

Tsuchida, Ayako 1997. Phonetics and Phonology of Vowel Devoicing. PhD dissertation, Cornell University.
 Vance, T. J. 1987. *An Introduction to Japanese Phonology*. Albany, NY: State University of New York Press.

Department of Linguistics
 Rutgers University
 18 Seminary Place
 New Brunswick, NJ 08903
 tsuchida@rci.rutgers.edu

Integrating Speech Perception and Formal Phonology*

Natasha Warner
University of California, Berkeley

1. Introduction

Formal theories of phonology have concentrated overwhelmingly on the production of speech, almost to the exclusion of its perception. All of the major recent theoretical approaches to phonology, from early generative phonology to current work in Optimality Theory, are theories of how to convert the underlying representation to the surface form, or an approximation of it. These theories do not address how a listener, who hears a surface form, knows which underlying representation the surface form belongs to. The task in spoken word recognition is to determine which lexical entry, which underlying representation, a surface form corresponds to. Some aspects of spoken word recognition, such as normalization for speaker specific variation, probably do not require the listener to use knowledge about the phonology of the language.¹ However, other aspects of matching a surface form to a lexical entry do require the listener to have, and use, knowledge about the phonology of the language. How does an English listener know that a voiceless unaspirated [t] in one environment is to be associated with the underlying form /t/ and in another with /d/? This, and much more complicated knowledge about the grammar of the language, is required. If working from the surface form back to the lexical entry requires the listener to use the grammar of the language, then this might be a process which formal theories of phonology can model.

Very few researchers have attempted to do this directly, but some have attempted to bring formal phonology and speech perception together in other ways. Lahiri and Marsten-Wilson (1991, 1992) use formal theories of phonology to shape theories of speech perception. They propose that speech is perceived by comparing it to a radically underspecified underlying representation, not by comparing it to a surface form. Steriade (1997) takes a different approach, using experimental findings about speech perception to

* The work on Japanese reported here was supported by a grant from the Vice Chancellor's Research Fund of the University of California, Berkeley. I would also like to thank John Ohala for helpful discussions of this material and Shawn Ying for technical assistance. Johnson's (1997) exemplar based model calls into question the separation of speaker normalization from other aspects of spoken word recognition, however.

inform formal theories of phonology. She does this by arguing that the relative availability of perceptual cues in different environments explains some phonological patterns better than syllable structure does. Similarly, Flemming (1995) proposes a set of perceptually based Optimality Theoretic constraints which model the choice of what segments contrast in a language.² Finally, Smolensky (1996) models young children's speech perception in an OT framework, but his example is quite preliminary, and will not be discussed here.

In other work, I have performed a gating experiment on English and Japanese which has implications for the Lahiri and Marslen-Wilson and Steriade approaches to integrating perceptual factors with formal phonology. In this paper, I will discuss these two proposals and present results from my experiment which bear on them.

2. Experimental methods

The results in this paper are a subset of the results from a larger gating experiment on English and Japanese. In a gating experiment (Hart and Cohen, 1964; Grosjean, 1980), listeners are presented with various truncated portions of a word and asked to identify the word. By comparing responses at different truncation points (gates), one can determine where in the signal crucial perceptual cues for a given segment become available. By asking listeners to identify the entire word (of which they have heard only a part), one can also collect information about spoken word recognition: the responses given by a group of listeners to a particular stimulus are assumed to represent the pool of lexical items listeners are considering as candidates for the word (Grosjean, 1980; Tyler and Wessels, 1985; Tyler 1984).

The most common method in gating experiments (at least for psycholinguistic purposes) is to end-truncate a word at successively earlier points separated by regular intervals. Each listener hears all of the stimuli produced from a word, in order from the most severely truncated to the least severely truncated (the entire word). Thus, the listener hears first the initial 50 ms of each word, then the initial 100 ms of the same words, then the initial 150 ms, then the initial 200 ms, etc., until the listener hears the complete words. (Shorter gating intervals, such as 10 ms or 20 ms, are also common.) This method of presenting all of the gating stages of each word to the same listeners is called successive presentation. Colton and Grosjean (1984) present an experiment showing similar responses from successive presentation and individual presentation (using different listeners for each gating stage of a word). Ohala and Ohala (1995), however, argue convincingly against successive presentation in gating experiments. Because the individual presentation method (in which no listener hears more than one gating stage of a particular word) multiplies the number of subjects required by the number of gating stages, most researchers use successive presentation.

In this experiment, I chose to use the individual presentation method in order to avoid effects of previously given responses on listeners' behavior for later stimuli. I also chose

² While Flemming's work focuses on the role of perceptual factors in phonology, and involves comparing surface alternants rather than working from an underlying form to a surface form (1995:120-121), it is not a model of speech perception. It is a model of production in that it attempts to account for what forms are allowed (can be produced) in a language, not for how a listener recognizes those forms.

³ Some researchers argue that gating allows excessive postperceptual processing because it allows listeners several seconds to respond, and that responses therefore do not represent the candidate words which are being considered online at a particular point. Tyler and Wessels (1985) address this issue.

to use a 20 ms gating interval so that relatively fine temporal details of how perceptual cues become available would be reflected in the results. In order to keep the necessary number of subjects feasible, I chose to gate not through the entire word, but only through a two segment sequence of each word. Each word was chosen to represent a particular two segment sequence. For example, "remedy" was chosen to test perception of the /em/ sequence. The shortest gate allowed the listener to hear from before the beginning of the word up to halfway through the [e], the next gate allowed the listener to hear from before the beginning of the word up to 20 ms beyond the endpoint of the first gate, the next gate again 20 ms longer, etc. The longest gate allowed the listener to hear from before the beginning of the word to halfway through the [m]. Each listener heard parts of many different words, but no listener heard more than one gating stage of any word. Different words require different numbers of gates, depending on the duration of the two segment sequence of interest.

127 English words and 76 Japanese words were chosen to represent a wide variety of two segment transitions, including CV, VC, CC, and VV sequences. Voicing, place, and manner of consonants was manipulated. Placement of stress in English and pitch accent in Japanese was also manipulated. The English words were recorded by one male native speaker of American English, and the Japanese words by one male native speaker of Tokyo Japanese. Subjects were asked to respond with a whole word the stimulus might have been the beginning of, in an open ended response format. Further details can be found in Warner (1998).

Eleven native speakers of English responded to each English stimulus, and twelve native speakers of Japanese responded to each Japanese stimulus. The entire experiment required 154 English listeners and 120 Japanese listeners.⁴ The responses can be analyzed in several ways. For example, one can analyze the number of different responses given by the entire group of listeners at each gate of a word. This reflects the timing of listeners' progress toward recognizing the word, as they converge on a smaller number of different responses. One can also analyze the percentage of responses at each gate which have the second segment of the two segment transition of interest (the target segment) correct. This shows the timing of how perceptual cues for that segment become available. Furthermore, one can analyze the percentage of responses at each gate which match the target segment in a particular feature, such as voicing, place, or manner, to investigate the perceptual cues for that feature.

3. Perception by comparison to the UR

3.1. Lahiri and Marslen-Wilson's proposal

Lahiri and Marslen-Wilson (1991, 1992) argue that because a given lexical entry can have many different surface forms depending on its environment and on speech rate, listeners could not possibly perform spoken word recognition by comparing the signal they hear to a surface form, since they could not know which of the many surface forms to use. Instead, they propose that listeners extract distinctive features from the signal, and compare these distinctive features to the underlying representations of lexical entries. Furthermore, they claim that the underlying representation listeners use for this purpose is a radically underspecified one. They believe this means that listeners can make use of non-distinctive phonetic cues only if they are for marked features, not if they are for unmarked

⁴ This is because the English word with the longest duration for its two segment sequence of interest required 14 gates, while the maximum number of gates required for any Japanese word was 10. Again, further details appear in Warner (1998).

features. That is, if the speech signal contains a cue for the unmarked value of a feature, this is of no use to the listener, because one cannot compare an acoustic cue to the lack of a specification in the lexicon, only to the presence of some specification (the marked value).

Lahiri and Marslen-Wilson (1991, 1992) present results from a gating study of English and Bengali on perception of distinctively and non-distinctively nasalized vowels in the two languages to support this proposal. They presented English listeners with CVC and CVN words, gated from the end of the word. Bengali listeners were presented with gated CVC, CVN, and CVC words. Responses were evaluated for whether the final consonant was oral or nasal, and in Bengali, for whether the vowel was distinctively nasalized. I will focus on the English portion of Lahiri and Marslen-Wilson's experiment. Their claim is that since English does not have distinctively nasalized vowels, and [+nasal] is the marked value for consonants, English listeners can use the non-distinctive cue of vowel nasalization to perceive that a following consonant is nasal (the marked value of the feature). However, they predict that English listeners cannot use lack of nasalization on a vowel to rule out a following nasal consonant as a possibility, since [-nasal] is the unmarked value of the feature.⁵ They present experimental results to support this hypothesis, showing that at least some English listeners respond with CVN words even when the vowel is not nasalized, and that many respond with CVN when the vowel is nasalized. However, their conclusions rest largely on the interpretation of minority responses, given by only five to fifteen percent of the subjects, and they do not present statistical tests for most results. Ohala and Ohala (1995) replicate this experiment and re-analyze both the predictions and conclusions at length, so I will not discuss the details of these results here.

3.2. Results on perception of voicing

The results of my experiment for words with postvocalic obstruents as the target segment have implications for Lahiri and Marslen-Wilson's (1991, 1992) proposal. I analyzed the responses listeners in my experiment gave to such stimuli to see how accurately listeners' responses at various gates matched the voicing of the postvocalic obstruent of the stimulus. This is certainly not the first test of listeners' ability to perceive the voicing of a postvocalic obstruent based on cues occurring during the vowel (see Nearey, 1997), but the methods used in my experiment are very similar to those used by Lahiri and Marslen-Wilson to test their theory for nasalization. Therefore, the perception of postvocalic obstruent voicing in my experiment provides a good test of Lahiri and Marslen-Wilson's proposal using a different distinctive feature from the one they tested.

In English, vowels are considerably longer before a voiced obstruent than before a voiceless obstruent (Peterson and Lehiste, 1960; Chen, 1970). (Many languages have this durational difference, and Kluender et al., 1988 suggest an auditory motivation for it.) Lahiri and Marslen-Wilson's proposal that listeners can only use non-distinctive cues to perceive a marked feature, not to rule out an unmarked feature, means that English listeners should be able to use the non-distinctive cue of a lengthened vowel to perceive a following voiced obstruent, but they should not be able to use a short vowel to perceive that the

⁵ It is important to note that Lahiri and Marslen-Wilson do not predict that listeners will choose the unmarked value of the feature in the absence of evidence to the contrary. They predict that even when there are cues for the unmarked value of a feature, listeners will respond with either the marked or the unmarked value of the feature, because they cannot use a cue for an unmarked

following obstruent is voiceless. This is because this vowel length difference is not distinctive in English, and voicing is marked for obstruents. (The vowel length difference under discussion is the difference in duration between the vowels of "bad" and "bat," for example, not the contrastive difference between /i/ and /ɪ/, although that is often also referred to as a vowel length difference in English. While the vowel duration difference in "bad" and "bat" may serve as an important perceptual cue to the identity of the following stop, few English speakers would feel that these are two different vowels, so the durational difference will be considered non-distinctive.) Other cues in addition to vocalic duration may be present during the vowel, but duration is expected to be an important one, and the same argument would apply to other potential cues.

Lahiri and Marslen-Wilson's prediction for postvocalic obstruent voicing is quite clear, although they do not discuss this case specifically. They state that they are assuming radical underspecification, in which "the feature array for a given segment will not contain a specification for any feature, distinctive or not, that has the unmarked value. Consequently, the only specifications in the underlying representation, on this account, are those for features which are (a) distinctive, and (b) have the marked (or non-default) value" (1991:253). There can be no doubt that they would posit underlying representations of English postvocalic obstruents in which only [+voice], and not [-voice], is specified. They must therefore predict that listeners cannot use a short vowel to rule out a following voiced obstruent, and that when listeners hear a short vowel, they will choose randomly (within constraints of the cohort) between voiced and voiceless following obstruents, or perhaps also no obstruent at all. When listeners hear a relatively long vowel, however, they should be able to perceive that the following obstruent is voiced based only on the vowel length, before hearing the obstruent, since a long vowel is a cue for the marked value of voicing.

The English words in my experiment with postvocalic obstruents as the target segment appear in (1), with the two segment sequence of interest (through which the word was gated) transcribed after them.

(1)	a.	Voiceless obstruent:	Voiced obstruent:
		bucket /ʌk/	ravish /æv/
		master /æs/	judge /ʌdʒ/
		session /eʃ/	fade /eɪd/
		latches /æʃtʃ/	
		bite /aɪ/	
		doubt /aʊt/	
		oats /oʊ/	
		circle /ɜːk/	
	b.	citizen /ɪ/	muddy /ʌd/
		committee /ɪ/	soybean /oʊb/
		fitness /ɪ/	
		induction /ʌk/	
		leaf /i/	
		relief /i/	

For the evaluation of perception of voicing of postvocalic obstruents, the words in (1b) were excluded. "Citizen, committee, muddy" were excluded because their postvocalic obstruents are realized as flaps, and English dialects vary as to whether there is any vowel duration difference before flaps derived from /t/ vs. /d/. "Fitness, induction" were excluded because only one gate was short enough to end clearly before the beginning of the stop

obstruent vowel. "Soybean" was excluded because there are so few words beginning with /so/ ("soy, soymilk, soybean, soil"), so listeners' responses may be unduly affected by the small number of possibilities. Finally, "leaf, relief" were excluded because many listeners perceived them as "leave, relieve" even during gates ending during the /f/.⁶

The words in (1a) were evaluated for whether the responses contained a voiced obstruent, a voiceless obstruent, or no obstruent at three stages: the first gate of the word, which is located near the middle of the pre-obstruent vowel, the last gate before the obstruent begins, and the final gate, which is located in the middle of a fricative or just after the burst of a stop. An example of the responses given at these three gates, and the evaluation of the percent of responses with a voiced or voiceless obstruent, is shown in (2). The number of subjects giving each response is shown in parentheses after the response.

(2) Responses to:	"latches" /æf/ #	"fade" /e'd/ #
at first gate (middle of æ, early in e')	lap (3) Latin (2) laugh (1) laugther (1) lack (1) lactose (1) loud (1) box ⁷ (1) 91% voiceless	face (7) fate (3) fake (1) 100% voiceless
at last gate before closure	laugh (4) last (2) lap (1) lapse (1) blast (1) glasses (1) box (1) 100% voiceless	fade (8) phase/faze (2) faith (1) 91% voiced
at last gate (just after stop burst, or during frication noise)	latch (4) latched (1) latchkey (1) match (1) lattissimus dorsi (1) Latvia (1) Latimer (1) latter (1) 82% voiceless ⁸	fade (11) 100% voiced

⁶ Some idiosyncratic feature of the speaker's production of these words may have caused these misperceptions. The decision to exclude these words from the evaluation is post hoc.

⁷ Some responses such as "box, loud" may appear anomalous. Open response data, particularly for whole word responses, is known to be noisy. I believe the value of open response data for studying spoken word recognition outweighs this disadvantage.

⁸ "Latimer, latter" are not counted as voiceless responses because the /l/ in them is realized as a flap. It is possible that listeners have recognized the postvocalic obstruent as voiceless, identified it as the phoneme /l/, and then given responses in which the /l/ is not realized as voiceless, but I am not willing to assume that based on a pool of 11 listeners. If 100 listeners

For the word "latches," the voicing of the postvocalic obstruent is perceived correctly by most listeners at all three time points. The voiced obstruent in "fade," however, is perceived as voiceless by all listeners early in the vowel, but most listeners perceive it as voicing correctly by the last gate before the closure of the /d/. All listeners perceive it as voiced once they have heard the /d/. The average of such results for all the words in (1a) appears in (3).

(3) Average percentage of each type of response to words in (1a)

VOICELESS STIMULI			
Gate	Voiceless responses	Voiced responses	No obstruent
First	72%	6%	23%
Last pre-obstruent	89%	6%	6%
Final	97%	0%	3%

VOICED STIMULI			
Gate	Voiceless responses	Voiced responses	No obstruent
First	79%	21%	0%
Last pre-obstruent	27%	73%	0%
Final	0%	100%	0%

Overall, listeners tended to give responses with a postvocalic voiceless obstruent when the vowel was relatively short regardless of what the stimulus was. That is, for the voiceless stimuli at both pre-obstruent gates and for the voiced stimuli at the first gate, they gave a majority of voiceless responses. However, when the vowel was longer, as in the voiced stimuli at the last gate before the obstruent, they switched to a majority of responses with voiced obstruents. The shift for the voiced stimuli from 79% voiceless responses when the vowel is cut off early to 73% voiced responses when it has approximately its natural duration is striking. At the final gate, by which point listeners had heard part of the obstruent itself, both types of obstruents were perceived rather accurately.

I tested the difference between the voiceless and voiced stimuli in percentage of listeners giving voiced responses at the last pre-obstruent gate, using an ANOVA with analysis of weighted means to correct for the unequal number of words in the two groups. The difference between the two types of stimuli was significant ($F(1,9)=22.8, p<.005$). This shows that English listeners are able to use the cue of preceding vowel duration (or whatever other cues might be available during the vowel) to perceive the voicing of both voiced and voiceless obstruents, even though [-voice] is the unmarked value for obstruents. This result conflicts with Lahiri and Marslen-Wilson's proposal.

I also investigated the perception of voicing for the Japanese words with postvocalic obstruents. The voiceless obstruents had approximately the same results as in English, with a large percentage of listeners giving voiceless responses at all three points in time. However, the voiced obstruents did not show the shift which was present in the English data between the first gate and the last pre-obstruent gate. One Japanese word, /kanada/ 'Canada,' did have a higher percentage of voiced responses at the last gate before the /d/ heard this stimulus, and many gave responses with /l/ realized as a flap, but none gave responses with /d/ realized as a flap, then one could be sure these responses should be counted as voiceless.

closure than earlier in the vowel, but the increase was relatively slight. There were a majority of voiced responses at all three time points, probably because the corresponding word with a voiceless obstruent, /kanata/ 'yonder', is somewhat archaic and of low frequency. The other Japanese words with postvocalic voiced obstruents showed no increase at all in the percentage of voiced responses between the first gate and the last pre-obstruent gate.

I suspect the lack of this effect in Japanese reflects a smaller difference in vowel duration before voiced and voiceless obstruents than is present in English. English has phonologized this difference, making it considerably larger than in many other languages (Chen, 1970). Even in English, the difference in vowel duration is much larger for autosyllabic vowel-obstruent sequences than when the obstruent is the onset of the next syllable. In Japanese, the obstruent must always be the onset of the next syllable, since geminates were not under consideration in this comparison. In investigations of duration compensation regarding the hypothesis of mora timing, researchers have shown that Japanese vowels are longer before voiced obstruents than before voiceless ones (Port et al., 1980; Homma, 1981; Beckman, 1982), but this difference is much smaller than in English. Thus, the difference which forms the probable cue is not as large in Japanese, and Japanese listeners may be unable to use this cue to distinguish obstruent voicing. The Japanese listeners' relatively accurate judgments about postvocalic voiceless obstruents may stem not from the use of any durational cue, but from the low frequency of voiced obstruent phonemes in Japanese.

In sum, English listeners are able to distinguish the voicing of a postvocalic obstruent based on non-distinctive cues during the preceding vowel, whether the voicing of the obstruent is marked or unmarked. This provides evidence against Lahiri and Marslen-Wilson's (1991, 1992) proposal that speech is perceived by comparing it to a radically underspecified underlying representation. Warren and Marslen-Wilson (1988) did find results very similar to these, showing that listeners judge stimuli with short vowels to have voiceless following obstruents, whether the stimulus was an early gate of a word with a voiced obstruent or actually had a voiceless obstruent. However, they give an explanation for this finding which directly contradicts the position laid out in Lahiri and Marslen-Wilson (1991, 1992), saying that a short vowel is compatible with either a following voiced or voiceless obstruent (since the vowel could continue and be longer than the listener has heard yet), but listeners choose the unmarked value of the feature when the signal is compatible with either the marked or the unmarked value. That is, listeners choose voiceless responses because that is the unmarked value of voicing for obstruents, rather than failing to rule out voiced obstruents and therefore giving a mixture of voiced and voiceless responses, as Lahiri and Marslen-Wilson (1991, 1992) would predict.

A more plausible interpretation is that when listeners hear a gated stimulus in which the gating process alters the cues to a segment by cutting its duration short, they parse the part of the signal they have heard as *is*. They do not leave open the possibility that the segment they heard might have been longer, for example. When the vowel /a/ is gated, English listeners will perceive it as /a/ (Lang and Ohala, 1996), and will give responses with /a/ until a longer gate allows them to perceive it as /a/. They will not keep open the possibility that the vowel might have gone on for longer than they were allowed to hear and therefore be /a/. Similarly, I believe that my results for perception of postvocalic voiced obstruents show that when listeners hear a short vowel, they do not keep open the possibility that it might be the beginning of a longer vowel, but rather assume it is the

cue for a following voiceless obstruent.⁹ Therefore, the results of both Warren and Marslen-Wilson's (1988) experiment and mine do not mean that listeners assume the default specification of a feature when perceptual cues are compatible with either the marked or unmarked value, but mean that English listeners are able to use the cue of vowel duration to perceive both voiced and voiceless following obstruents, regardless of their status as marked or unmarked.

4. Perception as the basis for formal constraints

4.1. Steriade's proposal

Steriade (1997) investigates the neutralization of voicing distinctions in many languages, a phenomenon often analyzed as neutralization in coda position, with voicing distinctions maintained in onset position. She lists the perceptual cues to voicing distinctions, and evaluates a variety of environments for which of the potential perceptual cues are available in each environment. She ranks the possible environments for obstruents, for example "before another obstruent," "after a sonorant and at the end of the word," "between two sonorants," etc., for how many of the perceptual cues to voicing are present in each environment. She then shows that there is an implicational hierarchy based on the availability of perceptual cues: if a language has a voicing distinction in a certain environment, it also has a voicing distinction in all environments with more cues available. If it neutralizes its voicing distinction in a certain environment, it also neutralizes it in all environments with fewer cues available. She further shows that although many of the voicing neutralizations she discusses have been described as neutralization in coda position, some of them in fact cannot be adequately described by reference to syllable structure, and require reference to availability of perceptual cues. Thus, she shows that perceptual factors are the cause of some patterns which have usually been analyzed as effects of syllable structure.

Steriade further proposes that information about which environments have sufficient perceptual cues available is incorporated in the speaker's grammar of the language. She models this in Optimality Theory through the ranking of a constraint which preserves underlying voicing specifications and several constraints forbidding the voicing distinction in particular environments. Individual languages differ as to the position of the "preserve [voice]" constraint relative to the particular constraints forbidding voicing, and thus neutralize voicing in different environments. Flemming (1995) takes a similar approach by positing OT constraints which reflect perceptual factors as part of the speaker's grammar.

Stevens discusses in several publications (Stevens, 1980; Stevens and Blumstein, 1981; Stevens and Keyser, 1989) the importance of fast changes in the speech signal, particularly those taking less than approximately 30-40 milliseconds, as regions of the signal which carry a large amount of information. He suggests that some distinctions can be perceived during these short time windows of rapid spectral change, and that these are the distinctions which languages are most likely to use to distinguish their consonant inventories. These distinctions are the ones most commonly found in the world's

⁹ I do not mean that listeners, when hearing the vowel /a/ online, initially parse it as /a/ and then change their perception to /a/ as the vowel becomes longer. I believe that cases in which gating alters perceptual cues (here by altering duration), rather than simply cutting later perceptual cues out, do not represent listeners' online processing of the signal. However, gating can still be used in these cases to determine when sufficient cues become available for

languages, and a language which does not use very many distinctive features for its consonants (and therefore has a small consonant inventory) is likely to use only these distinctions. Presumably, only languages with larger numbers of distinctions would use the ones which take longer to perceive, and they would have the quickly perceived distinctions in their systems as well. These ideas are discussed further in Lang and Ohala (1996). An example is the fact that even languages with small inventories use the feature [continuant] and distinguish stops from some other manner of articulation. Continuity versus the lack of it can be perceived over a very short time window. Secondary articulations, such as distinctive palatalization, are likely to take longer to perceive, and these are usually only found in languages with large consonant inventories (Stevens and Keyser, 1989; Lang and Ohala, 1996).

Stevens discusses the perceptual saliency of various distinctions, but does not discuss how a segment's environment might influence its perceptual saliency, since he applies these ideas to entire distinctions regardless of environment.¹⁰ Considering Steriade's work on the differential perceptual saliency of the voicing distinction in various environments, it seems likely that other distinctions would also be more salient in some environments than in others. Perhaps some segments can be perceived during a very short time window in some environments, but not in others. Steriade's argument is that more contrasts are licensed where more perceptual cues are available for them, so one might find that more contrasts are licensed where cues which can be perceived over a short time window are available for them.

4.2. Results with regard to the preference for onsets

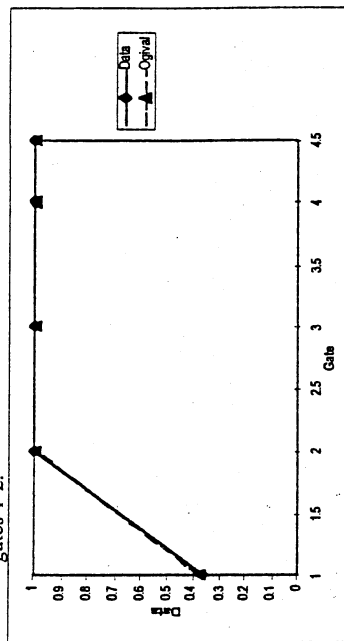
In order to address this issue, I evaluated the duration of the window over which listeners became able to perceive stops in word initial, coda, and postvocalic syllable onset position. I first calculated the percent of responses with the stop correct at each gate for all words with a stop as the target segment (the second segment of the transition of interest) or with a word initial stop as the first segment of the transition of interest. I then fit an ogival curve¹¹ to the data for each word. I located the point immediately before the curve exceeds 10% of its total rise and the point immediately after it exceeds 90% of its rise, and found the difference in number of gates between those two points (the rise time). This represents the number of gates which listeners require to go from not perceiving the stop correctly to perceiving it correctly. The figure in (4) shows the data (percent of responses with the stop correct), the fitted curve, and the rise time of the perceptual curves for two stops, one in word-initial position and one in coda position.

¹⁰ He does discuss how combinations of features (independent of environment) can enhance or detract from the saliency of a feature which is overall relatively salient, however (Stevens and Keyser 1989).

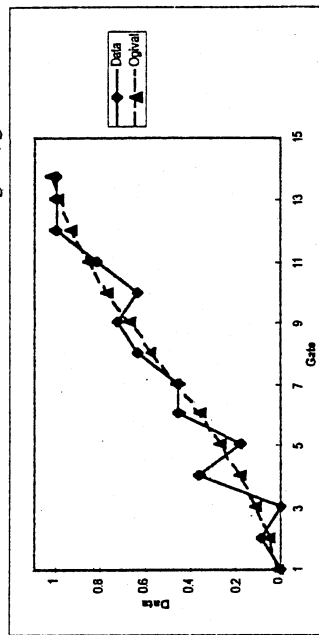
¹¹ Ogival curves are flat on the ends and quickly rising or falling in the middle. They are commonly used for investigations of categorical perception. They were used here to compensate for the noisiness of the open response data. The motivation for using these curves and the process of fitting them are described in Warner (1998).

(4) % of responses with stop correct, with fitted ogival curve overlaid, for the word-initial stop in "custom" and the coda stop in "fade."

a. Percent /k/ correct in "custom." Rise time is 1 gate, taking place between gates 1-2.



b. Percent /d/ correct in "fade." Rise time is 10 gates, gates 2-12.



I calculated the rise time in this way for all word-initial, postvocalic coda, and postvocalic onset stops. For English, the average rise time for the word initial stops was 1.62 gates, for coda stops it was 5.20 gates, and for postvocalic onset stops it was 2.43 gates. The difference in rise time among these three categories was statistically significant (using ANOVA of weighted means for unequal sample sizes, $F(2, 21)=11.75$, $p<.0005$). One can group the postvocalic onsets and the coda stops together, since these both have their VC transition gated through, and listeners in this experiment are therefore perceiving them from their VC cues. Stops perceived based on VC cues have a significantly longer rise time than the word initial stops, which are perceived based on CV cues ($F(1,22)=8.42$, $p<.01$). The large difference in rise time between the coda stops and the postvocalic onset stops is probably a result of chance factors in the word list: most of the coda stops followed diphthongs or long vowels and were in monosyllabic words ("fade, bite, oats"), so their inherently long vowels were subject to utterance final lengthening. Thus, these words had a larger number of gates over which perception of the stop could improve than the postvocalic onset stops, whose preceding vowels were of course not utterance final, and by chance were usually shorter vowels such as /t, ʌ/. Coda stops and postvocalic onset

stops are both perceived through their VC cues in this experiment, though, since the gates end before CV cues become available. The confounding factor of vowel duration does not affect the rise time of the word-initial stops, so comparing the initial stops to the combined group of coda and postvocalic onset stops is valid.

I also tested the rise time of the word initial stops against the postvocalic onset stops for Japanese. Japanese has no stops in coda position except when they are the beginning of a geminate, and the experiment did not include any transitions into geminates, so these could not be tested. For Japanese, the average rise time was 1.5 gates for word initial stops and 2.0 gates for postvocalic onsets. This difference was not statistically significant ($F(1,7) < 1$), but there were only three word-initial stops for Japanese, so there is not enough data to draw any conclusions on this subject.

When English listeners perceive a stop based on its CV cues (in this experiment, where it is word initial), the entire improvement in perception, even reaching 100% correct, often happens between the first and second gates. This is within less than 40 ms after the release of the stop. A postvocalic stop, especially one after a long vowel, often shows a very gradual improvement in perception of the stop occurring through a large portion of the vowel. Thus, it appears that stops can be perceived from their CV cues over a significantly shorter time window than when they are perceived from their VC cues. In consideration of Stevens' focus on distinctions which can be perceived over a short time window, this implies that stops are more perceptually salient when CV cues are available for them than when only VC cues are available. There are additional reasons for considering CV transitions as more important than VC transitions in perception of many distinctions, such as the presence of a burst in a stop-vowel transition. Stops which are in the onset of a syllable are likely to have these stronger and faster CV cues present, whereas stops in coda position may only have the slower VC cues present.¹²

There are well known cross-linguistic tendencies favoring consonants in onset position rather than in coda position. In many languages, onsets are maximized, so that as many intervocalic consonants as possible are included in the onset of the following syllable rather than in the coda of the preceding syllable. Furthermore, many languages have restrictions on what consonants are allowed in coda position, and make more distinctions among consonants in onset position than among consonants in coda position. In Japanese, for example, codas must consist of the mora nasal or be part of a geminate obstruent, and no place or voicing distinctions are made in coda position. Blevins (1995) documents and summarizes these and other cross-linguistic tendencies favoring onsets over codas.

The results of my experiment, showing that CV transitions contain faster cues for stops than VC cues do, provide a possible perceptual motivation for the cross-linguistic tendency to allow more distinctions in onset position, where faster cues are available, than in coda position. This is similar to Steriade's (1997) argument that languages maintain laryngeal distinctions in environments which supply more perceptual cues for them: languages allow more contrasts in onset position because consonants in onset position have faster and stronger cues available for them. Thus, the result with regard to the time

window over which stops are perceived in various environments helps to provide a perceptual motivation for patterns of syllable structure.

5. Conclusions

I have shown that the results from the gating experiment discussed here contradict Lahiri and Marslen-Wilson's (1991, 1992) claim that listeners perceive speech by comparing it to a radically underspecified underlying representation. English listeners are able to use the non-distinctive cue of vowel length to perceive the voicing of both voiceless and voiced postvocalic obstruents correctly, although Lahiri and Marslen-Wilson's conclusions about nasalization would predict that listeners would only be able to use vowel length to perceive a voiced (marked) postvocalic obstruent, not a voiceless (unmarked) one. Thus, listeners performing spoken word recognition do have access to information which is not present in a radically underspecified underlying representation.

I have also shown that cues to a stop are concentrated in a shorter time window in CV transitions than in VC transitions. Through Stevens' claim that distinctions which can be perceived over a short time window are perceptually salient, and Steriade's argument that more contrasts are allowed where more perceptually salient cues are available for them, I have argued that the quick recognition of stops from CV cues provides a perceptual motivation for the cross-linguistic tendency to allow more contrasts in syllable onset position than in coda position. This provides support for Steriade's approach of using knowledge about speech perception to improve the formal modeling of phonological patterns. Investigations of the timing of perception of distinctive features, such as this one, can help us discover perceptual motivations for patterns of syllable structure.

References

- Beckman, Mary. 1982. Segment duration and the 'mora' in Japanese. *Phonetica* 39. 113-135.
- Blevins, Juliette. 1995. The syllable in phonological theory. In *A Handbook of Phonological Theory*, John Goldsmith (ed.). Cambridge: Blackwell. 206-244.
- Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22. 129-159.
- Colton, Suzanne and François Grosjean. 1984. The gating paradigm: A comparison of successive and individual presentation formats. *Perception and Psychophysics* 35. 41-48.
- Flemming, Edward. 1995. Auditory representations in phonology. PhD dissertation, UCLA.
- Fujimura, Osamu, M.J. Macchi, and L.A. Streeter. 1978. Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech* 21. 337-346.
- Grosjean, François. 1980. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* 28. 267-283.
- Homma, Yayoi. 1981. Durational relationship between Japanese stops and vowels. *Journal of Phonetics* 9. 273-281.
- Johnson, Keith. 1997. The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics* 50. 101-113.
- Kluender, Keith R., Randy L. Diehl, and Beverly A. Wright. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16. 153-169.

¹² An onset stop which is part of an onset cluster and not the final member of the cluster, as /t/ in "train," might seem not to have the CV cues. However, such stops are usually released into a sonorant (at least in English), and are likely to have most of the same cues as would be present in a stop-vowel sequence. Stops in coda position, however, will have exclusively VC cues if they are unreleased.

- Lahiri, Aditi and William Marslen-Wilson. 1991. The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition* 38. 245-294.
- Lahiri, Aditi and William Marslen-Wilson. 1992. Lexical processing and phonological representation. In *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, Gerard J. Docherty and D. Robert Ladd (eds.), 229-254.
- Lang, Carrie and John J. Ohala. 1996. Temporal cues for vowels and universals of vowel inventories. *Proceedings of the Fourth International Conference on Spoken Language Processing, October 3-6, 1996, Philadelphia*.
- Nearey, Terrance M. 1997. Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101. 3241-3254.
- Ohala, John J. and Manjari Ohala. 1995. Speech perception and lexical representation: The role of vowel nasalization in Hindi and English. In *Phonology and Phonetic Evidence, Papers in Laboratory Phonology IV*, Bruce Connell and Amalia Arvaniti (eds.) Cambridge: Cambridge University Press. 41-60.
- Peterson, Gordon E. and Ilse Lehiste. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32. 693-703.
- Port, Robert F., Salman Al-Ani, and Shosaku Maeda. 1980. Temporal compensation and universal phonetics. *Phonetica* 37. 235-252.
- Steriade, Donca. 1997. Phonetics in phonology: The case of laryngeal neutralization. Manuscript, June 1997, UCLA.
- Stevens, K.N. 1980. Discussion. *Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, 1979*, 3. 185-186.
- Stevens, K.N. and S.E. Blumstein. 1981. The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J. Miller (eds.), *Perspectives on the Study of Speech*. Hillsdale, NJ: Erlbaum. 1-38.
- Stevens, K.N. and J. Keyser. 1989. Primary features and their enhancement in consonants. *Language* 65. 85-106.
- Smolensky, Paul. 1996. On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27.
- t'Hart, J. and A. Cohen. 1964. Gating techniques as an aid in speech analysis. *Language and Speech* 7. 22-39.
- Tyler, Lorraine K. 1984. The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics* 36. 417-427.
- Tyler, Lorraine K. and Jeanine Wessels. 1985. Is gating an on-line task? Evidence from naming latency data. *Perception and Psychophysics* 38. 217-222.
- Warner, Natasha. 1998. The role of dynamic cues in speech perception, spoken word recognition, and phonological universals. PhD dissertation, University of California, Berkeley.
- Warren, Paul and William Marslen-Wilson. 1988. Cues to lexical choice: Discriminating place and voice. *Perception and Psychophysics* 43. 21-30.

Department of Linguistics, UC Berkeley
 Dwinelle Hall, Mail Code 2650
 Berkeley, CA 94720-2650
 nwarner@uclink.berkeley.edu