Working Papers of the Cornell Phonetics Laboratory
No. 1 (December 1983), pp. 29-42

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 [Soli 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 1968]. 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 1982].

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]:

$$\overset{V}{[+\text{high}]} \longrightarrow \emptyset \;/\; [-\text{voice}] \;\_\; \left\{ \begin{matrix} [-\text{voice}] \\ \# \end{matrix} \right\}$$

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 [ʃ], which occurs with no obvious allophonic variation before both short /i/ and short /u/. The first

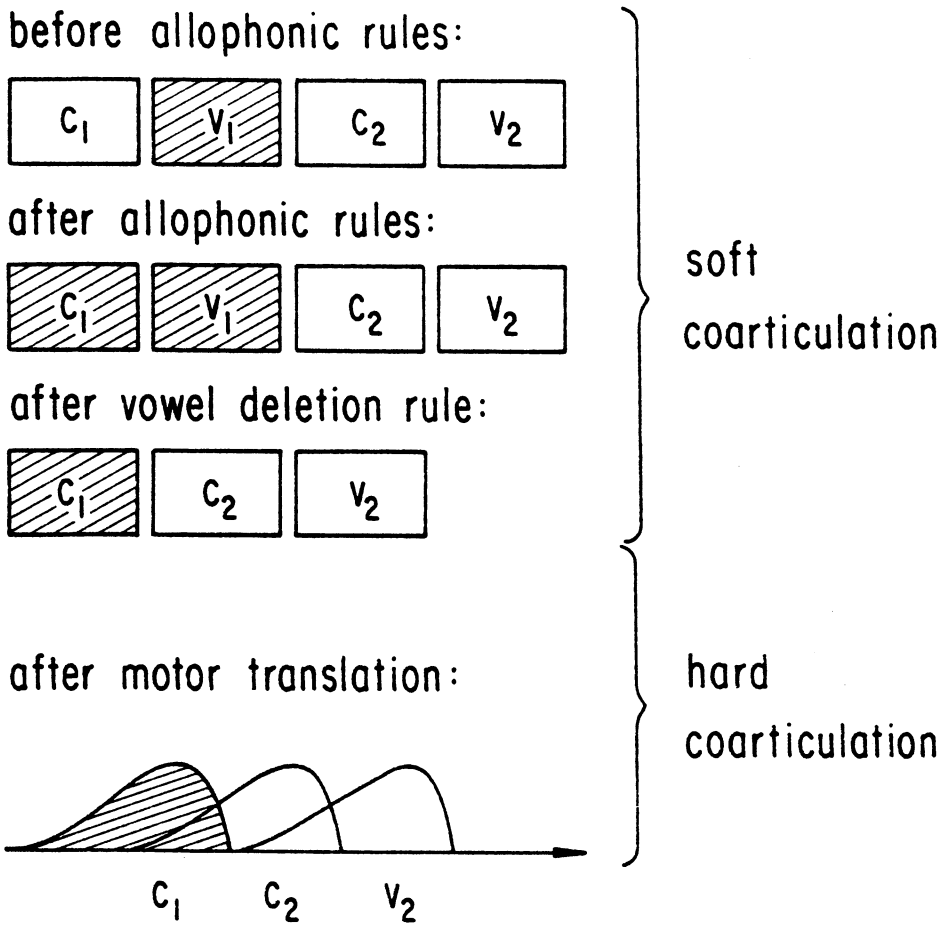

Figure 1. Traditional translation models account for the recoverability of the deleted vowel by ordering the vowel-deletion process after other high-level processes that produce vowel-dependent contextual variation in the syllable's consonant.

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 I. 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 o 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/
/kaˈsi/	/kaˈsyu/

/N/ is the "moraic" nasal and /ˈ/ 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 [ʃ]'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/ ([ʃk^{h}aã]) using the SRS rules for Japanese [Hertz and Beckman 1983], and then modified the OVE IIId 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.

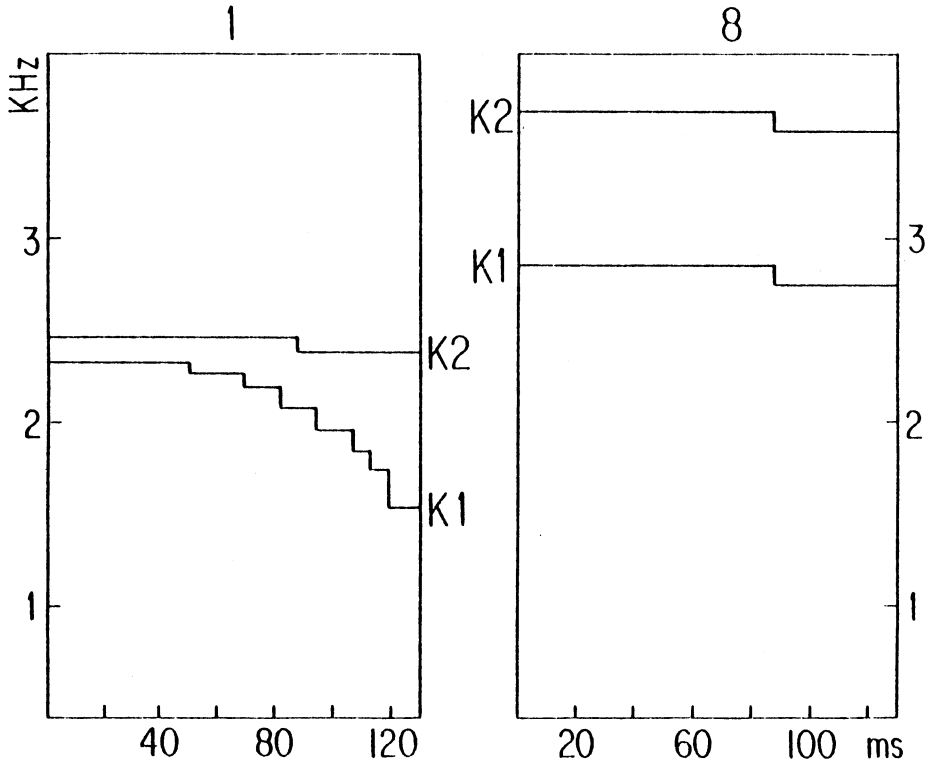

Figure 2. K1 and K2 values for synthesized syllabic fricatives in endpoint stimuli modeled on "best" natural tokens.

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

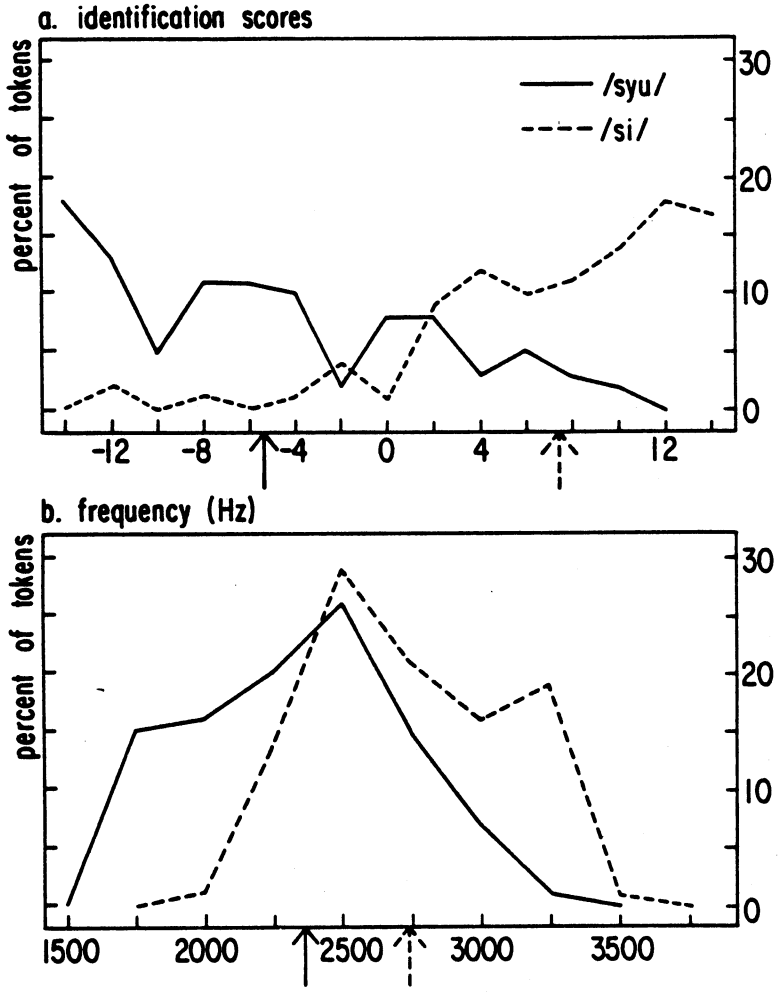

Figure 3. Relative frequency distributions of identification scores (upper figure) and of fricative's lower edge values (lower figure) in tokens of words with no voicing in target /syu/ or /si/ syllable. Arrows below abscissa indicate mean values.

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_4 are the speakers. The coefficient of determination for this function is 44.1%.

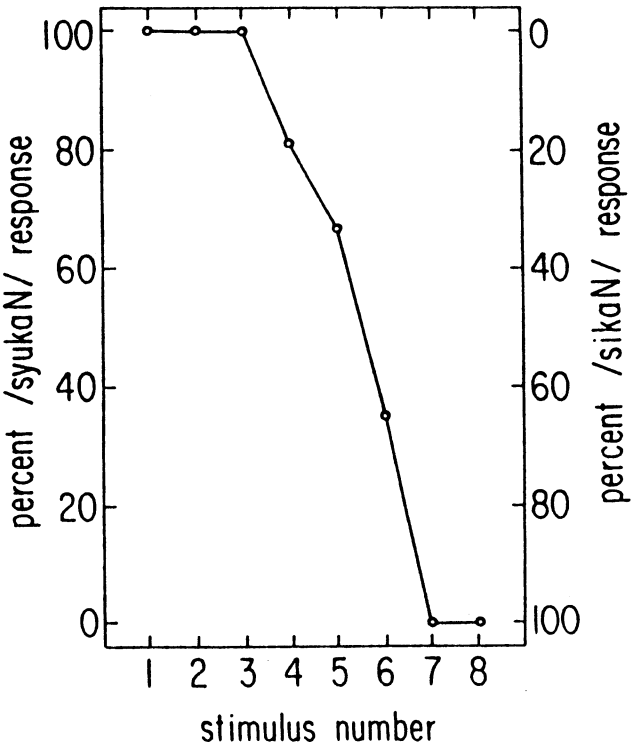

Figure 4. Percentage of /syukaN/ or /sikaN/ responses for each stimulus in second experiment.

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 consonsant 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

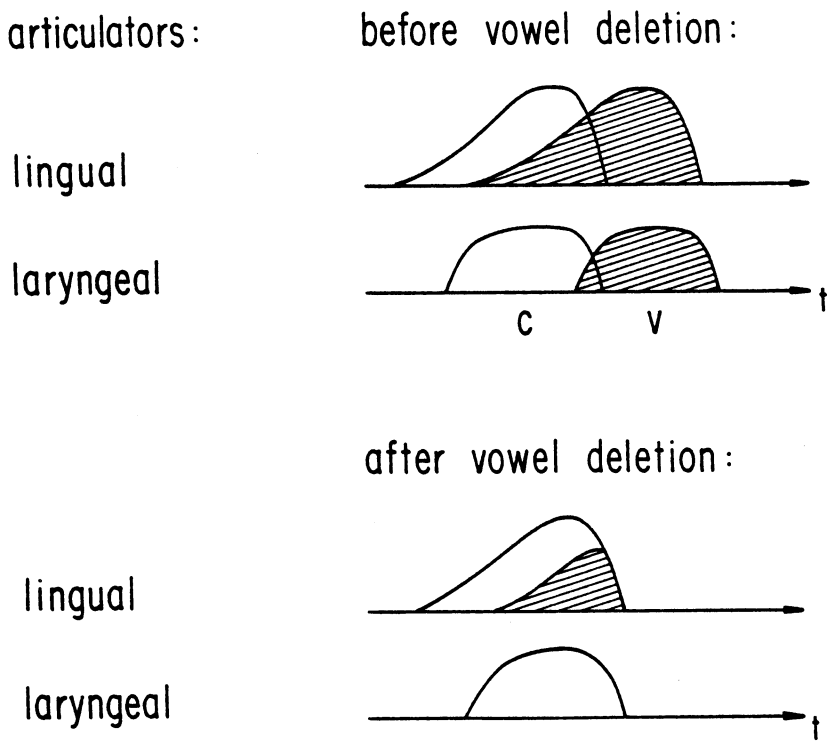

Figure 5. A model positing underlyingly dynamic segments might explain devoiced syllables as the result of removing only the portion of the syllable associated with the laryngeal gesture for the vowel.

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

Beckman, Mary. 1982. Segment duration and the "mora" in Japanese. Phonetica 39: 113-135.

Bell-Berti, Fredericka, and Katharine S. Harris. 1981. A temporal model of speech production. Phonetica 38: 9-20.

Fowler, C.A., P. Rubin, R.E. Remez, and M.T. Turvey. 1980. Implications for speech production of a general theory of action. In Butterworth, ed., Language production, pp. 373-420 (New York: Academic Press).

Fujimura, Osamu, and Julie Lovins. 1978. Syllables as concatenative phonetic units. In Bell and Hooper, eds., Syllables and segments, pp. 107-120 (Amsterdam: North-Holland).

Hertz, Susan R. 1982. From text to speech with SRS. J. Acoust. Soc. Am. 72: 1155-1170.

Hertz, Susan R., and Mary E. Beckman. 1983. A look at the SRS synthesis rules for Japanese. In Proc. 1983 Int. Conf. ASSP, pp. 1336-1339 (New York: IEEE).

Hirayama, Teruo. 1979. Zen Nihon no hatuon to akusento. In Nihongo hatuon akusento ziten, App., pp. 103-138 (Tokyo: Japan Broadcasting Corporation).

McCawley, James D. 1968. The Phonological Component of a Grammar of Japanese (The Hague: Mouton).

Ohso, M. 1973. A phonological study of some English loan words in Japanese. Ohio State University working papers in linguistics 14: 1-26.

Shibata, T. 1955. Museika. In Kokugogaku jiten, p. 899 (Tokyo: Kokugogaku Gakkai).

Soli, Sigfrid D. 1981. Second formants in fricatives: acoustic consequences of fricative-vowel coarticulation. J. Acoust. Soc. Am. 70: 976-984.

Yeni-Komshian, Grace H., and Sigfrid D. Soli. 1981. Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. J. Acoust. Soc. Am. 70: 966-975.