1072

1072

Journal of Phonetics (1997) 25, 501–528

Lexical effects in the perception and production of American English /p/ allophones

D. H. Whalen

Haskins Laboratories, 270 Crown St., New Haven, CT 06511, U.S.A.

Catherine T. Best

Haskins Laboratories, 270 Crown St., New Haven, CT 06511, and Department of Psychology, Wesleyan University, Middletown, CT 06459, U.S.A.

and

Julia R. Irwin

Haskins Laboratories, 270 Crown St., New Haven, CT 06511, and Department of Psychology, University of Connecticut, Storrs, CT 06269, U.S.A.

Received 1st April 1996, and in revised form 26th June 1997

Although modern theories of phonology tend not to include the phoneme as a theoretical object, positional variants, traditionally known as allophones, clearly play a role in language. In English, medial voiceless stops typically occur with aspiration in stressed syllables but without aspiration in unstressed ones. In the experiments reported here, the aspirated and unaspirated allophones of /p/ were studied both for discriminability and for perceptual preference in appropriate and inappropriate stress contexts in the second syllable of disyllabic targets. In tests with non-words, aspirated [p] was consistently preferred without regard to which syllable was stressed. Additionally, the unaspirated stop was somewhat more difficult to discriminate from the other allophones in a categorial AXB task than the aspirate was. When subjects were asked to repeat and imitate both varieties, nonetheless, they were usually able to mimic the inappropriate allophones. In real words, however, subjects did generally prefer the correct allophone perceptually, and they found the inappropriate ones difficult to imitate. The allophonic contrast, then, may require contact with the lexicon in order to affect production and perception, a finding that needs to be explicitly accounted for in phonology.

1. Introduction

The distribution of conditioned variants, or allophones, is central to modern phonology. Although the term "allophone" itself derives from a time when segment-sized phonemes were assumed as the basis of phonology (Bloch, 1948), the range of phenomena covered by the term still constitutes a major focus of phonology. In the *Handbook of Phonological Theory*, for example, Goldsmith (1995) outlined five degrees of distinctiveness of

0095-4470/97/040501 + 28 $25.00/0/jp970058

segments: contrastive segments, segments with some asymmetry in use, not-yet-integrated semi-contrasts, just barely contrastive segments, and allophones in complementary distribution (p. 12). He added that "it is the case of 'allophones in complementary distribution' that is kept qualitatively distinct from the other four" (p. 12). That the term "allophone" does not appear in the remainder of the volume is an indication not of its irrelevance but instead of its having become an implicitly assumed background of the field.

Modern generative phonologies take the perceptibility of allophones for granted. In pre-generative phonology, one formulation incorporated segments into the phonetic description only if their unique properties are perceptible to the observing linguist: "Note that we are speaking of perceptible articulations only; we assume nothing here about non-perceptible articulation" (Bloch, 1948:12). If it were not possible to detect and systematically transcribe the conditioned variants of contrastive segments, then there would be no allophonic variation to describe. Yet those same theories are silent on the perceptual implications of the allophones for the speakers of the language. Again from Bloch (p. 11), we find that the "use of the term 'perceptible' refers to the observer's perception, not to the speaker's." Other than to verify that changes in allophones cannot distinctively signal a change in meaning, phonological theory has by and large been silent as to whether speakers of a language should be able to perceive the differences between allophonic variants.

Generative phonologies have replaced the phoneme with feature bundles (Chomsky & Halle, 1968) or nexes of features (Goldsmith, 1990). The conditioned differences that the phonology is required to describe are just those that make use of the combinations of distinctive features found in the world's languages. The phones that combine the various features (whether distinctive or not) have been called members of "phonetic categories" by Keating (1990), and it is these phonetic categories that are still assumed to constitute the possible allophones in modern phonology. Although generative phonology is not usually intended to make predictions for the perception of speech, the most natural expectation would be that allophones should be difficult for native speakers to isolate and identify, just because they do not carry meaningful distinctions. Since speakers of any language can become linguists, though, the further implication is that it is possible to focus attention in such a way that the differences among allophones can be perceived.

The articulatory phonology of Browman & Goldstein (1986, 1989, 1992, 1995) makes no such distinction between the perceivable and the unperceivable. Those authors have pointed out that the gradients of the actual productions of phonological forms are transcribed in a segmental, categorical way by the linguist. This can lead to cases where a single articulatory effect is described as an allophonic change in some cases and as a loss of a segment in others. They give the example of the lack of aspiration for stops in English in an unstressed syllable (such as the "p" of "vapor") and the apparent absence of an "h" in a similar environment (such as in "vehicle"); in both cases, a glottal opening gesture has been minimized in articulation of the unstressed syllable (Browman & Goldstein, 1992). Indeed, articulatory phonology has even less of a corresponding entity to the traditional phoneme than do the variants of generative phonology. Distinctions among words are made solely on the basis of presence versus absence of gestures and/or differences in the "phasing" or timing relationship among the gestures associated with a word. This theory is also uncommitted as to the perceptibility of changes in gestures or gestural phasing on the basis of their environment (i.e., the equivalent of allophones). These changes are assumed to be regular, though potentially language-specific. Even

from an articulatory phonology view, it is clear that certain traditional allophonic variations cannot be handled both automatically and universally. Final stops in Thai are always unreleased (Abramson, 1972) while those in German are almost always released (Martens & Martens, 1965), a difference that is unlikely to derive from other aspects of the language and thus must be specified. So some level of language-specific learning and specification are required for at least some cases. In the present report, though, we examined a case in English that is claimed by articulatory phonology to derive automatically from the relation of stop consonant gestures to the syllable structure (Browman & Goldstein, 1992).

The general perceptual implications of articulatory phonology have yet to be worked out, as well. Since gestures form the basis of the phonology, it could be assumed that gestures are the elements perceived in speech as well. This is explicit in direct realism (Best, 1984, 1993, 1994, 1995; Fowler, 1984, 1986; Fowler & Smith, 1986; Fowler, Best & McRoberts, 1990), whose proponents usually assume gestural phonology as the basis for the generation of the sounds that are perceived by speakers of a language. Since gestures overlap, the acoustic signal is "parsed" by listeners into acoustic consequences of the gestures that could have given rise to such a signal. That is, the "distal event" (the sound producing source) rather than the acoustics *per se* is perceived. Since the perceptual recovery of gestures is predicated to be automatic (and indeed not a process that is special to speech sounds), this theory must explicitly cope with the way in which learning a language constrains perception. Infants are sensitive to (apparently) all potentially linguistically significant differences early on, but by their first birthday lose their sensitivity to many contrasts that are not supported by their native language (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, 1993, 1994, 1995). Adult listeners also have great difficulty with foreign contrasts, with many of them being virtually unlearnable late in life (see Flege & Fletcher, 1992 for a review). It is clear that the higher levels of language affect its perception, even if the exact mechanism is less clear (Fowler, 1990).

Although phonologies do not make explicit predictions about the perceptibility of allophones, our own expectation was that an allophone in an inappropriate contexts would be noticeably worse than that same allophone in its appropriate context. That prediction is explored here in the perception of the aspirate and inaspirate allophones of the English voiceless stop /p/. Word initially, the aspirate occurs, while after /s/, the inaspirate occurs. In medial position before stressed vowels, these stops are aspirated whereas medially before unstressed vowels, they are unaspirated (Lisker & Abramson, 1967; Kahn, 1976). (Word final position includes unreleased allophones, which need not be considered here.) In generative phonology, the combinations of features needed for this variation are provided by adding the [± spread glottis] feature to what is otherwise distinguished by [± voice] (Keating, 1990). In articulatory phonology, the difference in aspiration due to differences in stress may be an automatic consequence of the way that stress is implemented. If all the gestures associated with the stressed syllable are greater in magnitude than those in unstressed syllables, then the degree of aspiration should increase as a result of an increase in magnitude of the glottal devoicing (opening) gesture (Cooper, 1991; Browman & Goldstein, 1992). As such, in this theory, there is no more reason to posit a different mechanism for the allophonic difference in aspiration than for the multitude of other contextual articulatory adjustments that fall below the level at which we can detect them. Articulatory phonology makes an implicit prediction about what will happen if the range of automatic variation is exceeded. Listeners should perceive a difference if the variation is sufficient to constitute a lexical difference, and they

should be relatively insensitive to the difference if it is not enough to signal a meaningful distinction in the language.

Logically speaking, allophones must be treated by the speakers of a language in two ways at once. In phonological terms, different allophones realize the same phoneme, and thus they should count as a single perceptual object; at the same time, the allophones must be distinct in production. The production evidence is clear enough, since linguists are able to consistently identify and describe various allophones, but the perceptual status of allophonic differences for naive speakers is not as obvious. However, a pair of experiments by Jaeger (1980) found evidence that speakers of English do treat various allophones of [k] as instances of a single category. For those subjects who showed conditioning of the galvanic skin response (GSR) for electric shocks paired with the aspirated allophone, the GSR generalized to the inaspirate allophone after [s] and to the unreleased final [k] allophone. Jaeger's category formation task (further elaborated in Jaeger, 1986) yielded similar results, although it was subject to some orthographic interference as well. This is the most direct evidence we have that allophones truly belong to a single perceptual category. But it does not tell us whether the allophones are perceptually distinguishable as well, much as we would easily be able to discriminate a green Granny Smith from a red Macintosh even if a GSR conditioned to "apple" responded to both.

The question is an open one, since not every systematic difference in production is perceptually discriminable to those who produce it. Many instances of "near-merger" have been found in recent years (Labov, Yaeger & Steiner, 1972; Labov, 1975; Milroy & Harris, 1980; Nunberg, 1980; Janson & Schulman, 1983; Harris, 1985; Di Paolo, 1988, 1992; Di Paolo & Faber, 1990; Labov, Karen & Miller, 1991). In such situations, two phones (which are often phonologically distinct in other dialects of the language) are pronounced in a very similar way and result in patterns that look as if they are no longer contrastive. They are so similar, in fact, that the speakers themselves cannot easily tell them apart. Speakers of other dialects, however, often do a better job of discriminating the two. Rather than representing a complete merger, in which all distinctiveness is lost, this situation exhibits a near-merger, in which other factors may force the two sounds apart in production even though they are perceptually merged for the speakers. This situation is similar to that of allophones in that there are phonetic variations that are consistently produced and yet apparently do not constitute a phonemic contrast (for the speakers of that dialect).

The perception of allophones is also, in some ways, equivalent to the perception of unfamiliar contrasts from another language. Such a contrast, e.g., the dental versus retroflex stop distinction of Hindi, may be difficult for speakers of another language, such as English, to discriminate because they assimilate to a single native phonological category (in this case, alveolar stop). Since an allophonic difference cannot be one that will distinguish words, there may be a similar tendency to hear allophonic variants as instances of the same phonetic category. A formal model of the way in which listeners perceive non-native contrasts is the Perceptual Assimilation Model, or PAM (Best, McRoberts & Sithole, 1988; Best, 1993, 1994). Briefly, this theory posits that some perceptually assimilated non-native phonetic contrasts will be harder to discriminate than others. The most difficult of all is posited to be the Single Category (SC) contrast, in which listeners assimilate two non-native sounds to a single native category without perceiving any difference in their goodness as members of that native category. If one non-native sound is a better representative of that category than is the other non-native

sound, then the non-native contrast constitutes a more discriminable Category Goodness (CG) difference. The treatment of a non-native contrast as SC by speakers of one language does not entail that speakers of a different language will have difficulty with it. The perceived relationship of the contrasting non-native sounds to the native phonological system is the determining factor in ease or difficulty of discrimination.

This model can be extended to allophones of one language, as we will show in the present experiments. If speakers treat context-conditioned allophones as truly being equally good members of a phonological category (in either appropriate or inappropriate contexts), then we would expect the allophones to elicit the poor discriminability exhibited by SC non-native contrasts. If, however, the inappropriate context makes the allophone a poorer variant of the phoneme, then perception of contextually appropriate versus inappropriate allophones should resemble the CG contrasts, with much better discrimination, despite the use of a single category in classification.

In a previous study with English /l/ allophones, we found a pattern resembling SC categorization (Whalen, Best & Irwin, in press). Subjects had no preference for dark or light /l/ on the basis of its appropriateness to the context (initial and final syllable position), and they were scarcely able to discriminate them. However, the realization of the final /l/ allophone in natural conversational speech in American English includes large changes in vowel quality, which we had held constant in our stimuli in order to ensure that the discrimination was based on the /l/ and not on the vowel. Thus, our initial /l/s provided the better comparison, since the light /l/s were closer to their normal production. However, it may be that the phonetic difference is too small with this pair of allophones. English light /l/ in initial position is noticeably darker than the light /l/ of other languages, and the dark /l/ not as dark as in some languages. Therefore, even though there is a consistent productive (and acoustic) difference in English (and one that our subjects showed in their productions as well), it may not have been a large enough difference to be discriminable, especially without the concomitant natural variations in vowel color. This aspect of degree of phonetic difference may be crucial to the ability to treat allophones differentially. The present series of experiments examined a more clear-cut case, the allophones of voiceless stops, whose phonetic realization is less spread across the syllable and whose allophonic variants are widely used distinctively in the world's languages.

Another aspect of language that sometimes affects speech perception, and therefore may affect perception of allophonic variants, is lexicality. Perceptual differences between words and non-words have been found in several paradigms (Samuel, 1986; Pitt & Samuel, 1993). Words are more easily recognized in noise (Miller, Heise & Lichten, 1951), though only at certain noise levels. Phoneme detection is faster for words than for non-words (Rubin, Turvey & van Gelder, 1976), though it seems likely that there is an effect of orthography in such cases. Non-words behave differently than words in the phonemic restoration paradigm, even when subjects are asked only whether the speech has been replaced by noise or instead had the noise added to it (Samuel, 1987). And phonetic boundaries tend to shift toward the form that would make a word (Ganong, 1980). Elman & McClelland (1988) report a coarticulatory shift in a phoneme boundary based on a segment that could only have been generated lexically, since that influencing segment's acoustic realization was replaced with noise. While none of these experiments is individually conclusive, together they suggest that the lexicon may play a direct role in phonetic perception. While many other experiments find that lexical status does not affect perceptual results, it can be expected that there will be different results for words

and non-words in some circumstances. The perception of allophones may be one of those cases. We examined this possibility in the final two experiments reported in this paper; the first three experiments focused on allophonic variants in non-words.

To summarize, previous work has shown that listeners perceive contextually appropriate allophones as instances of a single phonemic category. The predictions for inappropriate allophones, however, are not as clear. Since speakers keep allophones separate in production, we might expect that inappropriate allophones would sound awkward or even sound like a different phoneme altogether. However, since the two allophones can never make a semantic distinction, they might still be considered exemplars of the same category and therefore be difficult to distinguish perceptually. We explored this issue with the perception of native, context-conditioned allophones, both by determining whether the allophones could be discriminated when both of them occurred (unnaturally) in the same context and by assessing whether the correct allophone would receive a higher preference rating than the incorrect one. If producing the correct allophone is part of being a native speaker, then native speakers should perceptually prefer the correct allophone to the incorrect one. However, the allophonic difference must be available to consciousness to make such rating differences. So if the allophones truly sound "the same" to the speakers of that dialect (even in the wrong context), then it may be impossible for them to rate the allophones according to the difference in their own productions. We assessed the degree to which the allophones are truly perceptually identical by comparing the results of our allophone discrimination tasks to the discrimination performance levels that have been reported for non-native distinctions of differing degrees of similarity to native contrasts. The ability to imitate these inappropriate native allophones was also examined.

2. Experiment 1

For the first experiment, we wanted to see if appropriate allophones were preferred perceptually to inappropriate ones in nonsense syllables. Phonological processes should apply to any form, whether or not it is a word, and using non-words avoids issues like word frequency that can influence the results. The allophones chosen were the aspirate and inaspirate allophones of medial /p/. Although the phonetic distinctiveness of /p/ allophones is more striking in initial position, it is not possible to compare the allophones there. An inaspirate stop in initial position will sound like a voiced stop, and so it would rate as a very poor or unacceptable exemplar of a voiceless stop. Even after /s/, there is the possibility of contrast, at least for casual speech. The unaspirated [/spɪt/] should be perceived as *spit*, but the aspirated [/spʰɪt/] could be a reduced form of *It's Pitt*. In medial position, there is less likelihood of a stop being misperceived as an unintended voicing category.

The voiced, voiceless unaspirated and voiceless aspirated labial stops were used. Each occurred medially between two vowels, one stressed and one unstressed. Medial voiceless stops before stressed vowels in American English are aspirated whereas those before unstressed vowels are unaspirated (Lisker & Abramson, 1967; Kahn, 1976). Thus, the aspirated version should be preferred before stressed vowels, and the unaspirated version before unstressed vowels, if listeners are sensitive to the phonetic difference as well as to the conditioning imposed by its context, that is, if the allophonic difference behaves perceptually like a category goodness (CG) difference in the PAM model (Best *et al.*,

1988). If the two allophones are members of one perceptual category no matter what the context, that is, if they behave like a single category (SC) assimilation, it may not be possible to distinguish them.

2.1. *Method*

2.1.1. *Subjects*

The subjects were 13 undergraduate students at the University of Connecticut. All were young adults who were native speakers of American English, and they received course credit for their participation. None reported any hearing problems. Three were male and 10 were female. One often incorporated his answers into a drawing, making it difficult to determine the response; his data were excluded from the analysis.

2.1.2. *Stimuli*

The vowels [oʷ] and [ɚ] were used, because they change their quality very little under different stress conditions; thus the perception of the medial stop should be virtually uninfluenced by perception of vowel color variation. A phonetician with over 40 years of experience produced 20 tokens of [óʷbɚ], [oʷbɚ́], [óʷpɚ], [oʷpɚ́], [óʷpʰɚ], and [oʷpʰɚ́]. From these, six tokens of each type were selected by the experimenters to be as closely matched in intonation, duration, and loudness as possible. The mean VOTs (in this case, defined as length of voicelessness after release) for the stops in these six utterance types were 7.9, 12.7, 10.8, 15.3, 42.9, and 54.8 ms, respectively. VOTs were measured from the acoustic waveform, beginning with the first evidence of the release burst and ending with the first voiced pitch period of the vocalic segment following the stop. The closures were fully voiced for [b], so the positive VOT value represents the duration of the release burst, which can be compared with the unaspirated stop. Closures were voiceless for more than half of their duration for the other two stops. Many voiceless stops have some residual voicing after the attainment of closure, which does not affect their perception as voiceless.

2.1.3. *Procedure*

Preference was measured with a rating task, but it was also necessary to include an identification task, since the unaspirated [p] in the inappropriate position ([opɚ]) might be heard not as a bad [p] but rather as a bad [b]. Therefore, the subjects were instructed to write both whether they heard "b" or "p" and give it a rating from 1 (a bad example) to 7 (a good example). Initial stress and final stress disyllables were presented in separate tests, because pretesting had indicated that subjects were confused by stress differences within a single experiment, and gave essentially random results when the two stress conditions were combined. There were 100 trials, the first 10 of which were warm-ups that were not tabulated. The remaining 90 trials consisted of five repetitions of each of the six tokens of each of the three consonant types. Half of the subjects began with the initial stress and half with the final stress stimuli.

For the discrimination task, a categorial AXB paradigm was used. In this task, the "A" and the "B" stimuli are always taken from two different categories (though with the same stress pattern). The "X" stimulus belongs to one or the other of those categories, but is a different token than the A or B target it matches categorially. There were two

categories that could be the non-matching one; for example, the aspirate could be paired with either the voiced stop or the inaspirate. The non-matching stimulus was chosen equally often from the two possible categories. There were 190 trials in each block, the first 10 of which were unscored warm-ups. The remaining 180 trials consisted of 10 repetitions of each of the six tokens of each of the three types. The X matched the A or B equally often.

Initial and final stress stimuli were presented in separate blocks. The order of the blocks for half of the subjects, then, was: rating, initial stress; rating, final stress; discrimination, initial stress; and discrimination, final stress. For the other half of the subjects, the order of the stress conditions was reversed.

2.2. *Results*

2.2.1. *Ratings*

The mean ratings for [b], [p], and [p^h] are presented in Table I and Fig. 1 (solid lines). These means do not include the 15 cases out of the 360 presented (4.2%) for final stress and 14 cases (3.9%) for initial stress in which the inaspirate was identified as "b" rather than "p." There were also four cases (0.6%) for final stress and one case (0.1%) for initial stress in which other allophones were misidentified, and these were excluded from the calculation of the means as well. When the stress was on the final syllable, the aspirate stop is the expected allophone, and it received the highest rating. The aspirate stops were also preferred, however, when the stress was on the initial syllable, where the inaspirate is the expected allophone at onset of the unstressed, second syllable. An analysis of variance based on the ratings for the six categories, with the within factors Stress (initial or final) and Allophone (b p p^h), showed that the allophones did differ in their mean ratings ($F(2,22) = 18.88$, $p < 0.001$). Although the initial stress elicited somewhat higher ratings than the final stress (5.50 *vs.* 5.06), this difference was only marginally significant ($F(1,11) = 4.06$, $p < 0.10$). There was a tendency for the inaspirate to receive a higher rating in the initial stress context than in the final stress context, as seen in the marginal interaction of the two factors ($F(2,22) = 2.60$, $p < 0.10$). If we compare the two ratings that we expected to differ, we find that the context-appropriate aspirate (in stressed syllables) is rated 1.69 units higher than the context-inappropriate inaspirate, but the appropriate inaspirate (in unstressed syllables) is rated 0.94 units *lower* than the inappropriate aspirate. The mean of these predicted differences is not significantly different

TABLE I. Ratings for the voiced, inaspirate, and aspirate stimuli, Experiment 1. Ratings were from 1 ("very poor example") to 7 ("very good example"). Standard deviations follow (in parentheses). Only the responses corresponding to the intended category are included. Inaspirate [p] is the expected allophone with initial stress, aspirate [p^h] with final stress. Cells that were expected to be rated highest are in bold face

	[b]	[p]	[p^h]
Final stress	**5.38** (1.2)	4.05 (1.7)	**5.74** (1.2)
Initial stress	**5.76** (0.9)	**4.91** (1.3)	5.84 (1.0)

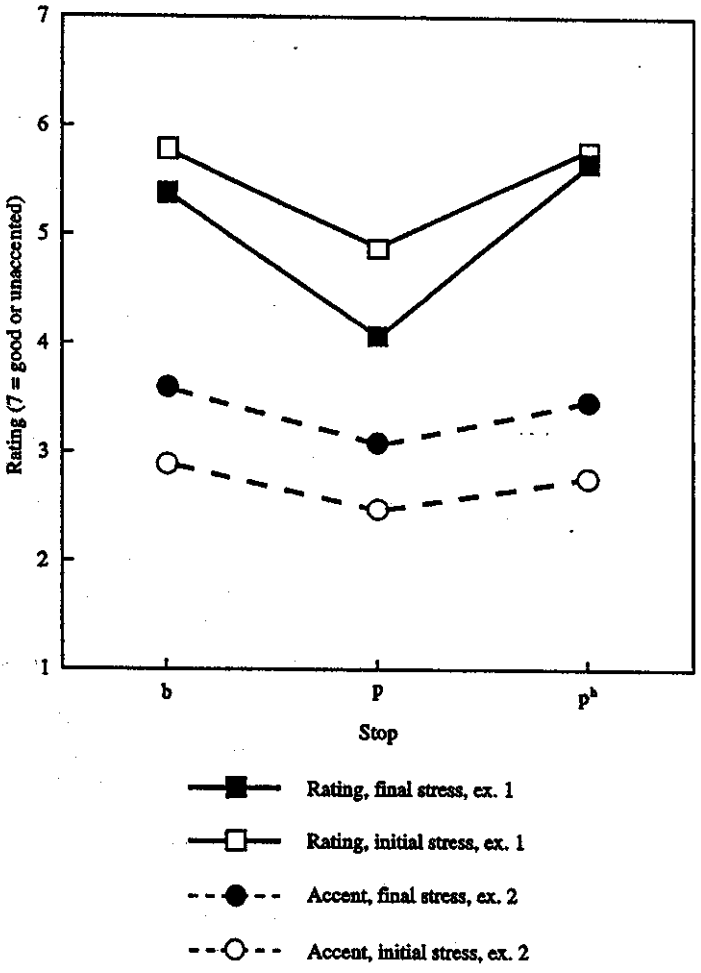


Figure 1. Rating results from Experiment 1 (——). Values range from 1 (a "bad" example) to 7 (a "good" example). Rating results for Experiment 2 (– – –). The values range from 1 (an "accented" example) to 7 (an "unaccented" example).

from 0 ($F(1,11) = 3.19$, n.s.) indicating that the predictions for context-conditioned category goodness differences in the perception of the /p/ allophones are not supported. The two predicted values themselves, being strongly positive in the one case and negative in the other, were significantly different ($F(1,11) = 21.05$, $p < 0.001$), however, indicating a category goodness difference does exist but is not determined by stress context, at least for these non-word stimuli.

2.2.2. *Discrimination*

The three phonetic categories were discriminable, though the inaspirates were less discriminable (i.e., less often matched correctly in AXB) than the other two (see Table II).

TABLE II. Discrimination for the voiced, inaspirate, and aspirate stimuli, Experiment 1. Discrimination is in percent correct, with chance at 50%. Standard deviations follow (in parentheses). Inaspirate [p] is the expected allophone with initial stress, aspirate [p^h] with final stress

	[b]/[p^h]	[b]/[p]	[p]/[p^h]
Final stress	97.2 (3.1)	93.5 (6.2)	69.7 (8.9)
Initial stress	97.8 (2.5)	96.1 (3.1)	75.8 (10.8)

An analysis of variance with the factors stress (initial and final) and allophone pair (b/p, b/p^h, p/p^h) was performed. The two stress conditions elicited different levels of performance, with the final (86.8%) being slightly but significantly more difficult than the initial (89.9%) ($F(1,11) = 13.90$, $p < 0.01$). The three allophone pairs also differed, with the aspirate/inaspirate pair being clearly more difficult than either pair with /b/ ($F(2,22) = 69.38$, $p < 0.001$). There was no interaction of stress and pair type ($F(2,22) = 2.53$, n.s.). A separate ANOVA examined the responses to the first two pairs. Even though each involved a shift in category (from /b/ to /p/), there was a somewhat lower level of performance with the pair involving the inaspirate (94.8%) than the one with the aspirate (97.5%), a significant difference ($F(1,11) = 12.30$, $p < 0.01$). (Stress did not affect overall performance in this subanalysis ($F(1,11) = 2.13$, n.s.), nor was there an interaction ($F(1,11) = 1.07$, n.s.).) The last pair type consisted of two varieties of /p/. For half of those comparisons, the "X" token was the inaspirate, while for the other half, the "X" token was the aspirate. Subjects performed better when the aspirate occurred as the "X" token of the AXB trial (77.2% correct) than when the inaspirate was the "X" (68.3%). (These numbers are collapsed in the table.) This difference in identity of the "X" token was significant in a separate ANOVA ($F(1,11) = 8.05$, $p < 0.05$). In this analysis, the difference due to stress was significant ($F(1,11) = 9.51$, $p < 0.05$), while the interaction of "X" token and stress was not ($F(1,11) < 1$, n.s.).

2.3. *Discussion*

Subjects preferred aspirated stops in both stress contexts despite the fact that in American English, stops are produced with little or no aspiration at the onset of an unstressed syllable. There was a trend toward higher ratings for the inaspirate in its appropriate (post-stress) position than in its inappropriate (pre-stress) position, but its rating was always lower than that of the aspirate. In the rating task, then, these allophones are perceptually distinct (i.e., show a perceived difference in category goodness for /p/), but perceptual preference does not behave according to contextual expectations.

The discrimination results indicate that these two sounds are members of the same category, in our extension of the PAM assimilation type SC (see Introduction). The 72.8% accuracy level is within the range found empirically for non-native contrasts that are perceptually assimilated to a single category in the native language (Best, 1990, 1995). This level was attained even though, in every comparison, one allophone or the other was inappropriate and thus might be assumed to be, by definition, a poor member of the category. This should have led to a category goodness (CG) difference, which, in

non-native perception studies, typically yields discrimination values in the range of 80–90% correct. In discrimination, then, these allophones truly appear to be members of a single native phonetic category.

This pattern of results, while seemingly contradictory, has a parallel in another field, that of taste. MacRae & Geelhoed (1992) found that subjects could consistently rate one water sample more highly than another even when they could not reliably pick it out in a discrimination task (in their case, an oddity judgment, which indicates which of three samples is unlike the other two). While concerns about which is the best measure of discrimination are important, they should not obscure the fact that the comparison of the present levels of allophonic discrimination with those found for non-native SC contrasts are based on the same kinds of discrimination tests and thus should be telling us about the same level of processing.

Aspirated stops occupy the extreme on the scale of voicing distinctions used by languages of the world. Certainly if a speaker of English wishes to emphasize the voicelessness of a stop, adding (inappropriate) aspiration is an effective way to do that. This fact about emphatic speech may account for the lack of an effect of contextual appropriateness in the ratings obtained in Experiment 1. Since subjects seemed to be willing to allow for an exaggerated articulation in the present experiment, an attempt was made in the second experiment to shift their attention to the critical facts of the contextual appropriateness of the allophones. This was accomplished by addressing one salient feature of allophones, namely, that mastering them is part of being a native speaker.

3. Experiment 2

In the first experiment, subjects were asked to rate the goodness of a speech cue. In this second experiment, subjects were instead asked to rate the tokens for degree of foreign accent. It is possible that the subjects perceived the aspirated allophone as being the most salient and/or frequent allophone, and so gave it high goodness ratings whether it was appropriate to the context or not. Since command of allophonic production is part of native competence, violations of the allophonic rules should constitute evidence of a foreign accent. Such accents have been found to be detectable from speech segments as short as a stop release burst (Flege, 1984). Thus, a task focusing on the accentedness of the speech may allow the contextual appropriateness of the allophone to dominate the ratings.

3.1. *Method*

3.1.1. *Subjects*

The subjects were 12 undergraduate students at the University of Connecticut. All were native speakers of American English and were given course credit for participating. None reported any hearing problems. Six were male and six were female. None had participated in Experiment 1.

3.1.2. *Stimuli*

The stimuli of Experiment 1 were used.

3.1.3. *Procedure*

The same tests (rating and discrimination) as in Experiment 1 were used, except that the rating task did not include identification of the stop (since there had been so few misidentifications in Experiment 1) and the ratings were described as ranging from 1 (a foreign accented example) to 7 (a native English example).

3.2. *Results*

The ratings are shown in Table III. An ANOVA, with the factors Allophone ([b], [p] or [p^h]) and Stress (initial or final) found that the ratings did not differ significantly by Allophone ($F(2, 22) = 2.23$, n.s.). Stress affected the ratings only marginally ($F(1, 11) = 3.34$, $p < 0.10$). The interaction was not significant ($F(2, 22) < 1$, n.s.). If we compare the two ratings that we expected to differ, we find that the context-appropriate aspirate (in stressed syllables) is rated 0.38 units higher ("less accented") than the context-inappropriate inaspirate, but the appropriate inaspirate (in unstressed syllables) is rated 0.38 units *lower* than the inappropriate aspirate. If we analyze the differences between the appropriate and inappropriate contexts, we find that the mean of the two is not significantly different from 0 ($F(1, 11) < 1$, n.s.) indicating that the prediction of context-conditioned category goodness differences in perceptual preference was again not supported. The two difference values themselves were significantly different ($F(1, 11) = 6.14$, $p < 0.05$), confirming that the aspirate received a higher rating overall than the inaspirate.

These phones were also discriminable, though there was some difficulty with the inaspirate/aspirate pairing (see Table IV). An analysis of variance with the factors stress (initial and final) and allophone pair (b/p, b/p^h, p/p^h) was performed. The stress conditions were statistically indistinguishable ($F(1, 11) = 1.72$, n.s.), but the three allophone

TABLE III. Ratings of the voiced, inaspirate, and aspirate stimuli, Experiment 2. Ratings were from 1 ("very accented example") to 7 ("very unaccented example"). Standard deviations follow (in parentheses). Inaspirate [p] is the expected allophone with initial stress, aspirate [p^h] with final stress. Cells that were expected to be rated highest are in bold face

	[b]	[p]	[p^h]
Final stress	**3.56** (1.5)	3.08 (1.2)	**3.46** (1.5)
Initial stress	**2.92** (1.3)	**2.45** (0.6)	2.83 (1.1)

TABLE IV. Discrimination for the voiced, inaspirate, and aspirate stimuli, Experiment 2. Standard deviations follow (in parentheses). Discrimination is in percent correct, with chance at 50%. Inaspirate [p] is the expected allophone with initial stress, aspirate [p^h] with final stress

	[b]/[p^h]	[b]/[p]	[p]/[p^h]
Final stress	84.9 (15.6)	83.5 (13.2)	64.6 (10.6)
Initial stress	87.9 (16.1)	88.2 (14.9)	64.5 (10.7)

pairs elicited different levels of performance $F(2, 22) = 65.23$, $p < 0.001$). The interaction was not significant ($F(2, 22) = 1.96$, n.s.). This replicates the results of Experiment 1 for the responses to the three allophone pairs, although the level of performance was an average of 9.4% lower in Experiment 2 (see Fig. 1, dashed lines).

A separate ANOVA examined the responses to the first two pairs, those that involved a difference in category assignment. The factors were allophone pair (b/p or b/p^h) and stress (initial or final). The difference in discrimination performance for the two pairs was not significant in this experiment ($F(1, 11) < 1$, n.s.) as compared with the significant difference found in Experiment 1; there was less than a percentage point of difference in Experiment 2. Stress ($F(1, 11) = 3.29$, n.s.) and the interaction with allophone pair ($F(1, 11) < 1$, n.s.) were also nonsignificant.

A separate analysis that divided the p/p^h pair into the cases with the aspirate as the "X" token versus ones with the inaspirate as the "X" token was also conducted. Both allophones were discriminated above chance ($F(1, 11) = 27.12$, $p < 0.001$), and equally well for the two stress conditions ($F(1, 11) < 1$, n.s.). As in Experiment 1, pairs with the inaspirate as the "X" were more difficult to discriminate (68.3— correct for the inaspirate as the "X" token *vs.* 77.2% correct for the aspirate; $F(1, 11) = 8.05$, $p < 0.05$).

3.3. *Discussion*

Shifting the task to a rating of foreign accentedness still failed to yield a context-conditioned pattern of ratings for the inappropriate allophones in these utterances. Subjects rated all the allophones as equally accented. Several differences between these results and those of Experiment 1 can be seen in Fig. 1 (comparing solid and dashed lines). In Experiment 1, the ratings were almost as high as one can expect, since subjects rarely use the two extremes of a rating scale. It can be inferred that they perceived the stops as appropriately produced. When asked to rate the foreign accent of those same utterances, however, another group of subjects was quite critical, assigning ratings that indicate a slight foreign accent for all the stimuli. It is clear that the subjects in the two experiments were doing something different. It is not quite clear what that something was. The accent ratings (Experiment 2) were equally low for all three stops, while the goodness judgments were lower for the inaspirate than the other two regardless of stress context (Experiment 1).

Despite the difference in the meaning assigned to the numbers in the two experiments, it is of interest to see whether the pattern was similar. An ANOVA was performed which compared the ratings in the two experiments (that is, the solid versus dashed lines in Fig. 1). The one between factor was Group [goodness (Experiment 1) *vs.* accentedness (Experiment 2)], and the two within factors were Allophone ([b], [p], and [p^h]) and Stress (initial and final). The two groups, as could be expected from Fig. 1, differed in their overall level of rating ($F(1, 22) = 29.75$, $p < 0.001$). Stress was not significant as a main effect ($F(1, 22) < 1$, n.s.), but the interaction with group was significant ($F(1, 22) = 6.89$, $p < 0.05$). The subjects in Experiment 1 rated the initial stress tokens more highly, while the Experiment 2 subjects rated the final stress tokens more highly (i.e., more native-like). Allophone was a significant main effect ($F(2, 44) = 15.94$, $p < 0.001$), with aspirates rated more highly than inaspirates. The interaction with Group was significant as well ($F(2, 44) = 4.03$, $p < 0.05$), supporting the presence of an effect of Allophone in Experiment 1 but not in Experiment 2. The other two interactions were not significant.

These results indicate that it is possible for listeners to detect, at some level, the difference between the voiceless aspirate and inaspirate stops. However, they do not treat this difference in a context-sensitive manner, as had been expected according to a Category Goodness difference assimilation pattern. Even when determining degree of foreign accent, where a lack of mastery of contextual allophonic variation might be a salient indicator, subjects did not take context into account. In these stimuli, it seems that the contextual allophones truly represent a single category in a manner such that the contextual determination of allophony is unavailable to consciousness. Nonetheless, the two allophones were discriminated in the AXB paradigm at levels well above chance (yet still well below the two-category, b/p pa irs). This indicates that in medial position, the phonetic distinction is only partially available either for detection of differences, or for judgments of quality, by the listener.

The change in the rating task also seems to have affected discrimination, even though the discrimination tasks were identical in the two experiments. The subjects in Experiment 1 had accuracy rates over 9% higher than those of Experiment 2. Apparently, focusing on foreign accent and/or the absence of a phonetic judgment on the stop made the subjects less sensitive to the phonetically important aspects of the signal. This can be seen especially in that the easiest discrimination ([b]/[p^h]) was also impaired under the "foreign accent" instructions. The pattern of difficulties was different as well (see Fig. 2, solid versus dashed lines). The differences in discrimination performance were analyzed by means of an ANOVA with the between factor of Group (Experiment 1 *vs.* Experiment 2) and the within factors of Stress (initial or final) and Allophone Pair (b/p, b/p^h, p/p^h). The overall performance of the Groups did differ, as they appear to in Fig. 2 ($F(1,22) = 6.50$, $p < 0.05$). The stimuli with initial Stress were discriminated better ($F(1,22) = 7.13$, $p < 0.05$). The Stress factor did not interact with Group ($F(1,22) < 1$, n.s.). The Allophone Pairs differed in the accuracy they elicited ($F(2,44) = 134.39$, $p < 0.001$), without interacting with Group ($F(2,44) < 1$, n.s.). Stress and Allophone Pair did not interact ($F(2,44) < 1$, n.s.), but the three-way interaction with Group was significant ($F(2,44) = 3.91$, $p < 0.05$). The subjects in Experiment 1 had somewhat more difficulty with the final stress stimuli for the inaspirate/aspirate pairs, while the subjects in Experiment 2 had somewhat more difficulty with final stress for the other two pairs. For the most difficult pairs, those within the /p/ category, the subjects in Experiment 1 were also more accurate than those in Experiment 2 (in a separate ANOVA, $F(1,22) = 4.54$, $p < 0.05$).

The better discrimination of the aspirate/inaspirate pair for the "goodness" rating subjects (Experiment 1) when they judged the initial stress stimuli as compared to the final stress suggests an effect of allophonic appropriateness. This is the condition in which the inaspirate is appropriate, and it is the one in which the inaspirates received their highest rating (see Fig. 1, solid lines). If that superiority translates into a more memorable stimulus, then it would be expected that discrimination would improve as well, which can be seen in Fig. 2. Thus, subjects may not be completely insensitive to allophonic appropriateness. However, it is still a weak effect, in that the inappropriate aspirated allophone received a higher rating than the appropriate inaspirate allophone. The inaspirate/aspirate discrimination was also easier when the aspirate was the "X" of the AXB triad (see Fig. 3). So this minor indication of some sensitivity to allophones must be weighed against the other indications of relatively low sensitivity. In particular, the 64.6% accuracy level is even more strongly consistent with a single category (SC) pattern of perception than the 72.8% of Experiment 1.

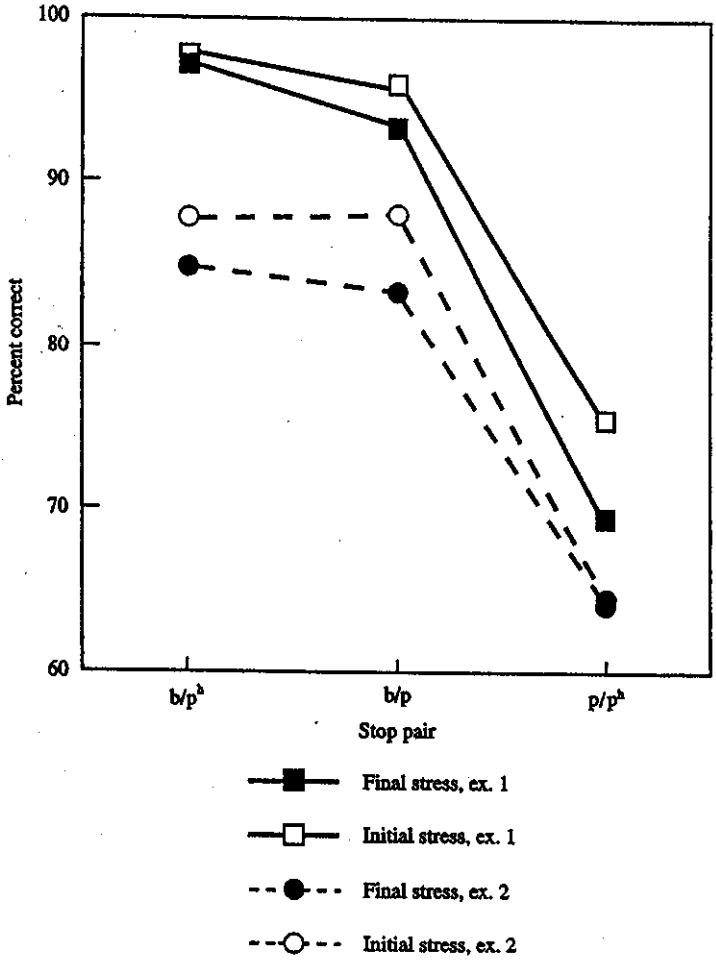


Figure 2. Comparison of the discrimination performance from Experiments 1 and 2. Tasks and instructions were identical in the two experiments.

4. Experiment 3

At this point, it became clear that we needed to know how well the subjects could imitate these different VOT values. If the allophones are well-established in speakers' production patterns, then repetition of the tokens should be influenced both by the model speech and by the allophony. Therefore, we re-called as many of the subjects in Experiment 2 as were available to collect their productions of the target items. In one condition, subjects were simply told to repeat each item presented to them auditorily. In the second condition, they were instead asked to imitate each target item as well as possible. The latter instruction was expected to produce a different VOT pattern than the first set of instructions, with the imitation expected to show less influence of the typical American English allophonic pattern and a greater adherence to the VOT of the model.

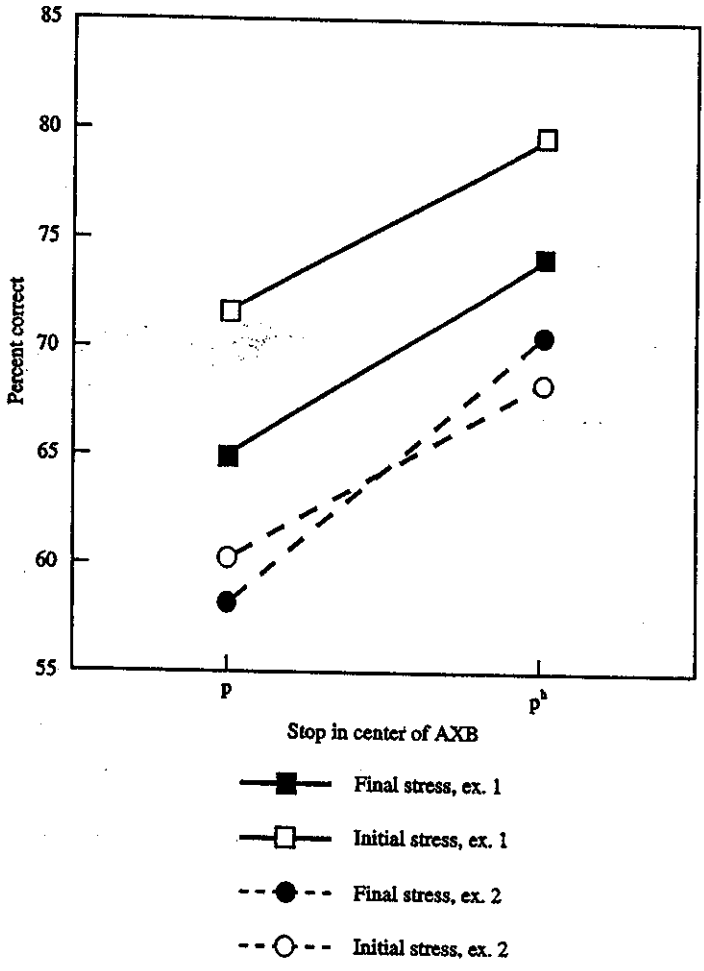


Figure 3. Comparison of the discrimination performance from Experiments 1 and 2 based on which sound occurred in the "X" position. Tasks and instructions were identical in the two experiments.

4.1. *Method*

4.1.1. *Subjects*

Six of the 12 subjects of Experiment 2 were available for this follow-up study. Four were females and two were males.

4.1.2. *Stimuli*

The stimuli of Experiments 1 and 2 were used.

4.1.3. *Procedure*

In the first condition (Repetition), subjects were told that they would hear one utterance from the set on each trial, and that they were to repeat each item into the tape recorder.

No further instructions were given as to the nature of the repetition. In the second condition (Imitation), they were explicitly asked to imitate each item as closely as they could. We maintained this constant order for these conditions on the assumption that some subjects would spontaneously adopt an imitation strategy in the Repetition condition while others might just use the voiceless stop category of their normal speech (and thus show appropriate allophones only). After the Imitation instructions, it seemed as though even repetitions would be likely to be imitations as well. Each of the six tokens of each stimulus type was presented twice, with all tokens in one randomized list. Thus, subjects repeated (or imitated) each category 12 times in each of the stress conditions.

4.2. *Results*

Mean values for the 12 responses to each category for each subject are given in Table V. (Individual values are given in the Appendix.) There was no systematic VOT difference between the Repetition and Imitation conditions. The three types of stop were produced with different VOT values, with the inaspirate being between that of the voiced and the aspirate, in both conditions. The inaspirate's mean value was not as small as that of the model (30 ms *vs.* 13 ms in the model), but the aspirates were also somewhat longer than the model (53 ms *vs.* 49 ms). Subjects were expected to be more successful with the inaspirates in the imitation than the repetition condition. Instead, individual speakers failed to show any single, consistent response pattern, instead showing one of three patterns of responses: (1) consistently short (unaspirated) VOTs even for the aspirate models, i.e., replacement of aspirates with inaspirates; (2) consistently long VOTs, i.e., replacement of inaspirates with aspirates; and (3) good imitation. There were two subjects of each pattern in each condition (though two subjects switched from one pattern to another). Thus, the production results are ambiguous. There was no overwhelming tendency to use the correct English allophones even in the repetition condition, but on the other hand, neither was there great accuracy (for four of the six subjects) in imitating the inaspirate VOTs.

It is possible that the level of accuracy in the subjects' discrimination would be related to their ability to imitate. Although the number of subjects available to test this relation is small, we correlated the two measures of accuracy. The discrimination measure was the average accuracy for the p/p^h pairs collapsed over stress condition. The imitation accuracy was the *r* value obtained by correlating the attained VOT (in both imitation and repetition) with the model VOT. The correlation between these two values was 0.39, which, for six subjects, is not significant. It does, however, suggest that a larger study might find a significant relationship between allophonic perception and production.

TABLE V. Mean VOT values (in ms) for repetitions and imitations of the three phones, Experiment 3. Standard deviations are given (in parentheses)

	[b]		[p]		[p^h]	
	Repetition	Imitation	Repetition	Imitation	Repetition	Imitation
Initial stress	13.2 (8.7)	11.7 (5.7)	32.7 (26.0)	32.9 (24.6)	50.9 (26.9)	49.5 (32.8)
Final stress	11.7 (5.8)	9.8 (4.6)	32.6 (24.3)	24.5 (19.6)	56.9 (23.8)	55.1 (28.2)

4.3. *Discussion*

Although the ratings of Experiments 1 and 2 indicate that subjects are not highly sensitive to the context-conditioned differences in these /p/ allophones, the production results here indicate that they perceive the VOT differences to at least some degree, and that they are modestly able to reproduce those VOTs accurately. The accuracy of the repetitions and imitations is somewhat at odds with the discrimination results, which indicate that these sounds are fairly difficult to tell apart, but they are more in agreement with the rating results, which suggest the two allophones are perceptually distinct to some degree.

It is relevant to note that second language learners (L2 learners) often shift their VOTs toward the norm for the language, but end up with a value intermediate between their native (L1) values and the values used by native speakers of the L2 (Caramazza, Yeni-Komshian, Zurif & Carbone, 1973; Williams, 1980; Flege & Hillenbrand, 1984). Here, we found that the unusual VOT values were sometimes imitated quite well, but there were also some subjects who showed a tendency to produce something in between the normal VOT for that position and the incorrect one presented.

5. Experiment 4

In the first and second experiments, subjects rated the /p/ allophones in nonsense utterances. It may be the case that the allophonic difference does not apply strongly to non-words; that is, it might be a lexical characteristic. Several paradigms have found that words and non-words have different effects on perception (Miller *et al.*, 1951; Rubin *et al.*, 1976; Ganong, 1980; Samuel, 1986; Pitt & Samuel, 1993). If non-words are typically produced and perceived as stylized or hyperarticulated, then the preference for the aspirated stops in both allophonic contexts may be correct in terms of the (non-)lexical status of the items presented. In this experiment, subjects rated a new set of items, real English words this time, for their degree of foreign accent, depending on the contextual appropriateness of aspirate *vs.* inaspirate /p/ allophones.

5.1. *Method*

5.1.1. *Subjects*

The subjects were 12 undergraduate students at the University of Connecticut. All were native speakers of American English and were given course credit for participating. None reported any hearing problems. Five were male and seven were female. None had participated in Experiment 1, 2, or 3.

5.1.2. *Stimuli*

Six real words, approximating the syllables of the previous experiments as closely as possible, were selected for this experiment: *coper*, *doper*, *Hopi*, *soapy*, *opaque*, and *opine*. The electronic version of the *Shorter Oxford English Dictionary* (Coltheart, 1981) was searched for words in which the vowel of the first syllable was /o/ and the medial consonant, /p/. This left very few lexical items to choose from, and none where a final *-er* would not be a morpheme. The words *open* and *opal* were rejected because of the

multiple ways that stops can be released into syllabic nasals and liquids. Several more (*groper*, *sloper*, and *toper*) were excluded since they did not add anything beyond the two words ending in *-er* that we had already selected. This left no instances of an initial-stress monomorphemic word, so *Hopi* was added to the list. In all likelihood, there should be no difference in effects between one and two morpheme targets. It is true that individual phonemes that constitute a morpheme have been found to be longer in duration than homophonous nonmorphemic phonemes, as in *rapped* and *rapt* (Losiewicz, 1992). But the difference in that study was only 3.6%, which is unlikely to be perceptible. Indeed, Bond (1973) found that listeners are unable to detect which of two homonyms is intended when one is bimorphemic and one is monomorphemic (e.g., *wade* and *weighed*.) Still, having instances of each will at least indicate whether one or the other category has a complete absence of the allophonic effect. For final stress, *opaque* and *opine* were the only words meeting those criteria in the database. The only words among our six targets to occur in the Kucera & Francis (1967) frequency count were *soapy* (2 occurrences) and *opaque* (6 occurrences). We judged that having words of similar phonetic composition was more important than having higher frequency words. If there is a difference between these items and the nonwords of the earlier experiments, we can expect that the differences would only be accentuated if words of higher frequency were used.

The first four test words have stress on the first syllable and thus should be produced with the unaspirated allophone, while the last two, having stress on the second syllable, should be produced with the aspirated allophone. In order to have both allophones occurring with every word, the speaker from Experiments 1 and 2 produced fifteen repetitions of the two variations of each word. Each word was presented to the speaker in its standard spelling, and the desired allophone was indicated by appending "-asp" or "-unasp" to the end. From these, six tokens of each version of each word were selected, matched as far as possible for pitch contour, amplitude, and duration. (The mean VOT values across the six tokens can be found in Table VII, the results for the next experiment, in the "Model" column).

5.1.3. *Procedure*

Each word was presented in a separate sequence, with both allophonic versions (aspirated and unaspirated) randomized together. There were five repetitions of each token, six tokens of each word for a total of 60 items for each word. With the six words, there were 360 items presented altogether. The order of the word lists differed for each subject. Subjects were instructed to listen for a foreign accent and the ratings were described as ranging from 1 (a foreign accented example) to 7 (a native English example), as in Experiment 2.

5.2. *Results*

In this experiment, unlike the previous three, the appropriate allophone was preferred over the inappropriate one in both stress contexts (Table VI), with the item *coper* being an exception. An analysis of variance was performed with the within factors Word (six levels) and Appropriateness of Aspiration (two levels). Appropriateness rather than aspiration *per se* was analyzed, so that any interactions with word would indicate a difference in the magnitude of the preference, not the difference in stress pattern that is obvious from the means. Word was a significant factor ($F(5, 55) = 2.91, p < 0.05$),

TABLE VI. Ratings of the inaspirate and aspirate real word stimuli, Experiment 4. Standard deviations follow (in parentheses). Ratings were from 1 ("very accented example") to 7 ("very unaccented example"). Inaspirate [p] is the expected allophone with initial stress, aspirate [p^h] with final stress. Cells that were expected to be rated highest are in bold face

	Initial stress		Final stress	
Word	[p]	[p^h]	[p]	[p^h]
coper	**4.63** (1.3)	4.69 (1.3)		
doper	**4.98** (0.8)	4.65 (0.9)		
Hopi	**5.36** (0.7)	5.03 (1.3)		
soapy	**5.45** (1.2)	4.79 (1.2)		
opaque			3.79 (1.3)	**4.80** (1.4)
opine			3.16 (0.8)	**4.93** (1.1)
mean	5.11	4.79	3.48	4.87

indicating that some of the words were rated, overall, higher than others. ("*Hopi*" and "*soapy*" tended to receive the highest ratings.) Appropriateness was also significant ($F(1, 11) = 24.45$, $p < 0.001$) with the appropriate versions having significantly higher ("less accented") mean ratings. The interaction of the two was also significant ($F(5, 55) = 8.20$, $p < 0.001$), indicating that the difference was not consistent across the words. In general, there seemed to be a larger preference difference for the iambs (*opaque, opine*) than for the trochees.

5.3. *Discussion*

In the context of real words, the appropriate allophones of /p/ were preferred over the inappropriate ones in perception. This contrasts with what was found for non-words, namely, either a lack of preference with the accentedness rating (Experiment 2) or the consistent preference for aspirates in both contexts with the goodness rating (Experiment 1). The acoustic difference was no larger in the present real-word stimuli than in the previous non-word stimuli (VOTs of 15.5 *vs.* 13.1, respectively, for the unaspirated, 45.9 *vs.* 48.9 for the aspirated), and yet the perceptual pattern was quite different. This result may have a parallel in Flege & Fletcher's (1992) finding that listeners found sentences to be more accented after they became familiar with them. Once the context was familiar, they argue, the easier it was to focus on those phonetic features that made a production accented. In the same way, the words used in this experiment are intrinsically more familiar than the non-words used in the previous experiments, which may have allowed the subjects to focus in on the phonetic differences of the /p/ allophones more easily. The non-words were deemed more accented than the words overall (3.05 for the non-words, 4.69 for the words, with a lower rating indicating greater accentedness), which makes it appear that the less familiar items are, overall, perceived as more accented, which is the opposite of the explanation for the allophone effect. However, the unfamiliarity of the non-words may be of a different kind than the unfamiliarity of meaningful sentences. A "foreign" word (one way of looking at non-words) may intrinsically sound foreign, and the degree to which its allophones are correct may not influence that judgment. (It is also

the case that the non-words were presented over 100 times in the course of the experiment, so they became familiar in a way, but not as lexical items, since there was no meaning attached to them.) These issues need further treatment for a definitive answer, but they are consistent with our interpretation, that the real words were familiar enough to allow the subjects to focus on the part of them that was indicative of a foreign accent, namely, the inappropriate allophones.

Even within this preference, however, it is certainly the case that the inappropriate allophone does not receive a "bad" rating. If we consider any value below the midpoint (4, in this case) to be a bad rating, then only the inappropriate unaspirated allophones in the final stress words were actually bad; the aspirated allophones in the initial stress words were simply not as good as their appropriate counterparts, a finding quite consistent with Experiments 1 and 2. This indicates that the tendency for the aspirates to be rated highly is very strong.

6. Experiment 5

Since the subjects of Experiment 3 were fairly accurate in their imitations of the VOTs of the inappropriate allophones, we needed to know how well the subjects could imitate the VOTs in words. If lexical status affects perception of allophones by making the context-conditioned phonetic distinctions easier to detect, it should, by rights, make imitations more difficult. That is, it should be more difficult to use atypical VOT values for real words than it was for non-words.

6.1. *Method*

6.1.1. *Subjects*

All subjects of Experiment 4 participated.

6.1.2. *Stimuli*

The stimuli of Experiment 4 were used.

6.1.3. *Procedure*

As in Experiment 3, each item was presented individually for repetition (in the first condition) or imitation (in the second condition). All words were randomized together, with two repetitions of each of the six tokens.

6.2. *Results and discussion*

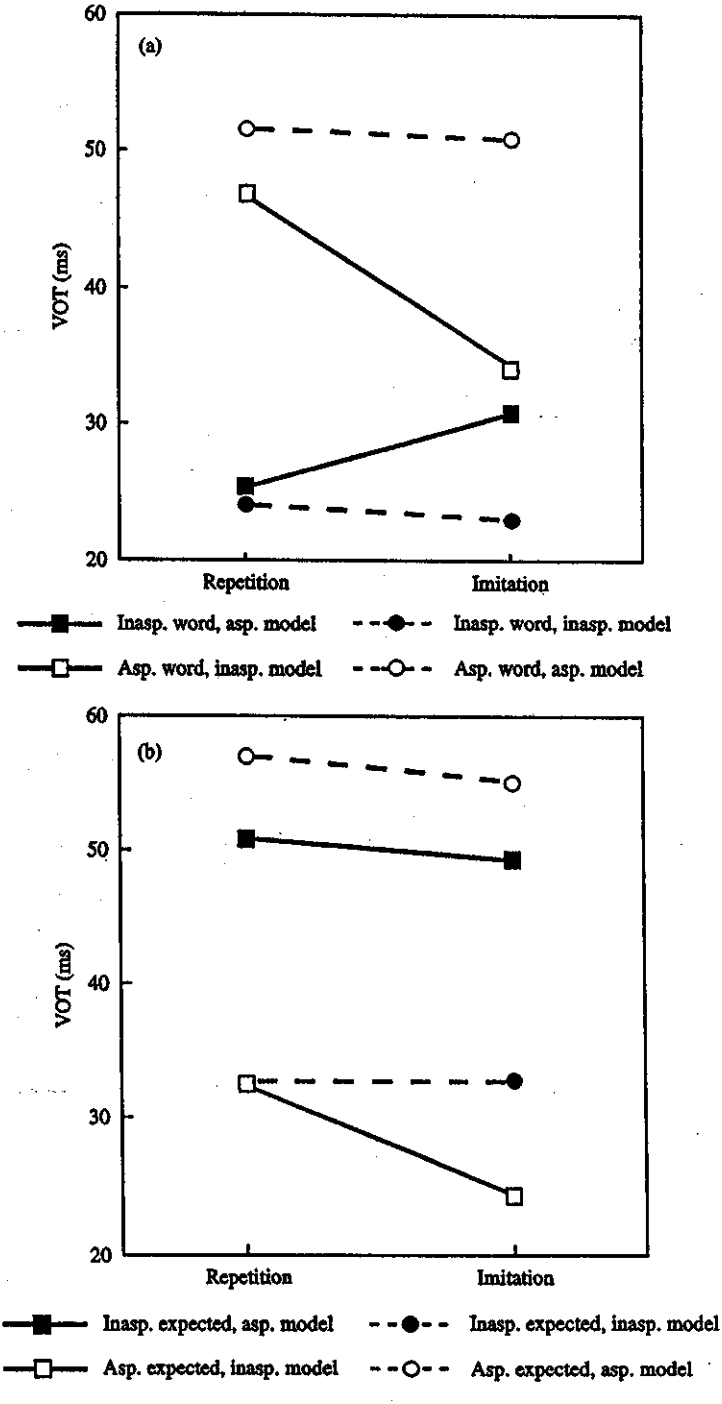
In the repetition condition, both the aspirated and unaspirated versions of the words elicited essentially the same VOT value from subjects (see Table VII). That is, as far as we can tell, subjects used the correct allophone of the word in question. (The VOTs were longer than those of the model, but the model may have had shorter VOTs than the subjects preferred.) In the imitation condition, subjects were somewhat more successful at producing the presented version of the incorrect allophone rather than the lexically correct one (see the top panel of Fig. 4). However, they were still closer to the lexically determined allophone than they were to the allophone presented. As for the appropriate

TABLE VII. VOT values for repetitions and imitations of the inaspirate and aspirate real word stimuli, Experiment 4. Standard deviations follow (in parentheses). The expected allophone is in bold face

		Model	Repetition	Imitation
coper	-aspirated	44.0 (9.7)	24.5 (5.6)	36.3 (13.5)
	-unaspirated	12.7 (3.0)	28.8 (14.7)	25.4 (10.9)
doper	-aspirated	42.4 (6.9)	29.4 (11.5)	36.5 (11.6)
	-unaspirated	18.5 (3.6)	27.5 (11.4)	27.1 (10.8)
Hopi	-aspirated	54.0 (5.7)	22.0 (8.0)	21.1 (8.4)
	-unaspirated	18.0 (3.0)	18.8 (6.2)	20.2 (8.4)
soapy	-aspirated	51.1 (11.8)	26.3 (12.7)	30.7 (13.0)
	-unaspirated	13.2 (5.9)	21.6 (8.6)	20.3 (6.8)
opaque	**-aspirated**	40.3 (15.3)	49.7 (15.0)	46.6 (11.5)
	-unaspirated	16.2 (2.9)	43.0 (10.6)	30.9 (14.8)
opine	**-aspirated**	43.8 (12.2)	53.7 (13.6)	55.6 (16.9)
	-unaspirated	14.4 (3.9)	51.2 (14.9)	37.8 (17.0)

allophones, repetitions and imitations were, on average, about 10 ms longer in VOT than the models, but otherwise very accurate. By comparison, the repetitions and imitations of Experiment 3 (Fig. 4, bottom panel) are reasonably accurate for both appropriate *and* inappropriate allophones. The results with real words (Experiment 5) extend those found in imitations of a VOT continuum (Flege & Eefting, 1988), in which speakers of English, Spanish, and of both tended to use values associated with their language's categories when imitating arbitrary VOT values. In the present case, the values that were used were typical for English, but not always in the correct position. Thus, the imitators should have had a better chance at successful imitation than Flege and Eefting's. But even in the imitation condition, where the subjects' average VOT came closer to the model, the subjects were not alternating between long and short VOTs; there was a single distribution of values which tended to be skewed toward the higher values (for the aspirate models of inaspirate words) or the lower values (for the inaspirate models for aspirate words). Thus, these subjects seemed to be using a somewhat different strategy in mimicking these VOTs than were Flege and Eefting's subjects or than the subjects of Experiment 3, involving non-words.

Figure 4. (a) VOT values for imitations of appropriate and inappropriate allophones in real words, Experiment 5. The four functions are for the inaspirate words with aspirate models (mean VOT for the models was 47.9 ms); the inaspirate words with inaspirate models (mean VOT for the models was 15.6 ms); the aspirate words with inaspirate models (mean VOT for the models was 15.3 ms); and, the aspirate words with aspirate models (mean VOT for the models was 42.1 ms). (b) VOT values for imitations of appropriate and inappropriate allophones in nonsense items, Experiment 3. The four functions are for the initial stress non-words (where inaspirate is expected) with aspirate models (mean VOT for the models was 54.8 ms); the initial stress non-words (where inaspirate is expected) with inaspirate models (mean VOT for the models was 15.3 ms); the final stress non-words (where aspirate is expected) with inaspirate models (mean VOT for the models was 10.8 ms); and the final stress non-words (where inaspirate is expected) with aspirate models (mean VOT for the models was 42.8 ms).


(a)
60
50
40
30
20
VOT (ms)
Repetition
Imitation
Inasp. word, asp. model
Inasp. word, inasp. model
Asp. word, inasp. model
Asp. word, asp. model
(b)
60
50
40
30
20
VOT (ms)
Repetition
Imitation
Inasp. expected, asp. model
Inasp. expected, inasp. model
Asp. expected, inasp. model
Asp. expected, asp. model

The difference between the words and the non-words can be further seen in a comparison of how well the subject VOTs correlated with the models for both Experiment 3 and 5. For the voiceless stops in Experiment 3, subjects had a respectable correlation of 0.35 between their productions and the model presented to them. This was about the same in magnitude in the repetition condition (0.33) and the imitation condition (0.36); the match between model and production was basically unaffected by condition. In Experiment 5, by contrast, the overall correlation of production and model was, though significant, very small in magnitude (0.09). In the repetition condition, it actually fell to 0.00, and improved to 0.17 in the imitation condition. Thus, although there was a significant correlation (due to a larger n) between model and production for the imitations, that relationship accounted for much less of the variance here ($r^2 = 0.03$) than in the imitation condition of Experiment 3 ($r^2 = 0.13$). In short, subjects were fairly successful at imitating allophonically incorrect VOTs in non-words but fairly poor at imitating them in words.

7. General discussion

In five experiments, subjects' perception and production of aspirated and unaspirated [p] were assessed. Since only one or the other allophone is typically used in each of the environments studied, it was expected that the rating task would elicit a preference for the appropriate allophone over the inappropriate one. In discrimination, however, the fact that the two allophones realize a single category was expected to lead to poor performance (if there is no perceived difference in category goodness) or moderate performance (if one allophone is a better exemplar than the other). The pattern in production was expected to follow that in perception, with poor imitation of inappropriate allophones. The prediction for the rating task was borne out when the stimuli were English words (Experiment 4), but when the stimuli were non-words, the aspirate allophone received a higher rating in both appropriate and inappropriate contexts (Experiments 1 and 2). While this indicates that the allophones were somewhat discriminable, they were not relatable to the appropriateness of the context. In the discrimination task, the inaspirate [p] allophones elicited a very low level of accuracy when they were the target "X" item in the AXB discrimination trials. In fact, their discrimination level is similar to performance on the most difficult non-native phonetic distinctions, those which perceptually assimilate to equally good tokens of a single category (SC) in the native language. Although the above-chance discrimination indicates that subjects were able to detect some phonetic difference between inaspirates and the other two categories, the level of performance indicates that the two sounds truly were realizations of one category to the listeners and did not result in context-conditioned differences in category goodness (CG). On the production side, only a few subjects were able to imitate the different VOTs of the non-words fairly well. When the same allophonic distinction was presented in real words, though, subjects produced the contextually appropriate allophone (not the one presented) and were thus less able to imitate the VOTs. This lexical effect indicates that subjects were able to focus on the phonetic details more easily when the items were non-words, at least when the task was a production one that could, in principle, avoid any categorization. By contrast, they only perceptually preferred the correct allophone with those same word stimuli, indicating that they could better detect the difference in the two categories with words than with non-words.

Two possible accounts may be offered for the pattern of results found with the perception of the non-words. On the one hand, it may be that non-words require the aspirated allophone because they are not established as consistent phonetic strings. The allophone in which the constituent parts (here, place and voicelessness) are signaled most unambiguously might be called for, in that case. For stops in English, the presence of aspiration is a sure sign of voicelessness, while the absence of aspiration occurs with both voiced and voiceless stops. Whether such "hyperarticulation" would occur under more natural conditions might be testable by presenting subjects with an unfamiliar technical vocabulary that contains relevant word shapes, to see whether they tend to aspirate all voiceless stops regardless of stress context. On the other hand, it may be that the allophonic difference should normally be present even for non-words, but the unfamiliarity of the non-words makes it difficult to focus on that aspect of their realization. The only way of alleviating that unfamiliarity would be to present the non-word multiple times, but even eight repetitions in a meaningful context have been found sufficient to make a non-word act like a word in shifting the identification curve on a VOT continuum (Remez, Rubin, Katz & Dodelson, 1985). The present stimuli were presented multiple times in a nonsense context, which seems not to have been successful at making them familiar in the necessary way. Thus, it is difficult to determine from our experiments just how new phonological strings may be entered into the lexicon.

Both gestural phonology and generative phonologies currently provide phonetic interpretations for words given by the lexicon and syntax. Since the lexicon, by definition, does not contain non-words, there is some reason to think that the difference in behavior between the words and non-words makes this organization of phonology the correct one. The continual addition of lexical items throughout the speaker's life, however, indicates that there must be some mechanism for handling the phonological structure of previously unlexicalized strings, so that the new words end up with the correct phonological interpretation. If new words are treated as non-words initially (as phonological descrip tions currently require), then there might be a period in which their allophonic realization is atypical. For new words that fit the pattern of the stimuli from the present experiments, we might expect that the /p/ in a new word like "zóper" would have a longer VOT the first few times it is pronounced than it would after it became a more settled member of the lexicon. This could be tested by, say, introducing *ad hoc* technical terms and requiring subjects to use the terms in conversation.

Although it is difficult to compare different ratings and different experiments, it was the case that the overall rating was much lower ("more accented") when subjects were judging nonwords for degree of foreign accent (Experiment 2) than for phonetic acceptability ("less good") (Experiment 1). The reliability of such a difference emerges in the discrimination results for the two experiments, even though the stimuli and procedures for evaluating discrimination *per se* were identical. It is as if it is only possible to rate accent on words, and having made those ratings, the subjects had turned these nonwords into some kind of word and were less able to listen to the phonetic aspects that would allow them to discriminate the items. If in rating accent subjects had to envision the forms as being real English words rather than meaningless phonetic forms, then there should be some indication that the correctness of the allophone was available in some way. And, indeed, there is a minor indication of this in the discrimination data for Experiment 2, in which the appropriate allophone was somewhat easier to discriminate than the inappropriate one (Fig. 3, bottom two functions). In Experiment 4, of course, the

real words gave clearer, direct evidence of the expected difference in accent ratings for the different allophones.

Modern phonology has dispensed with the phoneme as a theoretical construct, yet the classification of phones into appropriate and inappropriate realizations of certain phonological distinctions remains at the core of phonological descriptions. The current results allow us to assess the degree of perceptual equivalence of allophones of a single segment. The discrimination results indicate that the two allophones can be distinguished, but only at a much lower level than with native and non-native contrasts that correspond to two separate phonemes (i.e., phonologically distinctive). The non-word rating results also show that the allophones are discriminable, but not in a way that is systematically related to the context. If listeners did not treat the allophones as members of one category, then presumably they would have perceived the phones as members of separable categories, and so their discrimination performance would have been much better. Instead, it was difficult for the subjects even to notice the inappropriateness of the wrong allophones in non-words, despite the presence of just that pattern in the subjects' own production of real words. Thus, allophones do seem to represent consistently different realizations of a single underlying phonological category. Moreover, consciously varying the allophonic difference in production is a difficult task when the targets are lexical items. Taken together, the results indicate that there is a complex relationship between phonological rules and the existence of an entry in the lexicon, a relationship that needs to be addressed explicitly in phonology and explored in phonetics.

This research was supported by NIH grants DC-00403, DC-02717 and HD-01994 to Haskins Laboratories. Additional help with the stimuli was provided by Michele Sancier, Larry Brancazio, and Sonya Sheffert. We thank Carol A. Fowler, Michael Studdert-Kennedy, Arthur Abramson, Alice Faber, Patrice Speeter Beddor, and two anonymous reviewers for helpful comments.

References

Abramson, A. S. (1972) Word-final stops in Thai. In *Tai phonetics and phonology* (J. G. Harris & R. B. Noss, editors), pp. 1–7. Bangkok: Central Institute of English Language

Best, C. T. (1984) Discovering messages in the medium: speech and the prelinguistic infant. In *Advances in pediatric psychology* (H. E. Fitzgerald, B. Lester, & M. Yogman, editors), pp. 97–145. New York: Plenum Press

Best, C. T. (1990) Adult perception of nonnative contrasts differing in assimilation to native phonological categories, *Journal of the Acoustical Society of America*, 88, S177–S178.

Best, C. T. (1993) Emergence of language-specific constraints in perception of non-native speech: a window on early phonological development. In *Developmental neurocognition: speech and face processing in the first year of life* (B. de Boysson-Bardies, editor), pp. 289–304. Dordrecht: Kluwer Academic

Best, C. T. (1994) Emergence of native-language phonological influences in infants: a perceptual assimilation model. In *The development of speech perception: the transition from speech sounds to spoken words* (J. C. Goodman & H. C. Nusbaum, editors), pp. 167–224. Cambridge, MA: MIT Press

Best, C. T. (1995) A direct realist perspective on cross-language speech perception. In *Cross-language speech perception* (W. Strange & J. J. Jenkins, editors), pp. 171–204. Timonium, MD: York Press

Best, C. T., McRoberts, G. W. & Sithole, N. M. (1988) Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants, *Journal of Experimental Psychology: Human Perception and Performance*, 14, 45–60

Bloch, B. (1948) A set of postulates for phonemic analysis, *Language*, 24, 3–46.

Bond, Z. S. (1973) The perception of sub-phonemic phonetic differences, *Language and Speech*, 16, 351–355

Browman, C. P. & Goldstein, L. (1986) Towards an articulatory phonology, *Phonology Yearbook*, 3, 219–252.

Browman, C. P. & Goldstein, L. (1989) Articulatory gestures as phonological units, *Phonology*, 6, 201–251

Browman, C. P. & Goldstein, L. (1992) Articulatory phonology: an overview, *Phonetica*, 49(3/4), 155–180

Browman, C. P. & Goldstein, L. (1995) Gestural syllable position effects in American English. In *Producing speech: contemporary issues. For Katherine Safford Harris* (F. Bell-Berti & L. Raphael, editors), pp. 19–34. Woodbury, NY: AIP Press

Caramazza, A., Yeni-Komshian, G., Zurif, E. & Carbone, E. (1973) The acquisition of a new phonological constrast: the case of stop consonants in French-English bilinguals, *Journal of the Acoustical Society of America*, **54**, 421–428

Chomsky, N., & Halle, M. (1968) *The sound pattern of English*. New York: Harper & Row

Coltheart, M. (1981) The MRC psycholinguistic database, *Quarterly Journal of Experimental Psychology*, **33A**, 497–505

Cooper, A. (1991) *An articulatory account of aspiration in English*. Unpublished Ph.D. dissertation, Yale University

Di Paolo, M. (1988) Pronunciation and categorization in sound change. In *Linguistic change and contact: proceedings of the sixteenth annual conference on New Ways of Analyzing Variation in Language* (K. Ferrara, B. Brown, K. Walters, & J. Baugh, editors), pp. 84–92. Austin, TX: Texas Linguistic Forum

Di Paolo, M. (1992) Hypercorrection in response to the apparent merger of (ɑ) and (ɔ) in Utah English, *Language and Communication*, **12**, 267–292

Di Paolo, M. & Faber, A. (1990) Phonation differences and the phonetic content of the tense-lax contrast in Utah English, *Language Variation and Change*, **2**, 155–204

Elman, J. L. & McClelland, J. L. (1988) Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes, *Journal of Memory and Language*, **27**, 143–165

Flege, J. E. (1984) The detection of French accent by American listeners, *Journal of the Acoustical Society of America*, **76**, 692–707

Flege, J. E. & Hillenbrand, J. (1984) Limits on phonetic accuracy in foreign language speech production, *Journal of the Acoustical Society of America*, **76**, 708–721

Flege, J. E. & Eefting, W. (1988) Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation, *Journal of the Acoustical Society of America*, **83**, 729–740

Flege, J. E. & Fletcher, J. (1992) Listener and talker effects on the perception of degree of foreign accent, *Journal of the Acoustical Society of America*, **91**, 370–389

Fowler, C. A. (1984) Segmentation of coarticulated speech in perception, *Perception and Psychophysics*, **36**, 359–368

Fowler, C. A. (1986) An event approach to the study of speech perception from a direct-realist perspective, *Journal of Phonetics*, **14**, 3–28

Fowler, C. A. (1990) Calling a mirage a mirage: direct perception of speech produced without a tongue, *Journal of Phonetics*, **18**, 529–541

Fowler, C. A. & Smith, M. (1986) Speech perception as "vector analysis": an approach to the problems of segmentation and invariance. In *Invariance and variability in speech processes* (J. Perkell & D. Klatt, editors), pp. 123–136. Hillsdale, NJ: Lawrence Erlbaum Associates

Fowler, C. A., Best, C. T. & McRoberts, G. W. (1990) Young infants' perception of liquid coarticulatory influences on following stop consonants, *Perception and Psychophysics*, **48**, 559–570

Ganong, W. F. (1980) Phonetic categorization in auditory word perception, *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 110–125

Goldsmith, J. (1995) Phonological theory. In *The handbook of phonological theory* (J. Goldsmith, editor), pp. 1–23. Oxford: Blackwell

Goldsmith, J. A. (1990) *Autosegmental and metrical phonology*. Oxford: Basil Blackwell

Harris, J. (1985) *Phonological variation and change: studies in Hiberno English*. Cambridge: Cambridge University Press

Jaeger, J. J. (1980) Testing the psychological reality of phonemes, *Language and Speech*, **23**, 233–253

Jaeger, J. J. (1986) Concept formulation as a tool for linguistic research. In *Experimental phonology* (J. J. Ohala & J. Jaeger, editors), pp. 211–237. New York: Academic Press

Janson, T. & Schulman, R. (1983) Non-distinctive features and their uses, *Journal of Linguistics*, **19**, 321–336

Kahn, D. (1976) *Syllable-based generalizations in English phonology*. Bloomington, IN: Indiana University Linguistics Club

Keating, P. A. (1990) Phonetic representations in a generative grammar, *Journal of Phonetics*, **18**, 321–334

Kucera, H. & Francis, W. N. (1967) *Computational analysis of present-day American English*. Providence, RI: Brown University Press

Labov, W. (1975) On the use of the present to explain the past. In *Proceedings of the eleventh international congress of linguists*. Bologna: Societa Editrice Il Mulino, pp. 825–851

Labov, W., Yaeger, M. & Steiner, R. C. (1972) *A quantitative study of sound change in progress* (Report No. on NSF contract GS-33287). US Regional Survey

Labov, W., Karen, M. & Miller, C. (1991) Near-mergers and the suspension of phonemic contrast, *Language Variation and Change*, **3**, 33–74

Lisker, L. & Abramson, A. (1967) Some effects of context on voice onset time in English stops, *Language and Speech*, **10**, 1–28

Losiewicz, B. L. (1992) *The effect of frequency on linguistic morphology*. Unpublished Ph.D. dissertation, University of Texas, Austin

MacRae, A. W. & Geelhoed, E. N. (1992) Preference can be more powerful than detection of oddity as a test of discriminability, *Perception and Psychophysics*, **51**, 179–181

Martens, C. & Martens, P. (1965) *Phonetik der deutschen Sprache*. München: Max Hueber Verlag

Miller, G. A., Heise, G. A. & Lichten, W. (1951) The intelligibility of speech as a function of the context of test materials, *Journal of Experimental Psychology*, **41**, 329–335

Milroy, J. & Harris, J. (1980) When is a merger not a merger? The MEAT/MATE problem in a present-day English vernacular, *English World-Wide*, **1**, 199–210

Nunberg, G. (1980) A falsely reported merger in eighteenth century English: a study in diachronic variation. In *Locating language in time and space* (W. Labov, editor), pp. 221–250. New York: Academic Press

Pitt, M. A. & Samuel, A. G. (1993) An empirical and meta-analytic evaluation of the phoneme identification task, *Journal of Experimental Psychology: Human Perception and Performance*, **19**, 699–725

Remez, R., Rubin, P. E., Katz, M. & Dodelson, S. (1985) On the influence of lexical status in phonetic perception. In *Twenty-sixth annual meeting of the Psychonomic Society*, pp. 29–30. Boston, MA: Psychonomic Society

Rubin, P. E., Turvey, M. T. & van Gelder, P. (1976) Initial phonemes are detected faster in spoken words than in spoken nonwords, *Perception and Psychophysics*, **19**, 394–398

Samuel, A. G. (1986) The role of the lexicon in speech perception. In *Pattern recognition by humans and machines: speech perception, volume 1* (E. C. Schwab & H. C. Nusbaum, editors), pp. 89–111, New York: Academic

Samuel, A. G. (1987) Lexical uniqueness effects on phonemic restoration, *Journal of Memory and Language*, **26**, 36–56

Werker, J. F. & Tees, R. C. (1984) Cross-language speech perception: evidence for perceptual reorganization during the first year of life, *Infant Behavior and Development*, **7**, 49–63

Werker, J. F. & Lalonde, C. E. (1988) Cross-language speech perception: initial capabilities and developmental change, *Developmental Psychology*, **24**, 672–683

Whalen, D. H., Best, C. T. & Irwin, J. (in press) Perception of American English /l/ allophones, *Haskins Laboratories Status Report on Speech Research*

Williams, L. (1980) Phonetic variation as a function of second-language learning. In *Child phonology, vol. 2: Perception* (G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson, editors), pp. 185–215. New York: Academic Press

Appendix

VOT values (in ms) for repetitions and imitations of the three phones, Experiment 3, for the individual subjects. Standard deviations are given (in parentheses).

Subject	[b]		[p]		[p^h]	
	Repetition	Imitation	Repetition	Imitation	Repetition	Imitation
S1 Initial stress	10.9 (5.5)	10.2 (1.5)	17.0 (6.5)	18.0 (3.4)	19.4 (10.6)	25.3 (18.6)
Final stress	11.1 (4.2)	10.4 (4.6)	12.3 (3.6)	12.1 (3.2)	18.2 (7.5)	26.8 (23.5)
S2 Initial stress	8.4 (1.6)	12.0 (4.8)	27.1 (22.0)	57.1 (17.8)	66.4 (22.0)	82.3 (14.9)
Final stress	11.2 (2.7)	9.7 (4.7)	27.9 (19.5)	26.4 (17.2)	69.1 (7.8)	75.1 (10.4)
S3 Initial stress	12.2 (7.2)	6.3 (1.6)	15.0 (12.7)	10.0 (3.3)	12.2 (7.2)	15.0 (8.0)
Final stress	9.8 (6.8)	6.4 (1.2)	14.8 (8.8)	8.6 (3.5)	39.5 (10.1)	23.1 (15.5)
S4 Initial stress	19.9 (6.0)	15.2 (4.0)	36.3 (29.6)	15.7 (4.6)	46.2 (30.5)	38.9 (37.1)
Final stress	13.3 (4.6)	14.6 (2.8)	41.0 (31.3)	14.7 (4.1)	79.0 (11.8)	71.8 (29.9)
S5 Initial stress	21.8 (11.2)	17.7 (6.6)	43.6 (21.3)	29.9 (8.6)	66.9 (12.6)	50.5 (12.0)
Final stress	18.4 (9.0)	11.5 (4.9)	48.4 (21.7)	27.4 (15.5)	67.8 (16.5)	65.7 (15.0)
S6 Initial stress	5.7 (4.6)	8.7 (5.4)	56.9 (30.4)	66.8 (20.6)	74.8 (12.9)	85.6 (10.8)
Final stress	6.2 (4.0)	6.1 (2.3)	51.1 (18.7)	57.6 (12.0)	68.0 (9.9)	68.4 (8.2)