Working Papers of the Cornell Phonetics Laboratory

No. 1 (December 1983)

Editor: Mary Beckman
Associate Editor: Stuart Milliken

SRS Pitch Rules for Japanese
Mary E. Beckman, Susan R. Hertz, Osamu Fujimura

The "Morphology" of English Spelling: A Look at the SRS Text-Modification Rules for English
Susan R. Hertz

Perception of Devoiced /si/ and /syu/ in Japanese: The "Segment" Reconsidered
Mary Beckman, Atsuko Shoji

Vowel Devoicing and Tone Recoverability in Cheyenne
Stuart Milliken

Production and Perception of the Voicing Contrast in Indian and American English
Katharine Davis, Mary Beckman
Cornell Phonetics Laboratory
Dept. of Modern Languages and Linguistics
Morrill Hall
Cornell University
Ithaca, New York 14853

Director: Joseph E. Grimes
Administrative Assistant: Theresa Ticknor
Technical Coordinator: Mark J. Sanford
Computer Programmer: Mark Pedrotti
Researchers and Students: Mary Beckman, Prathima Christdas, G. N. Clements, Katharine Davis, Susan R. Hertz, Christiane Laeufer, Elizabeth Leung, Stuart Milliken, Atsuco Shoji
SRS Pitch Rules for Japanese

Mary E. Beckman, Susan R. Hertz, and Osamu Fujimura

Abstract. SRS (Speech Research System) [Hertz 1982] has been used recently to write a set of phoneme-based rules for Japanese. This rule set tests several hypotheses about Japanese phonology. In particular, it tests the usefulness of a hierarchical phrase-structure analysis for generating pitch patterns. The user demarcates the phrases in the romanized input string with boundary symbols, which have associated boundary features. These boundary features, together with user-provided accent marks, are used to generate a segment-by-segment specification of pitch features, such as [low] (low-pitched) and [acc] (accented). Synthesizer parameter rules then refer to the features of the boundaries and of the intervening segments to produce various parts of an utterance's pitch pattern, such as phrase-final lowering or the post-accentual fall. Our SRS pitch rules for Japanese have produced natural-sounding intonational and accentual patterns in a large variety of utterances.

Introduction

Synthesis by rule of prosodic patterns is important not only because it strongly affects the quality of the resulting speech, but also because it offers an experimental means of testing phonological theories. In this paper we will describe our preliminary model for synthesizing pitch.

---

1A shorter version of this paper was presented at the 105th Meeting of the Acoustical Society of America, May 1983.

2Osamu Fujimura, Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974.
patterns in Tokyo Japanese. Japanese is particularly suitable for developing such a model because it allows us to study the effects of phrasal units on pitch patterns unobscured by such complications as the durational effects and large repertoire of pitch contours that can accompany stresses in, for example, English.

First we will describe quite generally our posited phrasal units and their effects. Then we will show how these effects are implemented by the specific rules we developed using SRS, Hertz's language-independent synthesis rule development system [Hertz 1982].

The Accentual Phrase

The smallest phrasal unit that affects the pitch contour is the accentual phrase, which we represent in the traditional manner as patterns of pitch highs and lows determined by the placement of an accent kernel [Miyata 1927, 1928; Hattori 1954]. (See Figure 1.) An accentual phrase

---

Our accentual phrase corresponds roughly to the unit that McCawley [1968] calls a minor phrase (Japanese bunsetsu or gosetsu—see Arisaka 1941, Hattori 1947, 1949). However, our treatment of accent suppression (see discussion under "Accent Suppression") often makes our accentual phrase a somewhat larger unit, encompassing two or more minor phrases.

There is an extensive body of literature describing accentual patterns at the level of the minor phrase, including various attempts to predict the accentual pattern of a phrase from the underlying patterns of the component words. The three main issues in this area are the patterns of compound words, the patterns of inflected forms such as verbs and adjectives, and the accentual characteristics of particles [cf. Akinaga 1958, Kawakami 1966, McCawley 1968]. However, these questions lie outside the scope of this paper, since we take as our input the surface forms resulting from any modifications to accent kernel placement made during the word-formation process.

We chose a hybrid analysis involving the high and low levels in addition to the accent kernel, because it is convenient to implement in our present segment-based rule set. Although the high-low analysis forces an oversimplified approximation to such phenomena as accent subordination, it allows for a surprisingly accurate differentiation of perceptually relevant phrasal types.
Figure 1. The Accentual Phrase

L H L
hasi'-ga 'bridge-(nominative)'

H L L
ha'si-ga 'chopsticks-(nominative)'

L H H
hasi-ga 'edge-(nominative)'

Figure 2. Accent Contrasts
has at most one pitch drop, located at the accent kernel. Everything that follows this pitch drop is low in pitch, and everything that precedes it is high, with one exception—if the phrase-initial syllable is both short and unaccented, it is low.\footnote{The limitation of this phrase-initial low to short syllables is discussed in Hattori [1954] and Fujimura [1966].}

Accent Contrasts

The placement of the accent kernel that determines the highs and lows varies depending on the particular lexical items in the phrase. For example, an accentual phrase consisting of the noun plus nominative particle \textit{hasi-ga} can have one of three contrasting accentual patterns, as illustrated in Figure 2. It can have accent on the second syllable, as in \textit{hasi'-ga}, 'bridge'; it can have accent on the first syllable, as in \textit{ha'si-ga} 'chopsticks'; or, it can have no accent, as in \textit{hasi-ga} 'edge.'

Accent Suppression

Not all lexical accents are realized, however. For example, in most contexts, when a short two-accent sentence, such as \textit{hasi'-ga a'ru}, is imbedded into another sentence as a relative clause, the second pitch fall, in this case the one on \textit{a'ru}, is suppressed (see Figure 3). We deal with accent suppressions of this sort by assuming the lexical item with the suppressed accent to be contained in the same accentual phrase as a preceding item with realized accent. All accent kernels but the first are then ignored.\footnote{There is some evidence that this suppression is not a complete phonological deletion, but a variable phonetic reduction [e.g., Kawakami 1977]. In addition, there seem to be some special peculiarities when the last element in the accentual phrase is a monosyllabic particle [see Kawakami 1966]. Our model, however, is a first approximation for testing only the gross effects.} In utterance 2, for example, we would analyze as one accentual phrase the entire sequence \textit{hasi'-ga-a'ru}, causing the accent on \textit{a'ru} to be ignored.
1. \underline{hasi'-ga} a'ru. ‘There is a bridge.’

2. \underline{hasi'-ga-a'ru} mati'-desu. ‘It's a town where there is a bridge.’

**Figure 3.** Accent Suppression

DECLARATIVE INTONATION:

UPDATE THIS

QUESTION INTONATION:

**Figure 4.** The Sentence
The Sentence

At the other end of the scale from the accentual phrase, we posit—as as the largest phrasal unit in our current scheme—the "sentence."[6] A sentence is the domain of such basic "intonational patterns" as declarative intonation or question intonation (see Figure 4). An intonational pattern has two components. The first is an overall declination; fundamental frequency starts high and gradually falls throughout the span of a sentence. The second is terminal effects, such as the sharp downward curve for a declarative sentence, or the slight upward curve for a question.

The Sentence [with examples]

For simple, short utterances, the two levels "accentual phrase" and "sentence" usually suffice. The highs and lows of the constituent accentual phrases are simply superimposed onto the overall intonational pattern, as illustrated in Figure 5.

The Major Phrase

The pitch contours of complex or longer utterances, however, cannot be generated without assuming an intermediate phrasal unit between the sentence and the accentual phrase.

We currently set up one level of intermediate unit—the "major phrase"—illustrated in Figure 6. The main effect of a major phrase boundary is to reset the basic declination line, starting it again sentence-medially at a fixed value slightly lower than the sentence-onset value. A major phrase, like a sentence, can have various terminal effects, such as a phrase-final rise, fall, or leveling-off. The sentence-final effects discussed above, then, may be considered a special class of such phrase-final adjustments.

6 In future implementations, we may add a unit larger than the sentence—perhaps the "paragraph"—in order to systematically control variations in starting pitch values and other parameters as they relate to relationships among sentences in a discourse.
L H L L L L H L L

hasi'-ga-a'ru mati'-desu.

'It's a town where there's a bridge.'

Figure 5. The Sentence

MAJOR PHRASE BOUNDARY

san-hurans'i'suko-wa; yuumei-na hasi'-ga-a'ru-mati'-desu.

'San Francisco is a town that's famous for its bridge.'

Figure 6. The Major Phrase
Annotated Input Text

The first step in generating the fundamental frequency pattern for an utterance is to annotate a romanized input representation with the appropriate phrasal boundaries and accent marks. In the sample sentence shown in Figure 7, for example, the apostrophes represent the accent kernels, the semicolon represents the end of a major phrase, and the spaces represent the boundaries of accentual phrases. The hyphens mark the boundaries of what might be called "words." These word boundaries affect durational and segmental patterns, but not the F0 contour.

Symbols and Features

The annotated romanized text is then rewritten by a set of "conversion rules" into a string of symbols, which are defined in terms of features to be used by later rules. For example, the semicolon after wa in our sample sentence is rewritten as the symbol [$], which has the associated feature [majp] (for "major phrase boundary"), as illustrated in Figure 8. Similarly, the boundary after na is defined by the feature [accp] (for "accentual phrase boundary"), and the segment [i] of hasi' is defined by the features [nuc] (for "syllable nucleus") and [vh.3] (for "third degree of vowel height")—that is, "high vowel"). Notice how the accent symbol following the [i] adds the prosody feature [acc] (for "accented") to the features inherent in the [i]. Note, too, that at this stage, the [a'] of a'ru is also accented.

Accent-Related Features

The output of the conversion rules is then passed to a set of "feature-modification rules." These rules first add the features of lower-level boundaries to those of higher-level boundaries, so that, for example, an accentual phrase boundary would acquire also the features for a word boundary, and similarly, a major phrase boundary would acquire also the features for an accentual phrase boundary. The rules then use these boundary features to assign accent-related features to the utterance segments. For example, Rule 1 in Figure 9 assigns the feature [rise] (for "pitch rise") to a syllable nucleus that is [-long] and [-acc], when it follows an accentual phrase boundary and an optional intervening consonant—that is, when it is phrase-initial. Rule 2 then propagates the feature [low] to all segments in the same accentual phrase to the right of the first accented segment. More specifically, a segment becomes low when it follows
Figure 7. Annotated Input Text

Figure 8. Symbols and Features
another segment that is either accented or low, so long as no accen
tual phrase boundary intervenes. Rule 3 uses the feature [low] assigned by Rule 2. It deletes the feature [acc] for any [low] syllable, in this case, suppressing the accentual pitch drop in a'ru.7

The Topline

This completed segmental representation is then passed to a set of parameter rules, which generate the appropriate F0 patterns.8 The rules first produce a declining topline for each major phrase, by assigning a fixed percentage decrement to each segment, as shown in Figure 10.

Terminal Effects

The next rules produce the appropriate terminal contours for major phrases. The rule shown in Figure 11, for example, implements a phrase-final lowering in the last segment of non-sentence-final major phrases. It specifies a sharp linear descent that starts from the topline value at the beginning of the segment—that is, at (.0)—and terminates at the end of the segment—that is, at (.99)—with a value 90% of the topline value at that point.

7 As stated in note 5, this implementation of accent suppression may be an oversimplification. A more accurate rule might merely reduce the accentual effect of a kernel following another kernel in the same accen
tual phrase, with the degree of reduction a variable function of the distance between the kernels [see Sagisaka and Sato 1983].

8 The parameter rules described in this paper have been simplified somewhat so that irrelevant detail will not distract from the discussion of our pitch contour implementation. For example, the rule for the phrase-final lowering shown in Figure 11 collapses aspects of several more complicated rules that interact to handle all of the terminal effects. Also, the exact placement of the rise and fall illustrated in Figure 12 varies depending on the segmental composition of the affected syllables, and in some cases there is a slight boost on the accented syllable before the fall.
1. [nuc-long-acc] → [rise]/[accp][{con}]
2. [seg] → [low]/[{acc | low}] ~ [accp]
3. [low] → [-acc]

... n a # h a s i' + g a + a' r u ...

<table>
<thead>
<tr>
<th>accp</th>
<th>nuc</th>
<th>nuc</th>
<th>seg</th>
<th>seg</th>
<th>seg</th>
<th>seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>acc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td></td>
<td>-acc</td>
</tr>
</tbody>
</table>

Figure 9.Accent-Related Features

... w a $ y u H me J ña # h a s i' ...

Figure 10. The Topline
Figure 11. Terminal Effects

Figure 12. Accentual-Phrase Effects
Accentual-Phrase Effects

Next, the highs and lows of accentual phrase patterns are generated. (see Figure 12). First, the F0 of any syllable marked [rise] is lowered by 15% of the value assigned by the topline rules, yielding the initial rise from low to high for short, unaccented phrase-initial syllables. Then, the segments marked [low] are lowered by 20% of their topline values to produce the accentual pitch drop and subsequent sequence of low-pitched syllables.

The Pitch-Scale Transform

Finally, the values generated by the parameter rules are fed into a function that raises the lower end of the frequency scale, as illustrated in Figure 13. This pitch-scale transform prevents the F0 values from descending to an unnaturally low frequency range. It also has the effect of flattening out the rises and falls at later, lower F0 values.

Figure 13. The Pitch-Scale Transform
Figure 14. The Pitch-Scale Transform

Figure 15. Review
The Pitch-Scale Transform [with examples]

Consider, for example, the accentual drop in hasi'-ga. Before the transform applies, this drop is implemented as a 20% lowering from the original topline value, as illustrated in Figure 14. After the transform, however, because it occurs at the lower frequencies relatively late in the utterance, it is reduced to a fall of only 16%. The effect of the pitch-scale transform is especially evident when the utterance has a long major phrase.

Review

In summary, we will retrace the steps in our synthesis of Japanese pitch patterns (see Figure 15). First, the text of the utterance to be synthesized is annotated with accent marks and boundary symbols. Next, the annotated text is passed to a set of conversion rules, which rewrite the text as a string of symbols and associated features. Then, the utterance is passed to a set of feature-modification rules, which add accent-related features to the utterance segments. These features are then sent to a set of parameter rules, which use the features to produce a set of nominal F0 values for the utterance. Finally, the values generated by the rules are fed to a pitch-scale transform function that produces the final F0 values to be played on the synthesizer. The only step in this process that is not accomplished automatically by our rule system is the initial annotation of the input text that specifies prosodic information, absent in standard romanizations.9

9 We have not as yet formally tested how easy it is for users to learn our present annotation system or whether the prosodic units specified are intuitively obvious to native users.
References


----- 1949. "Bunsetsu" to akusento. (Reprinted in Hattori [1960]).

----- 1954. On'inron kara mita kokugo no akusento. (Reprinted in Hattori [1960]).


The "Morphology" of English Spelling:
A Look at the SRS Text-Modification Rules for English\(^1\)

Susan R. Hertz

Abstract. Any scheme that accurately converts English text to phonemes must be based on morphological analysis—i.e., an analysis of words into prefixes, suffixes, and roots. This morphological analysis is accomplished in the SRS synthesis rule set for English [Hertz 1982] by a set of text-modification rules, which divide words into morphological units on which subsequent phoneme-generating rules are based. Because these units are not always true morphemes (for example, fiber + er for fiber on analogy with bribe + er for briber), they are more accurately called "spelling morphs." By analyzing text into spelling morphs, the text-modification rules not only simplify phoneme and stress prediction, but the handling of exceptions as well. In addition, by marking particular characteristics of these morphs (for example, syllable structure), the text-modification rules also simplify the prediction of certain low-level effects, such as aspiration.

Introduction

One of the biggest problems facing the developers of text-to-speech systems for English is the fact that the spelling of an English word is often a poor reflection of its pronunciation. For example, the spelling of the words in such pairs as enable/tenable, hanged/changed, axis/taxis, and naked/baked are quite similar, but the letters common to them

\(^1\)This paper is a slightly revised version of a talk that was presented at the 105th Meeting of the Acoustical Society of America, May 1983.
are pronounced quite differently. The pronunciation of these words, however, is entirely regular if their underlying morphological structure is taken into account:

\[
\begin{align*}
\text{enable} & = \text{prefix (en-)} + \text{root (able)} \\
\text{tenable} & = \text{root (ten)} + \text{suffix (-able)} \\
\text{hanged} & = \text{root (hang)} + \text{suffix (-ed)} \\
\text{changed} & = \text{root (change)} + \text{suffix (-ed)} \\
\text{axis} & = \text{single morpheme} \\
\text{taxis} & = \text{root (taxi)} + \text{suffix (-s)} \\
\text{naked} & = \text{single morpheme} \\
\text{baked} & = \text{root (bake)} + \text{suffix (-ed)}
\end{align*}
\]

The relative success of the SRS text-to-phoneme rules for English over most other rule systems of comparable size [Hertz 1981] is primarily a result of a set of about 200 text-modification rules, which divide words into such constituent forms. This paper will focus on the nature of these forms and on their importance for predicting pronunciation.

Morphology and Phoneme Prediction

Consider, for example, the word divider, which the text-modification rules would rewrite as shown in Figure 1. The plus signs represent morph boundaries, the first hyphen marks the preceding letters as a prefix, and the second hyphen marks the following letters as a suffix. The morphological information inserted into this word allows a set of very general rules to predict its pronunciation. The rule shown in the figure, for example, applies to the i of vide, and generates the pronunciation [ay] for it because it precedes a consonant followed by a morph-final letter e.

Since English spelling is so conservative, reflecting older stages of the language, one might expect that the constituent forms important for predicting a word's pronunciation can be derived from an etymological analysis of the word. However, whereas this is frequently the case, as in the word divider, it is also often not the case, as shown in Figure 2.
Figure 1. Morphology and Phoneme Prediction

<table>
<thead>
<tr>
<th>SPELLING</th>
<th>SRS ANALYSIS</th>
<th>ETYMOLOGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. fiber</td>
<td>+fibe--er+</td>
<td>+fiber+</td>
</tr>
<tr>
<td>2. signal</td>
<td>+signal+</td>
<td>+sign--al+</td>
</tr>
<tr>
<td>3. need</td>
<td>+nee--ed+</td>
<td>+need+</td>
</tr>
<tr>
<td></td>
<td>+need--ed+</td>
<td>+need--ed+</td>
</tr>
</tbody>
</table>

Figure 2. SRS "Folk Etymologies"
SRS "Folk Etymologies"

The first example in Figure 2 illustrates how the er of the
word fiber is stripped, on analogy with words like
divider, even though the resulting form fibe is historically
incorrect. One advantage of this analysis is that the very
general rule shown in Figure 1 that generates the
pronunciation [ay] for the letter i of vide (in
di- + vide + -er) will do so as well for the i's of words
like fiber.

Whereas in this case, the SRS rules insert an
etymologically non-existent morph boundary, in other cases
the rules purposely fail to insert an etymologically correct
boundary, as illustrated in the second example in Figure 2.
Here, the etymologically correct ending -al is not stripped
in the word signal, since doing so would lead to the
incorrect pronunciation [s ay n ø l].

The final example in Figure 2 illustrates how words that
are morphologically related may be analyzed differently.
Both need and needed are handled by the same ed-stripping
rule, which produces etymologically incorrect forms for need,
but correct ones for needed.²

Because the forms generated by the SRS text-modification
rules are not always true morphemes as a linguist would use
the term, they will be referred to as "spelling morphs" in
this paper.

The ly-Stripping Rule

The three main types of spelling morphs--prefixes,
suffixes, and roots--are marked at the text-modification rule
level via a set of context-sensitive rules. The rule shown
in Figure 3, for example, marks the word-final sequence ly as
a suffix when it is not part of a one-syllable morph such as
sly or ply--that is, when it does not follow a morph boundary
and a consonant. Again, this rule, like the er-stripping
rule alluded to above, generates both historically correct
and incorrect roots.

² Of course this particular analysis for need (and
analogous divisions of similar words, such as seed and deed)
would not be appropriate in systems that require
etymologically correct word-class information in order to
predict, say, phrase-level stress and intonational patterns.
ly+ → --ly+ / ~+[con]_ __

HISTORICALLY CORRECT:
+time+-ly+  +state+-ly+  +sur+-ly+

HISTORICALLY INCORRECT:
+ho+-ly+  +simp+-ly+  +cur+-ly+

Figure 3. The ly-Stripping Rule

EXCEPTIONS:

crumbly, humbly, nimbly, assembly ~<du>mbly

EXCEPTION TO EXCEPTIONS:
dumbly

INCORPORATION INTO RULE:
ly+ → --ly+ / ~+[con] | ~<du>mb_ __

Figure 4. ...mbly Exceptions to ly-Stripping
...mbly Exceptions to ly-Stripping

The ly-stripping rule, like most morph-marking rules, has a number of exceptions. Consider, for example, the words ending in mbly shown in Figure 4. Stripping the ly of these words would yield an undesirable morph-final b that would be made silent on analogy with morphs like crumb and limb, resulting in such incorrect pronunciations as [kr a m b l i] for crumbly. It is not enough, however, to simply exclude the sequence mbly from the ly-stripping rule, since in the word dumbly the morph-final silent b is desired.

This complicated set of exceptions, however, like most such exceptions, can be described relatively compactly in SRS rule notation—in this case, as the class of words ending in the letters mbly, so long as these letters are not preceded by the letters du. This pattern can be incorporated directly into the ly-stripping rule, as shown in the figure.

Rule Ordering and Exceptions to ly-Stripping

In addition to these exceptions, there are other words, such as imply and comply, that will be exceptions to the ly-stripping rule only if it is ordered before the prefix-stripping rules. If the prefix stripping rules were ordered first, for example, comply would be divided into the prefix com- and the root ply, as shown in Figure 5. Since this root has only one syllable, it would then automatically be excluded from the ly-stripping rule, as described in the above discussion for Figure 3. Unfortunately, however, the ly-stripping rule, like most suffix-stripping rules, must be ordered before the prefix-stripping rules, since the opposite ordering would produce many more exceptions.

Analyses Favoring ly-Stripping before Prefix-Stripping

Consider, for example, the analyses given in Figure 6, which favor placing the ly-stripping rule first. The first example illustrates how stripping the ly in the word surly leaves the root sur. Since this root has only one syllable, it will automatically be excluded from the later prefix-stripping rules, in this case from the rule that strips the prefix sur—in words like surprise and surpass. If, instead, the prefix-stripping rules were ordered first, this word would incorrectly receive the pronunciation [s r l a y], on analogy with words like comply.
Figure 5. Rule Ordering and Exceptions to ly-Stripping

Figure 6. Analyses Favoring ly-Stripping before Prefix-Stripping
The second example illustrates how stripping the ly of the word carelessly uncovers the suffix -ful, which is in turn stripped by a subsequent rule, leaving the root care. Again, this root, having only one syllable, will automatically be excepted from the prefix-stripping rules—in this case, from the rule that strips the di of words like divide and direct.

The final example shows how stripping the ly of the word intimately leaves the root-final ending -ate, an ending which causes third degree of stress (that is, highest stress) to be placed two syllables before it, in this case on the syllable in. Once marked as stressed, this syllable will automatically be excluded from the rule that strips the in in words like invent and intuitive.

There are many more analyses of this type that favor placing the ly-stripping and many other suffix-stripping rules before the prefix-stripping rules.

Revised ly-Stripping Rule

Thus words like imply and comply, which require the opposite ordering, must be treated as exceptions, even though they are entirely regular in any conventional sense. Either they must be placed in an exception dictionary or they must be excluded explicitly in the suffix-stripping rule itself along with words like crumbly and humbly, as shown in Figure 7.

Suffix-Stripping vs. Prefix-Stripping

Although there are exceptions to the suffix-stripping rules of the sort just illustrated, these exceptions are few in number. Because there are so many words which look and behave as though they contained a true historical suffix, even though they do not, the strategy of not respecting etymological accuracy allows for very general suffix-stripping rules.

Prefixes, on the other hand, are of such a nature that the rules that recognize them cannot benefit as much from failing to respect historical origins; the prefix-stripping rules usually must refer to a relatively large, but nevertheless manageable, number of word-specific letter sequences as exceptions or conditions. Contrast, for example, the ly-suffix-stripping rule shown in Figure 7 with the two dis-prefix-stripping rules given in Figure 8.
\[ ly+ \rightarrow +ly+ / \sim \langle [\text{con}] \rangle | \sim \langle \text{du} \rangle | \text{mb} | \langle o | i | \text{mp} \rangle \ldots \_ \]

**Exceptions:**
- ply
- humbly
- comply
- sly
- crumbly
- imply

**Compare:**
- deeply
- dumbly
- simply
- grisly

*Figure 7. Revised ly-Stripping Rule*

1. \(+\text{dis} \rightarrow +\text{dis}'2-+/_ \sim \langle \text{us}+ | \text{ipli} \rangle | \langle \text{a} -\langle \text{s} \rangle | \text{ric} \rangle \ldots \_ \)
   - Applies to: discuss, disciple, distasteful, ...
   - But not to: discus, discipline, distant, ...

2. \(+\text{dis}'2- \rightarrow +\text{dis}-/_ \sim \langle \text{c} [u | ov] | \langle \text{r} -\langle \text{ed} \rangle | g | d | \ldots \rangle \)
   - Applies to: discuss, disgust, disdain, ...
   - But not to: disconnect, discredit, discolor, ...

*Figure 8. The dis-Stripping Rules*
The dis-Stripping Rules

The first rule in Figure 8 marks the morph-initial sequence dis as a prefix with secondary stress when it precedes anything but one of the sequences in angle brackets. More specifically, the letters dis are stripped when they do not precede the letters cus followed by a morph boundary, the letters cipli, the letters ta when not followed by s, etc. This rule would apply, for example, to words like discuss, disciple, distasteful, and discover, but not to words like discus, discipline, and distant, excluded by the sequences in angle brackets. The figure shows the result of applying the rule to the word discover.

The second rule in the figure then applies to a subset of the words handled by Rule 1—namely, to words like discuss and discover, in which the s belongs to the second syllable. It rewrites the prefix, inserting a dot to represent the syllable boundary, and deleting the secondary stress mark to indicate that the vowel of the prefix is reduced. Again, the figure shows the result of applying this rule to the word discover. Note that after this rule applies to this word, the irregular morph cover is still delimited in the same way that it would be in words like recover, uncover, and cover (where the syllable and morph boundaries coincide), and it can thus be handled in the same way.

Prefix-Stripping and Aspiration

The information about syllable structure provided by these rules is used by the lower level rules that handle phonetic details—for example, by the rules that aspirate voiceless stops when they occur at the beginning of stressed syllables. Thus, after the two dis-stripping rules apply, the aspiration rules will correctly aspirate the syllable-initial voiceless stop of a word like discolor, but not the non-syllable-initial voiceless stop of a word like discover, as illustrated in Figure 9.

Prefix-Stripping and Vowel Reduction

Another kind of information that prefix-stripping rules provide for the later rules that handle phonetic details is information about the quality of reduced vowels. For example, the rules mark the prefix a— with an equals sign, rather than with a dash, in words like ascribe and ascetic (see Figure 10), in order to indicate that this prefix should be pronounced [ə] rather than [æ]. That is, it should not be
Figure 9. Prefix-Stripping and Aspiration

Figure 10. Prefix-Stripping and Vowel Reduction
the higher variant, as the reduced vowel often is before [s] (for example, in the second syllable of the word fantasy).

Note that although the stress and phoneme patterns of the word ascetic are predictable without stripping the prefix (the ending -ic almost always being preceded by highest stress), the prefix-stripping rule is nevertheless necessary in order to insure that the reduced vowel of the first syllable be pronounced correctly.

Conclusion

This paper has described how a small set of SRS text-modification rules analyze English words into spelling morphs. It has shown that such spelling morphs are important both for the prediction of stress and phoneme patterns, and for the prediction of low-level phonetic effects such as aspiration. In general, it is the failure of the text-modification rules to respect historical origins, together with the way that they handle exceptions, that is responsible to a large extent for the accuracy and compactness of the SRS text-to-speech rules for English.

References


Perception of devoiced /si/ and /syu/ in Japanese:

The "segment" reconsidered

Mary Beckman and Atsuko Shoji

Abstract. We examined devoiced /si/ and /syu/ syllables in Japanese to see if their spectral and perceptual characteristics conform to traditional accounts of speech production as a motor translation of discrete, static segments. Measurements of the lowest-frequency spectral prominences in the syllabic [ʃ] of these syllables showed a spectral coloring of the fricative by the deleted vowel segment similar to fricative-vowel coarticulation in other languages. Perception tests showed that Japanese listeners can use this spectral coloring as a cue to the identity of the underlying vowel, although identification was substantially less than perfect, varying with the strength of the spectral coloring. These results suggest that a supposedly lower-level motor interaction between the fricative and the vowel can occur before a higher-level process deletes the vowel, contradicting the order implied by traditional accounts.

Introduction

A central problem in the study of speech production and perception is the difficulty of reconciling linguistic representations of an utterance as a series of discrete, static phonetic segments with the lack of such units in the acoustic signal. The usual solution to this problem is what Fowler et al. [1980] call "translation theories." Translation theories assume that the sequence of phonetic segments is real at some higher, pre-motor level, whereas the overlapping and dynamic realization of the segments in the acoustic signal is a lower-level translation of these abstract cognitive units into the motor mechanisms of the vocal tract.
A consequence of this assumption is that two types of interaction among segments, corresponding to the two different levels, must be distinguished. Following Fujimura and Lovins [1978], we will call these two types "hard coarticulation" and "soft coarticulation." Hard coarticulation is any context-dependent variation in the realization of segments that is so minor as to be generally ignored in phonological descriptions. For example, the last few centiseconds of an [s] before an [i] in English will usually contain spectral peaks near the frequencies dominant in the spectra of an [ʃ] in addition to the high-frequency prominence characteristic of the dental fricative's noise pattern [Solé 1981]. Soft coarticulation, on the other hand, is any more obvious context-dependent variation that would be ignored only in the broadest transcriptions. For example, /s/ before /i/ in Japanese is a completely palatal [ʃ].

These two types of interactions must be differentiated in translation theories because hard coarticulation cannot be stated as all-or-none changes to the features of static, discrete, temporally unspecified phonetic targets, whereas soft coarticulation is most conveniently stated in this way. Translation theories explain this difference by assuming that hard coarticulation is a physically inevitable artifact of the motor translation, whereas soft coarticulation consists of language-specific phonological processes occurring at the earlier pre-motor level.

Because of this necessarily rigid differentiation between the two types of segmental interaction, however, translation theories do not accord with any case of segmental interaction in which there is an apparent mixing of levels. In this paper we will discuss one such case, involving devoiced syllables in Japanese.

Devoiced syllables occur in many Japanese dialects as variants of CV syllables, where C is any voiceless consonant and V is a short /u/ or short /i/, when the syllable precedes another voiceless consonant or occurs word-finally. This devoicing is apparently a process of soft coarticulation, because speakers can systematically manipulate it as a mark of the prestigious standard (Tokyo) dialect. The Japan Broadcasting Corporation, for example, advises its announcers that they should not devoice overmuch, but that "appropriate devoicing improves the coherence of words and phrases, giving a feeling of articulate crispness to one's speech"--the appropriate amount of devoicing being "the extent to which it is done in the contemporary standard language" [Hirayama 1979, p. 108].
The traditional phonological description of devoiced syllables is that they contain voiceless variants of the /i/ or /u/ in the corresponding CV syllables [e.g., Shibata 1955, McCawley 1958]. When the waveform of a devoiced syllable is examined, however, neither its spectral nor its temporal structure indicates the presence of a voiceless vowel. Spectrograms typically show only the frication noise characteristic of the syllable-initial consonant with no following formant-like bands of the sort seen in many voiceless vowels (in, for example, Cheyenne), and duration measurements typically show little difference between the length of a devoiced syllable and that of just the consonant in any corresponding syllable with a voiced vowel in a different token of the same word uttered by the same speaker [Beckman 1932].

An alternative phonological description that better captures this physical reality is that the vowels are not devoiced, but rather are deleted, as stated in the following rule from Ohso [1973]:

\[
\begin{align*}
\forall & \rightarrow \emptyset \\
[+\text{high}] & \rightarrow \emptyset \\
[-\text{voice}] & \rightarrow \{[-\text{voice}]\}
\end{align*}
\]

Note, however, that this description will conform to the perceptual reality of voiceless syllables only if the /i/:/u/ contrast is either perfectly maintained or perfectly neutralized when the vowel is deleted. Thus, for example, native speakers can accurately distinguish /ti/ and /tu/ syllables even when the vowel is deleted because /t/ is a palatal affricate before /i/ and a dental affricate before /u/ whether or not the conditioning vowel is present as any kind of vowel-like structure in the acoustic signal. The phonological description can account for the recoverability of the underlying vowel in devoiced /ti/ and /tu/ by ordering the vowel-deleting rule after the other soft-coarticulation rules that produce the different affricate allophones of /t/, as illustrated in Figure 1. On the other hand, if there is no such high-level modification of a given consonant in the environment of /i/ or /u/, the underlying vowel must not be recoverable in devoiced syllables containing that consonant. A theory that distinguishes high-level pre-motor soft coarticulation from low-level post-cognitive hard coarticulation requires that these two patterns be the only possibilities.

In light of this requirement we considered the palatal fricative [], which occurs with no obvious allophonic variation before both short /i/ and short /u/. The first
Figure 1. Traditional translation models account for the recoverability of the deleted vowel by ordering the vowel-deletion process after other high-level processes that produce vowel-dependent contextual variation in the syllable's consonant.

possible prediction is that the /i/::/u/ contrast will be completely neutralized in voiceless syllables containing [ʃ]. The other possibility is that we have overlooked an obscure allophonic process that allows the vowel to be recovered. The latter is, in fact, what Ohso [1973] claims. She says that Japanese speakers never confuse words like /kasi/ and /kasyu/ when the vowel is deleted because there is a rule "darkening" the fricative before /u/. That is, she says that devoiced /si/ can always be distinguished from devoiced /syu/ because there is a systematic velar or labio-velar coloring in the syllabic [ʃ] when /u/ is intended.
In order to test Ohso's claim, we did two experiments on the recoverability of the vowel in devoiced /si/ and /syu/ syllables. The results of the experiments showed, however, that the vocalic contrast is neither perfectly maintained nor perfectly neutralized. In this paper, we will describe those experiments and show how the results support a model in which segments are inherently overlapping and dynamic, as posited by Fowler et al. [1980] and Bell-Berti and Harris [1981], rather than the traditional translation theories with their inherently discrete, static segments.

**Methods**

**Experiment 1.** In the first experiment, we tested the perception of natural tokens of words containing /si/ and /syu/ syllables in positions where the syllables could be devoiced. The corpus was the eleven minimal pairs shown in Table 1. These twenty-two test words, along with twelve filler words not containing the target syllables, were written in Japanese orthography in the frame sentence kore c to yakusimasita ('I translated this as .' ) three times in three different lists. The order in any of the three lists was random except that no sentences containing the two members of a minimal pair occurred next to each other, and no test words occurred in the first four sentences at the top of a page or in the last four sentences at the bottom of a page. One male and three female native speakers of standard (Tokyo) Japanese read the lists in a sound-proof booth, producing 12 tokens (4 subjects X 3 lists) for each word type.

We then made a stimulus tape by splicing together in a random order all the utterance tokens containing test words. Fourteen native speakers, including the four who had produced the tokens, listened to the tape individually in a sound-proof booth. The subjects were instructed to score each token by circling the appropriate response on a five-point scale ranging from "definitely the word containing /u/" through "can't tell" to "definitely the word containing /i/." The point labels on the answer sheet were written in Japanese, and the instructions were recorded in Japanese directly onto the beginning of the tape. All fourteen subjects responded to all of the stimuli on the tape.

We then converted the responses to a numerical score by counting each "definite /u/" as -1 and each "definite /i/" as +1. Summing the responses for each utterance token, we calculated an "identification score" that could range from
Table I
Words used in Experiment 1.

/sikaN/  /syukaN/
/sikkoo/  /syukkoo/
/siteN/  /syuteN/
/sittoo/  /syuttoo/
/siseki/  /syuseki/
/sissiN/  /syussiN/
/sityoo/  /syutyoo/
/sittyoo/  /syuttyoo/
/sihaN/  /syuhaN/
/sippi/  /syuppi/
/ka'isi/  /ka'syu/

/N/ is the "moraic" nasal and /'/ represents accent.

-14 if all 14 subjects circled "definitely /u/" to +14 if all subjects circled "definitely /i/.

We then made wide-band spectrograms of the test utterance tokens, and noted whether the spectrograms showed any voicing in the target syllable. For each target syllable that showed no voicing, we measured to the nearest 250 Hz the frequency of the bottom edge of the [f]'s noise band as an indicator of the amount of velar or labio-velar coloring (i.e., lower frequency indicates more dorsal retraction or lip rounding). We made the measurements just before the fricative's cessation, where the greatest variation among the tokens occurred.

We then calculated, as a measure of the extent of a relationship between the recoverability of the vowel and the syllabic fricative's spectral characteristics, a multiple regression function for the tokens' identification scores against the fricatives' frequency values and the speakers' identities. We included the speaker's identity as a variable
in this equation because we assumed that the listeners would adjust for any differences among speakers due to their different vocal tract sizes.

Experiment 2. In order to further test the relationship between the recoverability of the vowel and the fricative's coloring while controlling for any other possible cues, we then did a second experiment. In this experiment, we tested the perception of synthesized variants of a word consisting of a voiceless syllabic [ʃ] followed by the syllable [kaN]. There were eight stimuli along a continuum between one that sounded clearly like the word /syukaN/ and one that sounded clearly like the word /sikaN/.

To make the stimuli we first synthesized the word /sikaN/ ([ʃkʰa]) using the SRS rules for Japanese [Hertz and Beckman 1983], and then modified the OVE IIIb fricative-branch parameter values for the initial syllabic [ʃ] using SRS editing routines [Hertz 1982]. We modeled the fricative in the two endpoint stimuli on the male speaker's tokens of /syukaN/ and /sikaN/ that had, respectively, the lowest (most /u/-like) and the highest (most /i/-like) identification scores in the first experiment. The values in these endpoint stimuli are shown in Figure 2. The six intermediate stimuli had K1 and K2 values ranging in equal logarithmic steps between those of the two endpoint stimuli.

We then made a stimulus tape consisting of nine blocks, each containing four stimulus tokens. The first block was a practice block containing only tokens of the endpoint stimuli. The other eight blocks contained in random order four tokens each of the eight test stimuli. A 200 ms pause separated the stimulus tokens within a block, and a 600 ms pause interrupted by a 200 ms orientation tone separated the blocks.

Fourteen native speakers of Tokyo Japanese listened to the tape individually in a sound-proof booth. They were instructed to respond to each stimulus token by making a check mark in the box in the appropriate column on the answer sheet. There were two columns, labeled in Japanese with the words /sikaN/ and /syukaN/. The instructions were recorded in Japanese at the beginning of the test tape. All subjects gave responses for all of the stimulus tokens.
Figure 2. K1 and K2 values for synthesized syllabic fricatives in endpoint stimuli modeled on "best" natural tokens.

Results

The results of the first experiment are shown in Figures 3a and 3b. Figure 3a displays relative frequency polygons for the identification scores of the utterance tokens that had no trace of voicing in their target syllables. The solid line is for tokens of words with /u/ intended and the dashed line for tokens of words with /i/ intended. The mean scores for the two groups are indicated by the solid and dashed arrows below the abscissa, and are significantly different (t=14.213, p<0.0001).

Figure 3b contains relative frequency plots for the measurements made of the fricatives on the spectrograms. The solid and dashed lines and arrows are as in Figure 3a. The
means of the two groups are again significantly different (t=7.652, p<0.0001).

The multiple regression equation calculated for the data in Figure 3a against those in Figure 3b was:

\[ Y = -203 + 37 \ln(X_1) - 4X_2 - 6X_3 + X_4 \]
where $Y$ is the identification score, $X_1$ is the fricative's frequency value, and $X_2$ through $X_{10}$ are the speakers. The coefficient of determination for this function is 44.1%.

![Graph showing percentage of /syukaN/ and /sikaN/ responses for each stimulus in the second experiment.](image)

**Figure 4.** Percentage of /syukaN/ or /sikaN/ responses for each stimulus in second experiment.

Figure 4 shows the results of the second experiment. This figure plots the percentage of responses identifying a token as /syukaN/ or as /sikaN/ against the number of the synthesized stimulus, where 1 is the stimulus modeled on the natural /syukaN/ and 8 is the stimulus modeled on the natural /sikaN/.
Discussion

As Figure 3a shows, the contrast between /i/ and /u/ is not perfectly neutralized in the tokens with devoiced /si/ and /syu/ syllables. The significant difference between the mean scores for the two groups indicates that the subjects could differentiate the words on the basis of the intended vowel at a level substantially better than chance.

Figure 3b shows an acoustic cue that the listeners could have used to identify the intended vowel. The significantly lower mean frequency of the fricative's noise band in the tokens with /u/ intended indicates that the speakers often produced a velar or labio-velar coloring of the sort posited by Ohso.

The relationship between the data in Figure 3a and those in Figure 3b as measured by the regression function strongly suggests that the listeners did use the coloring of the fricative as a cue to the identity of the intended vowel. The coefficient for the fricative's frequency value ($X_1$) in the regression function is positive and large, showing that lower (more /u/-like) identification scores were generally associated with lower frequency values and higher (more /i/-like) identification scores with higher frequency values.

Moreover, the results of the second experiment show that the fricative's coloring is a robust cue. When all other variables were kept invariant, variations in the fricative's frequency characteristics alone shifted the perception of the synthesized syllabic consonant from 100% identification as /syu/ to 100% identification as /si/. We can conclude from these results that the speakers could modify their production of the syllabic fricative in accordance with the underlying vowel and that the listeners could use the acoustic results of that modification as a cue to the identity of the underlying vowel.

However, these results do not support Ohso's claim that the fricative is systematically modified to prevent neutralization of the vocalic contrast. Figure 3b shows a large area of overlap between the values for the two syllable types, and Figure 3a shows a similar overlap in their identification scores. In other words, the speakers did not systematically distinguish the syllables with /u/ intended from the syllables with /i/ intended, and the listeners consequently could not systematically identify the intended vowel. It should be noted also that the biggest effect of an intended /u/ occurred toward the end of the fricative, where the influence of a following high vowel is seen most clearly.
in CV syllables in other languages [Soli 1981]. The better-than-chance but less-than-perfect identification of the intended vowel is likewise reminiscent of the identification of the coarticulated vowel in fricatives excised from CV syllables [Yeni-Komshian and Soli 1981].

Thus, instead of being a high-level systematic allophonic interaction of the sort posited by Ohso, the variation seen in the syllabic palatal fricatives in Japanese seems to be of the sort that translation theorists would call hard coarticulation; the velar or labio-velar coloring in the fricative when /u/ is intended looks like a low-level anticipation of a following vowel segment. However, the coloring in this case cannot be an artifact of the motor translation, because the anticipated segment must already have been deleted at a higher level. These results cannot be reconciled with the model illustrated in Figure 1.

Note, however, that the ordering of hard-coarticulation processes after all soft-coarticulation processes is necessary in translation theories only because the segments at the higher level are different from the segments at the lower level. The segments subject to soft coarticulation are static, discrete, and temporally unspecified, whereas those subjected to hard coarticulation overlap and change through time.

Fowler et al. [1980] and Bell-Berti and Harris [1981] have suggested that this representation of the higher-level segments is not correct. They propose that phonological units are inherently dynamic, that movement toward the articulatory target for a segment is as much a part of that segment's underlying form as is the target itself. In such a model, the difference between soft coarticulation and hard coarticulation need no longer be one of kind, but instead may be a difference merely of degree. Moreover, no specification of the ordering of the two types is necessary to the theory.

Another attractive aspect of this theory is that the timing of the articulatory movements making up a segment need not be the same for all contexts or even for all articulators. Such a model might represent some of the inherent features of a CV syllable such as /si/ or /syu/ in the manner shown in the upper portion of Figure 5. In this illustration, the vowel as a laryngeal gesture toward the goal of adducted vibrating vocal folds begins later than the vowel as a lingual gesture. When the vowel is deleted for a devoiced syllable, then, it may be only the time portion associated with the vowel as a laryngeal gesture that is removed. A portion of the lingual gesture, if it is
articulators: before vowel deletion:

lingual

laryngeal

Figure 5. A model positing underlyingly dynamic segments might explain devoiced syllables as the result of removing only the portion of the syllable associated with the laryngeal gesture for the vowel.

compatible with the gesture toward the following consonant, may be maintained. The intended vowel, inasmuch as it is still present to some extent in the lingual gesture, can sometimes be recovered from the acoustic signal, even though it may have triggered no gross allophonic modification to the consonant. Such a model of speech production is compatible with the results of our experiments on devoiced /si/ and /syu/ syllables, whereas translation models clearly are not.
References


Vowel Devoicing and Tone Recoverability in Cheyenne

Stuart Milliken

Abstract. Cheyenne has both vowel devoicing and tone. The resolution of the conflict between these two phenomena differs depending upon the position of the vowel in the word. The tone of a non-final devoiced vowel is lost completely (or, in some contexts, is inferable because of a particular tonal restriction on devoicing), whereas the tone of a final devoiced vowel is always recoverable. Leman [1981] claims that the mechanism of recoverability is special sandhi tones on the penult, which result from tonal assimilation rules applying prior to devoicing. In this paper I will present phonetic evidence that Leman's sandhi tones do not exist, and therefore that his analysis cannot correctly characterize the nature of tone recoverability in Cheyenne. An alternative autosegmental analysis will be proposed in terms of tone spreading, and an independently motivated accent will be suggested as an explanation for the difference in behavior between final and non-final tones.

Introduction

The phonology of Cheyenne\(^1\) has two prominent characteristics which are in apparent conflict with one another. On the one hand, there is extensive vowel devoicing, as described by Leman and Rhodes [1978], and, on

\(^1\)Cheyenne is an Algonquian language now spoken in Montana and Oklahoma. The group in Montana are called 'Northern Cheyennes' (NCh), and those in Oklahoma 'Southern Cheyennes' (Sch). The language of the two groups is essentially the same.
the other hand, a tone system in which a vowel bears one of two contrastive pitch levels.

The vowel devoicing results from two different devoicing processes. These processes may be described informally as follows:

(1a) **Final devoicing.** The vowel of the final syllable devoices unless followed by a glottal stop in the same word.

(1b) **Prepenultimate devoicing.** A vowel earlier in the word than the penultimate syllable devoices if (i) followed by a voiceless continuant, (ii) word initial or preceded by a consonant other than h, and (iii) if HIGH-toned, the vowel must be immediately preceded by another syllable with a HIGH-toned vowel.

(The vowel of the penultimate does not devoice.) These devoicing processes are exemplified by the following forms (where capital letters signify devoiced vowels):

---

2 Tone is not characteristic of the Algonquian family, but Frantz [1972b] has argued that the distinction between long and short vowels in proto-Algonquian was replaced by a contrast between HIGH and LOW tone when vowel length was lost in Cheyenne.

3 The vowel of the initial syllable of 'eagles' and of 'bead' does not meet condition (i), and so remains voiced. In 'twin girl' and 'twin' the vowel of the initial syllable does not devoice since condition (ii) is not met. Note that the a that devoices in 'twin girl' remains voiced when found in penultimate position in 'twin'. Condition (iii) is not met in the second syllable of 'bead', nor is the condition for devoicing the final vowel o in this word because of the following glottal stop (which I take to be [+sonorant]), following Clements and Halle [1983]. The superscript e is inserted by a relatively late epenthesis rule to be discussed below.) When the above conditions are met, however, devoicing does occur, and is completely regular.
The tone system is one in which a vowel bears one of two contrastive pitch levels ('HIGH' and 'LOW'). The following minimal pairs illustrate the tonal contrast: 4

(3) énaa?e 'he died'
énaá?e 'he doctored'
matana 'milk'
matána 'breast'
éstsema?e 'gopher'
éstsemá?e 'gophers'
tséhetanéveto 'you who are a man'
tséhetanéveto 'I who am a man'

These two phenomena of vowel devoicing and tone are in conflict because the phonetic correlate of tone is pitch, which is primarily a function of fundamental frequency (F0), which in turn is a function of the rate of vocal cord vibration invoicing. When a tone-bearing vowel devoices, therefore, the pitch of that tone can no longer be a property of the vowel.

4Examples not enclosed in brackets or slashes are given in Cheyenne orthography. An acute accent indicates HIGH tone, and LOW tone is left unmarked.
A language with both vowel devoicing and tone has recourse to four ways of resolving the conflict:

(4a) The tone may be realized on the devoiced vowel by some other 'secondary' phonetic cue, such as length or vowel quality.

(4b) The tone of a devoiced vowel, being unrealizable as the pitch of the vowel, may be lost.

(4c) Vowels with a certain tone may be blocked from devoicing, so that the tone of a devoiced vowel, although phonetically unrealized, can be inferred phonologically by a process of elimination.

(4d) The tone may be redistributed on the word, being realized in some way on a voiced vowel adjacent to the devoiced vowel.

Option (4a) is not attested in Cheyenne, as there is no noticeable difference in formant structure or duration between devoiced vowels that can be correlated with different underlying tones. Leman [1981:284] asserts that certain HIGH-toned vowels are also 'more tense', but he gives no examples or evidence, and moreover makes the claim only for voiced vowels.

Prepenultimate devoicing in Cheyenne involves in part option (4b) and in part option (4c). Option (4c) is involved because a HIGH vowel does not devoice except after another HIGH-toned vowel (condition (iii) in (1b)). Thus a devoiced vowel not preceded by a syllable with a HIGH-toned vowel can only be underlyingly LOW itself. This is the case for the prepenultimate devoiced vowels in the following examples:

(5)  [o?kOhomeh0]  'coyotes'

[héstAhkEHά?E]  'twin girl'

[kAsowááh]  'young man'

When a devoiced prepenultimate vowel is preceded by a HIGH tone, however, option (4b) is involved. The devoiced vowel may be underlyingly HIGH or LOW, and the tone is not recoverable on the basis of any phonetic or phonological information present. This is the case for the prepenultimate
devoiced vowels in the following forms:

(6)  [háMEškoNO]  'beetles'
     [náhkOheoʔO]  'bears'
     [heváfAhkeMA]  'butterfly'

As the above discussion suggests, the interaction of tone and devoicing in prepenultimate syllables is reasonably straightforward. The rest of this paper therefore will be concerned instead with the final syllables of words, where there is no restriction comparable to condition (iii) concerning the tone which a vowel undergoing devoicing may have.

The behavior of the tone of the final vowel appears to be quite different from that of tones earlier in the word. Leman [1981] posits a set of assimilation rules by which the tone of the final syllable affects the tone of the penultimate prior to final devoicing. This creates a special sandhi tone on the penult, which he claims renders the underlying tone pattern of the last two syllables phonologically recoverable, despite the lack of any phonetic realization of the final tone:

Cheyenne pitch sandhi is closely related to Cheyenne devoicing phenomena. From a functional viewpoint, the pitch sandhi indicates what the phonemic pitches of devoiced [final] vowels are [Leman 1981:293].

This analysis can be considered a variant of option (4b), with the twist that the unrealized tone of the devoiced vowel is nevertheless completely recoverable by virtue of its effect on an adjacent tone at an earlier stage of the derivation.

In this paper I intend to demonstrate that Leman's pitch sandhi analysis is phonetically inaccurate, and that his characterization of the interaction of tone and vowel devoicing in final syllables is thus inadequate both in terms of the extent of recoverability and in terms of the mechanism responsible for the recoverability. The following discussion will be divided into two major sections. In the first section, I will review Leman's claims and provide phonetic evidence against those claims. In the second section, I will present a phonological analysis of the devoicing and tonal
processes based on the phonetic facts. I will argue that final devoicing involves a variant of option (4d), and will propose an explanation for the different behavior of the tones of devoiced final and prepenultimate vowels.

To begin with, however, it is necessary to present a brief summary of Cheyenne segmental phonology, including the many processes that insert or delete segments which make it rather difficult for the uninitiated to interpret examples without confusion.

Overview of Cheyenne Segmental Phonology

Cheyenne has ten consonant phonemes (p, t, k, m, n, s, š, v, h, ?) and three vowels (e, a, o). These will be classified according to the feature system of Clements and Halle [1983], in which the vowels, nasals, h, and ? are [+son(orant)], h is distinguished as [+spread (glottis)], and ? as [+const(ricted glottis)]. Among the vowels, e is distinguished as [-back], and o as [+round].

Assimilation processes found throughout the examples include the following:

(7a) š becomes [x] when next to a back vowel (i.e., mirror Image), with an intervening glottal stop permissible.

(7b) v becomes [w] next to a rounded vowel.

(7c) t assimilates to ts before a front vowel.

(7d) h, when preceded by e, becomes [s] or [š] if followed by t or k respectively.

The consonants and vowels of Cheyenne are further subject to the rules of deletion, insertion, and devoicing given below. In the examples used to illustrate these, plural forms of nouns will often be given. Due to certain historical changes there have arisen two classes of plural suffixes, called 'e-class' and 'o-class'. There is a further division between plural suffixes for animate-gender and

------------------------

5For a more complete discussion of Cheyenne segmental processes, see Davis [1962], Frantz [1972a], and Leman [1979] and [1980].
inanimate-gender nouns. The o-class suffixes are /-o/ (animate) and /-ot/ (inanimate). The e-class suffixes are /-é/ (animate) and /-ét/ (inanimate). The tones given in these examples will be motivated later below.

Deletion of final nasal or h. A sonorant consonant, excluding glottal stop, deletes word finally. This process thus affects nasals and h. The underlying presence of these segments can be seen in the first and third examples below, and their deletion in the second and fourth.

(8)  
\[
\begin{array}{c}
\begin{array}{c}
C \\
\text{[+son]}
\end{array} \\
\text{[-const]}
\end{array} \rightarrow \emptyset / _- # #
\]

Thus: /hánñomah + o/ \rightarrow [hánñomah0] 'bees'
But: /hánñomah/ \rightarrow [hánñomA] 'bee'

Thus: /šéʔšenon + o/ \rightarrow [šéʔšenono] 'rattles'
But: /šéʔšenon/ \rightarrow [šéʔšenO] 'rattle'

Final e deletion. A word-final e deletes when preceded by a sonorant consonant other than glottal stop (i.e., nasals and h).

(9)  
\[
\begin{array}{c}
\begin{array}{c}
V \\
\text{[-back]}
\end{array} \\
\begin{array}{c}
C \\
\text{[+son]}
\end{array} \\
\text{[-const]}
\end{array} \rightarrow \emptyset / _- # #
\]

Thus: /hetane/ + [oʔ0] \rightarrow [hetaneoʔ0] 'men'
But: /hetane/ \rightarrow [hetaeN] 'man'

Thus: /mænehe/ + [oʔ0] \rightarrow [mæneheoʔ0] 'pelicans'
But: /mænehe/ \rightarrow [mæneh] 'pelican'

Note that rule (8) feeds rule (9):
Thus: /maʔšemen + o/  →  [maʔxemeNO]  'apples'

But: /maʔšemen/  →  maʔšeme  →  [maʔxeM]  'apple'

Vowel Copying. In a word-final sequence of two vowels, the second vowel is copied to the right with an intervening glottal stop. An obstruent may follow the vowel sequence:

\[
\begin{align*}
C & \quad \text{C} \\
V_i \ V_j \text{([-son])} \ # \ # & \rightarrow \ V_i \ V_j \ # \ V_j \text{([-son])} \ # \ # \\
\text{Where } V_i & \text{ may equal } V_j
\end{align*}
\]

Thus: /méhtanke + o/  →  [méstAhekoʔO]  'owls'
And: /vého + o/  →  [véhooʔO]  'chiefs'
And: /séot/  →  [seoʔ0tsə]  'ghost'
But: /séot + o/  →  [séot0]  'ghosts'

Note that in 'ghost' the following rule of _e_ epenthesis also applies.

_e_ epenthesis. An epenthetic voiceless _e_ is added after a final obstruent or glottal stop. Note that this process together with rule (9) has the effect of ensuring that final nasals and _h_ are never followed by _e_, and all other final consonants are always followed by _e_. Leman and Rhodes [1978] have argued convincingly that this _e_ is not present in underlying representation, and that the rule assigning it is fairly late so that it does not affect the tonal and devoicing processes. It does, however, trigger the rule assibilating a preceding _t_ to _ts_ (7c), as can be seen in the second example below. The epenthetic _e_ will be represented by a superscript to avoid confusion with an _e_ present underlyingly.

\[
\begin{align*}
\emptyset & \rightarrow e / \ C \\
\text{[-nasal]} & \text{[-spread]} \ # \ #
\end{align*}
\]
Thus: /hehtétat + o/ → [hêstsêtatO] 'kidneys'
But: /hehtétat/ → [hêstsetAtsE] 'kidney'
Thus: /éhtema? + é/ → [éstsemâ?E] 'gophers'

Final devoicing. A sonorant devoices word finally or before a word-final obstruent.\(^6\) It is this rule that accounts for the devoicing of the vowel of the final syllable. (Examples follow (14).)

(13) [+son] → [-voice] / ____ ([+son]) # #

Consonant devoicing. A consonant preceding a voiceless segment devoices. This rule is fed by final devoicing (13), as well as by prepenultimate devoicing (1b).

(14) C → [-voice] / ____ [-voice]

Final devoicing and consonant devoicing apply after the deletion rules (8) and (9) and vowel copying (11), but before e epenthesis (12).

(13) (14)
Thus: /ma?šemen + o/ → ma?šemenO → [ma?xemeNO] 'apples'
(11) (13) (12)
And: /sëot/ → sëot → së0t → [së0tsE] 'ghost'
(9) (14)
And: /hováhn + é/ → hováhn → [howáhN] 'animals'

\(^6\)In very rapid connected speech this sonorant sometimes remains voiced, but I find this quite uncommon, and not easily elicited without considerable arm-twisting.
Summary of Rules

Rules (8) through (14) are listed here for convenient reference. They are given in order of application.

(8) Deletion of Final Nasal or h:
\[ C \quad \begin{cases} (+\text{son}) \\ (-\text{const}) \end{cases} \rightarrow / \quad \# \quad \# \]

(9) Final e Deletion:
\[ V \quad \begin{cases} (-\text{back}) \end{cases} \rightarrow \emptyset \quad C \quad \begin{cases} (+\text{son}) \\ (-\text{const}) \end{cases} \rightarrow / \quad \# \quad \# \]

(11) Vowel Copying:
\[ C \quad V_i \quad V_j \quad \begin{cases} (-\text{son}) \end{cases} \rightarrow V_i \quad V_j \quad V_j \quad \begin{cases} (-\text{son}) \end{cases} \rightarrow / \quad \# \quad \# \]

Where \( V_i \) may equal \( V_j \)

(13) Final Devoicing:
\[ (+\text{son}) \rightarrow (-\text{voice}) \quad / \quad \begin{cases} \# \quad \# \end{cases} \]

(14) Consonant Devoicing:
\[ C \rightarrow (-\text{voice}) \quad / \quad \begin{cases} \# \quad \# \end{cases} \]

(12) e Epenthesis:
\[ \emptyset \rightarrow e \quad \begin{cases} (-\text{nasal}) \\ (-\text{spread}) \end{cases} \rightarrow / \quad \# \quad \# \]

1. Leman's Pitch Sandhi Analysis

Leman [1981] asserts that 'five phonetically distinguishable levels of pitch' can occur on a penultimate vowel in a word. These pitches are derived from the two underlying tones HIGH and LOW. The names Leman gives these variant tones are listed below from highest to lowest, along with the symbols and the diacritic vowel-marking he uses (shown here on [a]):
(15) RAISED HIGH  \hat{H}  [â]
HIGH  \ H  [á]
MID  \ M  [ă]
LOWERED HIGH  \breve{H}  [ă]
LOW  \ L  [a]

Leman describes RAISED HIGH as being 'slightly above high' with MID and LOWERED HIGH falling between HIGH and LOW. He analyzes these three tones as sandhi tones conditioned by the adjacent tones in the following ways:

LOW to MID raising [Leman 1981:294]. Penultimate LOW becomes MID if followed but not preceded by HIGH:

(16) \[
L \rightarrow M / \{L\} \ _ \ H \ # \ # \\
/kosán/ \rightarrow [kōsA] \ 'sheep (sg)'
/ne?e?é/ \rightarrow [ne?e?E] \ 'bullsnake'
\]

HIGH raising [Leman 1981:296]. Penultimate HIGH becomes RAISED HIGH if followed but not preceded by HIGH:

(17) \[
H \rightarrow \hat{H} / \{L\} \ _ \ H \ # \ # \\
/šé?šé/ \rightarrow [še?šE] \ 'duck'
/hotónké/ \rightarrow [hotónkE] \ 'star'
\]

The underlying tone patterns are from Leman [1979], and will be discussed later below.
HIGH lowering [Leman 1981:297]. Penultimate HIGH becomes LOWERED HIGH if followed by LOW and not preceded by HIGH:

\[(18) \quad H \rightarrow \hat{H} /\{L\} _{-} L \# \# \]

/phon?kon/ \(\rightarrow\) [honõ?kO] 'quilt'
/pẽ?e/ \(\rightarrow\) [pẽ?E] 'nighthawk'

As can be seen in these examples, the final vowel devoices following the tone sandhi, and so the final tone is not realized.

My reasons for questioning this analysis are, first, that I have never heard all five pitch levels, and second, that the phonetic description of them is implausible. Leman and Rhodes [1978:19] observe the following:

The musical distance between a high tone and a following low tone is approximately that of a major third. But the musical distance between a low tone and a following high is approximately that of a major second, but sometimes is as small as a minor second.

Given the above rules, however, RAISED HIGH, MID, and LOWERED HIGH can only occur on the penultimate vowel if they are preceded by a LOW antepenult, or are word initial. If, therefore, the antepenult is LOW, four different surface pitch sequences are possible over the antepenult and penult: LOW-RAISED HIGH, LOW-MID, and LOW-LOWERED HIGH, as well as LOW-LOW. It would be quite remarkable if consistent distinctions were made among four pitch levels within the range of a major second plus however much 'slightly above' HIGH is.

Testing the Analysis

In order to test Leman's analysis, I measured the difference in F0 above a preceding LOW tone of each of the five putative pitch levels, and compared the differences. My goal was to determine whether the five pitch levels are phonetically distinct from one another, and if not, how they
should be grouped into a smaller number of levels that are distinct.

Method. The data used in this study were taken from recordings made of three speakers, two in Oklahoma (L.L. and J.F.) and one in Montana (J.G.). No phonological differences were noticed between the speakers from the two areas. The recordings were made in the course of several different interviews with each of the subjects, and consisted of words elicited in random order with an oral English cue, to which the subject responded with the corresponding Cheyenne word in the frame:

(19) nawóomo (NOUN) 'I see the (NOUN)'

Forms were elicited in this frame rather than in running text because a preliminary study showed that in rapid speech there is a decrease in the already narrow pitch separation between tones. Furthermore, words in contrastive position within a sentence generally exhibit greater pitch separation. The frame 'I see the (NOUN)' tended to elicit a contrastive reading, and so was conducive to maximal pitch separation, at the same time providing a natural context for the word being examined.

In the analysis, the F0 of a tone was always measured in reference to that of another vowel in the same word. In all cases the reference vowel was a LOW-toned vowel in the immediately preceding syllable. The vowel of the target tone was always the penult for RAISED HIGH, MID, LOWERED HIGH, and LOW, and was the antepenult for HIGH (which does not occur penultimately following LOW, as described by Leman). Only forms that are segmentally identical or nearly identical in the environment of the tones being examined were compared.

Each test utterance was digitized at 8 kHz, and then analyzed using software developed in the Cornell Phonetics Laboratory for a PDP-11/40. The reference vowel was isolated on the wave form using a wave-form editing program, and was then fed to a companion F0-extraction program. F0 readings were taken once every 10 ms. Sharp transitional pitch changes of 30 ms duration or less in the reference vowel were disregarded (average vowel length was about 100 ms total), and the resulting readings for the reference vowel's pitch contour were then averaged to give a mean F0 value for the vowel.
Next, the target tone was segmented using the wave-form editor and analyzed by the F0-extraction program. This time, however, measurements were converted to semitone difference above or below the mean F0 value of the reference tone. The result was a 'semitone difference curve' for the target tone contour, with the semitone difference from the mean value of the preceding LOW as a function of time from the onset of the target vowel. The time scale was then normalized to the average duration of the target vowel for a given type, and the mean and standard deviation were calculated for the curves of each type. The standard deviation curves were then graphed for a visual comparison with the corresponding curves of other types.

Results. A representative sample of these curve comparisons is given below in Figures 1 to 6. In each of these figures, the horizontal axis represents 100% normalized time, and the vertical axis indicates semitones above or below the mean F0 value of an immediately preceding LOW tone. The dashed and solid lines represent one standard deviation above and below the mean semitone difference curve for the given type. The reference and target vowels are underlined in the phonetic transcriptions below the graphs. The tones marked are those claimed by Leman.

![Figure 1. Comparison of HIGH and RAISED HIGH (speaker L.L.)](image)

- **HIGH** [hetðhkeoʔ] 'stars' (6 tokens)
- **RAISED HIGH** [hetðhkeʔ] 'star' (9 tokens)
Figure 2. Comparison of HIGH and RAISED HIGH (speaker J.F.)

--- HIGH [hesamoNOts] 'his boats' (7 tokens)

--- RAISED HIGH [hesamo] 'his boat' (7 tokens)

Figure 3. Comparison of HIGH and MID (speaker J.F.)

--- HIGH [heskowo] 'porcupines' (7 tokens)

--- MID [hexowo] 'clam' (5 tokens)
Figure 4. Comparison of RAISED HIGH and MID (speaker J.G.)

--- RAISED HIGH [moʔeʔhāN] 'magpies' (6 tokens)
--- MID [neʔhāN] 'lake' (7 tokens)

Figure 5. Comparison of LOW and LOWERED HIGH (speaker L.L.)

--- LOW [maŋeŋ] 'pelican' (10 tokens)
--- LOWERED HIGH [hoʔeranŋ] 'judge' (9 tokens)
Figure 6. Comparison of RAISED HIGH and LOW (speaker J.F.)

--- RAISED HIGH [maʔəxə] 'eye' (8 tokens)

--- LOW [moʔeškə] 'finger' (9 tokens)
Discussion. The nearly complete overlap of area within the curves in Figures 1 and 2 indicates that HIGH and RAISED HIGH are not distinguishable. Figure 3 shows HIGH and MID likewise to be indistinguishable. Figure 4 shows MID to be as far above a preceding LOW tone as RAISED HIGH is. Thus MID, HIGH, and RAISED HIGH do not represent distinct ranges of F0 difference from a preceding LOW tone, and so cannot be different tone variants. In Figure 5 LOW and LOWERED HIGH are shown to be in an identical relationship to a preceding LOW. They also, then, should be considered to be the same tone. Finally, Figure 6 compares the distance of LOW and RAISED HIGH from a preceding LOW. It is evident from this figure that there is no room for three additional distinct pitch levels HIGH, MID, and LOWERED HIGH between these two.

These results indicate that RAISED HIGH, HIGH, and MID are not phonetically distinct, and that LOW and LOWERED HIGH are not phonetically distinct. We may therefore collapse the five spurious pitch levels as follows:

(20) RAISED HIGH
     \         \ HIGH
     /          / MID
    HIGH       HIGH
     \              \ LOWERED HIGH
         \          \     LOW
    MID

Since Leman's pitch sandhi account is thus not in accord with the phonetic facts as determined above, we are still left with the question of whether the tone of the devoiced vowel of the final syllable is phonologically recoverable, and if so, by what means.

2. An Alternative Analysis

In this section I will argue for an alternative analysis of tonal phenomena involving the final vowel. This analysis is most conveniently stated using the principles and
formalisms of autosegmental phonology.8

Tonal Patterns in LOW-LOW and HIGH-LOW Sequences

The clearest evidence for the underlying and corresponding surface tone patterns on the final and penultimate syllables is provided by nouns undergoing various types of suffixation. Consider first the following o-class words, in which all of the tones are LOW, both in the singular and when the plural suffix is added:

(21) [maʔxeM] 'apple' [maʔxemeNO] 'apples' (NCh)9
     [maʔxemeNots6] 'apples' (SCh)
     [heʔkO] 'bone' [heʔkoNots6] 'bones'
     [hoxeoʔO] 'sock' [hexeoNO] 'socks'

As Leman [1981] demonstrates, underlying tonal representations of noun stems can be determined from equative forms. These are denominal verbs formed by affixing the third person prefix /é-/ and the equative formative /-éve-/. When the equative formative is added, all of the noun stem falls to the left of the penultimate syllable, where tonal processes such as those discussed above are not claimed to occur. In the 'equative' forms corresponding to the nouns in (21), the stem remains LOW (stems are underlined here for easier reference):

(22) [émaʔxemenefE] 'It is an apple'
     [éheʔkonefE] 'It is a bone'
     [éhoxeonefE] 'It is a sock'

Thus the stems and o-class plural suffixes in (21) are

8 For an overview of autosegmental theory, see Goldsmith [1976], Van der Hulst and Smith [1982], and McCarthy [1982].

9 The word 'apple' is animate for speakers in Montana and inanimate for those in Oklahoma.
evidently LOW in tone on all syllables.

Contrast these o-class forms with the e-class words in (23) and (24). In the e-class words the final syllable of the stem has HIGH tone when in penultimate position in the plural form:

(23) [éstseməʔe] 'gopher' [éstsemáʔE] 'gophers'
[véʔhoʔe] 'white man' [véʔhóʔE] 'white men'
[onéhawoʔkE] 'bead' [onéhawóʔkEstsE] 'beads'

In the corresponding equatives, however, the stem final vowel is LOW:

(24) [ééstseməʔefE] 'It is a gopher'
[évéhoʔefE] 'He is a white man'
[écnéhawoʔkefE] 'It is a bead'

The evidence of the equatives indicates that the underlying form of the stem of 'gopher', for example, is /H L L/, that the animate e-class plural suffix is in some way contributing a HIGH tone in the surface plural form.

Compare also the tones on the equative formative (underlined) in the following o-class animate and e-class inanimate plural equative forms below. (The inanimate equatives always take the e-class inanimate plural suffix.)

(25) [émaʔxemenéveoʔO] 'They are apples' (NCh)
[éneʔhanévéNEstse] 'They are lakes'

In the first example in (25), the equative formative is [-vé-] when followed by the o-class plural suffix. This plural suffix was shown to be LOW in (21) above. When followed by the e-class plural suffix in the second form in (25), however, the equative formative is [-vé-]. This tone change is the same as that seen in (23).

Since the LOW-toned o-class suffixes do not induce this
alternation, we can attribute the tone change on the equative formative in inanimate plural equatives, and on the stem in the plurals in (23), to the presence of a HIGH tone on the e-class plural suffixes in underlying representation. Thus both the animate and inanimate o-class plural suffixes show evidence of being underlyingly LOW, and both animate and inanimate e-class suffixes give evidence of being underlyingly HIGH.

First Formulation of Tone Rules

The tonal alternations observed above can be summarized as follows: the penultimate vowel appears to acquire the HIGH tone of the following final syllable. This phonological patterning can easily be represented as tone spreading. The HIGH tone of the final syllable spreads to the penultimate syllable, as is expressed by the extension of an association line from the final HIGH tone to the vowel to the left in the representation below:

(26) Tone Spreading
(Tentative): L H
V C o V C o ## → L H
V C o V C o ##

or equivalently: L H
V C o V C o ##

The result is a LOW-HIGH contour on the single penultimate vowel. Such contours do not, however, appear on the surface, thus suggesting the following contour simplification rule:

(27) Contour Simplification
(Tentative): L H
V / \ → L H
V

or equivalently: L H
V / \ → L H
V

This rule states that the left member of a pair of tones associated to the same vowel disassociates.

A convention may also be proposed with regard to the association between a devoiced vowel and its tone. Since
tone is not realized by phonetic correlates other than F0 in Cheyenne, there is no useful sense in which a vowel can be understood as bearing a tone. It is probably best to maintain, therefore, that in Cheyenne only voiced vowels are specified as possible tone-bearing units. As an automatic consequence of this specification, the tone of the devoiced vowel will disassociate by the following convention:

(28) Disassociation Convention: \[ T \dagger V \] [-voice]

'Gopers' [éstsemá?E] and 'apples' (NCh) [maʔxemeNO] thus derive as in (29). At this point no claim is intended as to the order of devoicing (13) and disassociation (28) relative to the tone rules (26) and (27). Tone spreading as tentatively formulated in (26), with the HIGH associated in the structural description of the rule would, however, have to apply before devoicing and disassociation.

(29)

Underlying: \[ H L L H \quad L L L L \]
\[ \quad | | | | \quad | | | | \]
\[ ehtemaʔ + e \quad maʔšemen + o \]

Rule (26): \[ H L L H \]
\[ | | | | \]
\[ ehtemaʔe \]

Rule (27): \[ H L L H \]
\[ | | | | \]
\[ ehtemaʔe \]

Rules (13) and (14):

Rules (13):

\[ H L H \]
\[ | | | | \]
\[ ehtemaʔE \]
\[ L L L L \]

and (14):

\[ maʔšemeNO \]

Conv. (28):

\[ H L H \]
\[ | | /\dagger \]
\[ ehtemaʔE \]
\[ L L L L \]

\[ maʔšemeNO \]

Other:

[éstsemá?E] [maʔxemeNO]
Rules (26) and (27) thus account for the behavior of forms with underlying LOW-LOW and LOW-HIGH sequences in the penultimate and final syllables.

Generalization of Rules to Other Tone Sequences

The following singular/plural examples illustrate forms with another underlying tone pattern. Here the penultimate stem-vowel in the singular is HIGH, but the same stem-vowel is LOW when the plural o-class suffix is added.

(30) [hexówO] 'clam' [hexowoNO] 'clams'
     [matšškóM] 'raccoon' [matšškomehO] 'raccoons'

The equatives show that the penultimate stem vowel is underlyingly LOW, and that the stem final vowel is HIGH:

(31) [éhexcwóneveɔ?O] 'They are clams'
     [ématšškoméhevɔ?O] 'They are raccoons'
     [éma?pà?ónévëNEsts^e] 'They are backs'

The singulars in (30) thus can be seen as behaving exactly like the plural forms in (23). The plurals in (30), however, take the LOW toned o-class suffixes, and so give evidence of the LOW tone of the devoiced suffix displacing the stem final HIGH. This also follows from our analysis if rules (26) and (27) respectively are generalized as follows:

(32) Tone Spreading (revised): T  T
     |     |
     V CO V CO ##

(33) Contour Simplification
     (final form): T  T
     x /
     V

where 'T' represents any tone. Thus the singular and plural
of 'clam' derive as in (34):

(34)
Underlying: \[
\begin{array}{cccc}
L & L & H & L \\
\text{hešovon} & \text{hešovon + o} \\
\end{array}
\]

Rule (32): \[
\begin{array}{cccc}
L & L & H & L \\
\text{hešovon} & \text{hešovono} \\
\end{array}
\]

Rule (33): \[
\begin{array}{cccc}
L & L & H & L \\
\text{hešovon} & \text{hešovono} \\
\end{array}
\]

Other: \[
\begin{array}{cccc}
\text{[hexóW0]} & \text{[hexowoNO]} \\
'clam' & 'clams' \\
\end{array}
\]

The behavior of underlying penultimate-final tone patterns LOW-HIGH (seen in the plurals of (23)) and HIGH-LOW (seen in the plurals of (30)) are thus accounted for by the analysis presented so far. The pattern LOW-LOW shown in (21) was not problematic, and the forms may be understood as undergoing rules (32) and (33) with the final LOW displacing the penultimate LOW with no apparent change resulting in the surface LOW-LOW sequence. The plurals in the following sets of forms illustrate the last remaining pattern, HIGH-HIGH:

(35) \[
\begin{array}{cccc}
\text{[eʔeʔtA]} & \text{'thrush'} & \text{[eʔeʔtáhN]} & \text{'thrushes'} \\
\text{[maʔéNO]} & \text{'turtle'} & \text{[maʔenón]} & \text{'turtles'} \\
\text{[maʔtáNO]} & \text{'bow string'} & \text{[maʔtanóNEsts]} & \text{'bow strings'} \\
\text{[neʔhán]} & \text{'lake'} & \text{[neʔhanéNEsts]} & \text{'lakes'} \\
\end{array}
\]

Here there is an apparent rightward shift of the stem's HIGH tone when the e-class plural suffix is added. The shift must actually be leftward, however, and occur in the singular forms, since the equative forms indicate that the second stem vowel is underlyingly LOW, and the third stem vowel is
underlyingly HIGH:

(36) [ée?e?tánneveo?0] 'They are thrushes'
[éma?enóévéNEstse] 'They are turtles'
[éma?tanónévéNEstse] 'They are bow strings'
[éne?hanvévéNEstse] 'They are lakes' 10

The singular forms of 'thrush', 'turtle' and 'bow string' in (35) can be accounted for in exactly the same way the plurals in (23) were. The plurals 'bow strings' and 'lakes' in (35) also are accounted for by rules (32) and (33), with the final HIGH displacing the penultimate HIGH. Here the effect of the spreading is vacuous, just as it was when a final LOW displaced a penultimate LOW in (21). Words with penultimate-final LOW-HIGH and HIGH-HIGH sequences derive as follows:

(37)

Underlying: L L H L L H H

Rule (32): L L H L L H H

Rule (33): L L H L L H H

Other: [e?éta] [ma?tanóNEstse]

'thrush' 'bow strings'

Thus a process of tone spreading along with contour

10 The stem-final é of 'lake' coalesces with the initial é of the equative formative by a process that will be discussed briefly later below.
simplification accounts for the surface realization of all possible underlying sequences over the penultimate and final vowels.

The Cause of Tone Spreading

While the posited tone spreading rule can account for all of the patterns observed above, it would be even better motivated if we knew why the tone spreading occurs. A reasonable first guess is to suppose that the devoicing of the final vowel itself causes the leftward spreading of its tone by creating a floating tone that must (by a grammar-particular constraint) reassociate. This supposition cannot be easily maintained, however, because the tone of a devoiced prepennultimate vowel does not spread. Furthermore, in forms where vowel copying (11) has applied, the vowel that is copied (the last vowel) does not itself devoice, yet its tone nevertheless spreads, as illustrated in the following forms. (The corresponding equatives are given to demonstrate underlying tones.)

(38)  \[maʰheón/ \rightarrow [maʰhéO] \] 'god'

\[oomahóon/ \rightarrow [ooMhooO] \] 'lumber'

Compare with:

[émaʰheónefE] 'He is God'

[éooMAhóonefE] 'It is lumber'

There is, however, evidence for a different sort of phonological property accounting for tone spreading. That property is accent.

\[11\] It is likely that the 'copy' of this vowel should rather be considered an epenthetic e added by rule (13) after insertion of the glottal stop. It then assimilates to the quality of the preceding vowel by the application of an independently required rule (not discussed above). The final 'copy' would thus have no more effect on the preceding tones than the usual case of epenthetic e. Space does not permit a fuller discussion of the nature of vowel copying here.
Evidence for Accent

Three types of phonological and comparative evidence suggest that the penultimate vowel is accented.

First, the penultimate vowel never devoices. This fact would not be unexpected if an accent were present on this syllable, since accents by nature mark a syllable as prosodically significant. The presence of this accent would also simplify the form of a rule for prepenultimate vowel devoicing (not formalized here), since it would obviate the need for a lengthy string of segmental variables to ensure that the rule applies prepenultimately. Instead, the rule could simply specify an unaccented vowel in its structural description.

Second, there is a process of vowel coalescence in Cheyenne which can be seen operating in the equative form below.

(39) Thus: /é # netё + éve + o/ → [énetséveoʔo]

'They are eagles'

Compare: /netё + o/ → [netseoʔo]

'eagles'

The final e of the stem coalesces with the initial e of the equative formative /-évе-/. The plural nominal form demonstrates the underlying presence of the stem final e. The rule involved deletes an e following another vowel across a formative boundary (+).

This process does not apply, however, when the preceding vowel is penultimate, but instead an epenthetic n appears following the penultimate vowel, blocking coalescence:

(40) Thus: /é # voʔ + éve + ét/ → [éwоʔéveNEstsε]

'They are clouds'

This resistance to coalescence of the penultimate vowel is also a likely effect of an accent. Without the accent, it would be an unexplained accident that vowel devoicing and
vowel coalescence both cannot apply to the vowel of the penultimate.


Thus, although there has been as yet no controlled investigation yielding phonetic evidence for penultimate stress in Cheyenne, the phonological restrictions against vowel devoicing and coalescence in the penult suggest an abstract accent assigned to the penultimate syllable, and the comparative evidence is in accord with this hypothesis.

**Tentative Reformulation of Tone Spreading**

This hypothesized accent motivates the differential treatment of tones in final versus non-final syllables discussed above, because the tone spreading from the final syllable can be reformulated as a natural process of attraction to accent. Segmental variables are thus unnecessary, and the rule is considerably simplified:

$$\text{(41) Attraction to Accent: } \text{T} \quad \text{T} \quad \text{V} \quad \text{*}$$

where '*' represents the accent, and the tone to the right may or may not be associated with another vowel.

This formulation has the advantage of rendering the ordering relationship between the tone rules on the one hand, and the devoicing and deletion rules on the other hand, irrelevant. Consider the singular 'lake' [ne?háN] and the plurals 'thrusts' [e*e*táhN] and 'turtles' [ma?enóN] shown in (35) above. Note that in these forms the final vowel has not devoiced, but has deleted by rule (9). (This deletion occurred in the singular of 'raccoon' in (30) as well.) These forms nevertheless behave exactly as the other forms in (35).

This behavior of tones on deleted vowels might be taken as indicating that tone spreading must apply prior to the deletion rule. However, given the representation of tones as
autosegments on an independent tier, there is no need to assume that the deletion of a vowel implies the deletion of its tone. Thus the tone may spread prior to vowel deletion, or the vowel may delete first, leaving the tone floating (unassociated), whereupon the tone reassociates. In either case, the spreading or reassociation can be accounted for simply as attraction to accent by (41). 'Lake' could therefore derive in either of the following ways:

(42)

Underlying | | | Underlying | | |
(with *):  ne?hane  (with *):  ne?hane

Rules (9) |  L  L  H  |
and (13):  ne?haN  and (33):  ne?hane

Rule (41): |  L  L  H  |
ne?haN  and (13):  ne?haN

Surface:  [ne?háhN]  Surface:  [ne?háhN]

'lake'  'lake'

Recoverability

What, then, about the recoverability of the tone of the final vowel? Because of contour simplification (33), the tone on any given vowel can have only one of two values: HIGH or LOW. Since tone is not realized on voiceless vowels in Cheyenne, there is a possibility for only a two-way surface tonal distinction (HIGH or LOW on the penultimate) for the four possible underlying tone patterns on the last two syllables (LOW-LOW, LOW-HIGH, HIGH-LOW, HIGH-HIGH). The recoverability of these underlying tone patterns must therefore be limited.

As we have seen, recoverability is limited in the following way. Attraction to accent (41) provides for the complete recoverability of the tone of the final vowel, and
constitutes a case of 'option (4d)' proposed at the beginning of this paper. Contour simplification (33), however, results in the complete loss of the tone of the penultimate. The system thus does not function to render the underlying tonal configuration of the last two syllables of a word maximally recoverable, as is implied in Leman's analysis. Instead, it keeps the tone of the final syllable at the expense of the penult's tone, as shown in the following comparison of derivations claimed by the two analyses for all possible underlying final sequences of tones:

(43)  
<table>
<thead>
<tr>
<th></th>
<th>Leman</th>
<th></th>
<th>Milliken</th>
</tr>
</thead>
<tbody>
<tr>
<td>L L</td>
<td>L -</td>
<td>L L</td>
<td>L -</td>
</tr>
<tr>
<td>H L</td>
<td>Ĥ -</td>
<td>H L</td>
<td>L -</td>
</tr>
<tr>
<td>L H</td>
<td>M -</td>
<td>L H</td>
<td>H -</td>
</tr>
<tr>
<td>H H</td>
<td>Ĥ -</td>
<td>H H</td>
<td>H -</td>
</tr>
</tbody>
</table>

Conclusions

My purpose in presenting the above phonetic facts and phonological analysis has been to illustrate something of the fundamental nature of tone and tonal processes (in Cheyenne), and to show how tone interacts with devoicing, especially with regard to recoverability.

First, the instrumental analysis showed that a penultimate tone does not vary in pitch under the influence of the underlying tone of the following vowel. When low-level phonetic effects such as the interaction of vowel or consonant type and pitch are experimentally controlled, we find only two distinct surface F0 ranges. This phonetic evidence does not accord with the complex assimilation processes and sandhi tones posited by Leman for the penultimate and final syllables. While the tone changes that are actually attested could be interpreted in terms of a process of complete assimilation, a simple and straightforward analysis of the system is readily available if the tones are understood as autonomous units formally independent from the vowels with which they are associated. Vowels can provide a tone for reassociation only if the tone is not an intrinsic part of the vowel itself, as it would be if it were
merely a set of specifications within the vowel's feature matrix. The tone changes observed on the penultimate vowel then follow from principles and rule types well established in autosegmental theory.

Second, the complete recoverability of the tone of the final syllable is provided for by a redistribution of the tone on the word. The tone of the final vowel spreads to the penultimate vowel, which does not devoice. This spreading creates a tonal contour on the penultimate vowel, however, which is eliminated by a rule of contour simplification, resulting in the complete loss of the tone of that vowel.

Finally, by recognizing the independently motivated abstract accent assigned to the penultimate syllable, the asymmetry between the behavior of the tone of the final syllable (which spreads or reassociates) and that of the prepenultimate and penultimate tones (which do not) is explained. The resulting difference in recoverability between the tone of the final syllable and tones earlier in the word is thus accounted for in a principled way.

Acknowledgements

To the Cheyenne speakers, Jenny Flying Out, Josephine Gilmore, and Lillian Levi, who contributed both their speech and their friendship in the course of this study, I say 'ma'xeahóó'. I also thank Mary Beckman for her insightful criticisms and suggestions, and Mark Pedrotti for valuable technical assistance.
References


Production and Perception of the Voicing Contrast

in Indian and American English

Katharine Davis and Mary Beckman

Abstract. We measured VOT's of initial stops and tested stop identification in speakers of Indian English (Hindi bilinguals) and American English. Our measurements revealed significant inter- and intra-dialectal differences in distribution of VOT's for the speakers, but only an inter-dialectal difference in identification patterns. Thus production of VOT can vary considerably within a general dialect group without affecting the uniformity of the group's perception norms.

Introduction

The description of phonemic stop contrasts along such dimensions as voiced/voiceless and aspirated/unaspirated was simplified greatly when Lisker and Abramson [1964] discovered that these contrasts can be classified by voice onset time (VOT). The perceptual relevance of this laryngeal timing mechanism has since been confirmed in many experiments showing that VOT is a robust cue in the perception of the phonemic contrasts so classified [cf. Lisker and Abramson 1970, Fischer-Jørgensen 1972, Williams 1977a, Shimizu 1977, Keating et al. 1981].

Lisker and Abramson found that the categories differentiated by VOT seem to fall into three well-defined groups—namely, stops with negative VOT ("voicing lead"), stops with zero or small positive VOT ("short lag"), and stops with large positive VOT ("long lag"). Languages in their study with more than a two-way contrast along the VOT scale used all three of these categories, and those with two-way contrasts always seemed to use adjacent categories. For example, Dutch and Spanish voiced versus voiceless stop
phonemes contrast voicing lead with short lag, whereas Cantonese unaspirated versus aspirated stop phonemes contrast short lag with long lag.) The importance of this finding has been confirmed by experiments that suggest a general psychoacoustic basis for the divisions between any two adjacent categories [cf. Streeter and Landauer 1976, Miller et al. 1975, Pisoni 1977]. Thus voice onset time as used in the world's languages seems to be a three-step scale rather than a continuum.

We performed an experiment to see whether this finding can be extended to production and perception patterns in two dialects of English that apparently differ in the VOT categories used to contrast initial voiced versus voiceless stops. The two dialects that we studied were American English and the variety of English spoken by Hindi speakers ("Indian English").

Lisker and Abramson's pioneering cross-language study showed that American English initial stops, unlike the similarly labeled Dutch and Spanish phonemes, fall into the short lag versus long lag categories. Three of their four American English speakers consistently produced voiced /b d g/ with zero or short positive VOT, and all four speakers produced voiceless /p t k/ with long VOT's like those in the analogous aspirated stops in Cantonese. On the other hand, informal observations have suggested that Hindi speakers do not have as much aspiration in their English /p t k/ as do American speakers.

If the ostensibly identical phonemic categories in these two dialects of English do indeed differ in VOT, they should fall into different adjacent pairs of the three VOT categories suggested by Lisker and Abramson's study. Moreover, there should be consistent interdialectal perceptual confusion for the intermediate category, short lag, which should be identified with the voiced stops by American English speakers and with the voiceless by Indian English speakers.

We measured VOT in minimally contrasting English words spoken by American and Indian speakers, and then tested these same speakers' perception of the various word tokens produced by the speakers in both dialect groups.

The results of the production test revealed a great deal more idiolect variation than Lisker and Abramson's study would suggest. Many of the American speakers produced voicing lead for /b d g/ (as did the single "anomalous" speaker in Lisker and Abramson's experiment). Moreover, the
results for most of the Indian speakers suggested a slight influence from American English, although the overall trend confirmed our earlier informal observations.

The results of the perception test, on the other hand, revealed a surprising stability in the perception of these categories. They confirmed all earlier studies showing the perceptually relevant distinction for American English to be short lag versus long lag, even for the speakers who consistently produced voicing lead for /b d g/. Moreover, they revealed that the perceptually relevant distinction for Indian English is voicing lead versus voicing lag, even for those speakers whose production seemed to have been influenced by American English norms.

In this paper, we will describe the experiments that we performed, and discuss the implications of the results for second-language acquisition patterns and sound changes involving voice onset time.

**Methods**

Production Experiment. Three minimal pairs, differing only in the voicing of the initial stop consonant (namely, puck/buck, tall/doll, cull/gull) were embedded in the frame sentence "Please say _____ again." The sentences containing these target words were placed in a list along with twice as many other sentences containing filler words. The order of the list was random, except that no two sentences containing minimally contrasting target words appeared next to each other. The target sentences occurred five times each on the list.

Ten subjects participated in the experiment. Five were native speakers of American English (four from New York State, one from North Carolina), and five were native speakers of the Hindi dialect spoken in North Central India. The speakers in the latter group had learned Indian English in school between the ages of six and ten. Four of the Indian speakers were long-time residents of the United States, having lived in the country for three to five years. The fifth speaker had arrived only six months previous to the experiment. Nine of the ten subjects were graduate students in engineering at Cornell University; the other was the spouse of a student. All were between the ages of 22 and 32.

The subjects read the list of minimal pair sentences individually in a soundproof booth. The readings were recorded at 7 1/2 ips on a good quality tape recorder. If
any error in reading was made, the subject was asked to repeat the entire sentence.

The target utterances were then digitized at 20 kHz onto the Phonetic Lab's PDP-11/40 computer. VOT was measured from the digitized wave-form, using a wave-form editor developed by David Walter and modified by Mark Pedrotti. In tokens with voicing lead, negative VOT was measured from the drop in intensity signifying stop closure after the preceding vowel to the highest-intensity spike in the release. In tokens with voicing lag, positive VOT was measured from the highest-intensity spike in the release burst to the onset of voicing in the following vowel. When there was a double release (with two equally intense spikes), the measurements were made to the first spike (for negative VOT) or from the second spike (for positive VOT).

Perception Experiment. After all the measurements were completed, the target words were excised from their frames and converted back to an analog recording in a random order on another tape. The tokens were separated on the tape by 1.5-second long silent intervals.

The stimulus tape of 300 tokens was presented in a quiet room to all ten subjects who had participated in the production experiment. The subjects were instructed to identify each stimulus by circling on an answer sheet the word heard. The answer sheet consisted of a list of word pairs, corresponding to the order of presentation on the tape, with the word having the initial voiceless stop on the left and the word having the initial voiced stop on the right.

Results

Production Experiment. The results of the production experiment are shown in Figures 1 and 2. The graphs in these figures are histograms of the distribution of measured VOT values for the two different types of stops as produced by the speakers in the two dialect groups.

As Figure 1a shows, the VOT values for /b d g/ produced by American speakers had a clear bimodal distribution; three-quarters of these stops had voicing lead and the other quarter had short lag. There was no overlap in VOT between these two types of voiced stops. There was also virtually no overlap between the VOT's for /b d g/ with short lag and the VOT's for /p t k/, which had a clear unimodal distribution in
Figure 1. Distribution of measured VOT's in the tokens of words with voiced and voiceless initial stops produced by speakers of American English (upper figure) and by speakers of Indian English (lower figure).
the long lag region, with a mean value of 80 ms (standard deviation 22).

These distributions contrast sharply with the distributions of VOT's for the same stops produced by the Indian speakers. As Figure 1b shows, nearly all of the Indian English tokens of /b d g/ had voicing lead. Only five (7%) of the tokens had short lag. Moreover, the values for these /b d g/ with short lag fell completely within the range covered by the Indian English tokens of /p t k/. The latter group of phonemically voiceless stops had a mean VOT of 25 ms (st. dev. 8). These results indicate a substantial inter-dialectal difference in the distributional patterns for VOT's of initial voiced and voiceless stops.

There is also evidence of intra-dialectal variability, as shown in Figure 2. Figure 2a is a histogram for the VOT values produced by American speaker AB. (For comparison, the two types of dashed lines in the background show again the distribution of values for all five American speakers displayed in Figure 1a.) As can be seen from the figure, speaker AB was responsible for nearly all of the tokens of /b d g/ with short lag. She produced only one token of a phonemically voiced stop that had any voicing during the stop closure. In other words, the bimodal distribution of VOT values for American English /b d g/ reflects an apparent idiolectal split among the speakers.

There is evidence for an idiolectal split among the Indian English speakers as well. Figure 2b is a histogram for the VOT values produced by Hindi speaker SB. (Again, for comparison, the two types of dashed lines repeat the distributions for all speakers in the dialect group shown in Figure 1b.) SB was the Indian speaker with the least exposure to American English. Her tokens of /b d g/ had on the average significantly more voicing lead than did those produced by the other Hindi speakers (t=5.021, p<0.0001), and she was responsible for none of the five tokens in this phonemic category that had positive VOT. Her tokens of /p t k/, moreover, had significantly less voicing lag than did the voiceless tokens produced by the other Indians (t=5.126, p<0.0001). These results suggest that the other four speakers, all of whom had been in the United States for three or more years, had been influenced somewhat by American English. Their production of /b d g/ and /p t k/ seems to have shifted somewhat toward the American English norms.
Figure 2. Distribution of measured VOT's in the tokens produced by American speaker AB (upper figure) and by Indian speaker SB (lower figure). The two different types of dashed lines in the figures represent the distribution of VOT's for voiced and voiceless stops by all of the speakers in each dialect group.
Perception Experiment. The results of the perception test are shown in Tables I and II. Table I shows the error rate for the American and Indian listeners divided by the phonemic identity of the stop and the identity of the speaker who produced the token heard.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>American</th>
<th>Indian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener</td>
<td>over</td>
<td>all</td>
</tr>
<tr>
<td>a. /b d g/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Indian</td>
<td>41</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. /p t k/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Indian</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>c. overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Indian</td>
<td>21</td>
<td>5</td>
</tr>
</tbody>
</table>

As the next-to-last column in the table shows, the Americans did better in identifying American English tokens than they did in identifying Indian English tokens. Their overall error rates for these two categories were 1% versus 16%, a highly significant difference ($\chi^2(1) = 104.86$, p<0.005).

The reason for this difference becomes apparent when the token categories are further broken down by stop type. The Americans misidentified Indian English /b d g/ only 3% of the time (not substantially more often they they did American English tokens of this category), but they misidentified the Indians' /p t k/ 29% of the time. Moreover, when this error rate is even further broken down by the identity of the individual speaker, it is seen that SB's tokens of these phonemically voiceless stops produced significantly higher error rates from the Americans than did those of the other
Indian speakers ($\chi^2(1)=23.91$, $p<0.005$). Since the Hindi speakers generally produced short lag for /p t k/, and since among them SB produced the shortest lag, the breakdown of error suggests a confusion on the part of the American listeners caused by the nearly complete overlap of the Indian English VOT's for voiceless stops with the VOT's identified with /b d g/ in American English.

The error rates in Table I also show a complementary confusion on the part of the Indian listeners. As the last column in the table shows, their overall error rate for American English tokens was not significantly worse than that for Indian English tokens ($\chi^2(1)=4.51$, $p>0.025$). When the tokens heard are broken down by stop type, however, a substantial difference emerges. The Indians misheard 14% of the tokens of /b d g/ produced by the American speakers, as opposed to only 5% of those produced by the Indian English speakers. Moreover, a significant part of the higher error rate is accounted for by speaker AB. Her tokens of the phonemically voiced stops were misheard 41% of the time, significantly more often than those of the four other Americans ($\chi^2(1)=55.17$, $p<0.005$). Since AB was the American speaker who consistently produced short lag for these stops, the high error rate on the part of the Indian listeners suggests confusion caused by the overlap with VOT's identified with /p t k/ in Indian English.

These suggestions are confirmed by the results shown in Table II. This table gives error rates for the tokens broken down by stop type and VOT production category, where "long lag" is defined as 50 ms or more VOT. (50 ms was chosen as the cutoff point between short and long lag because it falls in the middle of the overlap between American English and Indian English /p t k/.)

The highest overall error rates in Table II are for Americans listening to /p t k/ with short lag (27%) and for Indians listening to /b d g/ with short lag. The Americans listeners also misidentified /b d g/ with short lag more often than they did /b d g/ with voicing lead, but their error rate for this category is still significantly better than the Indian listeners' error rate ($\chi^2(1)=29.93$, $p<0.005$).

A second intriguing result evident in Table II is that the identification patterns for SB and AB were not different from those for the other speakers in their respective dialect groups, even though their production patterns were idiosyncratic (as demonstrated above in the results of the production experiment). Thus SB's identification of /b d g/ with short lag was not significantly worse than that of the
Table II. Percent error by individual listener for various VOT production categories.

<table>
<thead>
<tr>
<th>Production category</th>
<th>/b d g/ with voicing lead</th>
<th>/b d g/ with short lag</th>
<th>/p t k/ with short lag</th>
<th>/p t k/ with long lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>0</td>
<td>13</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>BS</td>
<td>0</td>
<td>9</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>CC</td>
<td>2</td>
<td>13</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>GC</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>JB</td>
<td>2</td>
<td>22</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>overall</td>
<td>1</td>
<td>11</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>Indian:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SL</td>
<td>2</td>
<td>30</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>RB</td>
<td>6</td>
<td>43</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>SB</td>
<td>1</td>
<td>43</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>UM</td>
<td>6</td>
<td>48</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VB</td>
<td>4</td>
<td>48</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>overall</td>
<td>4</td>
<td>43</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

other Indian speakers ($x^2(1)=0.01$, $p>0.9$), even though she had had far less exposure to American English, and consequently had less American-like VOT production. Similarly, AB's identification of /b d g/ with short lag was not better than that of the others in her dialect group, even though she was responsible for most of the tokens in this category. Conversely, her error rate for /p t k/ with short lag was not significantly worse than that of the other Americans ($x^2(1)=0.46$, $p>0.3$), even though the VOT's in this category overlapped considerably more with her VOT's for /b d g/.

Discussion

Four conclusions can be drawn from these results. First, even within the same general dialect group, there can be marked variation in the VOT categories used for a given phonemic type. The differential treatment of /b d g/ by AB
and the other Americans confirms the results for the distribution of short lag versus voicing lead among the American English speakers in Lisker and Abramson [1964], except that in the present study, the speakers who used voicing lead for /b d g/ were in the majority, rather than in the minority as in Lisker and Abramson's group. Other studies of American English also show more speakers like AB and fewer speakers like the others [cf. Keating et al 1981]. On the other hand, the distribution in the present experiment is similar to that of Caramazza et al. [1973] for Canadian English speakers. Further study is necessary to determine whether this apparently idiolectal variation is due to some finer regional or social dialect variation within the general category "American English."

A second conclusion is that different general dialects of the same language can differ substantially in the VOT categories used for the same phonemic categories. In conformity with our earlier informal observations, the present study shows that Indian English speakers use short lag for /p t k/ where American English speakers use long lag. It is interesting to relate this cross-dialectal difference to the patterns in the native language of this particular group of Indian English speakers. Hindi has a three-way contrast along the VOT scale, so that the Indian English speakers have available a category of aspirated stops that they could identify with English /p t k/, and yet they do not. The Indian speakers in this group, at any rate, claim that the Hindi aspirated stops are very different from the American English initial voiceless stops. Although measurements of Hindi stops elicited from these same speakers show VOT's comparable to those in American English /p t k/; informal observations of the wave-form suggest that the amplitude of aspiration is higher in the Hindi stops, perhaps accounting for the perceptual dissimilarity [cf. Repp 1979].

A third conclusion is that production of VOT seems to be susceptible to influence by other dialects. The Indian English speakers who had been in the United States for several years had mean VOT's for /p t k/ that were significantly more positive than that for the recent arrival, and they even produced a few tokens of /b d g/ with short lag (instead of with the voicing lead that is the norm for this dialect group). On the other hand, these four speakers had not altered their VOT production so far as to produce /p t k/ with the long lag VOT's that are the norm for American English. This slight shift suggests that, in terms of production, VOT is a continuum rather than a three-step scale.
Finally, it can be concluded that perception patterns are not like production patterns for VOT. They seem not to be so variable within the general dialect group, and are far less susceptible to influence by other norms. Thus despite the idiolectal differences evident in the American speakers' production patterns, they all perceived the various VOT categories in the same way. Similarly, despite the differential influence from American English that was evident in the Indian speakers' production data, they all perceived the different tokens according to their group norm. Although VOT may be a continuum in terms of production, it does seem to fall into a small number of well-defined perception categories.

Thus in general, it seems that the perception of VOT is uniform and stable within general dialect groups, whereas production is variable from speaker to speaker within dialects, and susceptible to influence from other dialects. That production can apparently change or differ without an accompanying change or difference in perception is reminiscent of other situations in which the two types of behavior are independent. For example, small children often acquire the perceptual category for a "difficult" sound such as [s] or [θ], before they can produce the sound; a two-year-old that says [ʃi] for 'see' nevertheless may not accept the same pronunciation from an adult. Conversely, there is some evidence that bilinguals acquire distinct production patterns in speaking their two languages without differentiating their identification patterns in listening to them. This result is seen for VOT production and perception in a study of French-English bilinguals by Caramazza et al. [1973] and in a study of Spanish-English bilinguals by Williams [1977b]. Elman et al. [1977], on the other hand, found that "strong" Spanish-English bilinguals do identify stops in the short lag category differently depending on the context language of the identification task, suggesting that perception may eventually catch up with production.

-----------

1 Mikoś et al. [1978] observe an analogous situation in Polish. They say that some of their Polish speakers used long lag for voiceless stops, even though the general distribution for the language is short lag for voiceless versus lead for voiced. Despite this idiolectal difference, however, all of the speakers displayed the same category boundary appropriate to the language's production norm in identifying stimuli along a synthetic /ta/-/da/ continuum.
This result also has important implications for sound change. That the four Hindi speakers with the most exposure to American English had apparently shifted their production somewhat toward the American distribution pattern without shifting their perceptual categorization suggests that some sound changes can be accomplished by a gradual drift along a production continuum. This drift might go unnoticed within the lifetime of any given speaker until it crosses a discontinuity in the perceptual scale, and is reinterpreted by others acquiring the language for the first time. Such a model of one type of sound change is especially attractive in the case of VOT, which seems to be fairly continuous in terms of articulation but shows strong evidence of having discrete steps perceptually.

Acknowledgements

We thank our informants Sushma Banthia, Vinod Banthia, John Benci, Rajendra Bordia, April Brown, Craig Cameron, Sanjiva Lele, Umesh Mishra, Bill Schaff, and Glenn Swan, for their willing cooperation in this set of experiments. We also thank Mark Pedrotti for technical assistance.

References


