
The ranking UNARITY\textsuperscript{PrWd,Ft}, FTBIN \textgreater\textless UNARITY\textsuperscript{σ}, PARSE-\textgreek{σ} is shown in the tableau in (49). The relative ranking of UNARITY\textsuperscript{σ} and PARSE-\textgreek{σ} cannot be determined.

<table>
<thead>
<tr>
<th>(49) Candidates</th>
<th>UNARITY\textsuperscript{PrWd}</th>
<th>UNARITY\textsuperscript{Ft}</th>
<th>FTBIN</th>
<th>UNARITY\textsuperscript{σ}</th>
<th>PARSE-\textgreek{σ}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textgreek{cə} [cə(bəc)]\textsubscript{0}</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>[cə(bə)]\textsubscript{0}</td>
<td></td>
<td></td>
<td>* !</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>[(cəbə)]\textsubscript{0}</td>
<td></td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[cəN(bəc)]\textsubscript{0}</td>
<td></td>
<td></td>
<td></td>
<td>** !</td>
<td>*</td>
</tr>
<tr>
<td>[(cəN)(bəc)]\textsubscript{0}</td>
<td></td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UNARITY\textsuperscript{Ft} and FTBIN both appear to be unviolated in Burmese: there is no circumstance under which either a (L L) foot or a (L) foot is permissible. If my analysis of reducing compounds is correct, UNARITY\textsuperscript{PrWd} is certainly ranked above UNARITY\textsuperscript{σ}, and may be unviolated, as implied by the tableau in (49).

We are now in a position to determine the prosodic structure of superlong words. The tableau in (49) shows that the candidate [(cəN)(bəc)]\textsubscript{0} fails in comparison with [cə(bəc)]\textsubscript{0}. Therefore, a superlong word like (41b) 'storm' must be generated as [([(məuN)]\textsubscript{0}-[(dəiN)]\textsubscript{0})\textsubscript{0}], since a candidate [([məuN])(dəiN)]\textsubscript{0} would fail in favor of *[mə(dəiN)]\textsubscript{0} by virtue of UNARITY\textsuperscript{PrWd}.

6. Other constraints

In order to account fully for what is and is not a permissible word in Burmese, some other constraints must be discussed.

As mentioned in the discussion of major and minor syllables (cf. especially (27d)), no word may end in a minor syllable: *\textgreek{bwɛzə} is not a permissible word. The constraint ALIGN-R (50) says that the right edge of every PrWd corresponds to the right edge of some foot. M-PARSE (51) is violated when there is no output for a certain form generated by Gen. This absence of output is also called the null parse.

(50) \textbf{ALIGN-R} (McCarthy and Prince 1993b, 32)

Align(PrWd, Right, Foot, Right)
(51) **M-Parse** (McCarthy and Prince 1993b, 112)

Morphemes are parsed into morphological constituents.

Like FTBIN and UnarityR, ALIGN-R is undominated and unviolated. This means that every PrWd must end in a foot—in other words, must end with a heavy syllable. The null parse—the candidate in which there is no morphological structure, and no phonetic realization—violates no constraints except M-Parse, and is optimal only when all other candidates violate undominated constraints. The null parse is indicated by angled brackets (\(\langle\)). The tableaux in (52)-(56) show the interaction of these constraints. (52) shows that an open syllable must be made long, in violation of FILL, in order to meet FTBIN.

<table>
<thead>
<tr>
<th>(52)</th>
<th>/bê/</th>
<th>FTBIN</th>
<th>ALIGN-R</th>
<th>NF(ə)</th>
<th>UNFr</th>
<th>M-Parse</th>
<th>FILL</th>
<th>UNσ</th>
<th>PARSE-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{ər} [(bê)])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( [(bê)] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In (53), we see that because a syllable with ə cannot be footed, the null parse is the optimal candidate for a word whose only vowel is ə.

<table>
<thead>
<tr>
<th>(53)</th>
<th>/bə/</th>
<th>FTBIN</th>
<th>ALIGN-R</th>
<th>NF(ə)</th>
<th>UNFr</th>
<th>M-Parse</th>
<th>FILL</th>
<th>UNσ</th>
<th>PARSE-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( [(bə)] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( [(bəɛ)] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* !</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{ər} \langle bə \rangle )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(54) shows that a [L L] word cannot be footed (L L) because of UnarityR; since the last syllable contains a full vowel (i.e. not ə), it can be lengthened in violation of FILL in order to be footable.

<table>
<thead>
<tr>
<th>(54)</th>
<th>/zəbwɛ/</th>
<th>FTBIN</th>
<th>ALIGN-R</th>
<th>NF(ə)</th>
<th>UNFr</th>
<th>M-Parse</th>
<th>FILL</th>
<th>UNσ</th>
<th>PARSE-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{ər} [zə(bwɛɛ)] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
| \( [(zəbwɛɛ)] \) | | | | | | * ! | | | *
| \( [zə(bwɛ)] \) | | | | | | * ! | | | |
(55) shows that because of the unfootability of ə, a [L L] word in which both syllables contain ə can never surface; the null parse is optimal.

<table>
<thead>
<tr>
<th>(55)</th>
<th>/zəbə/</th>
<th>FTBIN</th>
<th>ALIGN-R</th>
<th>NF(ə)</th>
<th>UNF</th>
<th>M-PARSE</th>
<th>FILL</th>
<th>UNσ</th>
<th>PARSE-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>[(zəbə)]₀</td>
<td></td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>[zə(bə)]₀</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>[zə(bə)]₀</td>
<td></td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>*</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>[(zə)bə]₀</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>ə əzəbə</td>
<td></td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
</tbody>
</table>

In (56) we see that ALIGN-R means that any word that ends with ə can never surface; again, the null parse is optimal.

<table>
<thead>
<tr>
<th>(56)</th>
<th>/bwəzə/</th>
<th>FTBIN</th>
<th>ALIGN-R</th>
<th>NF(ə)</th>
<th>UNF</th>
<th>M-PARSE</th>
<th>FILL</th>
<th>UNσ</th>
<th>PARSE-σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>[(bwəzə)]₀</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>[(bwəzə)]₀</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>[bwə(zə)]₀</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>ə əbwəzə</td>
<td></td>
<td></td>
<td>!</td>
<td>!</td>
<td></td>
<td>!</td>
<td>!</td>
<td>!</td>
<td></td>
</tr>
</tbody>
</table>

7. Conclusions

We have discussed several aspects of the prosodic structure of Burmese, and have addressed several problems. The complementary distribution of diphthongs and mid monophthongs is explained by hypothesizing that the Coda Condition in Burmese applies to both vowels and consonants; this prohibits diphthongs from occurring in open syllables.

The constraint UNARITY° prefers a syllable to be monomoraic, but FTBIN forces a foot to be bimoraic. Since FTBIN dominates UNARITYσ, a footed syllable must be bimoraic. The constraint ALIGN-R forces a word to end with a foot. These facts explain the distribution of major and minor syllables.

The constraint UNARITYPr forces a foot to be monosyllabic, which is why (L L) feet do not occur in Burmese. UNARITYPrWd forces a PrWd to have only one foot; therefore, words with more than one foot—even the superlong ones seen above in (41)—have more than one PrWd.
The constraints we have discussed are listed in (57) according to their relative ranking. In (58) are listed the specific rankings that have been established, along with the example number where the ranking is proved.

(57) Unviolated constraints:

ALIGN-R,
CODA-COND,
FtBIN,
Lx=PR,
NON-FOOT(ə)
UNARITYR,
UNARITYPrWd

Violable constraints:

» M-Parse
» PARSE,
» FILL
» *COMPLEXNuc
» PARSE-σ,
UNARITYσ

(58) ALIGN-R » M-Parse (55)
CODA-COND » *COMPLEXNuc (23)
CODA-COND » PARSE (21)
FILL » PARSE (21)
FtBIN » FILL (11)
FtBIN » M-Parse (56)
FtBIN » UNARITYσ, PARSE-σ (49)
Lx=PR » FILL (11)
NON-FOOT(ə) » M-Parse (53)
NON-FOOT(ə) » UNARITYσ, PARSE-σ (54)
M-Parse » FILL (52)
PARSE » *COMPLEXNuc (23)
UNARITYR » M-Parse (53)
UNARITYR » UNARITYσ, PARSE-σ (49)
UNARITYPrWd » UNARITYσ, PARSE-σ (49)
8. References


Cohn, A.C. and J.J. McCarthy (1994) Alignment and parallelism in Indonesian phonology. Ithaca and Amherst: Cornell University and University of Massachusetts, Amherst, ms.


McCarthy, J.J. and A.S. Prince. (1986) *Prosodic morphology*. Amherst and Waltham: University of Massachusetts, Amherst and Brandeis University, ms.


McCarthy, J.J. and A.S. Prince (1993a) *Generalized alignment*. Amherst and New Brunswick: University of Massachusetts, Amherst and Rutgers University, ms.


Speaker normalization in the perception of Mandarin Chinese tones

Corinne B. Moore

This study investigated speaker normalization in perception of Mandarin Tone 2 (mid-rising) and Tone 3 (low-falling-rising) by examining listeners' use of F0 range as a cue to speaker identity. Two speakers were selected such that Tone 2 of the low-pitched speaker and Tone 3 of the high-pitched speaker occurred at equivalent F0 heights. Production and perception experiments determined that Turning Point (or inflection point of the tone), and ΔF0 (the difference in F0 between onset and Turning Point) distinguished the two tones. Three tone continua varying in either Turning Point, ΔF0, or both acoustic dimensions, were then appended to a natural precursor phrase from each of the two speakers. Results showed identification shifts such that identical stimuli were identified as low tones for the high precursor condition, but as high tones for the low precursor condition. Stimuli varying in Turning Point showed no significant shift, suggesting that listeners normalize only when the precursor varies in the same dimension as the stimuli. The magnitude of the shift was greater for stimuli varying only in ΔF0, as compared to stimuli varying in both Turning Point and ΔF0, indicating that normalization effects are reduced for stimuli more closely matching natural speech.

1. Introduction

This study examines the role of speaker-dependent F0 information in the perception of lexical tone. It is well known that the perception of segments requires listeners to normalize to reduce overlap among phonetic categories. A classic example of this overlap is illustrated by formant frequency data for vowels in which two phonetic categories from different speakers have similar formant values (Peterson and Barney 1952). Listeners adjust to this acoustic variability, caused by differences in speaker vocal tract size, in order to identify segments accurately. Consequently, segments that have identical acoustic characteristics may not be perceived identically.

1.1 Speaker Normalization: The use of speaker-specific acoustic information in vowel perception

Previous work on normalization has shown that listeners use acoustic information outside of the speech sound itself (extrinsic information) about speaker identity in order to classify vowels. For example, Peterson and Barney (1952) found that perception of vowel tokens produced by a wide variety of speakers exhibited confusion within the areas of overlap in the vowel formant data. Ladefoged and Broadbent (1957) provided evidence that listeners refer to extrinsic information specifying the vowel space of a speaker. In this study, six versions of the phrase please say what this word is were synthesized, in which the formant frequencies of each version were manipulated to represent different speakers. In addition, four test words of the form, "b_t" were
synthesized with F1 and F2 values of the vowel approximately corresponding to the vowels in the words "bit", "bet", "bat", and "but". Subjects were asked to identify the test words in one of the precursor phrases. Results demonstrated that identification of the vowel stimuli changed depending on which version of the precursor phrase preceded them, and the change was in accordance with predictions about the relationship between F1 and F2 in the target words and the precursors. For example, one test word was identified as bit by 87% of the subjects when it occurred after one particular version of the precursor phrase, but as bet by 90% of subjects when it was preceded by a version of the precursor with a lower F1 value. The change of identification is predictable if the F1 in the stimulus is perceived according to the F1 in the precursor, since a lower F1 would generate the percept of a higher F1 in the target stimulus, and [e] is known to have a higher F1 than [i]. Because identical formant values were perceived differently, according to the formant values in the precursors, this study has been used to support the hypothesis that vowels are perceived according to the vowel space of the speaker.

While Ladejoged and Broadbent's results indicate what acoustic information is used in speaker normalization, it remained unclear how this information is used. In particular, acoustic information may either be used to identify the speech sound directly, or it may be used as a cue to speaker identity, establishing a representation against which acoustic characteristics may be calibrated. This question was investigated by Johnson (1990), for vowel continua. In particular, acoustic information may either be used to identify the speech sound directly, or it may be used as a cue to speaker identity, establishing a representation against which acoustic characteristics may be calibrated. This question was investigated by Johnson (1990), for vowel continua. In Johnson's study, speaker identity was defined by F0, whereas Ladejoged and Broadbent (1957) manipulated formant frequencies to specify speaker identity. Johnson hypothesized that if acoustic information is used directly, then changing speakers should have no effect on identification of vowel stimuli. On the other hand, if perception shifts as a function of the speaker, these results may be taken to indicate that listeners are using the acoustic information about the speaker in perception of the vowels (as a cue to speaker identity). In a series of three experiments, Johnson examined perception of test words in isolation and in carrier phrases whose F0 manipulations signaled the same speaker, different speakers, or were ambiguous with respect to speaker identity. In a series of perception pretests, Johnson manipulated F0 in a synthesized "hood-hud" continuum, and also in the synthesized carrier phrase "this is ____". These pretests were designed to determine the relationship between F0 and speaker identity for both the vowel tokens and the carrier
phrases. In the first experiment, Johnson compared perception of vowels in isolation and in carrier phrases. The F0 levels of the vowel tokens were 100 Hz and 150 Hz, levels which had been shown by the pretests to correspond to different speakers. These vowels were also embedded in carrier phrases which had been attributed to a single speaker. Listeners were asked to label the vowel tokens. The results of the vowels in carrier phrases were then compared to results of those vowels in isolation. The hypothesis of this experiment was that shifts in vowel identification observed for the 150 Hz versus the 100 Hz tokens would be reduced if the carrier phrase signalled that they were produced by the same speaker. This hypothesis was borne out by the results. The second experiment compared the vowel shifts when both vowel tokens and carrier phrases were ambiguous with respect to whether they had been produced by the same or different speakers. The hypothesis of this experiment was that there would be no significant difference in shifts for vowel tokens in isolation as compared to those in the carrier phrases, since there was no conflicting information about the speaker. This hypothesis was also confirmed by the results of the experiment. Finally, Johnson examined whether perception of the vowels shifted when the carrier phrases indicated two speakers, but the vowel tokens were at a constant F0 level, corresponding to a single speaker. This experiment tested the hypothesis of Ladefoged and Broadbent (1957), which predicted that identical stimuli would be perceived differently if they were produced by different speakers. The results of Johnson’s experiment confirmed this hypothesis as well. The three experiments in Johnson’s study thus provide evidence that listeners use F0 as a cue to speaker identity, and that listeners normalize for this information in vowel perception.

1.2 Context Effects and Mandarin Chinese Tones

While the majority of studies have investigated normalization in the perception of vowels, virtually no experimental work has been done to examine speaker normalization in the perception of other types of speech sounds. The present study attempts to fill this gap by extending work on normalization from the segmental domain to the suprasegmental domain, focusing on lexical tone. The tone language used in this study is Mandarin Chinese, whose four lexical tones include a high level tone (Tone 1), a mid-rising tone (Tone 2), a low-falling-rising tone (Tone 3), and a high falling tone (Tone 4). Examples of the four tones for one speaker are shown in Figure 1. In particular, the present series of experiments examines whether F0 range, as a cue to speaker identity, influences perception of Tones 2 and 3.
Figure 1. F0 contours for each of the four tones, taken from one token spoken in isolation by one of the speakers in this study (segmental context ma).

As suprasegmentals, tones are perceived relative to other tones, though they are also distinguished by tone-internal (intrinsic) acoustic properties—primarily pitch height and contour (Gandour 1978; Coster and Kratochvil 1984). For tones which contrast in both of these dimensions, intrinsic F0 information may be sufficient for correct identification. To identify tones differing only in F0 height, however, listeners must refer to their knowledge of the speaker’s F0 range, and where tones occur within that range. For example, a low tone produced by a high-pitched speaker and a high tone produced by a low-pitched speaker may be acoustically identical. The process by which listeners adjust perception according to speaker-specific acoustic information is referred to as speaker
normalization. Few studies have investigated the role of extrinsic F0 in tone perception, however, and results from these studies have not provided convincing evidence for speaker normalization.

In a study specifically addressing speaker normalization for tones, Leather (1983) tested perception of Mandarin Chinese tone stimuli in two natural precursor phrases, one representing a low F0 range, the other one a higher range. Seven stimuli from two Tone 1-Tone 2 continua, each continuum representing the F0 range of one speaker, were embedded in the precursor phrases. Four steps in the middle of the continuum were identical in F0 height and contour, and were included in both continua. Test items (precursor + tone stimulus) were blocked by speaker and presented to five listeners in a labeling task. Individual subject responses were reported for the four pairs of mid-continuum stimuli (paired by speaker condition). These results showed at least one significant chi-squared value for each of the subjects, demonstrating that perception of at least one stimulus pair varied as a function of the speaker.

Unfortunately, however, Leather does not explicitly predict the type of responses he expects in each condition, nor do the reported data present this information. It is difficult, therefore, to take these data as conclusive evidence for speaker normalization without more detailed information. In particular, it is essential to know the direction of any shift in identification, whether it is consistent across speakers and across stimuli, and whether subject responses conform to predicted results. For example, the identification shifts reported by Leather may have been in different directions, such that subjects who identified two stimulus pairs differently depending on the speaker may have classified one in an assimilatory direction (the higher pitch range produced more high tone responses) but contrastive in the other case. In addition, the results show inconsistency across speakers and stimuli. For example, several subjects identified only one stimulus pair among the four according to the speaker condition, while another did so for non-adjacent steps in the continuum. These inconsistencies suggest that if the analysis compared the entire identification functions for each subject rather than for selected stimuli, differences between the two sentence conditions may not have been robust enough to sustain the effect.

Subjects may also have identified the stimuli in a way not predicted by the hypothesis. For instance, the hypothesis may have predicted that the high precursor would trigger more Tone 2 responses for ambiguous stimuli (since the onset F0 of these stimuli is low relative to the precursor F0). It is impossible to determine, based on the chi-squared results, whether there were more Tone 1 or Tone 2 responses in the high
precursor condition. Furthermore, it is not possible to relate subject responses with the F0 information they received; only two stimulus pairs fell within the overlap in F0 range between the two speakers, but only one of those was ambiguous in both F0 height and contour, and for those stimuli, the reported results do not specify how they were identified in the two precursor conditions.

Leather's use of the Tone 1 - Tone 2 continuum may also have been problematic. These two tones, which vary in both F0 height and contour, were synthesized without controlling for confounding acoustic parameters such as onset or offset F0. Moreover, both tones occur in the upper region of a speaker's pitch range, making it difficult to compare these tones in terms of F0 height. Finally, subjects in Leather's study responded to test items blocked by speaker, so it is uncertain if the normalization effect would still obtain in a mixed condition, which corresponds more closely to natural conditions.

Other studies have examined the role of extrinsic F0 information in tone perception, though they did not specifically address speaker normalization. Using an AX anchoring paradigm, in which the A element of the stimuli is constant, Lin and Wang (1985) presented subjects with pairs of Mandarin Chinese tones in which the first tone, representing a high level tone (Tone 1), was held at a constant 115 Hz, while the second tone, representing the high falling tone (Tone 4), varied onset F0 from 110 to 140 Hz in 10 Hz steps with an F0 fall of 40 Hz. Subjects were asked to label the first tone in each pair. Their results showed that as the onset F0 in the second syllable increased, identification of the first tone as a rising tone (Tone 2) increased. Thus, the higher onset F0 of the second syllable cued a wider pitch range, altering the relative F0 height of the first Tone 1 syllable to be perceived as low. Without a statistical analysis it is uncertain how robust these results are, but they nevertheless provide some evidence that tones are perceived relative to F0 range, such that this information contributes directly to the acoustic characteristics of the tone. While the study more broadly indicates that tone perception is affected by extrinsic F0, it does not address whether F0 information which serves to distinguish speaker identity may influence perception.

Using a similar anchoring paradigm, Fox and Qi (1990) investigated whether context F0 influences tone perception, and whether the influence occurs for both native and non-native listeners. Tone stimuli were presented in isolation and in pairs. In the isolated-token condition, listeners were asked to rate the stimulus according to how closely it resembled the Tone 1 or Tone 2 exemplar. In the paired-token condition, the first tone was either a Tone 1 or Tone 2, while the onset F0 of the second tone varied along a continuum from Tones 1 to 2; subjects were asked to rate the second tone in the pair,
according to the same rating scale as in the isolated-token condition. Results showed no significant difference between perception in isolation and in the context condition for either language group.

Following Leather's study, Fox and Qi presented chi-squared values for individual subject responses to four mid-continuum stimuli, showing inconsistent patterns of identification across subjects and stimuli. Among the 27 chi-squared values (9 subjects x 3 continuum steps), significant shifts were represented in only six of the Chinese subjects, all assimilatory, with five subjects showing no significant values. For the English subjects, five out of 27 chi-squared values were significant among three of the nine subjects, all but one assimilatory. This proportion compares to seven significant values out of 20 in Leather's study which compared responses to four mid-continuum stimuli for five subjects. Fox and Qi interpret these results as weak support for context effects from F0 on tone perception, in contrast to those of Lin and Wang (1985), who showed differences in identification as F0 range widened.

The reasons for the inconsistencies in Fox and Qi may be related to the methodology used. In Lin and Wang (1985), manipulating the onset of the second tone had the effect of modifying the pitch range, as in Fox and Qi, but listeners were asked to identify the first tone in the sequence, a tone which was constant throughout the experiment. In comparison, Fox and Qi asked listeners to identify the tone containing the modifications, the second tone. The anchor in Fox and Qi did not shift, but rather it was intended that listeners would use the anchor to identify the onset of the second tone as lower, as in a Tone 2, or higher, corresponding to a Tone 1. A shift in identification for Tone 1 anchors may have been expected, since listeners may not have had enough F0 range information against which to calibrate the tone stimuli. However, a Tone 2 anchor would provide the listener with adequate pitch range information against which to compare the F0 onset of the second tone. Since both anchors were included in one test, listeners had the relevant F0 range information throughout the test. Therefore it is not surprising that the results yielded no context effects.

Results from these earlier studies have not provided robust evidence that tone perception is affected by contextual acoustic cues, despite the assumption that tones, as suprasegmentals, are perceived according to surrounding information. In Leather (1983) as well as Fox and Qi (1990), shifts in tone identification did not occur reliably for all subjects, nor for a particular stimulus. Also, the direction of the shift, whether contrastive or assimilatory, was either not specified or was inconsistent across subjects and stimuli.
Some of these problems may be remedied by employing a different methodology. For example, in order to test for speaker normalization, precursors must vary in speaker identity. Precursors in Leather (1983) represented different speakers, but Lin and Wang (1985) and Fox and Qi (1990) limited their investigation to context effects, and so precursors consisted of one syllable which did not represent more than one speaker. Moreover, stimuli should reflect a situation in which normalization would be expected to occur, for example, in perception of different tones occurring within an area of overlap in F0 range among speakers. Although Leather (1983) examined tone perception for speakers with overlapping F0 ranges, both of the tones occurring in that range were high tones (Tones 1 and 2), and so may not have been sufficiently distinguished by F0 height. The experimental design should also present test items in a mixed condition, as opposed to a blocked condition as in Leather (1983), in order to more closely reflect the natural environment, and to reveal the robustness of any effect. In addition, subject data should be analyzed over the entire continuum, rather than for selected stimuli, to determine whether the identification functions for each subject have shifted reliably, and in what direction. The present study incorporates these issues into a new investigation of speaker normalization for Mandarin Chinese tones.

In this study, production and perception tests are used to examine Tone 2 and Tone 3. These tones were chosen, as opposed to Tones 1 and 2 used by Leather (1983) and Fox and Qi (1990), because they occupy distinct registers in a speaker's range; although both tones originate at the midpoint of the range, Tone 2 rises to cover the high region of the range, while Tone 3 is distinctly low, falling to the low region and ending with a rise (in pre-pausal position) near the middle of the range. This distinction more clearly demarcates F0 height as a perceptual cue. Tones 2 and 3 are also similar in contour when spoken in isolation, which may be the reason they cause the most confusion in perception tests (Kiriloff 1969; Chuang, Hiki, Sone and Namura 1972; Gandour 1978; Li and Thompson 1977). While overall F0 height may contribute to the distinctive phonetic characteristics of Tones 2 and 3, two additional acoustic dimensions are relevant: timing of the Turning Point, defined as the duration from the onset of the tone to the point of change in F0 direction, and also the decrease in F0 from the onset of the tone to the Turning Point, hereafter called ΔF0. These properties are schematized in Figure 2.
Figure 2. Turning Point and ΔF0 properties schematized for a contour tone.

Figure 2 illustrates how Turning Point and ΔF0 values for a tone are defined in this study. Perception studies of Mandarin Tones 2 and 3 have found that both timing of the Turning Point and ΔF0 are perceptually relevant for identification of the tones (Shen and Lin 1991; Shen, Lin and Yan 1993).¹

Using these two acoustic dimensions, the present experiment examines perception of stimuli in a Tone 2 - 3 continuum whose F0 levels fall within an area of overlap in F0 range for two speakers. In this scenario, speakers overlap in F0 range such that the low region of the high-pitched speaker overlaps with the high region of the low-pitched speaker. Within the area of overlap, tones may occur at equivalent F0 heights such that they would be low tones for the high-pitched speaker, but high tones for the low-pitched speaker. This scenario is schematized in Figure 3.

¹Duration differences between the two tones may also be perceptually relevant (Blicher, Diehl and Cohen, 1990), but will not be investigated in this study. Production data generally show that durations for both Tones 2 and 3 are longer than for other tones and that Tone 3 is longer than Tone 2 (Dreher and Lee, 1966; Ting, 1971; Chuang et al., 1972; Rumjancev, 1972; Lyovin, 1978; Nordenhake and Svantesson, 1983), perhaps because the non-prepausal form of Tone 3 is shorter than in isolation. An examination of duration differences between Tones 2 and 3 for each of the two speakers in this study found mixed results as well, showing no significant differences between the tone durations for one speaker, but significant differences for the other speaker.
**Figure 3.** F0 ranges for two speakers overlap such that the low region of the range for Speaker 1 overlaps with the high region for Speaker 2. Normalization may occur for acoustically identical tones in the region of overlap.

The scenario in Figure 3 illustrates how two speakers may overlap in F0 range. Within that region of overlap, a high tone for the low-pitched speaker (Speaker 2) may be acoustically identical to a low tone for the high-pitched speaker (Speaker 1). In the example just described, identification of the tone would be expected to shift (contrastively) depending on the F0 range of the precursor. If normalization occurs, stimuli will be identified by using F0 range information to "calibrate" ambiguous tones. On the other hand, if tone identification does not shift as a function of different F0 ranges, speaker normalization will be judged not to have occurred (subjects will not have referred to talker F0 range in order to identify tones).

To achieve the scenario conducive to normalization, production data were gathered in order to find two speakers whose F0 ranges and tones exhibited areas of overlap. Data from the production study also provided acoustic measurements of Tones 2 and 3, which were then used in synthesizing stimuli for the perception experiments.

Stimuli for the perception tests were synthesized to vary in either ΔF0, timing of the Turning Point, or both acoustic dimensions. These stimuli were presented to listeners in
isolation, and then embedded in both high and low precursor phrases. Perception of stimuli in these two conditions will be compared to determine whether changes in speaker identity effect changes in tone identification.

The structure of the paper is as follows. Production data for speaker F0 ranges and Mandarin Tones 2 and 3 are given in Section 2. Perception data for tone stimuli varying in Turning Point and ΔF0 are provided in Section 3. In Sections 4-7, methodology and results of the normalization experiments are described, followed by a general summary and conclusion.

2. Experiment 1: Production

This experiment was designed to provide acoustic information about speaker F0 ranges and Mandarin Tones 2 and 3. The experiment consisted of three reading tasks, the results of which established the mean F0 and overall F0 range of the speakers.

2.1 Method

2.1.1 Subjects

Four female and three male subjects aged between 19 and 30 produced the data for this study. The subjects are all from Mainland China, and are native speakers of Mandarin Chinese. Since all were graduate or undergraduate students at Cornell University, the subjects are all competent English speakers as well. None reported any speech disorders. Subjects were paid for their participation.

2.1.2 Materials

The data collected were from three reading tasks. The first of these asked subjects to read a long passage of text from a story, approximately four minutes long, entitled "Guo ji da shi he ta de qi zi" 'The World Master and His Wife' by Xiao Fu Xin (Hsu 1990). For the second task, subjects read minimal sets for each of the four Mandarin tones of the segmental contexts wu, yi, bi, and ma. These syllables were randomized and produced in the carrier phrase Zhe ge zi nian ___ ('This word is ___'). The third task consisted of subjects reading a randomized list of the minimal sets spoken in isolation. Test items in the carrier phrases and in isolation were produced three times. Both lists also included fillers at the beginning and end of every page to avoid list effects. All reading materials were presented to subjects in Chinese characters.
2.1.3 Procedure and Analysis

Subjects read the materials in an IAC sound-proof booth. They were recorded using an Electrovoice RE20 cardioid microphone and a Carver TD-1700 cassette recorder in the Cornell Phonetics Laboratory. The data were digitized on a Sun Sparcstation 2 computer using a sampling rate of 11 kHz with 16-bit resolution, and were analyzed using Entropics WAVES+/ESPS speech analysis software.

Mean F0 and overall F0 range were obtained from computer measurements of F0 over the long passage. A computer program sampled F0 every five milliseconds, then filtered out F0 values corresponding to a probability of voicing of less than 99%. This was done in case the program erroneously calculated F0 points for non-voiced portions of the passage. The F0 values were then organized into histograms showing number of samples as a function of F0. Mean and modal F0 values were then calculated for each speaker.

F0 measurements for the minimal sets in isolation and in carrier phrases, as well as the carriers themselves, were taken every five milliseconds. Average F0 as well as peak and valley F0 values for the carriers were calculated for voiced portions. Valley F0 values were taken to be the lowest F0 value in the tone, peak F0 the highest value. F0 data for tones in isolation and in carrier phrases were measured at 5 ms intervals, beginning with the onset of the vowel, or at the first full period of the vowel if the onset of voicing resulted in "artifact" F0 values which did not appear to be congruous with following F0 points. Ending F0 values were determined to occur at the offset of voicing (probability of voicing below 99%), or at the offset of the vowel (according to the waveform and spectrogram analysis) if the data showed F0 values inconsistent with the path of the tone to that point. Vowel and tone duration was measured from onset to offset of periodicity in the waveform in the yi and wu syllable types, from the onset of the vowel to its offset as determined by the waveform in the bi and ma syllables. Spectrograms provided additional help in locating vowel onset and offsets, where vowel onset was marked as the onset of F1, and the offset of F2 was taken to be vowel offset.

Two acoustic dimensions were measured in isolated tones and in tones embedded in the carriers: timing of the Turning Point, and the change in F0 from the onset of the tone to the Turning Point (ΔF0). Turning Point was defined from the pitch track as the duration between the onset of the tone and the point at which the tone changed F0 direction from falling to rising--this point was also the valley for both tones. In cases where the valley was constant for longer than 5 ms (exhibited by more than one measurement point on the pitch track), the rightmost point before the increase in F0
values was taken to be the offset of the Turning Point measurement. ΔF0 was calculated to be the difference in F0 between the onset of the tone and the Turning Point. Values were averaged for all instances of the tones, as well as for each syllable type. Tokens repeated in isolation and carrier phrases comprised a corpus of 24 instances of each tone per speaker (4 syllable types x 3 isolation repetitions x 3 carrier phrase repetitions). Several instances of the tones contained creak, including four Tone 2 and eight Tone 3 tokens, which made the relevant measurements impossible, and so these tokens were excluded from analysis.

The next section will present results of speaker F0 range analysis, followed by analysis of Tones 2 and 3.

2.2 Results

2.2.1 F0 range data

Analysis of F0 range data was conducted for the long passage, generating roughly 40,000 data points for each speaker. These data were organized in the histograms.

Among seven speakers analyzed, F0 range data for two of the four female speakers were found to meet the requirements of the normalization study. Figure 4 reports the F0 range data for these two speakers.
Figure 4. F0 range data from the long reading task for Speaker 1 (top panel), and Speaker 2 (bottom panel). The horizontal axis denotes F0 range; the vertical axis denotes number of data points for each F0 level. Mean F0 for Speaker 1 is 212 Hz, mode is 190 Hz. Mean F0 for Speaker 2 is 186 Hz, mode is 194 Hz.
Figure 4 shows mean, mode and F0 range data calculated between 120 Hz and 270 Hz for each speaker. F0 points above 270 Hz decreased gradually in number until 300 Hz; below 120 Hz a relatively small number of F0 points were recorded and the amount remained relatively stable. The histograms show a large peak at the mode F0 value, decreasing at the upper and lower regions. Speaker 1 (hereafter S1) shows a mean of 212 Hz and a mode of 190 Hz. For Speaker 2 (hereafter S2), the mean is 186 Hz and the mode 194 Hz. These data show a large area of overlap in F0 range, with a non-overlapping region at the low end of the distribution, corresponding to S2. Although the overlap extends to the high end of the region, the means reflect that S1 produces more consistently in a higher range than S2.

2.3 Discussion

Among the seven speakers tested, these two female speakers illustrate the F0 range characteristics most conducive to testing for speaker normalization. The data show a region of overlap in the F0 ranges for S1 and S2. Tones which occur in the overlapping region could conceivably fall in the low region of S1's range, but the high region of S2's range. The tones corresponding to those areas of the speaker ranges are Tone 2, the mid-rising tone, which typically occurs in the upper region of a speaker's range, and Tone 3, the low-falling-rising tone, which occupies the low region of a speaker's F0 range. If those tones are to be perceived correctly, listeners must adjust tone perception according to which speaker produces the tone.

To verify that tones for the two speakers also overlap in F0 level, the next section presents Tone 2 and Tone 3 data for the two speakers, and describes the acoustic properties for the tones that are used in subsequent experiments.

2.4 Mandarin Tone 2 and Tone 3 analysis

Figure 5 shows the F0 contours of Tones 2 and 3 for both speakers.
Figure 5. Averaged F0 contours of Tones 2 and 3 for the two female speakers (S1 and S2), across all syllable types.

The F0 contours in Figure 5 represent Tone 2 and Tone 3 average F0 onset, Turning Point and offset values for all syllable types in the isolation and carrier conditions. Although the figure does not include a time dimension to show realistic F0 contours, it shows that the two tones are very similar in F0 at onset, and have a similar falling-rising contour. The crucial observation in Figure 5 is that the F0 height of S1's Tone 3 falls somewhat below the Tone 2 of S2 in a relationship corresponding to the overlap in F0 ranges; the Tone 2 contour of S2 has an onset of 192 Hz, falling to 157 Hz at the Turning Point, and ending at 234 Hz; the Tone 3 of S1 has an onset of 187 Hz, falling to 143 Hz, and ending at 204 Hz. The other two tones for each speaker, Tone 2 for S1 and Tone 3 for S2, are produced outside of the region of tonal overlap.
Recall that the F0 range data in Figure 4 indicate that the two female speakers share a region of overlap which encompasses the lower range of S1 and the upper range of S2. The low tone of S1 and the high tone of S2 occur precisely in this region, a pattern that is predicted by the F0 range data. Figure 6, which contains both F0 and duration information, exemplifies how tones in this region are produced at similar F0 heights.

![Graph](image)

**Figure 6.** F0 contours for one [u] token each of Tone 2 of S2 and Tone 3 of S1, from the isolation production task.

The data in Figure 6 show both duration and F0 aspects of representative tone tokens. The similar F0 level of these two tokens attests to the likelihood that Tone 2 for S2 and Tone 3 for S1 may be acoustically identical. They are, thus, appropriate to use in subsequent tests for normalization. The following section describes them in more phonetic detail.
2.5 Acoustic characteristics of Tones 2 and 3

The preceding section established that Tone 2 for S2 and Tone 3 for S1 may be produced at equivalent F0 levels, and thus satisfy the scenario conducive to testing for normalization. Further acoustic analysis of these tones defined parameters used in subsequent tests using synthesized stimuli. The measurements taken included tone duration (taken to be equivalent to vowel duration), Turning Point (TP) and ΔF0.

Mean duration measurements taken for each tone according to the vowel in each syllable type are shown in Table 1.

<table>
<thead>
<tr>
<th>Tone2 (S2)</th>
<th>ma 'hemp'</th>
<th>yi 'move'</th>
<th>bi 'nose'</th>
<th>wu 'nothing'</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>278</td>
<td>356</td>
<td>331</td>
<td>363</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tone3 (S1)</th>
<th>ma</th>
<th>yi</th>
<th>bi</th>
<th>wu</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>288</td>
<td>388</td>
<td>320</td>
<td>377</td>
</tr>
</tbody>
</table>

Table 1. Average duration (ms) of Tone 2 for S2 and Tone 3 for S1, according to syllable type.

Table 1 lists average tone durations for each syllable type, including six tokens of each type (three tokens produced in isolation, three in carrier sentences), for a total of 24 tokens possible. One token of S1's ma was excluded because the presence of creak made location of vowel offset impossible. S2's Tone 2 durations range from 244 ms to 415 ms; S1's Tone 3 range is 324 to 485 ms. Average duration for the Tone 2 (S2) tokens is 332 ms, as compared with 346 ms for Tone 3 (S1) tokens. An unpaired, two-tailed t-test showed this difference not to be significant \( t(45) = -0.89, p > 0.38 \). A large area of overlap is thus represented in these Tone 2 and 3 tokens: from 324 to 415 ms.

Within each syllable type, statistical analyses between Tone 2 (S2) and Tone 3 (S1) show that the differences are not significant in each case (yi: \( t(10) = -1.73, p > 0.12 \); bi: \( t(10) = -0.49, p > 0.64 \); wu: \( t(10) = -0.48, p > 0.65 \); and ma: \( t(9) = -0.49, p > 0.64 \)).

Duration was also measured in terms of timing of the Turning Point for Tone 2 (S2) and Tone 3 (S1), over all syllable types. These data were analyzed in two ways: in absolute milliseconds, and also as a percentage of tone duration. Tone 2 Turning Point values averaged 66 ms, occurring at an average of 20% into the tone. Tone 3 showed an

---

2Other tokens of both Tones 2 and 3 exhibited some degree of creak as well, although vowel onset and offset points were undisturbed. In these cases the creak was located in the middle of the vowel, and the expected formant structure returned before vowel offset.
average Turning Point of 139 ms, occurring at 35% into the tone on average. Turning Point values for Tone 2 ranged from 25 to 96 ms, as compared to 105 to 200 ms for Tone 3, demonstrating a significant difference between the two tones \( t(27) = -6.22, p < .001 \). Calculated as a percentage of total tone duration, Tone 2 ranges from 7-24% of the tone, while the Tone 3 range is 28-41%. This difference between the two tones is also significant \( t(27) = -4.76, p < .001 \). The Turning Point regions for Tones 2 and 3 are therefore quite distinct. 3

In addition to Turning Point, the other acoustic parameter being observed is the decline in F0 from the onset of the tone to the Turning Point, or ΔF0. These data showed an average ΔF0 of 35 Hz for Tone 2 (S2), and 51 Hz for Tone 3 (S1). For Tone 2, ΔF0 ranged from 4 to 67 Hz, and from 24 to 106 Hz for Tone 3. There is, thus, an area of overlap occurring between 24 to 67 Hz. Unpaired two-tailed t-test results show that ΔF0 differences between Tone 2 (S2) and Tone 3 (S1) are not significant \( t(23) = -1.7, p > .09 \), but at a level suggesting a strong trend.

2.6 Discussion

Results of the tone analysis show Tones 2 and 3 to be similar in F0 height and contour. Both tones show a decline in F0 from the onset to the Turning Point, as well as a final rise. The data also show that Tone 2 for S2 and Tone 3 for S1 may be produced at a virtually equivalent F0 height in terms of onset and overall contour. Thus, Tone 2 for S2 occurs at roughly the same F0 height as Tone 3 for S1. Finally, duration differences between these two tones are not significant.

Differences were found between the tones, however. The intrinsic property of F0 showed that Tone 2 tends to have smaller decreases in F0 from the onset to the Turning Point than Tone 3, though these differences between the tones did not reach significance. The most conspicuous difference is in timing of the Turning Point: Tone 2 tokens showed significantly earlier Turning Points than Tone 3 tokens, both in absolute duration and as a percentage of tone duration.

The data in Experiment 1 show that the two female speakers share a region of overlap in F0 range, and that Tone 2 for the low-pitched speaker and Tone 3 for the high-pitched speaker also overlap, two conditions essential to test for normalization effects. The hypothesis of this study is that if listeners are to correctly identify these tones they must

---

3 Since Turning Point is one of the acoustic parameters that will be manipulated in the creation of a Tone 2 - Tone 3 continuum, it would be desirable to have an area of overlap, rather than the completely distinct regions observed. Nevertheless, the 4% difference between the regions is probably under the JND for duration in this case (Henry, 1948; Ruhm, Mencke, Milburn, Cooper, and Rose, 1966; Lehiste, 1970).
normalize for speaker identity (F0 range). To test this hypothesis, perception of tone stimuli in high and low F0 precursors is examined, where tone stimuli form a continuum from Tone 2 to Tone 3, varying in ΔF0 and Turning Point characteristics. The next section describes how test items for these experiments were created, beginning with the tone stimuli.

3. Experiment 2: Perception of Turning Point and ΔF0 in isolation

Results of Experiment 1 suggest that timing of the Turning Point and ΔF0 distinguish Tones 2 and 3. Earlier studies using Tones 2 and 3 have also considered timing of the Turning Point and ΔF0 to be perceptual cues for these two tones. Shen and Yan (1991) constructed two Turning Point continua, one with a fixed ΔF0 of 15 Hz, the other 30 Hz. They found that subjects' perception shifted toward Tone 3 earlier for the stimuli whose ΔF0 was 30 Hz, than for stimuli with a ΔF0 of 15 Hz. Blicher, Diehl and Cohen (1990) created a Tone 2 to Tone 3 continuum which manipulated three dimensions: timing of the Turning Point, ΔF0, and F0 offset. However, neither of these earlier studies have documented a systematic investigation of these parameters which addresses whether ΔF0 and Turning Point covary, whether perception based on each of these parameters was equally categorical, or what combinations of Turning Point and ΔF0 trigger shifts in identification from one tone to the other. The values of Turning Point and ΔF0 used in the previous studies were based on production data, but data for the present experiment show enough variability in the production of Tones 2 and 3 to allow either tone to possess the particular F0 and duration characteristics of the previous studies' exemplars. Therefore, a perception experiment was devised to determine the relative importance of timing of the Turning Point and ΔF0. The experiment tests perception of isolated synthetic stimuli in which timing of the Turning Point and ΔF0 have been systematically manipulated. Subject responses should clarify how these acoustic dimensions are perceived, their relative importance, and any ambiguity created by the combination of manipulations. In addition, results of this experiment will be used to more accurately model the synthetic stimuli used in subsequent tests.
3.1 Method

3.1.1 Subjects

Six subjects from Mainland China, three males and three females between the ages of 19 and 40, participated in Experiment 2. All were recruited from the Cornell University community. None reported any hearing disorders. Subjects were paid for their participation.

3.1.2 Stimuli

In order to reduce or eliminate speaker-specific or tone-specific cues in the stimuli, the syllable [u] was chosen for synthesis because of the relatively similar durations between speakers and tones. This syllable type was also used in the Shen et al. (1993) and Shen and Lin (1991) studies.

Stimuli were created using the Delta speech synthesis program developed by Hertz (Charif, Hertz and Weber 1992; Zsiga 1994) and a Klatt synthesizer (1980) in the Cornell Phonetics Laboratory.4 Formant frequency values for F1-F5 were averaged for the two speakers to create an ambiguous voice quality. Formant values for F1 and F2 included separate measurements for the onset and offset of the formant. Production data for F3 showed no substantial difference between onset and offset values for each speaker and thus F3 was held constant over the entire vowel. F4 and F5 were also held constant, based upon measurements from the steady-state portion of the vowel. The resulting composition of formant values is shown in Table 2.

---

4 There has been much concern about whether female voices can be successfully synthesized given the current design of the Klatt synthesizer. Parameters now considered to improve the naturalness of synthesized female voices include breathiness, open quotient, and glottal waveform (see Klatt and Klatt, (1990) for summary and experimental data). Stimuli synthesized for the present experiment relied largely on manipulating traditional parameters of fundamental frequency and formant frequencies. In addition, a more breathy quality was modeled by setting a Delta parameter which filters the upper frequencies relative to the lower frequencies. Subjects reported hearing a female speaking, and were often surprised to learn the stimuli were not produced naturally.
<table>
<thead>
<tr>
<th></th>
<th>onset</th>
<th>offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>345</td>
<td>304</td>
</tr>
<tr>
<td>F2</td>
<td>703</td>
<td>628</td>
</tr>
<tr>
<td>F3</td>
<td>2940</td>
<td>2940</td>
</tr>
<tr>
<td>F4</td>
<td>4320</td>
<td>4320</td>
</tr>
<tr>
<td>F5</td>
<td>4840</td>
<td>4840</td>
</tr>
</tbody>
</table>

**Table 2.** Formant frequency values (Hz) for synthesized stimuli

Duration of the stimuli was constant at 400 ms, well within the duration range for either Tone 2 or Tone 3. Amplitude of voicing began at 55 dB and declined to 53 dB over the duration of the token.

Based on the Turning Point data in Experiment 1, stimuli for the perception tests were designed to vary timing of the Turning Point along a continuum from 20 to 240 ms in 20 ms steps, for a total of twelve stimuli. This continuum contained Turning Point information which should trigger Tone 2 responses when the Turning Point occurs close to the tone onset, and Tone 3 responses when the Turning Point occurs late in the tone. ΔF0 was varied from 10 to 70 Hz in steps of 5 Hz, generating a continuum of thirteen stimuli. Because Tone 2 typically exhibits a shallower ΔF0, it was expected that tones with a ΔF0 equal to 10 Hz would produce more Tone 2 responses than tones with a ΔF0 of 70 Hz.

The two continua together allowed for testing of both timing of the Turning Point and ΔF0, in an effort to understand how these acoustic parameters are used in the perception of Tones 2 and 3. Figure 7 represents all combinations of these parameters which were included in Experiment 2.
Figure 7. Combinations of Turning Point and ΔF0 manipulations for synthesized stimuli. Turning Point manipulations are represented along the horizontal axis, ΔF0 manipulations on the vertical axis. The shaded region corresponds to predicted Tone 2 responses, and the lined region corresponds to predicted Tone 3 responses.

Based on the duration and F0 manipulations represented in Figure 7, predictions can be made about which regions of the graph might be expected to trigger Tone 2 and Tone 3 responses. According to traditional phonetic descriptions, Tone 2 is characterized by a short fall in F0 followed by a long rise, while Tone 3 has a deeper, longer fall followed by a long rise. The shaded region in Figure 7 which corresponds to the Tone 2 characterization contains stimuli with Turning Points from 20ms to approximately 100ms, along with lower ΔF0 values. It might also be expected that a high ΔF0 coupled with an early Turning Point (20 to 40 ms) would yield a Tone 2 percept, since the F0 rise of the stimulus is predominant. On the other hand, listeners would be expected to label as Tone 3 any stimulus containing a deeper F0 fall and a longer duration to Turning Point, stimuli marked in the lined region of Figure 7.
3.1.3 Procedure

Experiment 2 used a forced-choice labeling paradigm in which subjects heard each [u] stimulus in isolation and were asked to choose from two lexical items. There were 468 tokens in total (12 Turning Point x 13 ΔF0 x 3 repetitions). Stimuli were low-pass filtered at 5.2 kHz and played out on a 12-bit audio system using the BLISS software program (Mertus 1989) on a Swan 386 PC. Due to the number of stimuli (156 different manipulations), stimuli were presented in three blocks, one set of stimuli per block, with an intertrial interval (ITI) of 2 s.

One to four subjects at a time participated in the test. They were instructed to respond to each item as quickly as possible by pressing the button corresponding to the Chinese character for 'not' (Tone 2) or 'dance' (Tone 3). A practice session consisting of 23 test items preceded the test. These practice items provided listeners with endpoints for each parameter manipulated, as well as stimuli in between. Instructions were given in English, since most of the subjects were undergraduate or graduate students at Cornell and therefore highly proficient English speakers. However, for this and all subsequent tests, the few subjects who were not proficient in English were given instructions in Mandarin in addition to English. There were no differences in responses between the subjects instructed in Mandarin and those instructed in English. Subject responses were collected by computer using the BLISS software system. Responses for each stimulus were added across speakers.
3.2 Results

Figure 8 gives number of Tone 2 responses for all stimuli, arranged according to values for timing of the Turning Point and ΔF0.

![Figure 8](image)

**Figure 8.** Tone 2 responses for isolated stimuli varying in timing of the Turning Point and ΔF0. Eighteen responses were possible for each stimulus (6 subjects x 3 repetitions). Responses to stimuli from Experiment 2 determined which tone continua to use in Experiments 3a - 3c. These continua are enclosed in boxes: the diagonal boxes indicate stimuli varying along both Turning Point and ΔF0, the horizontal boxes represent stimuli varying only in Turning Point, and the vertical row of boxes shows stimuli varying only in ΔF0. An "x" denotes stimuli added after Experiment 2.

Figure 8 shows that, as expected, stimuli are clearly identified as Tone 2 in the region where Turning Point and ΔF0 values are low. Along the Turning Point dimension, unambiguous Tone 2 responses span virtually the entire continuum, up to a ΔF0 of 30 Hz.
Thus, Tone 2 appears to tolerate substantial delays (up to 240 ms) when the initial F0 fall is 30 Hz or less.

Along the ΔF0 dimension, decreases of up to 70 Hz are still identified as Tone 2. According to Kratochvil (1971), who tested perception of synthetic tones with durations from 90 - 240 ms, a duration of between 50 and 100 ms is required for perception of isolated Mandarin tones. Weber ratios for duration have been reported for 100 ms signals as .026 by Ruhm et al. (1966), and for 400 ms signals as .12 by Stott (1935), corresponding to approximately 10 - 60 ms for the 400 ms stimuli in Experiment 2. If the initial 20 to 40 ms portion of the tone is imperceptible (below threshold), subjects may hear only a rise, not the initial fall, since ΔF0 is equivalent to 0 at the earliest Turning Points. ΔF0 begins to trigger Tone 3 responses when the Turning Point reaches approximately 80 ms into the tone. At this point, Tone 3 responses increase as a function of both ΔF0 and timing of the Turning Point. When the Turning Point occurs as late as 200 ms into the tone, however, relatively low ΔF0 values (approximately 30 Hz or greater) elicit Tone 3 responses. Thus, the later the Turning Point, the easier it is for ΔF0 to effect a change, though even late Turning Points are resilient to the effects of a ΔF0< 30 Hz. On the other hand, ΔF0 appears to trigger more Tone 3 responses in stimuli with later Turning Points rather than early ones.

3.3 Discussion

The purpose of Experiment 2 was to determine how the acoustic dimensions of timing of the Turning Point and ΔF0 contribute to perception of Mandarin Tones 2 and 3. As for which of the two acoustic dimensions might be more important, the data suggest that there is an interdependency between ΔF0 and timing of the Turning Point. It appears that ΔF0 is more relevant as Turning Point increases, while Turning Point is more relevant for a ΔF0 of more than 30 Hz. For example, a tone with an initial fall of 30 Hz will be perceived the same as a tone with an initial fall of 10 Hz for any Turning Point, but a tone with a Turning Point at 160 ms will be perceived differently if its ΔF0 is 30 Hz or 45 Hz. Tone 2 perception seems to tolerate more variability overall, while Tone 3 requires a late Turning Point and a large fall in F0. This may be because the later Turning Point enhances the perceptual salience of the initial fall.

While Turning Point and ΔF0 appear to be interdependent, there are places where either dimension alone is sufficient to produce categorical functions. For example, stimuli with a constant Turning Point of 140 ms show all Tone 2 responses for a ΔF0 of 20 Hz, moving to 50% responses for a ΔF0 of 50 Hz, and all Tone 3 responses for the
largest $\Delta F_0$. Along the Turning Point dimension, Tone 2 responses move categorically from 100 to 0% for the continuum of stimuli with a constant 50 Hz $\Delta F_0$. While either of the two acoustic parameters are robust enough to trigger categorical identification functions, it is clear from both the production and perception data that timing of the Turning Point and $\Delta F_0$ operate in tandem as perceptual cues to Tones 2 and 3.

In summary, the results of this perception test show how tone stimuli which vary in $\Delta F_0$ and timing of the Turning Point are perceived. Subjects made categorical responses based on these two acoustic dimensions, such that identification functions may be obtained for either F0, Turning Point, or both parameters. Experiments described in the following sections test normalization effects using Tone 2 to Tone 3 continua based on the acoustic parameters of $\Delta F_0$ and Turning Point examined above.

4. Experiments 3a-3c: Perception of stimuli in precursor phrases

The following three experiments test the hypothesis that listeners perceive tones in part by normalizing for speaker F0 range.

4.1 Method

Experiments 3a-3c employ a design which compares how identical stimuli are identified in two contexts differing only in speaker identity. This type of test ensures that the effect is caused by normalization of different talker characteristics; the test uses naturally spoken carrier phrases from different speakers to serve as precursors. Normalization effects in this experiment would cause a shift in identification of stimuli as a function of which precursor phrase is heard, high or low F0.

As previously noted, earlier experiments on tones have provided evidence that external acoustic information may influence perception, though only Leather (1983) examined normalization effects due to perceived speaker identity. Experiments 3a-3c of the present study expand on Leather's work by testing two different Mandarin tones, Tones 2 and 3, and seek to provide more robust evidence of normalization. This will be done in several ways: by examining the direction of any shift in identification relative to the precursor, by measuring shifts in identification over the entire function, rather than arbitrarily selected points in the middle, and by presenting stimuli in a mixed condition, rather than blocked by speaker.
4.1.1 Tone stimuli for Experiments 3a-3c

The three experiments were conducted based on the perception data from Experiment 2. Each experiment is distinguished according to the stimuli used: Experiment 3a employed a continuum of stimuli containing cues about both timing of the Turning Point and ΔF0; stimuli for Experiment 3b included a continuum of stimuli varying only ΔF0, and Experiment 3c used a continuum varying only timing of the Turning Point. The continua marked by boxes in Figure 8 were used in these experiments. All three continua share a common midpoint which has a Turning Point of 120 ms, and a ΔF0 of 50 Hz. This midpoint stimulus received 50% Tone 2 responses for the corresponding identification functions resulting from Experiment 2.

5. Experiment 3a: Perception of stimuli varying in ΔF0 and Turning Point

5.1 Method

5.1.1 Subjects

Eleven subjects, seven male and four female, aged between 19 and 40, participated in this experiment. All are native speakers of Mandarin Chinese, eight from Mainland China and three from Taiwan, with no known hearing disorders. Because there are many dialects spoken in Mainland China and Taiwan, the subject population in this study was restricted to those speaking only one of the Mandarin dialects according to Norman (1988, p. 191). This restriction provided a more homogeneous subject group, although not as strict as if they had been limited to Beijing Mandarin only. Examples of Chinese languages not represented by subjects included in the study were Shanghai, Cantonese and Taiwanese.

5.1.2 Stimuli: Tone continuum

Stimuli for this experiment were synthesized [u] syllables which formed a continuum from Tone 2 to Tone 3, varying in both timing of the Turning Point and ΔF0. This continuum is the diagonal set of stimuli shown in Figure 8. A schematized version of these stimuli is shown in Figure 9.
Figure 9. F0 contours of the Experiment 3a stimuli which varied in both the Turning Point and ΔF0 acoustic dimensions.

The schematized tone contours in Figure 9 illustrate manipulations in both the Turning Point and ΔF0 acoustic dimensions. One additional step was created on the Tone 3 end of the continuum to provide an equal number of stimuli on either end of the crossover stimulus. Timing of the Turning Point varied from 20 to 220 ms, in steps of 20 ms. ΔF0 ranged from 25 Hz to 75 Hz, in steps of 5 Hz.

5.1.3 Stimuli: Precursor phrases

In addition to these stimuli, two natural precursor phrases spoken at a normal speaking rate were chosen from the production data discussed in Experiment 1, one from each the high-pitched speaker (S1) and the low-pitched speaker (S2). Analysis revealed that phrases for the two speakers were not sufficiently distinct in average F0, in contrast with the data from the reading task. Instead, a phrase from another female participant in the study was selected to replace S2. The mean F0 for the carrier phrase of the substituted speaker was 187 Hz, which was very similar to the overall mean of 186 Hz for S2. Table 3 summarizes the acoustic information for the two precursors.

<table>
<thead>
<tr>
<th>speaker</th>
<th>duration</th>
<th>average F0</th>
<th>peak</th>
<th>valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>high F0</td>
<td>718 ms</td>
<td>226 Hz</td>
<td>272 Hz</td>
<td>192 Hz</td>
</tr>
<tr>
<td>low F0</td>
<td>722 ms</td>
<td>187 Hz</td>
<td>229 Hz</td>
<td>170 Hz</td>
</tr>
</tbody>
</table>

Table 3. F0 and duration information for precursors used in Experiments 3a-3c.
Table 3 presents duration and F0 information for each precursor, high and low. The peak and valley F0 points represent boundaries of the F0 range for each speaker, showing a shared region of 192 to 229 Hz. The two phrases differ by 39 Hz in average F0, but are further distinguished by the range; the high precursor spans 192-272 Hz, as compared to 170-229 Hz for the low precursor. In order to visualize how the stimuli are situated with respect to these F0 ranges, recall that the synthesized stimuli have a fixed onset and offset of 188 Hz and 212 Hz, respectively, levels which were based upon production data. The ΔF0 value decreases from 163 to 113 Hz in the continuum containing both ΔF0 and Turning Point cues, and also in the ΔF0 continuum.

Because they are naturally produced, the phrases differ in voice quality as well as F0 range. (Formant frequencies of the target stimuli, however, were synthesized to be ambiguous relative to the precursors to control for an effect of voice quality (see Table 3.2).) Other than these differences, the phrases were identical. Each contained the segmental context Zheige zi nian _____ ('This word is ___'), each had preceded a high tone syllable (Tone 1 or Tone 4) in the production task, and each matched in duration. Synthesized stimuli from the pretest were appended to the precursors, leaving a 50 ms silence between the precursor and the test word.

As additional controls, the carrier phrase contained no instances of Tones 2 or 3 or [u]. There were two advantages of limiting the phrases this way. One advantage was to eliminate the environment for tone sandhi effects, which particularly affect Tones 2 and 3 (Chao 1968). The other advantage is that subjects only hear one instance of the test tones, rather than possibly comparing precursor examples of the test tones with the stimuli.

---

5This restriction to a High tone context served as a control for tonal coarticulation cues which may have been present in carriers preceding Tones 2 or 3. However, a study on coarticulation in Mandarin tones by Shen (1990) shows that there is no anticipatory effect on F0 height or direction from Tones 2 or 3, and particularly no effect from those tones on a preceding Tone 4. The similar F0 onset of Tones 2 and 3 probably obviates coarticulation, since it would be in anticipation of the onset F0 height that anticipatory coarticulation would occur (Shen, 1990).
5.1.4 Procedure and Analysis

The experiment was conducted in the Cornell Phonetics Laboratory. Test items were presented to subjects by way of the PC-based software program BLISS, which randomized and played the stimuli via a D/A converter (12-bit resolution, 11 kHz sampling rate, low-pass filtered at 5.2 kHz). Eleven stimuli were preceded by each of the high and low precursor phrases, creating a total of 22 sentences. Subjects first heard 12 test items in a practice session. The test consisted of a total of 220 trials (22 sentences x 10 repetitions), with an inter-trial interval of 2250 ms. One to four subjects at a time listened to test items over headphones in separate booths. Subjects were instructed to respond by pressing one of two buttons, which were labeled using the Chinese characters for either 'not', or 'dance', corresponding to the Tone 2 or Tone 3 lexical item, respectively.

Responses were recorded and tabulated by computer. For each subject, the crossover points for stimuli in both the high and low precursor conditions were determined by using a probit statistical analysis (Finney 1971). This statistical method takes into account the subject's responses over the entire continuum of manipulations.

5.2 Results

Results of Experiment 3a are summarized in figure and table form below. Figure 10 shows the percentage of Tone 2 responses for stimuli in the two presentation types (high and low precursor conditions), averaged across subjects. Table 4 lists probit values for each subject as a function of the precursor conditions.
Figure 10. Experiment 3a Turning Point/ΔF0 continuum identification functions for high and low precursor conditions, averaged across subjects. Stimulus 1 corresponds to predicted Tone 2 responses.
<table>
<thead>
<tr>
<th>subject</th>
<th>high precursor</th>
<th>low precursor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.58</td>
<td>5.21</td>
</tr>
<tr>
<td>2</td>
<td>4.64</td>
<td>4.51</td>
</tr>
<tr>
<td>3</td>
<td>5.71</td>
<td>6.2</td>
</tr>
<tr>
<td>4</td>
<td>6.34</td>
<td>7.11</td>
</tr>
<tr>
<td>5</td>
<td>5.88</td>
<td>5.81</td>
</tr>
<tr>
<td>6</td>
<td>4.36</td>
<td>5.65</td>
</tr>
<tr>
<td>7</td>
<td>5.75</td>
<td>6.54</td>
</tr>
<tr>
<td>8</td>
<td>5.3</td>
<td>5.44</td>
</tr>
<tr>
<td>9</td>
<td>6.18</td>
<td>6.99</td>
</tr>
<tr>
<td>10</td>
<td>6.39</td>
<td>6.71</td>
</tr>
<tr>
<td>11</td>
<td>5.44</td>
<td>5.63</td>
</tr>
</tbody>
</table>

mean 5.60 5.99

Table 4. Experiment 3a probit values for Turning Point/ΔF0 stimuli in high and low precursor conditions.

The probit values in Table 4 represent category boundaries for the listeners who participated in Experiment 3a. As shown in Figure 10, subjects perceived more Tone 3 (low tone) responses when stimuli were preceded by a high precursor than when stimuli were preceded by a low precursor. The category boundary for the high precursor was earlier for eight of the eleven subjects, at 5.60, as compared to 5.99 for the low precursor. A paired two-tailed t-test shows this difference between boundaries to be significant [t(10) = -2.57; p<.03]. Subjects thus appear to refer to the F0 range of the precursor in perception of the tones. Moreover, the normalization effect is robust enough to be obtained in a mixed block condition, in comparison to Leather (1983) in which stimuli were blocked by speaker.

The shift away from Tone 2 responses in the high precursor condition demonstrates a contrast effect; the high F0 context causes a shift toward low tone (Tone 3) responses. While this result is to be expected given the assumption that F0 height of the tone is interpreted relative to a speaker's F0 range, it differs from earlier findings by Fox and Qi (1990), who instead found primarily assimilatory shifts for paired-token identification tasks.
The next section reports results of subject responses for stimuli in which only ΔF0 was manipulated.

6. Experiment 3b: Perception of stimuli varying in ΔF0

6.1 Method

6.1.1 Subjects

22 native speakers of Mandarin Chinese participated in this experiment. Twelve of these were excluded from the results on the basis of criteria outlined in section 6.1.3. The ten remaining subjects included 5 males and 5 females. One of the subjects is from Taiwan, and nine are from Mainland China. None reported any hearing disorders.

6.1.2 Stimuli

Test items were sentences composed of the two precursors used in Experiment 3a, followed by a test word taken from the ΔF0 continuum in the pretest. This continuum varied ΔF0 in 11 steps of 5 Hz from 163 Hz. The timing of the Turning Point was fixed at 120 ms.

6.1.3 Procedure and Analysis

The test procedure and analysis of data were identical to those used in Experiment 3a. However, this experiment seemed to be more difficult for subjects than Experiment 3a, judging both by number of missed trials and failure to achieve categorical identifications at continuum endpoints. Because of these two problems, it was decided that a subject's responses would be included in the results only if they met the following criteria: 1) they responded to more than 90% of the total trials for each continuum, and 2) they achieved at least an 80% correct response rate on continuum endpoints. Failure to meet these criteria led to the disqualification of 12 subjects.

6.2 Results

The averaged identification functions for the ΔF0 continuum in the high and low precursor conditions are presented in Figure 11, and the probit values are listed in Table 5.
Figure 11. Experiment 3b ΔF0 identification functions for high and low precursor conditions, averaged across subjects.
<table>
<thead>
<tr>
<th>subject</th>
<th>high precursor</th>
<th>low precursor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.47</td>
<td>6.18</td>
</tr>
<tr>
<td>2</td>
<td>4.74</td>
<td>6.08</td>
</tr>
<tr>
<td>3</td>
<td>6.02</td>
<td>7.76</td>
</tr>
<tr>
<td>4</td>
<td>4.29</td>
<td>4.09</td>
</tr>
<tr>
<td>5</td>
<td>3.33</td>
<td>4.75</td>
</tr>
<tr>
<td>6</td>
<td>4.49</td>
<td>4.56</td>
</tr>
<tr>
<td>7</td>
<td>3.32</td>
<td>6.28</td>
</tr>
<tr>
<td>8</td>
<td>4.79</td>
<td>6.57</td>
</tr>
<tr>
<td>9</td>
<td>5.06</td>
<td>7.24</td>
</tr>
<tr>
<td>10</td>
<td>4.10</td>
<td>5.59</td>
</tr>
</tbody>
</table>

mean 4.46 5.91

Table 5. Experiment 3b probit values by subject for ΔF0 stimuli in high and low precursor conditions.

The data in Table 5 show that for the Tone 2 to Tone 3 continuum varying only in ΔF0, there is an earlier shift to Tone 3 responses in the high precursor condition for nine of the ten subjects: 4.46 as compared to 5.91 in the low precursor condition. This difference is significant [t(9) = -4.69, p<.001]. The shift is one of contrast—the high precursor prompts more low tone responses and vice versa. If an assimilatory effect had been observed, there would have been more high tone responses when stimuli were preceded by the high precursor. This result supports the hypothesis that subjects refer to extrinsic F0 as a frame of reference for tone perception.

The magnitude of the shift in this experiment is much greater than the shift in Experiment 3a. These differences, computed as the difference between the low and high precursor probits, were shown to be significant in a two-tailed, unpaired t-test of the shifts for each subject [t(19)=-3.33, p<.003]. These results indicate that listeners relied more on speaker F0 range to disambiguate the tones when the stimuli provided less intrinsic acoustic information. The implications of this finding will be discussed in section 8. The next section reports results of subject responses for stimuli in which only the Turning Point dimension was manipulated.
7. Experiment 3c: Perception of stimuli varying in Turning Point

7.1 Method

7.1.1 Subjects

20 native speakers of Mandarin Chinese participated in Experiment 3c. Eight subjects were disqualified according to the criteria outlined in section 6.1.3. The twelve remaining subjects included six males and six females. Four subjects are from Taiwan, and eight are from Mainland China. None reported any hearing disorders.

7.1.2 Stimuli

Again, stimuli used in this experiment were part of the set used in the isolation pretest, appended to the end of the natural precursors used in Experiments 3a and 3b. The continuum from Tone 2 to Tone 3 varied only timing of the Turning Point, from 20 ms to 220 ms into the tone. The decrease in ΔF0 was constant at 50 Hz.

7.1.3 Procedure and Analysis

Test procedure and analysis of results are identical to those of Experiments 3a and 3b.

7.2 Results

Figure 12 displays average percent Tone 2 responses for the Turning Point continuum in the high and low precursor conditions. Table 6 lists probit values for each subject in both the high and low precursor conditions.
Figure 12. Experiment 3c Turning Point continuum identification functions for high and low precursor conditions, averaged across subjects.
<table>
<thead>
<tr>
<th>subject</th>
<th>high precursor</th>
<th>low precursor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.83</td>
<td>5.15</td>
</tr>
<tr>
<td>2</td>
<td>5.54</td>
<td>6.12</td>
</tr>
<tr>
<td>3</td>
<td>5.25</td>
<td>4.73</td>
</tr>
<tr>
<td>4</td>
<td>4.55</td>
<td>4.86</td>
</tr>
<tr>
<td>5</td>
<td>4.21</td>
<td>5.39</td>
</tr>
<tr>
<td>6</td>
<td>4.0</td>
<td>5.08</td>
</tr>
<tr>
<td>7</td>
<td>4.56</td>
<td>4.44</td>
</tr>
<tr>
<td>8</td>
<td>3.24</td>
<td>5.63</td>
</tr>
<tr>
<td>9</td>
<td>4.13</td>
<td>5.12</td>
</tr>
<tr>
<td>10</td>
<td>3.61</td>
<td>4.01</td>
</tr>
<tr>
<td>11</td>
<td>6.08</td>
<td>6.34</td>
</tr>
<tr>
<td>12</td>
<td>6.4</td>
<td>5.49</td>
</tr>
</tbody>
</table>

**mean** 4.78 5.20

**Table 6.** Experiment 3c probit values for Turning Point stimuli in high and low precursor conditions.

Table 6 lists probit values for subjects in the high and low precursor conditions. The average boundary in the high precursor condition as compared to the low is 4.78 vs. 5.20. Only eight of the twelve subjects show a shift in the direction predicted by the normalization hypothesis, and the difference between the probits in the two conditions is not significant [t(11)= -1.55, p>.15]. These results suggest that stimuli varying only in a duration dimension will not cause a normalization effect for contexts that vary in an F0 dimension.

**8. General Summary and Discussion**

The hypothesis of this study is that listeners use acoustic information about the speaker in the perception of lexical tones. In particular, the study investigates whether listeners use speaker F0 range in perception of intrinsic acoustic properties of Mandarin Tones 2 and 3. The hypothesis predicts that tone identification is affected by changes in speakers. If speaker information is not relevant in tone perception, on the other hand, changes in speakers should cause no significant shift in identification.
To examine the hypothesis, a series of production and perception experiments were conducted. First, production analyses from Experiment 1 located two speakers who share a region of F0 range overlap. The analysis revealed that within the area of F0 range overlap, a low tone for a high-pitched speaker and a high tone of a low-pitched speaker may occur at equivalent F0 heights. Further investigation of intrinsic acoustic properties of Tones 2 and 3 showed that the two tones may be distinguished in both F0 (ΔF0) and temporal (Turning Point) dimensions. Experiment 2 demonstrated that while both dimensions are used in production, either ΔF0 or Turning Point cues alone are sufficient to distinguish the two tones in perception. Isolating intrinsic cues was essential to the subsequent experiments in determining whether these cues were mediated through speaker F0 range, exhibiting a shift, or whether they were used independently of speaker F0 range, showing no effect of speaker condition.

The study then investigated whether changes in speaker identity affect tone perception by presenting tone continua in precursor phrases from two speakers, and observing whether identification shifted as a function of speaker F0 range. Results of Experiments 3a and 3b, which examined perception of both F0 and temporal properties of Tones 2 and 3 in high F0 and low F0 precursor phrases, showed a significant shift in tone identification, in the direction expected if tone stimuli were perceived according to the F0 range of the precursor; that is, identical stimuli were perceived as high tones in the low F0 precursor phrase, but as low tones in the high F0 precursor phrase. These findings thus support the hypothesis that tone identification is influenced by changes in speaker identity, demonstrating that this information is used as a frame of reference according to which ambiguous tones may be interpreted.

No significant shift was observed for the tone continuum in Experiment 3c, however, which varied the temporal dimension of Turning Point. The stimuli in Experiment 3c differed in only one aspect from the stimuli in Experiments 3a and 3b: they did not vary in ΔF0. These results suggest that normalization is triggered only when both stimuli and precursors vary in the same acoustic dimension. A context differing along a temporal dimension, such as speaking rate, may trigger normalization processes in perception of the Turning Point stimuli. This hypothesis is investigated in Moore (1995).

If the temporal dimension was not relevant for normalization in the Turning Point stimuli, it is tempting to assume that temporal information may not have contributed to the normalization effect in Experiment 3a, where stimuli varied along both dimensions. However, the larger magnitude of the effect in Experiment 3b as compared to Experiment 3a contradicts this assumption. This difference in the magnitude of the effect for stimuli
varying only in the F0 dimension as compared to stimuli varying in both the F0 and temporal dimensions supports the hypothesis that listeners utilize contextual information to a greater degree when intrinsic acoustic information is degraded. Such differences between effects have been observed in rate normalization work on vowel perception by Gottfried, Miller and Payton (1990), as well as rate effects in the perception of [b] - [w] continua in Shinn, Blumstein and Jongman (1985). Both of these studies show reductions in normalization effects as stimuli more closely resemble natural speech. Thus, it is possible that when listeners are given accompanying temporal information in Experiment 3a, they do not refer to speaker identity as much as when intrinsic tonal cues are restricted, as in Experiment 3b. Further work is needed to understand the relative contribution of temporal and F0 cues in contexts that also vary in both of these dimensions.

Although this investigation contributes additional data and addresses several inadequacies of Leather's study, findings of this study support the conclusions of Leather (1983). First of all, the present study observes normalization effects for tones which differ in F0 height; Tone 2 is an upper register tone, compared to the lower register Tone 3. Leather used two upper register tones whose contours are more dissimilar than Tones 2 and 3. Second, this study shows normalization effects robust enough to be obtained in a mixed block condition; Leather's subjects were trained on one speaker's voice before hearing stimuli embedded in precursors for that particular speaker. Third, analysis methods for the present study compare crossover boundaries based on the entire identification function, so that reliable shifts may be observed based on responses to all stimuli in each condition. The analysis of responses to only selected stimulus pairs rather than analysis of crossover boundaries may have led to the appearance of inconsistent results reported in Leather (1983) as well as Fox and Qi (1990). Fourth, while Leather did not report whether changes in perception are assimilatory or contrastive, or whether changes were consistent for all speakers, the present study provides conclusive evidence that shifts in identification are contrastive—in a direction opposite to the precursor F0—and that this shift is consistent enough across subjects in Experiments 3a and 3b to be statistically significant.

Although the context effects shown here are contrastive in direction, these results differ from the direction of shifts reported in Fox and Qi (1990). Their findings, for paired-token identification tasks, instead showed assimilatory shifts in all but one case. Their study focused on context effects from one preceding tone, rather than on speaker normalization, however. Fox and Qi further argue that assimilatory shifts are evidence
for auditory, rather than phonetic, processing of the acoustic signal, based on experimental work by Fujisaki and Kawashima (1971), Pisoni (1975), Shigeno and Fujisaki (1979) and Shigeno (1986). In these models of perception, assimilatory shifts occur for stimuli which do not undergo a category-level perceptual identification, such as for continua whose endpoints do not represent different phonemes, or for non-speech stimuli. For continua whose endpoints represent phonemic distinctions, or for complex tone continua, a categorical memory process is employed, generating contrastive shifts in identification. Shigeno (1991), however, provides evidence that both assimilatory and contrastive effects may occur within the process of phonetic judgment. From the standpoint of these two-stage perceptual processing models, results of the current study would suggest that higher-level phonetic processes are involved in speaker normalization for tones. Notwithstanding the different methods employed in Fox and Qi as compared to the present study, the opposite shifts in identification reported in the results raise the question of whether contextual F0 information is processed differently depending upon whether it was used as a cue to tone identity, as in Fox and Qi, versus as a cue to speaker identity, as in the present study.

Results of this study also support those of Johnson (1990), who found that both intrinsic and extrinsic F0 contributes to vowel perception. Johnson found that when precursor F0 corresponded with two different speakers but intrinsic F0 corresponded with one speaker, a shift in identification occurred, indicating that listeners were mediating intrinsic F0 differences through perceived speaker identity. As the results of Experiments 3a and 3b from the present study demonstrate, extrinsic F0 significantly influences tone perception by serving as a cue to speaker identity, causing intrinsic F0 cues (ΔF0) to be perceived relative to the extrinsic cues (F0 range). In other words, extrinsic F0 enabled listeners to construct a representation of F0 range, against which intrinsic acoustic characteristics of the tones were calibrated.

9. Conclusion

The conclusion of this series of experiments is that perception of tones is a talker-contingent process. Evidence was provided to show that listeners use extrinsic F0 information corresponding to speaker identity in perception of lexical tones, supporting the hypothesis that intrinsic acoustic information is mediated through a representation of speaker identity, rather than contributing to tone identification independent of speaker information. These results suggest that the same normalization processes participate in perception of suprasegmentals as well as segments.
Speaker normalization has been assumed to occur as a response to acoustic variability which derives from vocal tract differences among speakers. This variability is exhibited when different speech sounds are acoustically identical, as illustrated in this study, or when the same speech sound exhibits different acoustic characteristics. Further research on normalization in perception in the latter instances of variability would further clarify the relationship between acoustic variability and normalization.

Other research on the effects of speaker variability on perception indicates that speech perception is more difficult, and not as accurate, in multiple-talker conditions as compared to single-talker conditions (Mullenix, Pisoni, and Martin 1989; Sommers, Nygaard, and Pisoni 1992), blocked conditions (Strange, Verbrugge, Shankweiler, and Edman 1976; Assmann, Nearey and Hogan 1982) or when listeners have increased familiarity with the talkers' voices (Verbrugge, Strange, Shankweiler, and Edman 1976; Nygaard, Sommers, and Pisoni, 1994). These studies suggest that there is a "cost" associated with the process of normalizing for speaker differences. While the costs of normalizing for contextual information may be expected for segments, which are perceived highly accurately given only intrinsic cues (Verbrugge, Strange, Shankweiler, and Edman 1976), it is not as straightforward in the case of suprasegmentals, where context is assumed to be more intimately connected with identification. In the case of Mandarin Chinese, contour differences between the tones also yield high identification rates in isolation (Howie 1976). The more relevant case for establishing differing degrees of interdependence on context may be to examine normalization in perception of tones which contrast only in F0 height, such as the level tones in Cantonese (Fok 1974). To the extent that tone perception uses identical perceptual processes as segments, the observation in the present study that normalization effects obtain in a mixed block condition suggests that speaker normalization is a robust process, even for Mandarin tones.

This study has illuminated the dual nature of tones as suprasegmentals in that both extrinsic and intrinsic acoustic information contribute to the description of a tone. Tones do not depend on absolute acoustic values to gain their identity. Rather, they contrast with other tones in the utterance as well as speaker F0 range to attain a relative identity. These assumptions are consistent with the results of this study showing that listeners use speaker F0 range in tone identification. Despite their intimate relationship with context, however, lexical tones also exist as independent phonological units, contrasting intrinsic acoustic characteristics such as Turning Point and ΔF0 for Mandarin Tones 2 and 3. Thus, it is possible that in addition to contextual information specifying speaker F0 range,
intrinsic F0 may also enable listeners to establish a representation of speaker identity. This hypothesis is consistent with findings by Slawson (1968), and Johnson (1990) for vowel perception, and Mullenix et al. (1989), for word recognition. The use of both extrinsic and intrinsic acoustic information in identifying speaker F0 range avoids the "bootstrap" problem (Nearey, 1989; 2092), which confronts the issue of how listeners are able to establish a representation of speaker identity without precursor acoustic information.

Finally, this investigation raises another issue concerning the degree to which the speaker normalization process is conditioned by language experience. Language differences in tone perception have been reported in terms of the importance placed on intrinsic cues (Gandour and Harshman 1978). In the present study, speaker normalization was demonstrated to occur for native speakers of Mandarin Chinese, a language which uses F0 to distinguish lexical items. It may then be hypothesized that for non-native listeners, or those whose native language does not use F0 to make lexical contrasts, there may be no utilization of speaker F0 range. This hypothesis assumes that normalization is dependent on the presence of phonetic categories for tones. In an examination of this hypothesis, Moore (1995) finds evidence for normalization of speaker F0 range for tone stimuli by English listeners. However, results of that study indicate that the pattern of normalization is different between native and non-native listeners. The difference in patterns derives from the native listeners' greater degree of dependence on contextual F0 range information for less natural stimuli, a pattern not exhibited by the non-native listeners. That is, Mandarin listeners show the greatest effect of speaker F0 range for stimuli varying one acoustic dimension, rather than two (the magnitude of effect differences observed between Experiments 3a and 3b in the present study); English listeners show effects only when stimuli varies two acoustic dimensions. The findings of Moore (1995), therefore, demonstrate that language experience influences normalization.

One of the interesting observations exhibited in the present experiments has been a pattern of speaker normalization which occurs only when both precursors and stimuli vary in the F0 dimension. Normalization for stimuli varying in Turning Point may occur if precursors then vary in a temporal dimension. This hypothesis is also investigated in Moore (1995) for tone stimuli identical to those used in the present study. Findings from Moore (1995) indicate that when precursors vary in a temporal dimension (in this case, speaking rate), listeners normalize for rate by shifting category boundaries for the temporal cue (Turning Point). These results point to the importance of temporal
information in tone identification, information that is overlooked when discussions of tone focus on F0 properties alone.

10. References


    Working Papers of the Cornell Phonetics Laboratory 8, 82-117.
    Cornell University.
    contingent process. Psychological Science 5, 42-46.
Peterson, G. E., and Barney, H. L. (1952) Control methods used in a study of the
    3, 7-18.
    Differential sensitivity to duration of acoustic stimuli. Journal of Speech and Hearing
    Research 9, 371-384.
Rumjancev, M. K. (1972) Ton i Intonacija v Sovremennom Kitajskom Jazyke [Tone and
    Intonation in Modern Chinese] (Izdatel'stvo Moskovskogo Universiteta, Moscow).
    and Speech 34, 145-156.
Shen, X., Lin, M., and Yan, J. (1993) F0 turning point as an F0 cue to tonal contrast: A
Shigeno, S. (1986). The auditory tau and kappa effects for speech and nonspeech
    Acoust. Soc. Am. 90, 103-111.


1. Introduction

Words borrowed from one language into another may be subjected to a number of adjustments. It has been claimed that such adjustments are induced solely by the well-formedness conditions in the host language (Silverman 1992; Yip 1993). It has also been argued that well-formedness conditions in the native phonology may be relaxed, but no new constraints may be added, for the purposes of loanwords (Ito and Mester 1993).

In this paper, I argue that loanword-specific conditions do play a role in the phonology of Japanese. This paper provides an account of the phonological adjustments of English words borrowed into Japanese, focusing crucially on the conditions under which obstruents in English are realized as geminates in Japanese. Consonant Gemination (CG) is generally observed when the original English obstruent appears word-finally, following a short vowel. For example, the English word cut is realized as [katto] in Japanese. The English source does not have any geminates, thus it is puzzling that the Japanese counterpart is realized with a geminate. Well-formedness constraints in native Japanese phonology are not sufficient to account for the geminates. Since Japanese does not allow word-final obstruents, obviously phonological adjustments are necessary to realize the word-final obstruents in the English source: that is, the word cut cannot be realized as *kat. This form would violate the Japanese coda constraint, which prohibits a syllable-final obstruent unless it is part of a geminate: Coda (Ito 1989). Violation of Coda can be avoided if a vowel is inserted after the obstruent, yielding *ka.to., yet this perfectly
legitimate form is not the output form. As was mentioned before, the form with a geminate is the surface form. This suggests that constraints other than the well-formedness conditions in Japanese are necessary to account for the realization of obstruents in loanwords. What is also called for, I argue, is a set of constraints requiring that output forms mirror the foreign source as closely as possible. I propose an analysis utilizing both the Japanese well-formedness conditions and loanword specific constraints to account for CG. The analysis is framed in Optimality Theory (Prince and Smolensky 1993).

The structure of this paper is as follows. In section 2, I discuss several adjustments observed in English loans in Japanese. In section 3, I summarize the consonant gemination (CG) data. In section 4, I review previous analyses and propose constraint-based account in section 5. I conclude the paper in section 6.

2. Adjustments Observed in English Loans in Japanese

The following table shows the Japanese inventory (Nakajou, 1989):

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>t</th>
<th>k</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ф</td>
<td>s</td>
<td>f</td>
<td>ç</td>
<td>h</td>
</tr>
<tr>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ts</td>
<td>tʃ</td>
<td>dʒ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dz</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>ɲ</td>
<td>ɲ̃</td>
<td>ɲ̃</td>
</tr>
<tr>
<td>j</td>
<td>w</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| e | o | a |

Japanese syllable structure is (C)(G)V(V)(C), where G is the glide [j] (Itô 1986; Vance 1987). Coda consonants must be either the first half of a geminate or a nasal. Only the following consonants can be geminated: [p, t, k, s, f, ts, tʃ]. Voiced geminates do not exist in the native lexicon. It should be noted that syllables with more than two moras do not occur in monomorphemic words in the native lexicon. This means that syllables of (C)(G)VVC, (C)(G)VVN or (C)(G)VNC form, where N is a nasal and C is a non-nasal, does not occur¹. We can attribute this to the undominated constraint *SUPERHEAVY

¹ Polymorphemic words may contain superheavy syllables: e.g. [təʊ,tə] 'passed' < toor 'pass' + ta 'past-tense marker', [rɒn,dɒŋ,kəu] 'Londoner' < rɔndən 'London' + ko 'native'.
SYLLABLE, which prohibits extraheavy syllables in monomorphemic words. Some examples with different syllable types are given in (2):

(2) [kja.ku] 'guest' [ho.x] 'book'
    [kit.te] 'stamp' [ku.w.ki] 'air'

English and Japanese have different sets of well-formedness conditions, thus when an English word is borrowed into Japanese, naturally a number of adjustments take place (the description of adjustments below are based on National Language Research Institute (1990).

English consonants which have no counterparts in Japanese are realized as Japanese consonants with the same (or similar) place, voicing and manner (except [v], which is realized as [b], due to the lack of voiced bilabial fricative [β] in Japanese).

(3) English    Japanese
    [f] →   [φ]   fur     [φaa]
    [v] →   [b]   view    [bju.w]
    [θ] →   [s]   Ithaca [isaka]
    [ð] →   [z]   mother  [mazaa]
    [l], [ɾ] → [ɾ]   fly, fry [fuaɾai]

Note that both English [l] and [ɾ] are realized as [ɾ], since it is the only liquid in Japanese.

English vowels which have no Japanese counterparts are realized as Japanese vowels with a similar height and backness. English tense vowels are realized as long vowels in Japanese, while most lax vowels are realized as short. Some lax vowels are realized as long vowels in Japanese, presumably since they are phonetically long.

(4) Lax vowels
    [l] →   [i]   pin     [pin]
    [ɛ] →   [e]   pen     [pen]
    [æ] →   [a]   rally   [raɾii]
    [ɔ] →   [aa]  turn    [taan]
    [ə] →   [oo]  call    [koɔri]
Tense Vowels  
\[
\begin{align*}
[i] & \rightarrow [ii] \quad \text{key} \quad [kii] \\
[u] & \rightarrow [u\text{w}] \quad \text{cue} \quad [kju\text{w}] \\
[oU] & \rightarrow [oo] \quad \text{zone} \quad [zoon]
\end{align*}
\]

Japanese does not have consonant clusters other than a consonant followed by [j] in onset position. When an English word with a consonant cluster is borrowed into Japanese, vowels are inserted to break the cluster. It should be noted that no member of consonant clusters is deleted:

\[(5) \quad \text{star} \rightarrow [su\text{t}a\text{a}] \\
\text{grill} \rightarrow [gu\text{r}i\text{mu}] \\
\text{friend} \rightarrow [fu\text{re}ndo]\]

In general, [w] is the inserted vowel except after dentals and palatal affricates (for a more detailed discussion, see Lovins (1975)).

It was mentioned before that Japanese allows two types of coda consonant: a placeless nasal and the first half of a geminate. The coda nasal appears in loanwords when an English source has [n] in the coda position. This is not the case for other English nasals. [m] is realized as [m\text{w}], while [ŋ] becomes [ng\text{w}]. Examples are given below:\n
\[(6) \quad \text{skin} \rightarrow [su\text{k}i\text{n}] \\
\text{run} \rightarrow [ra\text{n}] \\
\text{skim} \rightarrow [su\text{k}i\text{mu}] \\
\text{rum} \rightarrow [ra\text{mu}] \\
\text{sing} \rightarrow [si\text{ng}u] \\
\text{lung} \rightarrow [ra\text{ng}u]\]

The adjustments discussed so far follow directly from constraints on Japanese segments and syllable structure.\n
---

2 It seems that the coda nasal also appears when [m] is preceded by a homorganic consonant:
\[
\begin{align*}
\text{hamburger} & \rightarrow [ha\text{g}ba\text{ga}] \\
\text{impact} & \rightarrow [ipakwuto]
\end{align*}
\]
Yet when the English source contains two [m]'s orthographically, the coda nasal may or may not appear:
\[
\begin{align*}
\text{hammer} & \rightarrow [ha\text{g}ma\text{a}] \\
\text{summer} & \rightarrow [sama\text{a}]
\end{align*}
\]

3 Although only segmental adjustments are discussed in this section, it should be noted that there is another type of adjustment in loanwords in Japanese: the pitch accent assignment. Since Japanese is a pitch-
3. Consonant Gemination: Data

Let us now turn to the Consonant Gemination data. As was mentioned in section 1, obstruents in English words are often realized as geminates in their Japanese counterparts, even though the corresponding English words do not contain any geminate (the description of Consonant Gemination (CG) below is based on Lovins (1975), Ono (1991) and National Language Research Institute (1990).

3.1 Word-final Obstruents

Word-final voiceless obstruents in the English source are usually geminated:

(7) [p] tulip → [tʃu̯u̯iŋpu̯u̯]
[t] blanket → [bu̯u̯aŋketto̯]
[k] kick → [kiŋku̯u̯]
[tʃ] pitch → [piŋti̯u̯]
[f] finish → [fiŋi̯i̯u̯]

It is important to note that voiced geminates do not exist in the native Japanese lexicon nor do they result from morpheme concatenation, yet they may result from word-final CG in loanwords:

(8) [d] bed → [beddo]
[g] dog → [doggu̯u̯]
[dʒ] edge → [eddʒi̯i̯]

The following voiced consonants, however, are never realized as geminates:

(9) [b] Bob → [bo̯u̯u̯] *[bo̯bu̯u̯]
[z] fizz → [fiʒu̯u̯] *[fiʒzu̯u̯]
[ʒ] vision → [biŋon] *[biʒon]
[l] bill → [bi̯i̯u̯] *[bi̯i̯u̯]

accented language, all loanwords must receive a certain pitch accent pattern. Traditionally (Akinaga 1966; McCawley 1968; National Language Research Institute; Haraguchi 1991), it has been argued that the accent in loanwords falls on the antepenultimate mora (e.g. *vanilla* is realized as [ba̯i̯i̯a], where [ba] is accented), or on the mora which corresponds to the English stressed vowel (e.g. *system* is realized as [fiŋu̯e̯m], where [fi] is accented). Older loans may be unaccented: e.g. pistol is realized as [pi̯u̯e̯toru̯u̯] with no accent. But see Kubozono (1994) for syllable-based accent assignment rule.
These consonants, except [b], are all voiced continuant segments. One may assume that CG does not apply to a voiced [+cont] segment. Spirantization of [b] may explain its irregular behavior. Vance (1987) notes that intervocalic [b] in Japanese is often changed to the bilabial fricative [ϕ]. We can posit that CG does not apply to [b], since it is treated as a voiced continuant.

### 3.2 Word-medial obstruents

Word-medial voiceless obstruents in English are realized as geminates when they are preceded by a stressed vowel:

(10) [p] háppy → [happii]
[t] cóttón → [kotton]
[k] hóckey → [hokkee]
[s] éssay → [essei]
[f] fáshion → [faffon]

However, no medial gemination is observed when the English source is preceded by an unstressed vowel:

(11) [p] suppórt → [sapooto] *[sappooto]*
[t] guitár → [gitaa] *[gitta]*

In addition, it seems that word-medial CG is affected by the English orthography. When a word-medial obstruent in the English source is orthographically a single consonant, CG does not occur even when the obstruent is preceded by a stressed vowel:

(12) [f] édition → [edifon] *[edifon]*
condítion → [kondifon] *[kondifon]*

Interestingly, word-medial voiced obstruents in English are never geminated regardless of the stress position:

(13) [d] Chéddar → [tʃɛddaa] *[tʃɛddaa]*
[g] búggy → [bagii] *[bagii]*
3.3 Further issues

There are several departures from the above generalization of CG. CG does not occur when an English obstruent follows a long vowel:

(14) repeat → [ripiito] *[ripiitto]
debate → [dibeeto] *[dibeetto]
grape → [guromeppu] *[guromeppu]

This follows from *SUPERHEAVY SYLLABLE, which prohibits syllables with more than two moras.

In addition, no loanwords are realized with two geminates. When there are two possible targets for CG, only the word-final obstruent is realized as a geminate:\(^4\):

(15) råcket → [raketto] *[rakketto], *[rakketto]
påcket → [poketto] *[pokketto], *[pokketto]
kåetchup → [kettappu] *[kettappu], *[kettappu]

3.4 Consonant Clusters

So far we have seen the CG of a single consonant. When an English word with a consonant cluster is borrowed into Japanese, CG is usually not observed:

(16) [pt] kept → [kepmto] *[kepmtoo], *[kepmtoo]
[kt] perfect → [paafekwtoo] *[paafektwtoo], *[paafektwtoo]
[ft] gift → [gifwtoo] *[gifwtoo], *[gifwtoo]
[kst] text → [tekiswtoo] *[tekiswtoo], *[tekiswtoo], *[tekiswtoo]
[sp] grasp → [gurasswppu] *[gurasswppu], *[gurasswppu]
[st] cast → [kjaswtoo] *[kjaswtoo], *[kjaswtoo]
[sk] task → [taswkw] *[taswkw], *[taswkw]

\(^4\) Words may contain more than one geminate only when they consist of two free morphemes:

- råpertopppu, "laptop"
- hattifikkw, "hatchback"
- bukkwretto, "booklet"
- pikkwappu, "pickup"

We can see that above words consist of two free forms: laptop < lap + top, hatchback < hatch + back, etc.
However, some clusters do show CG, in addition to the vowel insertion which breaks up the cluster (cf. (5)). -CI clusters geminate unless C is an alveolar stop:

(17) [pl] apple → [appɛʁw]  
nipple → [nɪpɛʁw]  
[kl] tackle → [tækɛʁw]  
knuckle → [nækɛʁw]  
[fl] waffle → [wæfɛʁw]  
[sl] castle → [kjæsɛʁw]  
hustle → [hæsɛʁw]  
but: [tl] bottle → [bʌtɔʁw] *[bɔtɔʁw]  
kettle → [ketɔʁw] *[ketɔʁw]

Interestingly, no CG is observed when the C in -CI clusters is a voiced obstruent:

(18) [dl] middle → [mɪdɔʁw] *[mɪdɔʁw]  
[gl] struggle → [swɔtɔʁɡɛʁw] *[swɔtɔʁɡɛʁw]

In addition, CG does not occur when the C in -CI clusters is orthographically a single consonant:

(19) [pl] triple → [tɔɾɪpɛʁw] *[tɔɾɪpɛʁw]  
[kl] article → [aɑtɪkɛʁw] *[aɑtɪkɛʁw]

It should be noted that the gemination pattern in -CI clusters is similar to that of word-medial gemination described above.

Word-final -Cs clusters are also realized with a geminate. CG in -Cs clusters occurs regardless of the stress position or English orthography:

(20) [ps] pops → [pɔpɛʁw]  
chips → [ʧɪpɛʁw]  
[ks] box → [bɔkkwɛʁw]  
wax → [wækkwɛʁw]  
őrhödox → [oosodɔkkwɛʁw]
We can see that CG occurs even when the preceding vowel is not stressed (as in *orthodox*) and the C is orthographically a single consonant (as in all the examples above).

The CG data seem to be quite complicated. As we have seen, they do not follow directly from constraints on Japanese segments and syllable structure. Many researchers have attempted to account for this phenomenon, yet their analyses are far from complete. Different researchers, however, have captured different aspects of CG, and it is helpful to see what they have found. In the next section, some of the previous analyses are reviewed and their problems are pointed out. I propose a constraint-based analysis in section 5.

4. Previous Analyses

4.1 Stress

Akasaka (1972, cited in Ooe 1991) claims that English obstruents following a stressed vowel are realized as geminates. This explains the word-medial CG of voiceless obstruents: e.g. *happy* [happii]. However, as was shown in (13), voiced obstruents are not always realized as geminates in this environment: e.g. *Cheddar* [tʃedaa]. In addition, when the stressed vowel is long, there is no gemination whether the obstruent is voiced or not: e.g. *repeat* [ripiito] (see (14)). Finally, it should be noted that in words like *racket* [raketto], it is the obstruent following an unstressed vowel that is realized as a geminate (see (15)). Thus, it is clear that English stress is not the only condition for CG.

4.2 English Orthography

Ooe (1968) and Imai (1980) claim that orthographically doubled consonants in English are realized as geminates in Japanese. According to them, the gemination in the following words is attributable to their spellings:

\[
(21) \begin{array}{ll}
[p] & \text{happy} \rightarrow [\text{happii}] \\
[t] & \text{cotton} \rightarrow [\text{kotton}] \\
[k] & \text{kick} \rightarrow [\text{kikku}] \\
\end{array}
\]

This analysis accounts for the irregularity of gemination of [ʃ] in word-medial position: *fashion* → [faʃʃon] vs. *edition* → [ediʃon], *[ediʃon* (cf. (12)). However, note that words like *lip* and *cut* are realized as [ripi̯u] and [kaʃto], respectively. Also, in *racket* [rakettò], an orthographically single consonant is realized as a geminate, while a doubled one in the same word is realized as a single consonant (see (15)). Therefore, there must be other factors that condition CG.
4.3 Phonetic Closeness

In English, word-final obstruents following a stressed vowel are considerably lengthened. Ohso (1973) claims that word-final CG occurs due to an attempt to make the Japanese counterparts phonetically close to the English source. Since a Japanese geminate is phonetically a long consonant, geminating the final consonant makes the Japanese output phonetically close to the original English word. She further claims that word-medial CG may be affected by English orthography.

It is hard to believe that Japanese speakers recognize the duration difference between a word-final consonant in a stressed syllable and those in an unstressed syllable. In fact, gemination occurs even if the last syllable is not stressed: e.g. *édit* [editto].

Ohso gives a slightly different view in her earlier work (Ohso 1971, cited by Lovins 1975). She claims that CG is the result of an attempt to preserve the English closed syllable. Word-medial CG may or may not occur, since the syllable boundary in word-medial position is not clear to Japanese. This analysis seems plausible, except that it does not account for the difference between CG of word-medial voiceless and voiced obstruents. Note that word-medial voiced obstruents are not geminated in their Japanese counterparts, while voiceless ones do. Lovins (1975) also states that word-medial gemination seems to be conditioned by other factors, such as English stress. She notes that word-medial CG usually occurs when the preceding vowel is stressed, while word-final gemination occurs whether the preceding vowel is stressed or not.

5. Proposal

We have seen that previous analyses capture subsets of the data, but not all the data. In this section, I will propose an analysis of Consonant Gemination in Optimality Theory (Prince and Smolensky 1993; McCarthy and Prince 1993). In this framework, phonological and morphological representations are not derived by rules. Rather, a set of ranked constraints on well-formedness evaluates the set of all possible output candidates for a given input and selects the most preferred output. A candidate is considered to be optimal when it best satisfies the whole constraint set. This does not mean it has to meet all the constraints. Constraints are violable in this framework. Yet violation must be minimal. The optimal output is the form that satisfies the higher ranked constraints with minimal violations of lower-ranking constraints.

Following Silverman (1992), I assume that the non-linguistic acoustic signal of a foreign language is the input to the loanword phonology. This indicates that Japanese speakers do not have an access to the prosodic structure of English.
5.1 Word-final CG of voiceless obstruents

I propose that there are four constraints that are playing a crucial role in the word-final CG cases:

**No Place in Coda**

Only placeless consonants can appear in the coda (Itô 1986, 1989)

**Loanword Correspondence**

Segmental content of a foreign source has to be preserved in the host language.

**Align** (PrWordE, R, SyllableJ, R)

The right edge of the English word must be aligned with the right edge of the Japanese syllable.

**Fill**

Syllable positions are filled with segmental material (Prince and Smolensky 1993; McCarthy and Prince 1993)

**No Place in Coda** restricts the set of possible codas to be the first half of a geminate or a nasal. This constraint is undominated in both native Japanese and loanwords. The next two constraints, which I propose, are loanwords specific. **Loanword Correspondence** ensures that segmental content of a foreign input be preserved in the host language. This is also an undominated constraint. **Align** requires that the right edge of the English word be aligned with the right edge of the Japanese syllable. It should be noted that this constraint aligns the English word edge with the Japanese syllable edge. I assume that there is another kind of **Align** constraint which requires the right edge of the English word be aligned with the right edge of the Japanese word. However, it is not always possible to align the English word edge with the Japanese word edge, since Japanese syllable structure is much more restricted than English syllable structure. In such cases, the alignment with the next highest prosodic structure is required and that is induced by the **Align** (PrWordE, R, SyllableJ, R).

Let us consider the word-final CG observed in the word *cut* [katto]:

---

5 This constraint should be distinguished from the constraint **Faithfulness** that Yip (1993) proposes. Yip's constraint ensures that the input and the output be identical. Thus, an output with an epenthetic vowel, for example, violates **Faithfulness**. However, such an output is not a violation of **Loanword Faithfulness**, as long as it contains every segment in the input.
(22) cut → [katto]

<table>
<thead>
<tr>
<th></th>
<th>Correspond</th>
<th>Coda</th>
<th>Align</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>ka.</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kat.</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ka.to</td>
<td>*!</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>katto</td>
<td></td>
<td></td>
<td>*</td>
<td>**</td>
</tr>
</tbody>
</table>

Since Japanese does not allow word-final obstruents, the word cut cannot be realized as *kat. This form violates NO PLACE IN CODA. Violation of the coda constraint can be avoided if a vowel is inserted after the obstruent, yielding *ka.to. However, this is not the surface form. Obviously, well-formedness constraints in native Japanese phonology are not sufficient to account for the surface form. Violation of NO PLACE IN CODA can also be avoided if the word-final [t] is not preserved in the Japanese counterpart (thus *ka), yet this is not the optimal form, either. I attribute this to the undominated constraint LOANWORD CORRESPONDENCE. The actual output is kat.to, and this is due to the higher ranking of ALIGN than FILL.

5.2 Word-medial CG

I assume that word-medial CG is induced by a different constraint than word-final CG, following Ohso (1971). I argue that word-medial CG results from the pressure to retain the prominence of the original stressed syllable, and is due to the following constraint:

**TONIC SYLLABLE**

English stressed syllable must be heavy in the corresponding Japanese form\(^6\).

**TONIC SYLLABLE** accounts for the CG in cotton [kotton]:

(23) cotton → [kotton]

<table>
<thead>
<tr>
<th></th>
<th>Tonic</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>ko.ton</td>
<td>*!</td>
<td></td>
</tr>
<tr>
<td>kotton</td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

The geminated form is preferred, although it violates FILL, since FILL is ranked lower than TONIC SYLLABLE. One may ask why the Japanese syllable becomes heavy by gemination

---

\(^6\) As was mentioned before, this constraint seems to be affected by the English orthography. When the obstruent in the English source is spelled with a single consonant, the pressure to make the corresponding Japanese syllable heavy seems to be weakened.
rather than vowel lengthening. As Fukui (1986) and Poser (1988) have shown, a floating mora in Japanese is always filled by leftward spreading in the native Japanese lexicon. The same strategy seems to apply in loanword phonology.

5.3 CG of Voiced Obstruents

It was noted before that no voiced geminates are permitted in the native Japanese lexicon. However, voiced [-cont] geminates are observed in loanwords. I assume that the constraint \textbf{*Voiced [-cont] Geminates} is relaxed in loanword phonology and it is ranked as follows: \textbf{Align} \textgreater \textbf{*Voiced [-cont] Geminates} \textgreater \textbf{Tonic Syllable}. CG of word-final voiced obstruents does occur, due to the lower ranking of \textbf{*Voiced [-cont] Geminates} than \textbf{Align}. Consider the following:

\begin{center}
\begin{tabular}{|c|c|c|c|}
\hline
 & Coda & Align & *Voiced [-ct] Gem & Fill \\
\hline
bed. & *! & & & \\
\hline
be.do. & *! & & & \\
\hline
\textit{er} bed.do. & & & * & ** \\
\hline
\end{tabular}
\end{center}

The form \textit{bed.} cannot surface since it violates the undominated coda constraint. The form \textit{be.do.}, which does not contain a geminate but violates \textbf{Align}, is not the surface form, either. The third candidate with a voiced geminate is the optimal form.

CG of word-medial obstruents, on the other hand, does not occur due to the higher ranking of \textbf{*Voiced [-cont] Geminates} than \textbf{Tonic Syllable}.

\begin{center}
\begin{tabular}{|c|c|c|c|}
\hline
 & Align & *Voiced [-ct] Gem & Tonic & Fill \\
\hline
\textit{er} ba.gii. & & * & & \\
\hline
bag.gii. & *! & & & \\
\hline
\end{tabular}
\end{center}

Both candidates satisfy \textbf{Align}. The second candidate satisfies \textbf{Tonic Syllable} as well, yet it is not the surface form. It violates \textbf{*Voiced [-cont] Geminates}, which ranks higher. The first candidate is the optimal form, since it satisfies \textbf{*Voiced [-cont] Geminates}. 
5.4 Further Issues

We saw before that no gemination occurs following a long vowel. We posited that this is due to the undominated constraint \*SUPERHEAVY SYLLABLE, which can be formally stated as:

**SUPERHEAVY SYLLABLE**

Syllables with more than two moras are prohibited in monomorphemic words.

This constraint is actually relaxed in loanwords, since syllables of the form (C)(G)VVN does occur: e.g. scene [jiiN]. Only those syllables of (C)(G)VVC or (C)(G)VNC form is prohibited in loanwords. We can assumes that **SUPERHEAVY SYLLABLE** can be divided into two parts, **SUPERHEAVY SYLLABLE** - N and **SUPERHEAVY SYLLABLE** - C. The former prohibits superheavy syllables ending in a nasal and the latter prohibits those ending in a non-nasal consonant (this is similar to the division of **VOICED GEMINATES** into **VOICED [-CONT] GEMINATES** and **VOICED [+CONT] GEMINATES** shown above). Only **SUPERHEAVY SYLLABLE** - C is undominated in loanword phonology. This constraint accounts for the realization of the word *beat*.

(26) beat → [biito]

<table>
<thead>
<tr>
<th></th>
<th>*Super heavy - C</th>
<th>Align</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>[w]bi.to</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>biit.to</td>
<td>*!</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

Gemination does not occur when the preceding vowel is long, since it creates a superheavy syllable.

In (15), it was noted that word-medial gemination never occurs if word-final gemination is possible. CG occurs only once per word. I assume that this is due to the following constraint:

**TWO GEMINATES**

More than one geminate per word is prohibited.
This constraint restricts the number of geminates in a word. No loanword has more than one geminate, thus we can assume that this is an undominated constraint. Let us consider the realization of the word \textit{racket}.

\begin{table}[h]
\begin{tabular}{|c|c|c|c|}
\hline
 & \textbf{*Two Gem} & \textbf{Align} & \textbf{Tonic} & \textbf{Fill} \\
\hline
\textit{ra.ke.to.} & *! & * & * \\
\textit{rak.ke.to.} & *! & * & ** \\
\textit{ɾɾ ra.ket.to.} & *! & * & ** \\
\textit{ɾɾak.ket.to.} & *! & & *** \\
\hline
\end{tabular}
\end{table}

The first two candidates violate \textbf{ALIGN}, thus they are not optimal. The fourth candidate, which satisfies both \textbf{ALIGN} and \textbf{TONIC SYLLABLE} is ruled out since it violates \textbf{*TWO GEMINATES}, which is undominated and higher ranked than the two. The third candidate, which has one geminate and satisfies \textbf{ALIGN}, is the optimal form.

It was pointed out that a voiced continuant is never geminated both word-finally and word-medially. Although the constraint which bans voiced [-cont] geminates is relaxed in the loanword phonology, the one which prohibits voiced [+cont] geminates is not:

\textbf{*VOICED [+CONT] GEMINATE}

Voiced [+cont] geminates are prohibited.

This is an undominated constraint, and thus a voiced continuant in the English source is never realized as a geminate:

\begin{table}[h]
\begin{tabular}{|c|c|c|}
\hline
 & \textbf{*Vc [+ct]} & \textbf{Align} & \textbf{Fill} \\
\hline
\textit{ɾɾ fi.zu} & * & * \\
\textit{fiz.zu} & *! & ** \\
\hline
\end{tabular}
\end{table}

\footnote{I do not know any literature that talks about this kind of constraint in the native lexicon. Yet it is plausible that the core lexicon has \textbf{*TWO GEMINATES} for monomorphic words. It is known that Old Japanese did not have any geminates. Geminates were introduced when Chinese words were borrowed into Japanese. Geminates appeared at the morpheme boundary in Chinese. Thus it may be the case that the number of geminates in a monomorphic word is limited in the core lexicon. This is not to say geminates are completely prohibited in monomorphemic words in the core lexicon. It is known that the consonant [p] always appears as a geminate in Yamato and Sino-Japanese (McCawley 1968; Ito and Mester 1993). Therefore, the constraint on the number of geminates per word should allow at least one geminate.}
The second candidate satisfies ALIGN, yet it violates the undominated *VOICED [+CONT] GEMINATE. The first candidate is preferred, although it violates ALIGN, since it satisfies *VOICED [+CONT] GEMINATE.

5.5 Consonant Clusters

It was shown before that CG does not occur in most consonant clusters. I assume that this is due to an undominated constraint which bans two consecutive epenthetic segments:

**CONSECUTIVE UNFILLED POSITION**

Consecutive unfilled structural positions are prohibited.

This constraint can be understood as one of the faithfulness constraints, which ensures that the input and the output differ minimally (McCarthy and Prince 1994). *CONSECUTIVE UNFILLED POSITION seems to be undominated in the native lexicon as well. This constraint accounts for the realization of the word task (underlined segments are epenthetic):

\[(29)\] task → [taswku]

<table>
<thead>
<tr>
<th></th>
<th>Cons</th>
<th>Coda</th>
<th>Align</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>task.</td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tas.ku</td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tas.su.ku</td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ta.su.ku</td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ta.su.ku</td>
<td></td>
<td>*</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

The first and second candidates cannot be the optimal form, since they violate the undominated NO PLACE IN CODA. The fourth candidate satisfies ALIGN (note that unless [k] is in the coda position, ALIGN is violated), yet it has two consecutive epenthetic segments. Between the third and the fifth candidates, the non-geminated form is preferred, since it has fewer violations of FILL.

We have seen that when the English source contains -C(l) and -C(s) clusters, C geminates (see data in 3.4).

I assume that Japanese listeners perceive -C(l) clusters not as clusters, but as a consonant followed by another syllable. In final -C(l) clusters, [l] is syllabic in English. I assume that it enters the Japanese phonology as the syllable /ru/. If we posit that these clusters are considered to be a consonant followed by a syllable, the C in -C(l) is not word-final. Thus
we expect the gemination pattern to be the same as in word-medial gemination. And indeed that seems to be the case. In (18), we have seen that voiced stops do not geminate. It was also shown that the CG in -Cl clusters seems to be affected by the English orthography. Thus -Cl clusters can be accounted for by the TONIC SYLLABLE constraint:

(30) apple → [appurw]

<table>
<thead>
<tr>
<th>Align</th>
<th>Tonic</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>ā.pu.ruw</td>
<td>*!</td>
<td>*</td>
</tr>
<tr>
<td>áp.pu.ruw</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

Both candidates satisfy ALIGN, since the word-edge of the input is assumed to be /ru/. The second candidate is preferred over the first one, since it satisfies TONIC SYLLABLE. CG of -Cl clusters does not occur when the C is a voiced stop, due to the constraint *VOICED [-CONT] GEMINATES. Consider the following:

(31) struggle → [sw toragurw]

<table>
<thead>
<tr>
<th>*Voiced [-ct] Gem</th>
<th>Tonic</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>ē sw.to.ra.gurw</td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>sw.to.rag.gurw</td>
<td>*!</td>
<td>****</td>
</tr>
</tbody>
</table>

The form with a geminate is not the surface form contrary to (30), since it violates *VOICED [-CONT] GEMINATES, which ranks higher than TONIC SYLLABLE.

As for the [-tl] cluster, which does not show CG, I assume that flapping is blocking the gemination. Note that in the word ‘bottle’, for example, the [t] becomes a flap, which is a voiced continuant. I assume that gemination does not occur, due to *VOICED [+CONT] GEMINATES, which is undominated.

-Cs clusters do not show the characteristics of the word-medial CG, unlike -Cl clusters. The vowel preceding the C does not need to be stressed. The C can be orthographically a single consonant. It seems that they should be treated as word-final CG cases. But then, it is not clear why -Cs clusters show CG while other consonant clusters do not. It might be the case that the English orthography is playing a role here. The [ks] clusters which is spelled with a single x may be treated as a single segment, resulting in CG. For other -Cs clusters, morphological information may be related. Many words with final -Cs clusters seem to be polymorphemic words, such as plural forms. In those cases, the C in -Cs clusters may be treated as the word-final consonant, yielding the gemination. Further
investigation is necessary, however, to determine the relationship between the English orthography and CG.

5.6 Constraints Ranking

In (32), I give all constraints that we have seen:

(32) Ranking of Constraints in Loanword Phonology

Undominated       No Place in Coda, *Superheavy syllable - C,
                   One Geminate, Loanword Correspondence,
                   *Voiced [+cont] Geminate, Consecutive Unfilled Position
                   »

Dominated         Align »
                   *Voiced [-cont] Geminate »
                   Tonic Syllable »
                   Fill

Obviously, **Loanword Correspondence, Align** and **Tonic Syllable** do not exist in the native phonology and are loanword phonology specific. All other constraints are well-formedness conditions in Japanese and exist in the native lexicon. **Fill** and ***Voiced [-cont] Geminate** are undominated in the native lexicon (Zec 1994), but relaxed in loanword phonology. Other well-formedness constraints are undominated both in the native phonology and loanword phonology.

6. Conclusion

In this paper, an analysis of Consonant Gemination in English loans in Japanese was proposed, using Optimality Theory. The analysis shows that well-formedness conditions in Japanese are not sufficient to account for the realization of obstruents in loanwords. What is also called for is a set of constraints requiring that output forms be faithful to the foreign input. These constraints are loanword-specific.

As Yip (1993) argues, loanword phonology has two objectives: the first is to obey the well-formedness conditions in the core lexicon and the second is to mimic the foreign input as closely as possible. Although it has been claimed that a new constraint or a stronger version of an existing constraint is not added in loanword phonology (Ito and Mester 1993), it seems natural that loanword phonology is subject to a set of identity conditions, since the output form must mimic the properties of foreign input as closely as possible.
7. References


McCarthy, J. and A. Prince (1994) *The Emergence of the Unmarked Optimality in Prosodic Morphology*. Ms. University of Massachusetts, Amherst, and Rutgers University, New Brunswick, N.J.


Non-native Productions of Thai: Acoustic Measurements and Accentedness Ratings

Ratree Wayland

In this study, productions of 10 Thai words with all five tones, namely high, mid, low, falling and rising were elicited from 3 native Thai speakers and 6 native English speakers, who had learned Thai in adulthood. A variety of acoustic measurements including voice onset time of the initial consonant, vowel length, the first and second vowel formants (F1, F2), peak, valley and range of the fundamental frequency (F0) of all the five tones, were taken on both native and non-native productions. The two groups were found to differ in almost all acoustic parameters measured, but significantly in the F1, F2 and F0 valley values. Two native and all 6 non-native productions were rated for accentedness by three native listeners. The rating data suggested that the non-native production can be readily distinguished from the native production. Only some non-native production of some of the target words were considered to be native-like. More importantly, amount of the experience with Thai did not seem to affect the rating data. When the acoustic data were regressed on the means ratings, it was found that significant predictors varied from word to word (or from to tone).

1. Introduction

Pronunciation is an aspect that one tries to master when one engages in learning a second language (henceforth L2). In order to be understood, one would try to phonetically approximate the pronunciation of the target language. However, the empirical evidence now available suggests that only a few people who learn a second language beyond early childhood manage to pronounce L2 sounds in a native-like manner when their speech is carefully examined (McLaughin 1978). The reasons for this incompleteness of phonetic learning among adults are not well understood. The age of first exposure and the amount of second language experience have been claimed to be the two important factors; thus the focus of study, in the literature (Fathman 1975; Oyama 1982a; Ervin Tripp 1974; Cochrane and Sachs 1979).

Some researchers have tried to approach the foreign-accented phenomenon by looking at the relationship between sounds or phones in the sound systems of L1 and L2. One such approach is the Speech Learning Model developed by Flege (1987, 1988, 1991). According to this model, specific predictions can be made about which sounds or phones of L2 would pose difficulty for an L2 learner. Specifically, it is predicted that L2 phones that are 'similar' to existing L1 phones will be more difficult to learn, while 'new phones' (phones which fall outside the phonetic inventory of L1) will eventually be mastered.

* I would like to thank Professor John F. Hartmann from the Department of Foreign Languages and Literatures, Northern Illinois University at DeKalb for assisting me in getting the recordings of four non-native Thai speakers as well as conducting informal interviews to obtain their language background.
However, as pointed out by Munro (1993) "the successful evaluation of this or any other approach to the study of accented speech is contingent upon the availability of adequate descriptive data which characterize the patterns of "errors" made by L2 learners" (Munro 1993; p. 39). Information concerning which sounds are mispronounced as well as nature of the mispronunciation are equally crucial to researchers.

According to Flege (1988), there is the "phonetic norm" of a language. Unlike "pronunciation norm" which refers to "the collective judgment of native speakers concerning how a sound ought to be pronounced", "phonetic norm" is based on physical measurements of specific aspects of sound production" (p. 229). For example, a phonetic norm for stop consonants in English may be partially specified in terms of voice-onset time (henceforth VOT), that is the time interval between the release of the closure and the beginning of the vibration of the vocal folds. The VOT of English [t] in pre-stressed position might be 80 ms in a particular context, as compared to 15 ms in comparable speech material produced by monolingual native speakers of French (Flege 1988). Thus, a foreign accented speech can be defined as speech which differs acoustically (phonetic norm) from the native norms, and is auditorily detectable by native speakers (pronunciation norm). The speech productions of L2 learners can be compared to these norms to determine how native-like these productions are.

2. This Study

To my knowledge, no foreign-accented phenomenon of a tonal language has ever been investigated. Thus, this study is considered the first study involving an investigation of the production and perception of foreign-accented speech both on segmental and supra-segmental levels of a tonal language. Specifically, this study proposes to investigate the way in which two Thai vowels [a:] and [a:u],¹ the aspirated velar stop [kʰ], and five tones, namely low, mid, high, falling and rising, produced by 6 male native speakers of English differ from 3 male native speakers of Thai, and how these differences influence perceived degree of accentedness by 3 female Thai native listeners.

This study consists of two experiments: the production and the perception experiments. In Experiment 1, the acoustic analysis is performed to see which acoustic parameters differ in the two sets of data. The acoustic parameters examined are vowel formant frequencies (F1, F2), vowel duration, voice-onset time of the aspirated velar stop [kʰ], fundamental frequency (henceforth F0) peak, F0 valley and F0 range of all five tones. i.e. The F0 peak

¹ [a:u] can be phonologically analyzed as /a:/+final consonant /w/; however, here it is phonetically treated as two consecutive vowel segments (diphthong). Thus, measurements were taken from both [a:] and [u] segments.
is the highest point in the fundamental frequency contour of a tone. The F0 valley is the lowest point in the fundamental frequency contour of a tone. The F0 range, is the difference between F0 peak and F0 valley.

In Experiment 2, three female native speakers of Thai listened to and rated the non-native speakers' production data on a five-point scale of accentedness. The purpose of this experiment is to investigate the relationship between acoustic parameters measured in Experiment 1 and the native judges' ratings in order to establish which properties of the non-native productions lead to a perception of accentedness. Rating of individual speakers' production are also examined to investigate individual differences in degree of accentedness.

3. Literature Review

A substantial number of researchers have investigated differences between native and non-native productions in terms of various acoustic parameters. For instance, studies by Flege (1980, 1987), Flege and Munro (1984), Flege, Munro and Skelton (1992), Caramazza, Yeni-Komshian, Zurif, and Carbone (1973), and Williams (1980) among others, have found that non-native speakers do not produce native-like voice-onset times. It was also found that second language learners of English produced very small differences in vowel durations before voiced and voiceless stops, (Flege et al. 1992; Crawther and Mann 1992; Port and Mitleb 1983; Mack 1982). Some studies also compared formant frequencies of vowels produced by native and non-native speakers (Flege 1987; Flege and Hillenbrand 1984). Results indicated that the formant values of 'similar' L2 vowels produced by non-native speakers differed significantly from those produced by native speakers.

However, as pointed out by Munro (1993), one of the drawbacks of studies using only one approach to investigate the 'foreign-accented' phenomenon, i.e. only comparison of native and non-native production is that it does not reveal which characteristics of the non-native productions cause native listeners to hear them as having a foreign accent. Detailed acoustic analysis may reveal differences between native and non-native speeches on various parameters. However, these differences may influence native speaker's perception of accentedness to a varying degree. Thus, acoustic data should be related to perceptual data.

In the literature, several methods have been employed to evaluate degree of accentedness. In some studies (Flege and Hillenbrand 1984), native listeners were asked to identify a single phoneme excised from natural speech or to identify an entire word produced by non-native speakers. It was found that non-native words were less well
identified in comparison to native tokens. Flege (1984) asked native speakers to judge whether utterances of various durations were produced by native or non-native speakers of English. These data indicated that the rate of correct detection of accent tended to be higher for relatively long intervals of speech (i.e. phrases) than short ones (single sounds), yet remained at significantly above-chance levels as the intervals were reduced from syllable-sized to segment-sized intervals, and finally to just the first 30 ms of /t/ (roughly, the burst portion of the sound /t/)" (Flege 1988, p.233). The capability of native listeners to detect "accentedness" in such a short interval of speech "indicates a high degree of sensitivity to divergences from native speaker norms" (Munro 1993, p. 41).

Various other techniques were also employed by researchers. In some cases, native listeners were asked to directly rate utterances on a scale of accentedness. However, no standard scale has been developed. Suter (1976) had native listeners score degree of accentedness of a two-minute utterance using a six-point scale. Snow and Hoefnagel-Höhle (1982) used a five-point scale to obtain rating of individual sounds in a list of words repeated after a recorded model. In order to obtain ratings for phones, intonation, rhythm and an overall pronunciation, Schneiderman et al (1988) used multiple five-point scales. Munro (1993) used a one hundred-point scale to obtain the accentedness rating for English vowels produced by Arabic native speakers.

The advantage or disadvantage of using one rating scale over another is not obvious in the literature. In research concerned with assessment of speech and voice quality, however, it was found that intrarater reliability was not related to the scale type or statistic used (Kreiman, Gerratt, Kempster, Erman and Berke 1993).

While the studies mentioned above used linguistically-trained judges, others have used naive native listeners (Flege and Eefting 1987; Flege 1988; Flege and Fletcher 1992). It has been shown that ratings obtained from linguistically-trained and naive listeners are fairly consistent to one another (Brennan and Brennan 1981; Cunningham-Andersson and Engstrand 1989).

4. Thai

Thai or Siamese is the national language of Thailand spoken by approximately 60 million people. The dialect spoken in the capital city of Bangkok is considered the standard one. However, outside Bangkok and the central plains, other languages of the Tai and Mon-Khmer family coexist with this standard dialect.
4.1 Consonants

An interesting feature of the Thai consonant system is the existence of three classes of stop consonants, namely voiced, voiceless aspirated and voiceless unaspirated. All consonants can occur initially, but only voiceless unaspirated stops, nasals and semi-vowels can occur finally. All final consonants are phonetically unreleased.

<table>
<thead>
<tr>
<th>Bilabial</th>
<th>Inter-dental</th>
<th>Alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stops</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced</td>
<td>b</td>
<td>d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vls.unasp</td>
<td>p</td>
<td>t</td>
<td>c</td>
<td>k</td>
<td></td>
</tr>
<tr>
<td>Vls.asp</td>
<td>pʰ</td>
<td>tʰ</td>
<td>cʰ</td>
<td>kʰ</td>
<td>h</td>
</tr>
<tr>
<td>Fricatives</td>
<td>f</td>
<td>s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sonorants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasals</td>
<td>m</td>
<td>n</td>
<td></td>
<td></td>
<td>η</td>
</tr>
<tr>
<td>Laterals</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trill/tap</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semi-vowels</td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
</tr>
</tbody>
</table>

4.2 Vowels

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Back-unrounded</th>
<th>Back-rounded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>i</td>
<td>i</td>
<td>u</td>
</tr>
<tr>
<td>Mid</td>
<td>e</td>
<td>ə</td>
<td>o</td>
</tr>
<tr>
<td>Low</td>
<td>e</td>
<td>a</td>
<td>ɔ</td>
</tr>
</tbody>
</table>

Diphthongs ia ia ua (Hudak, 1990)

Each of these 9 vowels may occur phonemically as short or long, for example short [a] and long [aː]. Phonetically, the long vowels are approximately twice as long as the short vowels. All 18 nuclei may occur alone, with an initial consonants, with a final consonant or with both initial and final consonants.

4.3 Tones

Each syllable in Thai carries one of the five phonemic tones, namely a mid tone ([kʰaː] 'to be stuck or lodged in'); a low tone ([kʰaː] 'galonga, a kind of aromatic root used in Thai cooking'); a falling tone ([kʰaː] 'I, servant'); a high tone ([kʰáː] 'to engage in trade'); a rising tone ([kʰāː] 'leg'). These five tones may be characterized in terms of pitch contour,
pitch height, and glottalised or non-glottalised voice quality. This can be schematically presented as below (adapted from Hudak 1990, p. 34).

<table>
<thead>
<tr>
<th>Tones</th>
<th>Tone mark</th>
<th>Pitch contour</th>
<th>Pitch Height</th>
<th>Voice quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>mid</td>
<td>unmarked</td>
<td>level</td>
<td>medium</td>
<td>non-glottalised</td>
</tr>
<tr>
<td>low</td>
<td>~</td>
<td>level</td>
<td>low</td>
<td>non-glottalised</td>
</tr>
<tr>
<td>falling</td>
<td>~</td>
<td>contour</td>
<td>low to high</td>
<td>glottalised</td>
</tr>
<tr>
<td>high</td>
<td>~</td>
<td>level</td>
<td>high</td>
<td>glottalised</td>
</tr>
<tr>
<td>rising</td>
<td>~</td>
<td>contour</td>
<td>low to high</td>
<td>non-glottalised</td>
</tr>
</tbody>
</table>

5. The Experiment

5.1 Experiment 1: Acoustic Measurements

This experiment is conducted to establish the differences between native and non-native Thai productions. Acoustic parameters measured including vowel formants (F1, F2, F1-F2), VOT, vowel duration, F0 peak, F0 valley and F0 range

5.1.1 Methods

Subjects: The participants were three male native Thai speakers, all recruited from the student population of Cornell University, and 6 male native English speakers. Four were recruited from Northern Illinois University at DeKalb, and the other two from Cornell University. An informal interview was conducted to obtain information on each English-native speaker’s language background. All three native Thai speakers are from Bangkok area and are between 19-24 years of age. Two have been living in the United states for the past three years, and the other for six years. They are designated as speakers 1, 2 and 3 in this study. The non-native group consisted of 6 native English speakers, between 20 and 40 years of age. They are designated as speakers 4 to 9 in this study.

5.1.2 Language Background

In this section, language background of 6 native speakers of English obtained from an informal interview is presented. They are arranged according to number of years of experience with Thai.

Speaker 6 Speaker 6 has been speaking Thai for more than 12 years. 30-40% of his daily conversation is carried out in Thai. He spends approximately 6 hours per week
reading Thai. He also was a teacher of Thai, and had spent approximately 6 weeks in Thailand.

**Speaker 4** Speaker 4 received language training through the U.S. Peace-Corps in Thailand, and spent an additional 3 years working there. He has been speaking Thai for the past 9 years, and approximately 20% of his daily conversation is carried out in Thai. He also understands and speaks some Lao and Khmer (Cambodian).

**Speaker 7** Speaker 7 has been speaking Thai for almost 8 years. He spends approximately half an hour per day reading Thai, but currently does not have much opportunity to speak. He understands some Lao, and had lived in Thailand for two years.

**Speaker 9** Speaker 9 has been speaking Thai for 6 years. His daily conversation at home is carried out in Thai. He has studied Pali, Sanskrit and Arabic. Altogether, he has spent approximately one year in Thailand.

**Speaker 8** Speaker 8 has been learning and speaking Thai for three and a half years. He spends around 12 to 15 hours reading Thai, but does not have much opportunity to speak, except for 3 hours per week in the Thai class that he is currently taking. He also knows Spanish and some Tamil.

**Speaker 5** Speaker 5 also received training through the U.S. Peace-Corps in Thailand, and spent approximately 2 years working there. He has also been auditing the advanced Thai class offered at Cornell University. However, he hardly has any opportunity to speak Thai on a daily basis, yet has been reading on a regular basis.

### 5.1.3 Materials and Procedure

Recordings for all three Thai-native speakers and two English-native speakers recruited from Cornell University were made in a sound-proof booth, using a cardioid microphone (Electrovoice, model RE 20) and high quality cassette recorder (Marantz, Model PMD 222). Subjects were asked to read a list of [KHA:U], and [NA:] words, with all five tones in a sentence frame "I said the word.....". (see word list below). Each word was repeated twice in a random fashion. The words, as well as the carrier phrase were written in Thai, using standard Thai orthography.

The recordings of four English-native speakers from Northern Illinois University at DeKalb were made at the main sound-proof studio of the University, using a cordioid
microphone (Electrovoice, Model RE 20) and a cassette recorder (Tascam Model 122). The recording process was carried out under the supervision of the studio technician. The same word list was used.

### 5.1.4 Word list

<table>
<thead>
<tr>
<th>ມະ [kʰaːu]</th>
<th>‘fishy smell’</th>
<th>ປາ [naː]</th>
<th>‘rice field’</th>
</tr>
</thead>
<tbody>
<tr>
<td>ມະ [kʰàːu]</td>
<td>‘news’</td>
<td>ປາ [nâː]</td>
<td>‘a nick name’</td>
</tr>
<tr>
<td>ມະ [kʰáːu]</td>
<td>‘rice’</td>
<td>ປາ [nâː]</td>
<td>‘face’</td>
</tr>
<tr>
<td>ມະ [kʰáːu]</td>
<td>‘a kind of fish’</td>
<td>ປາ [nâː]</td>
<td>‘mother’s younger siblings’</td>
</tr>
<tr>
<td>ມະ [kʰâːu]</td>
<td>‘white’</td>
<td>ປາ [nâː]</td>
<td>‘thick’</td>
</tr>
</tbody>
</table>

The recordings were digitized on a SUN 3/160 computer at 11 kHz, with a low-pass filter setting of 6 kHz and stored as files to be processed by the commercial software package WAVES+. This speech analysis package enabled us to simultaneously examine wave forms and (wide-band) spectrograms of each token. The target words were extracted from the carrier phrase and stored as separate files. These edited token were then submitted to LPC (Linear Predictive Coding) analysis using Hamming window of 25.6 ms, with 16 poles and preemphasis of .98. F1, F2 values were measured from the steady states of the vowels by placing the cursors on the formant tracks.

For the diphthong [aːu], F1 and F2 values were measured from these two steady state portions displayed on the wide-band spectrogram (e1, e2 and e3, e4 in Figure 1).

Vowel duration was measured by positioning the cursors at the vowel onset and offset (d1, d2). Vowel onset was taken to be the onset of the periodicity in the wave form. Vowel offset was indicated by the loss of F2 on the spectrogram.

The voice-onset time of the aspirated velar stop [kʰ] was also measured.
Figure 1. Segmented and labeled waveform and wide-band spectrogram of the word [kʰau] ‘rice’ as produced by Speaker 1. VOT is voice-onset time, d1 marks the beginning of the vowel, d2 marks the end of the vowel, e1 e2 marks the steady state of [aː] and e3 e4 marks the steady state of [u].

A pitch track (a representation of fundamental frequency over time) was derived for all of the tokens, producing a data point every 5 msec with a sampling frequency of 11 kHz and an LPC order of 16. Two F0 measurements were recorded as data for each token. These included the F0 peak and F0 valley. F0 peak was the highest F0 value on the F0 contour, and F0 valley was the lowest F0 value. F0 range was the difference between F0 peak and F0 valley.

Following Cooper & Sorenson (1981), to minimize the chance of including spurious F0 estimates in the data, F0 peak and F0 valley measurements were taken in the location where two or more contiguous occurrences of the same F0 value were observed (see Figure 2).
Figure 2. Waveform and pitch track of the word [kʰau] ‘rice’ as produced by Speaker 1. F0 peak marks the highest F0 value of the pitch contour (tone), and F0 valley marks the lowest F0 value.

5.1.5 Results and Discussion

5.1.5.1 Duration data

For the word [KHA:U] pronounced with all five tones, both vowel and voice-onset time duration values are log-transformed and submitted to a multivariate analysis with native language as an independent variable. The analysis reveals no significant language effect for either voice-onset time duration [p=.271], or vowel duration [p=.90]. Similar procedure is carried out for the vowel duration of all five tones of the word [NA:]. The analysis reveals no significant effect for native language [p=.145]. The mean durations in msec of both vowels (i.e. [a:]) in [NA:] and [au] in [KHA:U]), and voice-onset time durations for [kʰ] in [KHA:U] are given in Table 1 below.
Table 1. Mean vowel durations (in msec) and voice-onset time durations (in msec) for native Thai group and native English group. The means are based on 2 repetitions of each of the five tones from each speaker. The means of all five tones are shown in bold. Standard deviations are given in the parentheses.

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Tones</th>
<th>VOT (msec.)</th>
<th>Duration (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Native Thai</td>
<td>Native English</td>
</tr>
<tr>
<td>[aːu]</td>
<td>[kʰaːu] mid</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>[kʰəːu] low</td>
<td>88</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>[kʰəːu] falling</td>
<td>68</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>[kʰəːu] high</td>
<td>84</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>[kʰəːu] rising</td>
<td>69</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>80 (11)</strong></td>
<td><strong>88 (6)</strong></td>
</tr>
<tr>
<td>[aː]</td>
<td>[naiː] mid</td>
<td></td>
<td>371</td>
</tr>
<tr>
<td></td>
<td>[nəː] low</td>
<td></td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>[nəː] falling</td>
<td></td>
<td>343</td>
</tr>
<tr>
<td></td>
<td>[nːaː] high</td>
<td></td>
<td>366</td>
</tr>
<tr>
<td></td>
<td>[nːaː] rising</td>
<td></td>
<td>358</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>351 (23)</strong></td>
<td><strong>383 (17)</strong></td>
</tr>
</tbody>
</table>

5.1.5.2 Spectral Data

All spectral data for the word [KHAːU], namely vowel formants (F1, F2 and F2-F1) for the first and second elements of the diphthong [au] as well as F0 peak, F0 valley and F0 range were log-transformed and submitted to a multivariate analysis of variance with native language as an independent variable. The analysis revealed a significant effect on F2 [p<.023] and F2-F1 [p<.043] of the first element, i.e. [aː] of the diphthong [au], F1 of the second element i.e. [u] of the diphthong [au] [p <.003], and F0 valley [p<.006].
To examine individual differences, Student-Newman Keuls post-hoc analysis was performed. This analysis revealed that in general three native Thai speakers, namely speakers 1, 2 and 3 formed a homogeneous group, while differences among native English speakers existed in almost all acoustic parameters. No general pattern or conclusion could be drawn from the native English speakers' production.

Similarly, all spectral data measured from all five tones of the word [NA:] were log-transformed and submitted to a multivariate analysis of variance with native language as an independent variable. The analysis revealed a significant effect on F1 [p<.009], F2 [p<.000], and valley [p<.009].

Similar to [KHA:U], when student-Newman Keuls post-hoc analysis was performed, the results suggested that while native Thai speakers may form a homogeneous group, no general pattern emerged from the native English speaker group. Their production of Thai appeared to differ from one another in nearly all acoustic parameters measured.

The mean values of F1, F2 for vowel [a:] in [NA:] are given in Table 2, and the mean values of F1, F2 of the first and second element of the diphthong [au] are given in Table 3. Table 4 illustrates the mean values of F0 peak, F0 valley and F0 range for both [NA:] and [KHA:U].

**Table 2.** Mean F1 and F2 values (in Hz) of the vowel [a:] for native Thai and native English groups. The means were based on two repetitions of each of the five tones from each speaker. Standard deviations are given in parentheses.
Table 3. Mean F1 and F2 values (in Hz) of the first and second element of diphthong [au] (i.e. [a:] and [u]) of all five tones of the word [KHA:U] for native Thai and native English groups. The means were based on two repetitions of each five tones from each speaker. Standard deviations are given in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>Native Thai</th>
<th></th>
<th>Native English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[a:]</td>
<td>[u]</td>
<td>[a:]</td>
<td>[u]</td>
</tr>
<tr>
<td>Tone</td>
<td>F1</td>
<td>F2</td>
<td>F1</td>
<td>F2</td>
</tr>
<tr>
<td>mid</td>
<td>817</td>
<td>1408</td>
<td>575</td>
<td>946</td>
</tr>
<tr>
<td>low</td>
<td>819</td>
<td>1407</td>
<td>608</td>
<td>987</td>
</tr>
<tr>
<td>falling</td>
<td>804</td>
<td>1489</td>
<td>579</td>
<td>1006</td>
</tr>
<tr>
<td>high</td>
<td>822</td>
<td>1421</td>
<td>456</td>
<td>889</td>
</tr>
<tr>
<td>rising</td>
<td>827</td>
<td>1416</td>
<td>482</td>
<td>920</td>
</tr>
<tr>
<td></td>
<td><strong>818</strong></td>
<td><strong>1428</strong></td>
<td><strong>540</strong></td>
<td><strong>950</strong></td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>(34)</td>
<td>(67)</td>
<td>(48)</td>
</tr>
</tbody>
</table>


Table 4. Mean value of F0 peak, F0 valley and F0 range (in Hz) of both target words [KHA:U] and [NA:] spoken with five tones for native Thai and native English groups. The means are based on two repetitions of each of the five tones from each speaker. Standard deviations are given in parentheses.

<table>
<thead>
<tr>
<th>word</th>
<th>Native Thai</th>
<th>Native English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak</td>
<td>Valley</td>
</tr>
<tr>
<td>[kʰaːu]</td>
<td>128</td>
<td>97</td>
</tr>
<tr>
<td>[kʰaːu]</td>
<td>117</td>
<td>76</td>
</tr>
<tr>
<td>[kʰɔːu]</td>
<td>140</td>
<td>82</td>
</tr>
<tr>
<td>[kʰaːu]</td>
<td>137</td>
<td>113</td>
</tr>
<tr>
<td>[kʰɔːu]</td>
<td>147</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>134 (12)</td>
<td>91 (15)</td>
</tr>
<tr>
<td>[naː]</td>
<td>120</td>
<td>99</td>
</tr>
<tr>
<td>[nəː]</td>
<td>114</td>
<td>75</td>
</tr>
<tr>
<td>[náː]</td>
<td>135</td>
<td>82</td>
</tr>
<tr>
<td>[nəː]</td>
<td>127</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>128 (12)</td>
<td>91 (13)</td>
</tr>
</tbody>
</table>

5.1.6 Summary

In summary, the non-native production of Thai differ from the native-Thai production in several acoustic parameters measured. The average VOT of [kʰ] in [KHA:U], and vowel duration (both [aːu] and [aː]) produced by native English speakers were longer than those produced by native Thai speakers (Table 1). Vowel formants (F1, F2) of [aː] in [NA:] produced by native English speakers were lower than those produced by native Thai (Table 2). However, when a statistical analysis was performed, it was found that for [KHA:U], the two groups differed significantly in:
1. F1 value of the first element of the diphthong [aːu] (i.e. [aː]); the native English value was lower than that of the native Thai (783 vs. 818 Hz.).

2. F1 value of the second element of the diphthong [aːu] (i.e. [u]). The native English value was higher than that of the native Thai (540 vs. 633 Hz.).

3. F2-1 (the difference between F1 and F2) of the first element, i.e. [aː] of the diphthong [aːu]. The native English value was lower than that of the native Thai (559 vs. 610 Hz.).

4. F0 valley value of all five tones. The native English values were higher that those of the native Thai (107 vs. 91 Hz.)

For [NAː] the two groups differed significantly in:

1. F1 value of [aː]. Native English's value was lower than that of the native Thai (710. vs. 770 Hz.).

2. F2 value of [aː]. Native English's value was lower than that of the native Thai (1403 vs. 1542 Hz.).

3. F0 valley values. Native English's value was higher than that of the native Thai (105 vs. 91 Hz.).

However, individual differences also existed. In general, it was found that native Thai speakers, i.e. speakers 1, 2 and 3, form a homogeneous group, while significant differences in almost all acoustic parameters measured existed among native English speakers.

5.2 Experiment 2: Accentedness Judgments

In Experiment 2, the acoustic measurements from Experiment 1 were regressed on a set of accentedness ratings assigned by three judges in order to determine the properties associated with perceived accentedness in all of the native English data. As pointed out in Munro (1993), an implicit assumption in this kind of experiment is the idea that native speakers have (implicit) knowledge of a good exemplar of all or most of acoustic parameters of the language. For example, a native Thai speaker would know what a good exemplar of a high tone should sound like, what the appropriate VOT duration for the unaspirated [k] is, and so on. When s/he assigns accentedness rating, s/he would make reference to such knowledge. The exact nature of this process, however, is not well understood. One possible explanation is that tokens being rated are compared with an abstract prototype (Flege, 1984). Thus, tokens may be rated as having varying degrees of accentedness depending on the extent of the difference in all or some acoustic parameters between rated tokens and the 'prototypes'. If the prototype account of perceived
accentedness is correct, the relationship between acoustics properties and rating data might be observed by using multiple regression analysis.

5.2.1 Methods

Two replications of each repetition (from experiment 1) of all 10 target words, i.e. all five tones of the words [NA:] and [KHA:U] (10 words X 2 repetitions X 2 replications of each repetition X 8 speakers =320 in total) were recorded in random order. Data from all six English-native and two Thai-native speakers, namely speaker 1 and 3 were used. Ten practice trials were also included. The stimuli were presented in a block of ten replications with inter trial interval of 3 sec and an interblock interval of 5 sec. The native data were included to ensure that the listeners could distinguish between native and non-native speakers.

5.2.2 Procedure

Stimuli were presented to three female native-Thai judges recruited from the undergraduate student population of Cornell University. None of them were linguistically-trained. The three judges were from the central, southern, and northeastern regions of Thailand respectively. All reported normal hearing. The stimuli were played on the Aiwa cassette recorder (Model AD-F400), and were presented at a comfortable listening level over the Sony Dynamic Stereo Headphone MDR V6.

All three judges were given a print out of all 320 stimuli, printed in standard Thai script. A scale of one to five was also included to the right of each stimulus in the print out. A score of one indicated strongest degree of accentedness, and a score of five indicated native like production. To assign ratings, the judges were asked to give a score to each stimulus by circling number 1 to 5 on the scale, depending on degree of perceived accentedness.

To assess whether any of the judges differed significantly from the others in their rating scores assigned to the speakers, Pearson correlation coefficients were calculated for all the ratings for all possible pairs of the judges. Since the correlations obtained were in the range of 0.97 to 0.99 and were significant at the 0.01 level, it was concluded that the judges are significantly consistent with one another.

5.2.3 Results

The rating scores for [KHA:U] and [NA:] for both groups of speakers average across 3 judges are shown in Table 5. It is organized by words and native languages. Overall mean values of all five tones of both [KHA:U] and [NA:] are also given. The mean score for the
native Thai are 4.84 for both [KHA:U] and [NA:]. For the native English, on the other hand, the mean score for [KHA:U] is 3.52, and 3.97 for [NA:]. On the average, the mean score for the non-native group is around 1 point lower than the native group.

**Table 5.** Overall mean score of each of the target words for native Thai and native English groups by 3 judges. The mean scores across five tones of each target word are given in bold. Standard deviations are given in the parenthesis.

<table>
<thead>
<tr>
<th>Word</th>
<th>Native Thai</th>
<th>Native English</th>
<th>Word</th>
<th>Native Thai</th>
<th>Native English</th>
</tr>
</thead>
<tbody>
<tr>
<td>kʰaːu</td>
<td>4.88</td>
<td>3.44</td>
<td>naː</td>
<td>4.88</td>
<td>4.03</td>
</tr>
<tr>
<td>kʰàːu</td>
<td>4.71</td>
<td>3.35</td>
<td>nàː</td>
<td>4.63</td>
<td>3.69</td>
</tr>
<tr>
<td>kʰâːu</td>
<td>4.92</td>
<td>3.63</td>
<td>nāː</td>
<td>5.00</td>
<td>4.32</td>
</tr>
<tr>
<td>kʰáːu</td>
<td>4.79</td>
<td>3.28</td>
<td>náː</td>
<td>4.75</td>
<td>3.44</td>
</tr>
<tr>
<td>kʰǎːu</td>
<td>4.88</td>
<td>3.89</td>
<td>nǎː</td>
<td>4.94</td>
<td>4.38</td>
</tr>
<tr>
<td>mean</td>
<td>4.84 (.08)</td>
<td>3.52 (.25)</td>
<td>mean</td>
<td>4.84 (.15)</td>
<td>3.97 (.40)</td>
</tr>
</tbody>
</table>

Data in Table 5 also shows that mean scores of level tones, namely mid, low and high are lower than those of contour tones, i.e. falling and rising. More interestingly, for the English-native group, rating scores from low to high assigned to each individual tone are in the following order: high tone, low tone, mid tone, falling tone and rising tone. This pattern is consistent for both sets of words, i.e. [KHA:U] and [NA:].
Table 6. Overall mean score of each target words for native-Thai and native-English by 3 judges. Speaker 1 and 3 are native-Thai and speakers 4-9 are native-English. Native-English speakers are arranged according to number of years of experience with Thai.

<table>
<thead>
<tr>
<th>Speakers</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>4</th>
<th>7</th>
<th>9</th>
<th>8</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>Native-Thai</td>
<td>Native-English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kʰaːu</td>
<td>5</td>
<td>4.74</td>
<td>3.25</td>
<td>4.9</td>
<td>3.15</td>
<td>3.48</td>
<td>3.33</td>
<td>2.48</td>
</tr>
<tr>
<td>kʰàːu</td>
<td>4.8</td>
<td>4.55</td>
<td>2.55</td>
<td>4.83</td>
<td>2.97</td>
<td>2.4</td>
<td>4.25</td>
<td>3.09</td>
</tr>
<tr>
<td>kʰáːu</td>
<td>4.9</td>
<td>4.9</td>
<td>2.9</td>
<td>5</td>
<td>3.34</td>
<td>3.15</td>
<td>3.98</td>
<td>3.34</td>
</tr>
<tr>
<td>kʰáːu</td>
<td>4.75</td>
<td>4.84</td>
<td>2.34</td>
<td>4.4</td>
<td>3.34</td>
<td>3.09</td>
<td>3.08</td>
<td>3.34</td>
</tr>
<tr>
<td>kʰāːu</td>
<td>5</td>
<td>4.74</td>
<td>3.75</td>
<td>4.25</td>
<td>3.49</td>
<td>3.55</td>
<td>4.67</td>
<td>3.55</td>
</tr>
<tr>
<td>Mean</td>
<td>4.89</td>
<td>4.75</td>
<td>2.96</td>
<td>4.68</td>
<td>3.26</td>
<td>3.13</td>
<td>3.86</td>
<td>3.16</td>
</tr>
<tr>
<td>naː</td>
<td>5</td>
<td>4.75</td>
<td>3.65</td>
<td>4.24</td>
<td>4.03</td>
<td>4.55</td>
<td>3.4</td>
<td>4.24</td>
</tr>
<tr>
<td>nàː</td>
<td>4.67</td>
<td>4.65</td>
<td>2.99</td>
<td>4.59</td>
<td>3.55</td>
<td>2.9</td>
<td>4.7</td>
<td>3.33</td>
</tr>
<tr>
<td>nāː</td>
<td>5</td>
<td>5</td>
<td>3.24</td>
<td>4.65</td>
<td>4.58</td>
<td>4.49</td>
<td>4.25</td>
<td>4.67</td>
</tr>
<tr>
<td>náː</td>
<td>4.74</td>
<td>4.75</td>
<td>1.9</td>
<td>4.9</td>
<td>3.48</td>
<td>3.9</td>
<td>3.4</td>
<td>2.98</td>
</tr>
<tr>
<td>nāː</td>
<td>5</td>
<td>4.83</td>
<td>4.5</td>
<td>4.08</td>
<td>4.66</td>
<td>4.17</td>
<td>4.55</td>
<td>4.24</td>
</tr>
<tr>
<td>Mean</td>
<td>4.88</td>
<td>4.80</td>
<td>3.26</td>
<td>4.49</td>
<td>4.06</td>
<td>4.00</td>
<td>4.06</td>
<td>3.89</td>
</tr>
</tbody>
</table>

Table 6 shows rating scores for each individual speaker. Speakers 1 and 3 are native Thai and speakers 4 to 9 are native English. The native English speakers' scores are arranged in order of number of years of experience with Thai. The mean score of each word average across the native English group ranges from 2.96 to 4.68 for [KHAːU] and 3.26 to 4.49 for [NAːː]. For the native group, it ranges from 4.55 to 5 for [KHAːU] and 4.65 to 5 for [NAːː]. However, an examination of individual score reveals that for the native-Thai group, the score ranges from 4.25 to 5, and from 1.75 to 5 for the native-English group.

Since the rating scores of each native-English speaker is arranged according to the number of years of experience with Thai, it is easy to see that there does not seem to be a strong correlation between the scores received and the amount of experience with the target
language. Speaker 6, for example, had the greatest amount of experience with Thai, yet his scores were much lower than every other speaker in the group.

5.2.4 Multiple Regression Analysis

5.2.4.1 Predictor Variables

All acoustic parameters of the target words measured in the Experiment1 were transformed as described below and used as predictors in the stepwise linear multiple analysis. As pointed out in Munro (1993), the simple use of raw measurement values may not yield a correlation between the acoustic predictors variables with the rating data. A solution to this problem is to quantify the difference between the non-native data and the native data. The assumption is, of course, that the native data are close to the ideal values for all parameters measured. This, however, may not be the case since only a small sample of speakers were investigated.

In computing these variables, an attempt was made to characterize how much the rated non-native tokens differed acoustically from a good exemplar of Thai tokens. Thus, all predictors were calculated by subtracting mean values obtained from the native group (3 speakers) in Experiment1 from the corresponding mean values of each of the non-native speakers. the rated tokens. The values of vowel duration predictor for the word [KHA:U], for instance, were computed by subtracting the mean value of the native Thai group from the mean value of rated token of each speaker in the non-native group.

The remaining predictors were computed in the same manner. Each of the differences was then squared to yield a positive number in order that a relatively large value of any predictor would indicate a large discrepancy between it and the native mean, regardless of direction. The values of these predictors were taken to represent an assessment of the acoustic distance between the non-native tokens and a hypothetical, ideal native-like token.

5.2.4.2 Analysis

[KHA:U]

In a stepwise linear analysis, all 8 acoustic parameters measured from the target word [KHA:U], namely F1, F2, F2-F1, vowel duration, F0 peak, F0 valley and F0 range were used as predictors and the rating score was used as a dependent variable. The results indicated that for all five tones of this target word, only F0 valley and F2 of the second element, i.e. [u] of the diphthong [a:u] were significant predictors. F0 valley accounted for an additional 26% of the variance in the rating score, and F2 accounted for an additional 17%.
It is possible, however, that predictors for each individual tone or word will vary. Thus, a separate stepwise linear analysis was also performed for each individual tone of the word [Kha:U]. It was found that there was no single significant predictor for low tone [kʰâu] ‘news’, falling tone [kʰâu] ‘rice’, and rising tones [kʰâu] ‘white’.

For mid tone [kʰâu] ‘fishy smell’, the results indicated that F2-F1 of the second element of the diphthong [aːu] accounted for 75% of the variance in the rating data, vowel duration accounted for an additional 20%, F0 valley 4.8%, and F2-F1 of the first element of the diphthong [aːu], i.e. [aː] accounted for an additional 0.2%. These four predictors accounted for 100% of all the variance in the rating data.

For high tone [kʰâu] ‘a kind of fish’, the analysis revealed that F0 range accounted for 71% of the variance in the rating data, F0 valley accounted for an additional 25% and F2-F1 of the first element of the diphthong [aːu], i.e. [aː] accounted for an additional 4%. All of these three predictors accounted for 100% of the variance in the rating data.

Table 7. Results of stepwise multiple regression analyses of the words [kʰâu] ‘fishy smell’, [kʰâu] ‘a kind of fish’, [naː] ‘rice field’, [nâː] ‘face’ and [nâː] ‘a younger sibling of mother’. Predictors that accounted for a significant portion of the variance in the rating data are given. They were in the order shown in the table.

<table>
<thead>
<tr>
<th>Word</th>
<th>Predictors</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>[kʰâu] ‘fishy smell’</td>
<td>1. F2-F1 (of the second element of the diphthong [aːu])</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>2. Vowel Duration</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>3. F0 Valley</td>
<td>.998</td>
</tr>
<tr>
<td></td>
<td>4. F2-F1 (of the first element of the diphthong [aːu])</td>
<td>1.00</td>
</tr>
<tr>
<td>[kʰâu] ‘a kind of fish’</td>
<td>1. F0 Range</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>2. F0 Valley</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td>3. F2-F1 (of the first element of the diphthong [aːu])</td>
<td>.997</td>
</tr>
<tr>
<td>[naː] ‘rice field’</td>
<td>F0 Valley</td>
<td>.68</td>
</tr>
<tr>
<td>[nâː] ‘face’</td>
<td>F0 Peak</td>
<td>.68</td>
</tr>
<tr>
<td>[nâː] ‘younger sibling of mother’</td>
<td>F2</td>
<td>.89</td>
</tr>
</tbody>
</table>
Similarly, all acoustic parameters measured from the target word [NA:], namely F1, F2, F2-F1, vowel duration, F0 peak, F0 valley and F0 range were used as predictors in a stepwise multiple regression analysis. For all five tones, the results suggested that F0 valley was the significant predictor. It yielded relatively strong multiple R of .81 which indicated that it accounted for 65% of the variance in the rating data. A separate stepwise linear analysis was also carried out for each individual tone. The results indicated that for mid tone [na:] ‘rice field’, F0 valley yielded multiple R of .82 and accounted for 68% of the variance in the rating data.

For low tone [nː:] ‘a nick name’ and rising tone [nːː] ‘thick’, there is no single significant predictor. For falling tone [nːː] ‘face’, F0 peak yielded a moderate multiple R of .83 and accounted for 68% of the variance in the rating data. For high tone [nːː] ‘younger sibling of mother’, F2 yielded a strong multiple R of .95 and accounted for 89% of the variance in the rating data.

In summary, for all five tones of both [KHA:U] and [NA:] F0 valley predictor was moderately correlated with the judges’ rating. In addition, for [KHA:U], F2 of the second element of the diphthong [aːu] was also another predictor. The predictors, however, varied from word to word, or from tone to tone. For mid tone [kʰaːu] ‘fishy smell’, the F2-F1, vowel duration, F0 peak and F0 valley were significant predictors. For high tone [kʰaːu] ‘a kind of fish’, significant predictors were F0 valley, F0 range and F2-F1 of the first element of the diphthong [aːu]. For low, falling and rising tones, there was no significant predictor.

For mid tone [naː] ‘rice field’, F0 valley was the significant predictor. For falling tone [nːː] ‘face’, F0 peak was the significant predictor, and for high tone [nːː] ‘a young sibling of mother’, F2 was the significant predictor.

6. Conclusion

Unlike most previous studies on foreign-accented speech, this study investigated this phenomenon in a tonal language, i.e. Thai. Moreover, an attempt to relate production to perception data was also an important feature of this study.

When the production of Thai by 6 native-English speakers was carefully examined along several acoustic parameters in the first experiment, it was evident that the non-native production differed from the native production in all acoustic parameters measured. However, when a statistical analysis was performed, the results revealed that only differences in spectral data, namely vowel formants (F1, F2) and fundamental frequency data between the two groups were significant. Duration data, i.e. voice-onset-time and
vowel duration, on the other hand, was found to be comparable between two groups of speakers.

When the native and non-native production data were judged for degree of accentedness on a five-point scale by three female native-Thai listeners in Experiment 2, the results indicated that, on the basis of pronunciation, non-native Thai utterance can be readily distinguished from the native-Thai utterance. On the average, rating scores of the non-native group was approximately 1 point lower than that of the native group. This seemed to suggest that the non-native production was not dramatically different from the native production, and that in quite a few cases the non-native tokens were heard as native-like. A few tokens produced by speaker 4, for example, received a score of 5 indicating native-like production. The fact that only single words produced in a short carrier phrase were investigated may have accounted for this finding. An examination of longer intervals of speech, i.e., a paragraph or various durations of speech produced in a natural setting, (as opposed to a laboratory setting) may offer different results.

When the mean score of individual tokens was examined, it was found that it ranged from 4.25 to 5 for the native-Thai group, and from 1.75 to 5 for the native-English group. The wider range of scores assigned to the non-native token may lead one to hypothesize that amount of experience with the Thai language may be the source of explanation. The data on Table 6 in which the rating scores were organized according to the speakers' number of years of experience with Thai proved to the contrary. On the average, speaker 6 who had the most experience with Thai received lower scores than speaker 4 and 8. Thus, rating scores did not seem to correlate with number of years of experience with the target language. Similar observation was also found in other studies (Flege and Fletcher 1992; Flege, Munro and Fox 1993, (cited in Munro 1993); Munro 1993). The evidence in these studies suggested that "experience with a second language beyond one year does not guarantee improvement in L2 production and perception, particularly when such factors as amount and quality of the L2 are not controlled for" (Munro 1993, p. 61). Moreover, even after several years of experience, a group of Arabic-speaking learners investigated under the laboratory condition in Munro (1993)'s study were found to be unable to produce any vowel from English in a truly native-like way.

Another important finding from this present study was that five different tones in Thai seemed to pose different degrees of difficulty for the native-English speakers. In general, level tone, namely high, mid and low tones received lower scores than contour tones, namely falling and rising. The pitch contrast between the first and second portions of falling tone (from high to low) and rising tones (from low to high) may have provided non-native speakers with an internal point of reference in the process of production and
perception of these two tones. It was further found that among level tones, high tone received the lowest rating score, followed by low tone and mid tone. On the other hand, rising tone received higher score than falling tone.

In order to examine which characteristic of the non-native production investigated in Experiment 1 influenced native listeners to hear them as having a foreign accent, rating data were regressed on all acoustic parameter measured. Similar to Munro (1993)'s study, this analysis was found to be moderately successful. Significant predictors were found only for mid tone and high tone of the [KHA:U] words, and only for mid, falling and high tones for the [NA:] words. The 4 significant predictors found for [kʰa:u] 'fishy smell', the 3 significant predictors found for [kʰáːu] 'a kind of fish' accounted for 100% of the variance of the rating data of both words. Only one predictor each was found for [naː] 'rice field', [nâː] 'face' and [nâː] 'younger siblings of mother, and it accounted for 68%, 68% and 89% of the variance in the rating scores of these three words (Table 7).

Results on Table 7 also showed that significant predictors varied from word to word or from tone to tone. In general, spectral (as opposed to duration) characteristic of the non-native production, namely vowel formants and fundamental frequency were found to have greater influence on native listeners' judgment. This appears to agree with the production results (Experiment 1) whereby significant differences were found in the spectral, but not duration characteristic of the production of the two groups. From Table 7, 9 out of 10 significant predictors found for both [KHA:U] and [NA:] were spectral in nature.

It should be emphasized, however, that the process of trying to relate production and perception data in order to establish which aspects of the non-native production influenced the perceived degree of accentedness by native listeners is rather exploratory. Further investigation along this line is needed to verify the patterns of results found in this study. Furthermore, our knowledge on the differences between native and non-native speakers' production and perception, thus on the foreign-accented phenomenon will be enhanced by an investigation on the perception of both native and non-native production by non-native speakers.
7. References


