Perceptual rate normalization in naturally produced rate-varied speech a)

Kyoko Nagao b) and Kenneth de Jong c)

Department of Linguistics, Indiana University, Bloomington, IN 47405

Running Head: Perceptual rate normalization

a) Portions of this data were presented at the 146th meeting of the Acoustical Society of America.
b) Correspondence concerning this article should be addressed to Kyoko Nagao. E-mail: knagao@indiana.edu.
c) E-mail: kdejong@indiana.edu
ABSTRACT

The perception of voicing categories is affected by speaking rate, so that listeners' category boundaries on a VOT continuum shift to a lower value when syllable duration decreases [Miller and Volaitis, 1989; Volaitis and Miller, 1992]. Previous rate normalization effects have been found using artificially varied stimuli. This study examines the effect of speech rate on voicing categorization in naturally produced rate-varied speech. The stimuli contained natural decreases in VOT with faster speech rates so that VOT values for /b/ and /p/ overlapped at the fastest rates. Consonant identification results showed that the rate effects on the perceptual boundary between /p/ and /b/ very closely matched the effects of rate on the productions, though there was a small mismatch with fast rate productions whereby voiced stops were systematically miscategorized as voiceless. Another group of listeners judged the goodness of the consonant, indicating that best exemplars were rate-varied, and shifted away from the /p/-/b/ boundary. These results are discussed in light of exemplar-based and abstractionist models of speech perception.

PACS numbers: 43.71.Es, 43.70.Fq.
I. INTRODUCTION

How listeners cope with variability in the speech signal with respect to linguistic categories is a longstanding issue of speech perception. Speaking rate is one major source of such acoustic variability. Speaking rate affects acoustic properties of many kinds, such as consonant duration and vowel duration (Crystal and House, 1988), voice onset time (VOT) for voicing contrasts of syllable initial stops (Miller et al., 1986; Volaitis and Miller, 1992), closure duration for the voicing distinction of intervocalic stops (Lisker, 1957), transition durations for the manner distinction between stops and glides (Miller and Baer, 1983), and durations of adjacent segments (e.g. Repp and Lin, 1991 for the preceding fricative /s/; Newman and Sawusch, 1996 for the preceding and following consonants). It is also known that speaking rates vary both across speakers and within speakers (Miller et al., 1984c). Perceptual rate effects are not restricted to local syllable-internal temporal variation. Syllable external tempo estimates also have an effect on the perception of rate-sensitive information (Kidd, 1989; Summerfield, 1981; Wayland et al., 1994; Newman and Sawusch, 1996; Sawusch and Newman, 2000; Miller et al., 1984b).

Inter- and intra-speaker variability in speaking rate poses a problem in a theory of speech perception because many phonetic contrasts are differentiated by temporal differences. Unlike traditional views in which individual speech segments correspond to an invariant phonetic representation, the direct realism view (Fowler, 1977, 1980) and the exemplar-based or episodic models of categorization (e.g. Goldinger, 1996; Pisoni, 1997; Nosofsky, 1992) propose that listeners preserve fine phonetic details of their experiences with speech, and this would include contextual speaking rate variation. All of the fine detail is encoded in the perceptual structure that listeners use to categorize the speech. Since all of this information on temporal dynamics is
encoded and stored in the long-term memory, perceptual rate normalization should directly match rate variation effects in production.

However, there are cases of mismatch found between speech production and perception as we will show in detail below. The current paper proposes that certain of these cases of mismatch between speech production and perception are partially due to the use of artificially created stimuli in perception studies. More specifically, the current study focuses on the effect of speaking rate on production and perception of English voicing categories.

Among various measures, VOT is a critical acoustic property to differentiate the voicing contrast of syllable initial stops in English (Abramson and Lisker, 1985; Lisker and Abramson, 1964; Lisker, 1975). It has been known that speaking rate affects VOT; VOT increases as syllable duration increases. Furthermore, the rate effect is not symmetrical with respect to the voicing categories. VOT in voiceless consonants changes more with rate than in voiced consonants (Miller and Volaitis, 1989; Volaitis and Miller, 1992). Since VOT is rate sensitive, any model of the perception of linguistic contrasts must account for how listeners are affected by such variability. How can listeners distinguish voicing categories such as /p/ and /b/ in the presence of rate-induced variation in the speech signal?

The most extensive and comprehensive research on VOT and rate sensitivity has been conducted by Joanne Miller and her colleagues over the last two decades. They estimated the VOT values at the voicing category boundary based on production results and perception results. In addition, Miller and her colleagues conducted both production and perception studies to address the nature of voicing categories in English. There is clear evidence that speaking rate affects the location of voicing boundaries. However, close examination of their results also
revealed that there is a pervasive mismatch between the perceptual boundary and the production boundary reported in their studies.

Miller, Green, and Reeves (1986) explored rate effects in productions of the voicing contrast for initial bilabial consonants in English. They collected the syllables /bi/ and /pi/ produced by three native speakers of English, by employing a magnitude production technique to elicit the syllables across a wide range of speech rates. In this method, the subjects repeat each syllable at a normal rate, and then, they repeat the same syllable at a rate proportionally slower or faster than the normal rate. Using the same method with other stops, Volaitis and Miller (1992) replicated these findings, and noted that, although the large rate effect on VOT seen in /pi/ becomes smaller when the values were transformed into log scales, the difference between /bi/ and /pi/ remained. Similarly, Kessinger and Blumstein (1997) examined the relationship between VOT and speaking rate and found that the voiceless stops are produced with shorter VOT at fast speaking rates than at slow speaking rates, whereas the VOT for voiced stops does not change at different rates. These asymmetrical rate effects on voicing contrasts have also been found in other points of articulation with /t/-/d/ and /k/-/g/ contrasts (Kessinger and Blumstein, 1997; Volaitis and Miller, 1992) and in other languages (Kessinger and Blumstein, 1997; Magloire and Green, 1999).

Specific results for the production of English initial labial consonants are summarized in Table I. These studies are listed in terms of the rate of speech they examined (Fast, normal or medium, and slow), though some of the studies did not report the actual speaking rate. Speakers were usually asked to read the words at a small number of rates except Miller et al. (1986) and Volaitis and Miller (1992), and rates were controlled subjectively using labels such as normal, fast, and slow rates. Even in the most well designed studies by Miller and her colleagues, initial
speaking rate (normal) was determined by each individual speaker and other speaking rates were manipulated relative to the initial rate that each speaker employed. Some studies determined speaking rate in terms of subjective rates without specific acoustic measures, and some use either syllable duration or vowel duration as a reference of speaking rates. Even among those studies, there are slight contextual differences in the data collection such as words produced in isolation vs. words in a sentence. Also, there are various methodological differences between the studies such as elicitation methods, the following vowel quality, and linguistic context that each utterance was produced. Still, average VOT values for /b/ are relatively consistent across the studies, and appear at around 10 ms across different speaking rates. On the other hand, average VOT values reported for /p/ are quite different among those studies. Within normal or medium rates, average VOT values span from 28 to 80 ms. In addition, the range of VOT in the productions of /p/ is wider than VOT ranges in /b/ productions across studies.

VOT values at the boundary between /b/ and /p/ were estimated in some of the studies (Miller et al, 1986; Miller & Volaitis, 1989; Volaitis & Miller, 1992; Kessinger & Blumstein, 1997). As can be seen in the Table I, the estimated VOT values at category boundary between English /b/ and /p/ are also similar across the studies.

** Insert Table I about here. **

In Miller et al. (1986), category boundaries for production, again, in English, were estimated in terms of VOT. These optimal VOT boundaries, computed to differentially identify speakers’ intended consonant categories, classified the intended category (/b/ or /p/) with more than 90%
accuracy. When computed from their entire data set, the optimal VOT boundary was at 23.5 ms. They also computed rate-varied optimal VOT values for data within successive 50 ms syllable duration intervals from 100 to 700 ms. Rate-dependent optimal values improved categorization performance to 97.6%. In other words, allowing the VOT boundary to shorten at fast rates increased classification accuracy.

Such rate effects on the articulation of segments are reflected in perceptual responses to rate varied stimuli. Miller and Volaitis (1989) examined rate effects on voicing perception by systematically changing VOT values for /bi/ and /pi/. The VOT continuum was created to extend from English /b/ through /p/ to an exaggerated category beyond /p/ (called */p/). To simulate slow and fast speech, two different durations of syllables were synthesized with a serial resonant synthesizer. Listeners were asked to identify each of these stimuli as either /b/, /p/, or an exaggerated */p/. Listeners exhibited later VOT classification boundaries between /b/ and /p/ in long syllables than in short syllables. Volaitis and Miller (1992) extended the study to stops produced at other points of articulation (/g/, /k/, and */k/) and found similar results. Also, similar results of rate effects on the VOT boundary between voiced and voiceless stops have been found by other investigators (Kessinger and Blumstein, 1997; Summerfield, 1981; Sawusch and Newman, 2000).

To probe the internal structure of the listeners’ categories, Miller and Volaitis (1989) and Volaitis and Miller (1992) also asked listeners to rate the goodness of each stimulus using a 10-point scale ranging from 1 as the worst sounding /p/ to 10 as the best /p/. If stored categories do not contain rate-varied phonetic detail, the results would demonstrate no difference between ratings of short and long VOT stimuli as a function of rate. The authors found that listeners’ goodness judgments were also subject to rate changes. VOT values of the highest rated /p/ were
longer when syllable duration was longer. These results suggest that stored phonetic categories have detailed phonetic information encoding contextual speaking rate. Using a method of goodness judgment test, similar to the studies of Miller and Volaitis, Newman (2003) recently examined the correlation between speech perception and production within individual speakers. She found that the talkers who speak at a slower rate seem to have a perceptual prototype of /p/ with longer VOT values.

Table II summarizes prior perception studies of rate effects on voicing perception. Only results obtained for English initial bilabial stops are included in Table II to make a direct comparison with the current study. It reveals a striking common factor among these studies. Stimuli used in these perception experiments were either synthetic or edited speech. When edited speech was used, VOT was increased or decreased by adding or removing a portion of syllable onset, or cross-splicing the /p/-onset with the /b/-onset and vice versa. Although previous studies support the claim that the internal representation of phonetic categories contains detailed subcategorical information, much of the research into rate normalization employs synthetic or artificially manipulated speech. As far as we know, there is no perception study on rate effects conducted using naturally produced rate-varied speech. It is surprising that the characteristics of voicing categories examining VOT variation have been examined only with synthesized or modified speech, with sounds created in a manner that listeners have never perceptually encountered. While using synthetic or edited speech has the advantage of allowing researchers to manipulate only one variable (e.g. VOT) to create equally distributed samples ranging from one category to another, there are potential problems raised by using synthetic or manipulated speech as we will discuss further below.
We can see another commonality in these studies. Rate effects on perceptual judgments were explored with a small number of rate conditions. Furthermore, in some of the perception studies (e.g. Miller and Volaitis, 1989; Volaitis and Miller, 1992), stimuli were blocked by syllable duration when presented to the listeners. In these cases, if syllable duration is defined as speech rate, listeners did not need to calibrate rate of speech during each trial. In order to understand the on-line perceptual normalization mechanism, we need to explore how listeners perceive speech sounds at various rates.

Finally, comparing Table I and Table II reveals a systematic inconsistency between the production studies and the perception studies. Table I shows previously found mean and/or range of produced VOT values\(^1\) for initial bilabial stops and estimated /b/-/p/ boundaries, which are listed in terms of speech rate. Again, Table II lists the studies in terms of speech rate or stimuli duration. While the VOT values at the perceptual boundary become longer with longer stimuli durations, the actual reported values are not consistent with the production boundaries. The stimuli with 200 ms or shorter duration, i.e. fast rate stimuli, showed the perceptual boundaries at between 15 and 52 ms of VOT. The VOT at the production boundaries do not exceed 30 ms, as seen in Table I. For the stimuli with syllable duration between 300 and 500 ms, the VOT at the perceptual boundary appeared between around 27 and 45 ms, which is consistent with the production boundaries for medium rate speech, but exceeds the VOT production boundaries reported for slow speech rate.

Specifically comparing the results by Miller and her colleagues, we find that the estimated perceptual boundary (for example, reported in Miller and Volaitis (1989) and Volaitis and Miller
Nagao, JASA (1992) do not match with the estimated production boundaries in Miller et al. (1986). Fig. 1 superimposes the /b/-/p/ category boundary estimated in the production study (Miller et al., 1986) and the estimated boundary in the perception study (Volaitis and Miller, 1992) in the stimuli space employed in Volaitis and Miller (1992). Volaitis and Miller (1992) employed the two series of VOT continua. Filled diamonds in Fig. 1 represent their short stimuli (i.e. fast stimuli) with the syllable duration of 125 ms, and the filled squares represent their long stimuli (i.e. slow stimuli) with the syllable duration of 325 ms. Unfilled triangles on the thin line represent the rate-dependent optimal VOT values estimated as category boundaries in Miller et al. (1986). As is clearly seen in Fig. 1, perceptual VOT boundaries (thick line) were at much higher values than those estimated from the production studies (thin line).

**Insert Fig. 1 about here.**

Interestingly, a similar mismatch between production and perception has been found in a vowel study (Johnson et al., 1993). Johnson et al. (1993) compared speakers’ vowel productions and their perceptual vowel space. In order to examine listeners’ vowel spaces, a method of adjustment task was employed. Johnson et al. (1993) created synthetic vowels by changing the first and second formants (F1 and F2). A set of synthetic sounds for each vowel was arrayed in a two-dimensional grid of F1 and F2 on a CRT screen so that listeners could hear the sound associated with each cell in the grid. Prior to the perception experiment, subjects were asked to produce target vowels embedded in a word. Subjects were then asked to click a cell in the grid to locate the sounds that match a target vowel displayed on the same screen. The perceptual
responses were systematically more extreme in formant pattern than their productions. Based on the result that the perceptual vowel space is expanded similar to the vowel space in hyperarticulated vowels, Johnson et al. (1993) called this effect the hyperspace effect. Although the methodology and purpose of their study are quite different from the studies of Miller and Volaitis, the mismatch between production and perception is strikingly similar.

More recently, the interpretation of Johnson et al’s results as a hyperspace effect has been questioned (Whalen et al., 2004a; 2004b; but see Johnson et al., 2004). The crux of the criticism is that there is really no way of determining what the actual production space for a synthetic matrix of vowels is. Hence, perceptual responses could be a reflection of listeners’ expectations about indexical differences between the perceiver and their estimate of who (or what) might have produced the matrix of vowels. Whalen et al. (2004b; 2004c) replicated the study of Johnson et al. (1993) with speakers of a different dialect (Rhode Island), and found that back vowels showed more frontness (hypospace effect) rather than more backness (hyperspace effect) in perceptual vowel space.

However, a hyperspace effect has also been reported for English consonants using a somewhat different paradigm (Newman, 2003). Newman (2003) examined the correlations between speech production and perception for consonants within individuals. She examined the correlation between the average VOT value for /p/ that individual speakers produced and the VOT value of /p/ stimulus that each person rated as the best member of the category. The stimuli used in her Experiment 1 were created by modifying VOT values of natural speech to represent /b/ to /p/ to exaggerated */p/. A majority of the subjects (75%) showed the result that the highest rated /p/-stimuli had longer VOT values than their own average productions.
Comparing these studies suggests four possible reasons for the mismatch in production and perception in the rate normalization literature. First would be some sort of direct hyperspace effect induced by using synthetic or edited stimuli. Since synthetic speech and targeted editing generally fail to capture the many covarying aspects of the signal which indicate the voicing category, listeners may ‘load up’ on the VOT dimension to make up for the missing secondary cues. If we assume the best exemplars correspond to the hyperarticulated phonetic targets, the best exemplars should have longer VOT for /p/ and shorter VOT for /b/ than the VOT values typically found in speech. If both categories have hyperarticulated phonetic targets that are expanded symmetrically from a VOT category boundary, the hyperspace effect would not affect the location of the perceptual boundary. However, production studies show that the /p/ category varies more than the /b/ category, and hence the /p/ category is likely to have a larger hyperarticulated shift.

A second reason could be the choice of stimulus space with respect to what speakers actually produce. Miller and Volaitis (1989) and Volaitis and Miller (1992) employed quite extreme stimuli (as did Johnson et al., 1993 and Newman, 2003). Fig. 1 shows how wide a range of VOT was employed in Volaitis and Miller (1992) (the series of dots to the left and right). Stimuli with extremely long VOT were used as representations of an exaggerated */p/*. Hence the stimulus set for /p/ might be expanded to include this extreme category */p/ as a subcategory of /p/. If this is so, given the stimuli with a wide range of VOT as well as an expanded /p/ category, listeners’ phonetic targets might be altered for the specific task in the identification protocol and cause a shift in the perceptual boundary toward longer VOT. Then, it is possible that the mismatch between production and perception boundaries was due to a task space effect introduced by using such extreme stimuli.
Alternatively, the inclusion of these extreme stimuli could also create a perception-production mismatch because of the introduction of a three category system which differs from the naturally produced two category system of English. Subjects in Miller and Volaitis (1989) and Volaitis and Miller (1992) listened to all the stimuli once in order of increasing VOT (e.g. from /bi/ to /pi/ to */pi/). They were also told to pay attention to where the sound starts to change from /bi/ to /pi/ and from /pi/ to */pi/. As pointed out by Utman (1998), this process of familiarization might impose an external (three-category) system on the listeners so that the results might not actually reflect the structure of the listeners’ normal categories. Utman (1998) found that the familiarization process with an additional category of voiceless sounds does affect perceptual judgments on the normal voiceless sounds. As Utman (1998) noted, “the use of a third category in the identification task may have artificially narrowed the best-exemplar range for voiceless consonants by causing subjects to interpret these stimuli as falling into one of two categories”(p. 1647).

Finally, the perception-production mismatch might also be due to the way the stimuli were created. Unless a precursor phrase was given, speaking rates were defined only in terms of syllable duration in previous studies. In order to create VOT continua, syllable durations were fixed at either a short syllable duration or a long syllable duration, while VOT was modified in proportion to the total syllable duration. However, VOT is not the only change accompanying rate change. When speaking rate decreases, both vowel duration, and VOT become longer (Allen and Miller, 1999; Crystal and House, 1988; Peterson and Lehiste, 1960). Also, vowels following voiceless stops are shorter than vowel following voiced stops, and speakers maintain the difference in vowel duration between voiceless and voiced stops (Allen and Miller, 1999). Synthetic speech and artificial editing do not perfectly reflect the complexity of temporal
dynamics within a syllable; the relation between artificial rate changes and the natural temporal variation that listeners are accustomed to is unclear; and hence the effect of artificial modulation on perception expected on the basis of the listeners’ experience can often be unclear. For example, Kessinger and Blumstein (Kessinger and Blumstein, 1998) pointed out that VOT continua with fixed syllable duration could alter perception of the vowel following the initial consonant from /ɪ/ to /ι/ because the artificial vowel durations are not typical of the intended vowel /ɪ/ (cf. Allen and Miller, 1999).

The purpose of the current study was to examine speech rate effects on the perception of voicing contrasts in natural speech. Using natural speech eliminates most of the difficulties mentioned above. Natural speech reflects the dynamics of the production processes, and includes the general distribution of temporal variation that listeners are familiar with. Stimuli here were generated with a rate control paradigm to vary across an extreme range of rates found to be practical by previous speech research. Experiment I was conducted to examine the rate effects on productions of syllables /bi/ and /pi/. Experiment II explored the rate effects on perceptual identification of these rate controlled productions. The boundary estimates from Experiment I and the perceptual boundary estimates from Experiment II were compared with one another and with those of previous studies to determine if using naturally produced stimuli eliminates the production-perception mismatch found in previous studies. Experiment III was conducted to examine the effect of rate on listeners’ estimates of best exemplars, to determine if a hyperspace effect is present in the perception of naturally produced stop consonants.
II. EXPERIMENT I: ACOUSTIC ANALYSIS

Experiment I examined the speech rate effects on the VOT distributions of /b/ and /p/. Although acoustic analyses have been done for speech produced at various rates (Miller et al., 1986; Volaitis and Miller, 1992), rates were not externally controlled but manipulated by speakers. Miller and her colleagues collected syllables at various rates with a magnitude production procedure. In this method, speakers were asked to repeat each syllable six times first at normal rate, and then to produce two slower and two faster rates relative to this normal rate, and in addition a slowest and a fastest possible rate. In Miller et al. (1986), the syllable duration of their samples ranged from 111 to 700 ms. Although this range covers the average syllable durations observed in spontaneous speech, their samples can be thought as including a lot of slow speech, when compared with reported average syllable durations in American English, which ranges from 100 to 300 ms in spontaneous speech (Goldman-Eisler, 1968; Miller et al., 1984c; Roach, 1998). It is important to examine the rate effect with syllables produced at various rates, especially at fast rates, since self-controlled rates in laboratory speech tend to be slower than usual. The aims of Experiment I was to observe the VOT values for /b/ and /p/ productions at various rates systematically sampled, especially at fast rates, and to examine how the VOT values used for the /b/-/p/ boundary change as syllable duration increases.

A. Methods

1. Subjects

Two male and two female speakers participated in the recording at the Indiana University Phonetics Laboratory. All were native speakers of Midwestern varieties of American English and
were in their mid 30’s at the time of recording. All had no history of any speech and language disorders.

2. Speech materials

Stimuli were taken from a speech corpus originally collected for other purposes. The corpus includes productions of four native speakers of American English (2 male, 2 female). In this corpus (de Jong, 2001a; 2001b), speakers were asked to repeat the syllables /bi/, /pi/, /ib/, and /ip/ in time with a metronome. They repeated the same syllable for approximately twenty times with increasing, decreasing, or fixed rates of repetition. They were presented with a pulse train which began with a series of 6 clicks with a period of 400 ms. After the six, the period was decremented by 12.5 ms per period until reaching a period of 150 ms. The train concluded with a series of 6 more clicks at the fixed rate of 150 ms. Before the set was begun, the speakers were warned that the pacer would change rate. This elicitation technique has proved to be an easy task for naïve subjects. Only the CV syllables (/pi/ and /bi/) from the increasing rate condition were used for current study. The fastest 21 repetitions of each syllable were subject to the acoustical analysis. In order to compare our results with previous studies, open syllables whose onset consonant was a labial stop (/bi/ and /pi/) were used for acoustical analysis. The total number of tokens was 168 (= 4 speakers by 2 consonants (/bi/ and /pi/) by 21 rates).

3. Measurements

For each syllable, the beginning of the stop release, the beginning of the modal voicing for the vowel, and the end of the vowel were determined by a visual inspection of the waveforms and spectrograms (for more information on measurements, see de Jong 2001a, 2001b). Following Miller et al. (1986), the value of the VOT interval was defined as the time interval
from the consonant release to the beginning of the voicing, and syllable duration as the duration from the release to the end of the vowel. Note that these measurement criteria make our definition of VOT somewhat different from that sometimes used in the literature, where stops with prevoicing have negative VOT values. In previous work with this corpus of productions, the existence of prevoicing is considered as a separable acoustic attribute from VOT, which measures degree of gap between the burst and modal voicing. Systematically, the current stimuli have no voicing in the stop closure at slow rates, and both /pi/ and /bi/ stimuli gain it at faster rates (de Jong, 2001a).

In order to assess measurement reliability, six stimuli (three /bi/-stimuli and three /pi/-stimuli) from each of the four talkers were selected at random, and values of VOT and syllable duration for these twenty-four stimuli were measured by a different phonetician (the first author). Pearson correlation coefficients were computed for the two measurements for each stimuli type. The results of high r-values suggest that measurement reliability was adequate for this study [VOT for /b/-stimuli: \( r = .91 \); VOT for /p/-stimuli: \( r = .96 \); Syllable duration for /b/-stimuli: \( r = .95 \); Syllable duration for /p/-stimuli: \( r = .96 \)]. Average differences between the two measurements were small for both /b/-stimuli and /p/-stimuli [VOT: \( M = 1.1 \) ms for /b/-stimuli and \( M = 1.5 \) ms for /p/-stimuli; Syllable duration: \( M = 3.6 \) ms for /b/-stimuli and \( M = 2.7 \) ms for /p/-stimuli].

**B. Results**

In order to see the relationship between VOT and syllable duration, all the syllables except one were grouped into 50-ms bins in terms of the syllable durations, and the VOT values of the four speakers were averaged for each bin of syllable duration. Fig. 2 plots average VOT values
of /bi/ and /pi/ as a function of syllable duration. The relationship between VOT and syllable duration for /pi/ is similar to that reported in previous literature (Miller et al., 1986; Miller and Volaitis, 1989; Volaitis and Miller, 1992; Kessinger and Blumstein, 1997). VOT increases as syllable duration increases. As for /bi/, VOT does not change much as a function of syllable duration. Although the differences are small, the slope function for /bi/ is slightly negative, which means that VOT actually decreases as syllable duration increases.

** Insert Fig. 2 about here. **

In order to illustrate the detailed distributions of VOT values for /bi/ and /pi/, the VOT values of all the tokens produced by four talkers were plotted against their syllable durations in Fig. 3. In Fig. 3, a positive linear relationship between VOT and syllable duration is displayed for /p/-productions. Although the VOT values for /p/ increase as syllable duration increases, the VOT values for /b/ do not change much, as seen in Fig. 3, and, if anything, tend to shorten at slower rates. The actual slope values were 0.1685 for /p/ and –0.0413 for /b/. The VOT values of each consonant are well separated when syllables are relatively long (> 200 ms); however, they overlap considerably with each other when the syllables are short.

** Insert Fig. 3 about here. **

A binary logistic regression analysis (e.g. Menard, 2001) was performed with VOT and syllable duration (SYLLABLE) as predictor variables. The response variable was defined as whether the speakers’ intended syllables were either /p/ (=1) or /b/ (=0). The probability of /p/ was computed by the equation:
\[ \rho( / \ p \ /) = \frac{e^{\beta}}{1 + e^{\beta}} \]  \hspace{1cm} (1)

where \( e \) is the natural logarithmic base and \( \beta \) is given by the following equation:

\[ \beta = -4.564 + 0.232(VOT) - 0.007(SYLLABLE) \]  \hspace{1cm} (2)

Both VOT and Syllable variables were significant predictors of the speakers’ intended consonant \([ps < 0.001]\). The regression equation correctly classified 86.9% of /b/ tokens and 88.1% of /p/-tokens for a total predictive efficiency of 87.5%. The category boundary was estimated from VOT at the 50% point of the regression function; it is plotted as a solid line in Fig. 3. It is clear that the optimal criterion VOT is rate sensitive, with a positive slope as a function of syllable duration.

Miller et al. (1986) estimated the rate-dependent optimal /b/-/p/ boundaries by locating the VOT value to yield best categorization performance (greater than 90% for their data) for the data sets within successive 50-ms syllable duration between 100 and 700 ms. Their estimated VOT values for the /b/-/p/ boundaries are plotted as a dotted line in Fig. 3. We can see both boundary lines are close to each other at slower rates (to the right). Their estimated optimal boundaries successfully separate our speech samples into /b/ and /p/ categories at slower rates, but errors start to occur at faster rates. Their estimated boundaries appear in the middle of /b/ productions when the syllable has 200 ms or shorter duration. Comparing our boundaries to those in Table I, shows that our results are quite similar to previous results, and the estimated boundary actually is very close to that found by Kessinger & Blumstein (1997). Although our estimated /b/-/p/ boundary line is not as steep as the boundary reported in Miller et al. (1986), it is nevertheless clear that VOT at the category boundary increases with increasing syllable duration.
III. EXPERIMENT II: PERCEPTION EXPERIMENT I

Experiment II was conducted to examine speech rate effects on voicing identification. The aims were to examine whether the category boundary shift observed for synthesized and edited sounds occurs in the same way when listeners judge natural speech or not, or whether the perceptual category boundary matches the category boundary estimated from the production studies. We prepared the stimuli from the same speech corpus used in Experiment I. Our speech samples include faster rates of speech than the data set of Miller et al. (1986). In their studies, syllables with less than 100 ms duration were not observed. However, syllable durations of our speech samples ranged from 68 to 331 ms. All the syllables except one used in Experiment II were under 300 ms in duration. This matches with average syllable durations in American English, but could be considered in the range of the fast rate in the study by Miller et al. (1986).

There were also several methodological differences between the current study and the studies by Miller and Volaitis (1989, 1992). Miller and Volaitis (1989) and Volaitis and Miller (1992) employed three choices (/b/, /p/, and */p/) for the identification task. We assumed that the use of unnatural category */p/ was not necessary since the stimuli in Experiment II were all naturally produced /b/ or /p/, and */p/ was not an intentional category of the speakers. Also, it is not necessary to include a process of familiarization with the stimuli. In Miller and Volaitis (1989) and Volaitis and Miller (1992), stimuli with the same syllable duration were blocked, so listeners did not need to constantly normalize speech rate for every trial. The stimuli in Experiment II had gradiently varied syllable durations, and they were randomly presented to the listeners to see the effects of local rate on voicing categorization. Listeners were asked to listen to a portion of repetitive speech that we collected in Experiment I and identify the consonant that the speaker intended to produce.
The category boundary between /b/ and /p/ was estimated in terms of VOT and syllable duration based on the listeners’ identification in order to examine the consistency of production and perception. We compared the /b/-/p/ category boundary estimated from the current study with the one obtained for synthesized sounds (Miller and Volaitis, 1989; Volaitis and Miller, 1992) as well as with the one estimated in Experiment I. When the listeners judge natural sounds, if the boundary shift occurs in the same way as in previous studies, we can generalize the results of artificially manipulated or synthesized sounds to natural sounds. Unlike synthesized speech, natural speech preserves all the acoustic information during the transition from a consonant throughout the following vowel, so any relevant acoustic property can be used for voicing categorization when the rate changes. In addition, we presented the listeners repeated syllables as a stimulus in order to provide the listeners with rate information.

A. METHOD

1. Subjects

Twenty-three native listeners of American English from 18 to 24 years old participated as paid subjects in the study. Data from one subject was not included, because a hearing problem was self-reported. Four bilingual subjects were excluded from further analysis because the voicing distinction in the other language might influence their perception of the voicing contrast in English. Data from eighteen native listeners of American English were subject to further analysis.
2. Stimuli

Speech materials used for the stimuli in Experiment II were the same as used in Experiment I. In the production corpus from which the stimuli were drawn, the four speakers repeated the same syllable (/bi/, /pi/, /ib/, and /ip/) with increasing rate of speech. All four of these types were submitted to perceptual identification procedures; however, since previous studies of rate normalization focused on prevocalic stops, the current study only focuses on results for the prevocalic (/bi/ and /pi/) productions.

From each repetitive utterance, twenty-one stimuli were spliced. Each of these twenty-one stimuli consists of three repeated vowels. Three syllable nuclei were included in each stimulus in order to allow for perception of both post-vocalic and pre-vocalic consonants in the middle syllable. Splicing any of the abutting syllables out would inevitably remove information about the identity of the consonant. Analyses of perceptual results using different splicing locations, with and without consonant transients on both ends of the three-vowel stimuli, show that giving listeners a train of three syllables gives them enough context to allow them to abstract away from the edges created by the splicing technique and give consistent identifications across splice locations (de Jong et al., 2004). Fig. 4 displays waveforms of a repetitive utterance of /pi/ by a female speaker. The slowest stimulus was spliced from the vowel offset of the fifth slowest syllable to the vowel onset of the ninth slowest syllable, and the second slowest stimulus was spliced out from the vowel offset of the sixth slowest syllable (or the first syllable of the slowest stimulus) to the tenth slowest syllable in the same utterance. In this way, twenty-one stimuli were prepared from four types of repetitive utterances by four speakers. The total number of stimuli was 336. The current paper analyzed the responses for 168 out of 336 stimuli because responses for the stimuli /ip/ and /ib/ were excluded for the further analysis.
B. PROCEDURE

Listeners were instructed that they were to listen to a portion of speech whose speaker was originally repeating the same syllable over and over. They were asked to listen to each sample carefully and to answer what they think the speaker was repeating from the four choices (‘pea’, ‘bee’, ‘eep’, and ‘eeb’). Presentation of the stimuli was controlled by a customized Matlab protocol. Listeners listened to the stimuli through Sony MDR-CD280 Stereo Headphones in a computer laboratory and selected one of the four choices on a computer screen by clicking a mouse. They were allowed to listen to each stimulus as many times as they wished. Each listener was given each of the stimuli once in a different random order. Prior to the experiment, the listeners had eight trials as practice. Eight stimuli used in the practice session were not used in the experiment session. During the experiment, a short break was given every hundred trials. Most of the listeners completed the whole experiment in less than one hour.

C. RESULTS

Responses for the 15 stimuli identified as either /ip/ or /ib/ among more than 75% of the listeners were excluded for the further analysis. Twelve out of the fifteen excluded stimuli were /bi/-stimuli. All excluded stimuli were fast rate stimuli (i.e. average syllable duration was less than 125 ms) except one stimulus (the average syllable duration was 137 ms). Hence, the responses for the 153 stimuli from 18 listeners were analyzed.

Fig. 5 plots VOT values for /bi/-responses (filled circles) and /pi/-responses (unfilled triangles) against syllable duration. Both VOT and syllable durations were taken from the
averaged duration of the second and the third syllables in each stimulus. Stimuli with the
average syllable durations with 200-ms or longer will be categorized as slow rates, stimuli with
125-ms or shorter average syllable duration will be categorized as fast rates, and the rest of the
stimuli as mid rates. The size of markers indicates the performance levels on voicing perception;
the larger the marker, the more listeners correctly identified the consonant. In general, listeners’
consonant voicing identification was accurate for both /bi/- and /pi/-stimuli. There were 12.8 %
incorrect responses among the total of 2754 responses. Most of the errors (89%) occurred for the
/bi/-stimuli, and most of these occur at fast rates. The /bi/-stimuli with 125-ms or shorter average
syllable duration were responsible for 74.9% of the incorrect responses. Thus, while listeners
identified both consonants very accurately at slow rates (stimuli with 200-ms or longer average
syllable duration), performance of /b/ identification decreases at fast rates and performance of /p/
identification remained high. In other words, listeners tend to perceive the consonant in fast rate
syllables as /p/.

** Insert Fig. 5 about here. **

In order to estimate the perceptual boundary, a binary logistic regression analysis was
performed with average VOT and syllable duration of the second and the third syllables in each
stimulus as predictor variables. The response variable was whether the listener responded to the
stimulus as /p/ (=1), or as /b/ (=0). The probability of /p/ perception was computed by the
equation in (1). The equation for the relationship between the dependent variable and the
independent variables was:

24
\[ \beta = -2.394 + 0.225(VOT) - 0.016(SYLLABLE) \]  

Both VOT and syllable duration contributed significantly to the prediction of consonant identification \([ps<.001]\). The logistic regression equation accounted for 85.3% of the dependent variable’s variance. The percentage of correctly classified /p/-stimuli was 88.0%, and was greater than the correct classification of /b/-stimuli (80.7%).

Analysis based on the VOT and syllable durations of the middle syllable in each stimulus was also conducted. The results turned out to be similar to the analysis based on averaged durations. The equation was:

\[ \beta = -1.441 + 0.194(VOT) - 0.017(SYLLABLE) \]  

In the analysis based on the second syllable durations, the percentage of correctly classified stimuli slightly decreases to 87.2% and 79.2% for /p/- and /b/-stimuli, respectively.

VOT values at the perceptual /p/-/b/ boundary were estimated by calculating the VOT values to yield 50% /p/ responses for syllable durations between 50 and 325 ms. The estimated boundary is shown in Fig. 5 (solid line). The positive slope of the estimated boundary function indicates a strong rate-dependence such that as syllable duration increases, boundary VOT values increase. Fig. 5 also displays the perceptual category boundary between /b/ and /p/ reported by Volaitis and Miller (1992) (dashed line). It is clear that the perceptual boundary obtained with the current data has considerably shorter VOTs than that obtained from the listeners’ responses in the study by Volaitis and Miller (1992). Comparing the boundary obtained in Experiment II to those generally found with synthesized or edited speech summarized in Table II shows that, while not all previous studies obtain such high VOT values as Volaitis & Miller (1992), all of these previous studies (with the possible exception of Summerfield, 1981 and Miller et al., 1984a) obtain higher VOT boundaries than are obtained in Experiment II.
Comparing Fig. 3 to Fig. 5 shows that the production boundary, while somewhat greater in VOT than our current perceptual results, is still closer to the perceptual boundary found in Experiment II than to those found in previous studies. Examining the pattern for individual tokens shows that the production and perception boundary is remarkably close for syllable durations greater than 100 ms. The only systematic deviation obtained here is for the very short syllable durations where majority /b/ responses are virtually non-existent. The unidirectional misperception result that /p/ was not perceived as /b/ at very fast rate seems to be also support the claim that the hyperarticulated targets for /b/ and /p/ in English are asymmetrically distributed. This issue will be more directly addressed in Experiment III.

To summarize, speech rate affects voicing perception in naturally rate varied speech. This result is consistent with previous findings using synthesized and artificially edited sounds. However, the estimated perceptual boundary is closely matched with the category boundary estimated from productions in Experiment I. Deviations of the production and perception functions are mostly due to misidentification of /b/-stimuli with very short syllable durations. Aside from this discrepancy, the current results suggest that previous production-perception mismatches may be due to the use of synthetic or edited stimuli with unnatural distributions over the space of speech rates. Experiment III pursues the particular mechanism responsible for the mismatch by examining listeners’ estimates of best exemplars to determine if a hyperspace effect can be detected with the current stimuli.
IV. EXPERIMENT III: GOODNESS RATING

The results of Experiment I and Experiment II suggest that the perception-production mismatch found in the previous studies could be due to the use of artificial speech. In addition, the extreme VOT values employed in the previous studies could distort the distribution of each consonant category. Alternatively, if we assume the asymmetrical distribution of the hyperarticulated phonetic target for /b/ and /p/, the hyperspace effect should appear at slow rate speech even for the natural stimuli. The purpose of Experiment III was to examine the rate effects on estimates of the best exemplars of voicing categories for natural speech. Listeners were asked to rate the goodness of the consonant that they heard. We asked the listeners to identify the consonant, and then judge the goodness of the consonant in terms of the consonant category they perceived.

A. METHOD

1. Subjects

Seventeen native listeners of English were newly recruited and participated in the study. They were undergraduate students of Indiana University whose ages were from 18 to 27 years old. All of them reported that they had not been diagnosed as having any hearing problems or language disorder. They were paid for their participation.

2. Stimuli

The same stimuli used in Experiment II were used for this study. The 336 stimuli were presented to the listeners in random order, but the responses for the 168 coda stimuli (/ip/- and /ib/-stimuli) were excluded for the further analysis, as we did in the analysis in Experiments I and II.
B. PROCEDURE

Listeners were given similar instructions as in Experiment II; they were told that they were to listen to a portion of speech whose speaker was originally repeating the same syllable. There were two tasks in this study. One was to identify the consonant and the other one was to rate the goodness of the consonant. They were asked to listen to each sample carefully and to answer which consonant they thought the speaker was repeating from two choices (/p/ and /b/). This task differed from the four-way forced choice identification task in Experiment II to insure that the difference between the current results and previous results was not due to a task that integrated syllable affiliation and consonant voicing in the same task. Then, listeners were asked to rate the goodness for the consonant they selected using a scale from 1 (= “Terrible”) to 10 (= “Excellent”). Presentation of the stimuli was controlled by a customized Matlab protocol. As in Experiment II, listeners were to listen to the stimuli through headphones and provide their answers by a mouse click. They were allowed to listen to each stimulus as many times as they wished. Prior to the experiment, eight practice trials were given to each listener. Listeners had a short break every hundred trials. The whole experiment usually was completed in less than one hour.

C. RESULTS

The 15 stimuli excluded in Experiment II were also excluded in the analysis in Experiment III in order to compare the results of both perception experiments. Therefore the analysis of results were based on 2601 responses (=17 listeners by 153 stimuli). For the same reason, the values of VOT and syllable durations used for the analysis were the average values between the second and third syllables in each stimulus.
The consonant identification results replicated the results of Experiment II. Mean percent of /p/ responses for each stimulus in Experiment II and Experiment III were separately arcsine transformed (Studebaker, 1985). A two-tailed paired t test of transformed mean percent /p/ responses revealed that there was no significant difference between Experiment II and Experiment III for the overall consonant identification results [t(152) = 0.59, p = .56]. Overall, listeners’ consonant identification was accurate for both /bi/- and /pi/-stimuli. There were 9.8% incorrect responses among the total of 2601 responses. Among the incorrect responses, /b/-stimuli with syllable durations shorter than 125 ms were responsible for 83% of the incorrect responses. This result suggests that listeners tend to perceive the onset consonant in very fast syllables as voiceless. On the other hand, /p/-stimuli with 200 ms or longer syllable durations produced no misidentifications among the listeners. These results are virtually identical to those from Experiment II, indicating that the difference between ours and previous results were not due to the integrated judgments on syllable and voicing in our task in Experiment II.

Turning to the goodness judgments, Fig. 6 plots mean goodness ratings for the correct responses across all listeners against VOT. The VOT value and syllable duration for each stimulus are based on the average durations of the second and the third syllables. Responses were averaged for stimuli in 5 ms bins (for VOT < 40 ms) or 10 ms bins. The lowest mean rating (= 6.2) appeared for the stimuli with 15-20 ms VOT, which is the location of the category boundary. The highest mean ratings were found for the stimuli with the shortest VOTs and with the longest VOTs. Thus, the best exemplar of /p/ has an extremely long VOT, while the best exemplar of /b/ has an extremely short VOT. This general pattern is similar to the hyperspace effect found by Johnson et al. (1993), whereby best exemplars are far from category boundaries, and hence tend to be tokens which are extreme in phonetic attributes.
In order to see the relationship between rate and goodness rating, mean goodness ratings for correct responses were computed for three rates (Fast, Mid, and Slow). Three rates were defined by average syllable durations of the second and the third syllable in each stimulus. Stimuli with average syllable durations of less than 125 ms were grouped into a fast rate category (Fast), the stimuli with average syllable durations of longer than 200 ms were grouped into a slow rate category (Slow), and the stimuli with average syllable durations of longer than 125 ms but less than 200 ms were grouped into mid rate (Mid). Mean ratings are presented in Table III.

As is apparent in Table III, the consonants produced at slower rates received higher ratings. Also, the /p/ stimuli are more highly rated overall than the /b/ stimuli. A two-way ANOVA was performed on the mean goodness ratings across subjects for correct responses with rate (Fast vs. Mid vs. Slow) and consonant (/b/ and /p/) as independent variables. Significant main effects were found for both rate and consonant \[ F(2,147) = 81.4, p < .0001; F(1,147) = 63.1, p < .0001, \] respectively. The interaction effect was significant at \( \alpha = 0.05 \) level, but not significant at 0.01 level \( [p = .02] \). Thus for the stimuli in this study, regardless of voicing category, listeners judged that consonants produced at slow rates are better examples of each category. Although the reason is not clear, tokens in the voiceless category are consistently judged with higher ratings than those in the voiced category, and this rating difference was especially large at fast rates.
To examine in more detail the effects of rate on the relationship between VOT and goodness rating, Fig. 7 separates /b/-stimuli (top panels) from /p/-stimuli (bottom panels) and gives separate functions for the three different rates (Fast, Mid, and Slow). Fig. 7 also includes separate functions for correctly identified stimuli in the left panels and for incorrectly identified stimuli in the right. In order to see the small differences in VOT for fast /b/ stimuli clearly, 2.5ms-bins were employed to group the stimuli with VOT values less than 20ms. Among correctly identified /b/ syllables (top left panel), stimuli with shorter VOTs were rated higher, while stimuli with longer VOTs, which is closer to the boundary with /p/, received lower ratings. The one exception to this pattern is for the slowest rates (filled triangles), where the listeners preferred slightly longer VOTs. The ratings for /p/ syllables (bottom left panel) had the opposite trend, with stimuli with long VOTs generally receiving the higher ratings and stimuli with short VOTs closer to the boundary with /b/ receiving lower ratings (though for some reason the stimuli with the shortest VOT at fast rates and the mid-rate stimuli with 30–35 ms VOTs tended to be rated very high). The most frequent consonant misidentification occurred for the stimuli with 7.5-25 ms VOTs, which is again consistent with the critical VOT values that had been reported for the /b/-/p/ boundaries in previous literature (see Table II). On the whole, the listeners tended to rate /b/s with relatively short VOTs and /p/s with relatively long VOTs as better than their more moderate counterparts.
Furthermore, listeners gave higher ratings when they correctly identified the voicing of the consonant than when they misidentified it, except that the fast /b/-stimuli misperceived as /p/ were rated higher than those which were categorized as /b/. Also, note that the misperception occurs at 10-30 ms VOTs, where the category boundaries are expected. This result supports the findings in Experiment II that the listeners tend to perceive /b/ as /p/ at very fast rates.

In order to quantify these impressions and analyze the best exemplar for /b/ and /p/ for each listener and for each of the three rates, VOT values for highest rated /bi/- and /pi/-stimuli were tallied for each listener. When the highest rating was assigned to more than two stimuli with different VOT values, these VOT values were averaged. A two-way ANOVA was performed with consonant (/bi/ vs. /pi/) and rate (fast vs. mid vs. slow) on the VOT of the highest rated stimulus for each listener. The results revealed a significant main effect of consonant \( F(1,96) = 1057.6; p < .0001 \), such that best /p/s have higher VOT than best /b/s (mean VOT values for /b/ and /p/ were 14.2 and 38.8 ms, respectively). The main effect of rate was also significant \( F(2,96) = 102.3; p < .0001 \), such that preferred VOT values were generally higher for slower rates (mean VOT values for fast, mid, slow were 20.6, 25.2, and 33.7 ms, respectively). A significant interaction effect was also found \( F(2,96) = 119.0; p < .0001 \), wherein the rate-dependent VOT changes on goodness judgments are larger for /p/ than /b/.

The goodness judgment results also showed a consistent asymmetrical relationship between /b/ and /p/. Mean goodness rating for correctly identified /p/ was higher for /p/ (7.59) than for /b/ (6.79), and the standard deviation (SD) was smaller for /p/ than for /b/ (SD: 2.80 ms for /b/ and 2.47 ms for /p/). On the other hand, the range of average VOT values for the highest rated tokens was narrower for /b/ than for /p/ (/b/: 9-20 ms, /p/: 19-61 ms). The reason of perceptual asymmetry found in /b/ and /p/ is not clear, but it suggests that the listeners are more sensitive to
the VOT differences for the best /b/ than for the best /p/. The result that the best /b/ tokens appeared in a narrower VOT range than the best /p/ tokens may be related with temporal psychophysics. If temporal processing of VOT differences is not linear, being particularly sensitive to VOT variation near zero, it might yield the result that the best /b/ tokens appeared in a narrower range of VOT than the best /p/ tokens. Alternatively, differences in phoneme frequency might also explain the perceptual asymmetry. The words with onset /b/ occur more frequently than the words with onset /p/ in the CVC context (Kessler and Treiman, 1997). More frequent access to the exemplars of word initial /b/ than the exemplars of word initial /p/ might make listeners to be more attentive to the detailed acoustic information for the best exemplars of /b/ than for /p/. More fastidious judgments toward /b/ might lower the overall goodness ratings for /b/. A third possibility is that onset /p/s are simply more perceptually salient than onset /b/s. Although the consonants /b/ and /p/ are rarely confused with each other, and the consonant /p/ is more confused with other consonants such as /t/ or /k/ in general (Miller and Nicely, 1955), it may not be the case for highly ambiguous segments used in this study. Hence, subjects may simply find it easier to identify /p/s, and so rate them better. More work with different stop contrasts would be a likely way of pursuing these results further.

To summarize, then, what one finds in the goodness rating data is something like a hyperspace effect. The stimuli which were most often chosen as the best exemplars were not those which would most accurately characterize the distribution of the /p/ and /b/ categories, ones which fall in the center of the productions. Rather, the stimuli which are more extremely removed from other categories were rated more highly than their more moderate counterparts. Comparing the peak VOT values from Fig. 7 to the production distributions in Fig. 3 shows that listeners are selecting the top and bottom edges of the /p/ and /b/ distributions as best exemplars.
One other aspect of Fig. 7 is worth mentioning. The misidentified stimuli are consistently rated lower than the correctly identified ones except the fast /b/-stimuli. That is, it seems that when subjects make an identification error, they are sensitive to information in the signal that runs counter to their misidentification. Particularly of note is that such misidentified tokens sometimes have VOT values which would lead one to miscategorize them (/b/ with long VOT, and /p/ with short VOT), but sometimes do not have such error-inducing VOT values. Hence, sometimes the VOT information could be cause for both miscategorization and the low rating, and sometimes only for the low rating.

The main results for production and perception can be summarized in Fig. 8. Fig. 8 plots the category boundary based on the production results in Experiment I (thin line), and the perceptual category boundary based on the perception results in Experiment III (thick line). The average VOT values of the highest rated stimulus at fast, mid, and slow rates are also found in the same figure (filled circles for /b/ and unfilled triangles for /p/). The average VOT values of /b/- and /p/-stimuli at those three rates used in the analysis of Experiment III were computed separately for /b/-stimuli (filled diamonds) and /p/-stimuli (unfilled squares), and superimposed on the figure. The rate effects can be clearly seen for the category boundaries both in production and perception as well as the best members of each category. Perception and production boundaries between voicing categories appeared in the same regions because of their close match. However, the perceptual boundaries exhibited a steeper slope because the perceptual boundaries shifted to the lower values at the fast rate because of the systematic and consistent misperception of fast rate /b/ tokens. The average VOT values of the highest rated stimuli appeared on the both sides of the category boundaries in production. Similarly, the average VOT values of the best stimuli appeared on the both sides of the perceptual boundaries except the best /b/ at fast rate. This is
again due to the misperception of /b/ at fast rates. The average VOT values of the best /b/ were shorter than the average VOT of the /b/ productions just like expected. On the other hand, the average VOT values for the best rated /p/ were longer than the average VOT values for /p/ productions at mid and slow rates, but the average VOT for the best /p/ at fast rate was slightly shorter than the average VOT observed in the /p/ tokens at fast rates. The small VOT differences between average /p/ productions and the best /p/ were not expected. However, this was probably because the /p/-stimuli in this study did not include a lot of /p/ tokens with very long VOT. This might be a limitation of the use of natural stimuli. However, from the analysis of goodness ratings above, it was clear that the listeners selected the /p/-stimuli with longer VOT as the best /p/.

**Insert Figure 8 about here.**

**V. DISCUSSION**

Using natural speech, the current data replicates the finding in studies on rate effects in voicing perception with synthesized and edited stimuli that the perceptual /b-p/ boundaries shift to longer VOT values as syllable durations increase (Miller and Volaitis, 1989; Volaitis and Miller, 1992). Thus, rate effects on perceptual identification found with synthesized and edited speech can be generalized to natural speech.

However, there are notable differences here from previous results concerning rate effects on voicing perception. While there were great differences between the location of the perceptual /b-/p/ category boundary in Miller and Volaitis (1989) and Volaitis and Miller (1992) and the
boundaries to differentiate /bi/ and /pi/ productions in Miller et al. (1986), the perceptual category boundaries estimated in the current perception experiments (Experiment II and III) closely matched the production category boundaries estimated from the syllables in Experiment I.

There are at least two likely reasons why the production-perception mismatch in VOT was smaller in the current investigation. Both are related to the nature of the stimuli, and degree to which voicing and rate information are encoded in the stimuli. First, unlike the previous studies, the current study employed naturally rate-varied stimuli, and therefore the current stimuli include more accurate information relevant to voicing perception than the ones used in previous studies. Hence listeners were better able to determine the intended productions of the speakers. Therefore, the boundary that distinguishes the two categories is very closely aligned with actual produced differences except at fast speech rates. Another piece of evidence for the very close agreement of the speakers’ and listeners’ categories is found in the goodness judgment data. Even in cases where the listeners misperceived the intended voicing category of the speaker, the goodness evaluation data shows that their misperception is only partial; such erroneously labeled tokens received systematically lower evaluations. The perceptual mechanism these results suggest will be further discussed below.

With respect to rate information, another notable difference in the current stimuli from those in previous studies was that each stimulus included three repetitive syllables instead of one isolated syllable. Even if the stimulus contains an ambiguous syllable, listeners could identify the consonant if one of the syllables in a stimulus provided the information that the listeners could use to identity the consonant clearly. The VOT and syllable durations of the three sequential syllables in each stimulus are never drastically different from one another. For example, VOT and syllable durations for the second and the third syllables in each stimulus differ only 0.5 ms
for VOT and 6.8 ms for syllable duration. However, it is obvious that additional information was available to the listeners. The mismatch in VOT between production and perception boundaries might be reduced if the synthesized stimuli were repeated three times like the current study. It would be interesting to explore this possibility, but the authors expect that since VOT at voicing category boundary tends to shift to longer VOT in long stimuli (e.g. Summerfield, 1981), the VOT at perception boundary could become longer as the stimuli become longer as well.

One note at this stage should be mentioned concerning the use of VOT as a criterial dimension for the English pre-vocalic voicing category. Although VOT is the dominant cue for voicing perception, other acoustic properties are known to influence listeners’ voicing judgments (e.g. f0 shift (Abramson and Lisker, 1985), formant transitions (Stevens and Klatt, 1974; Lisker, 1975; Summerfield and Haggard, 1977), consonant to vowel ratios (Port and Dalby, 1982), and amplitude of aspiration noise (Repp, 1979). Especially when they encounter ambiguous tokens, tokens appearing at around the category boundary on a VOT continua, listeners’ identifications are demonstrably sensitive to some of these acoustic properties (Abramson and Lisker, 1985; Lisker, 1975; Whalen et al., 1993). For some of the stimuli used in the current study, especially at fast rate stimuli, the produced /b/ and /p/ categories overlap in VOT. The fact that listeners correctly identify the consonant of these stimuli with wildly aberrant VOT values at above chance levels indicates that they are sensitive to some other properties of the stimuli besides VOT. This conclusion is also supported by the category goodness data. Tokens produced with VOT values in the wrong range, such as /b/ with VOT values greater than 20 ms, often were misidentified as /p/, however, receiving generally very low goodness ratings. These low ratings could simply mean that these stimuli were ambiguous with respect to the consonant, but also
indicate some information conflicting with the listener’s consonant identification as /p/ besides the VOT that indicates the intended /b/ category.

Although it seems reasonable that we would observe closer perceptual and production boundaries in cases where the stimuli exhibited all of the characteristics of naturally produced speech, than in cases where the perceptual stimuli were artificial, this observation does not by itself indicate what sort of perceptual mechanisms might account for the listeners’ performance, nor does it explain why previous studies should observe a specifically upward shift in the VOT boundary, relative to the current results. Of the four possible reasons given in the introduction, the current results are compatible with the latter two, i.e. inclusion of an unnatural third category and the incompleteness of speech information. The current results suggest that some sort of hyperspace effect due to the use of artificial speech does not explain the production-perception mismatch in previous studies. The reason for this is that the current results for category goodness estimates tend to show the same sort of hyperspace effect as did Johnson et al. (1993). This hyperspace effect in the choice of best exemplars did not correspond to a shifted boundary in the identification task. Thus, we expect that the large mismatches in some of the previous studies were due either to task specific categories developed for the very wide range of VOT values for voiceless consonants, or due to the synthetic and editing techniques being not entirely effective at conveying realistic differences in VOT.

The current study explored the notion of a hyperspace effect to explain the behaviors across subjects. In previous studies of the hyperspace effect (Johnson et al, 1993; Whalen et al, 2004a, 2004b), the data analyses were conducted within individual subjects because the vowel space needs to be calibrated for each talker. This calibration problem would be reduced in the current analysis of English stops, because only the temporal dimension was examined. Newman (2003)
examined the production-perception relationship within subjects, and found a hyperspace effect for English stop consonant /p/ in a majority of subjects. Hence, as long as individuals exhibited a consistent and similar pattern in their behaviors, the hyperspace effect should appear in group data as a reflection of individual data. However, as we noted above, voicing contrasts are not solely determined by VOT. Analysis between the individual subjects’ productions and perceptions might answer a question whether the hyperspace effect remains when the analysis is conducted in multiple dimensions (e.g. in both spectral and temporal domains).

Finally, concerning the perceptual mechanisms for rate that the current study suggests, we would contend that the current results do not fit well with a heavily abstractionist model, such as a traditional prototype model. To get rate effects with such prototypes, one must have an algorithmic extension of the evaluation procedure that matches stimuli to stored prototypes. However, one of the current results is that the rate effect is much larger with /p/ than with /b/. No simple scaling and matching algorithm will achieve this segment-specific effect, suggesting that a prototype model would have to either include a number of rate-specific prototypes, or develop segment-specific scaling mechanisms. Note that switching from expressing VOT in absolute duration to a measure proportionate to the syllable duration will not solve this problem, since the rate variation in VOT is considerably less than proportional to the overall syllable duration (de Jong, 2001a), which would require reverse rate normalization. The good fit between production and perception obtained in the current study also highlights just how carefully these very specific algorithms would have to be tuned to the specifics of individual segments. In addition, the category goodness data suggests that such prototypes would have to be systematically shifted away from the distributional centers of the productions to get the hyperspace effect.
The very good match between production and perception suggests something more along the lines of “non-traditional” speech perception theories such as a direct realism view (Fowler, 1977, 1980) or an exemplar-based or episodic model of categorization (Nosofsky, 1986, 1992; Pisoni, 1997) where listeners retain fine phonetic details. Despite the high stimulus variability, listeners’ accuracy of consonant identification did not decrease with rate except for the very fast stimuli, suggesting that listeners have knowledge of /b/-/p/ categories in their rate-varied details including VOT. The hyperspace effect also suggests the mechanisms employed in exemplar models (e.g., Nosofsky, 1986, 1992) where attentional weighting effectively expands dimensions which encode contrasts. Listeners are not just simply taking in the distributions, but apparently are sensitive to the orientation of contrast differences and what makes for an easily identifiable sound. Thus, such models allow for having both specific perceptual tuning to the detailed variation in the voicing categories and a specific mechanism that is sensitive to the abstraction of a linguistic contrast in voicing. A perceptual model that has a combination of tuning and abstraction seems to be what the current data call for.

However, an exemplar-based account with VOT-weighting does not fully explain the current perception results. At very fast rates (syllable duration < 100 ms), most of the /b/-stimuli were perceived as /p/. Since the talkers did not produce /b/ and /p/ in the same way even at fast rates (see de Jong, 2001a), such systematic misperception of one category for the other under one condition would not be expected in our exemplar model, since there are plenty of exemplars in the region of space where the misperceptions are occurring. Yet, listeners persistently get the fast rate stimuli wrong, as though a decision boundary in the fast rate region was misplaced based on some independent motivation.
Such a misplacement might occur if the distributional patterns of the two categories are overgeneralized to the fastest rate conditions. One such possibility is that there is in fact some abstraction and normalization of rate effects. Since the distribution of rate effects on VOT values is more varied for /p/ than for /b/, the rate normalization effect for /p/ overestimates the shortening for /p/ at fast rates and induces a miscategorization of /b/. In favor of this sort of account, the slope of the categorization boundary from Experiment III (in Fig. 8) is very nearly parallel to the shortening of VOT in the /p/ productions in Experiment I. Hence, it is as if subjects were doing rate normalization without attending to the production of /b/’s. Since VOT does not vary as much by rate for /b/, the rate normalization effect would have to be somehow attached specifically to the /p/ category.

Though how to implement this model is not exactly clear, we should note that the general pattern of mismatch between production and perception is not likely to be limited to the voicing contrast in initial position. Our work with the rate varied data has also shown the same sort of systematic mis-categorization for stops in post-vocalic position as well, though in the opposite direction. Post-vocalic voiceless stops in fast speech tend to be heard as voiced (de Jong et al., 2002). Similarly, our work with repetitive speech shows a systematic mismatch in the syllabic affiliation of stops whereby post-vocalic stops tend to be categorized as pre-vocalic onsets (de Jong, 2001b, among others). While the main pattern of the current research indicates that the perceptual system is very closely tuned to productions, productions in the margins along dimensions such as speech rate exhibit pervasive mismatches with what the perceptual system is expecting.
VI. CONCLUSIONS

We examined rate normalization effects on voicing contrasts in prevocalic position using natural speech and found that rate normalization effects found with synthesized or modified speech also occur in the perception of natural speech. Speech rate affects both production and perception in a very similar manner. The perceptual identification system is neatly tuned to the distributions found in production. The results of goodness ratings also indicate that rate affects listeners’ internal representations. Furthermore the results of goodness ratings indicate that listeners store fine-grained information to distinguish voicing contrasts, and are sensitive to the existence of contrasts along a particular dimension, such as VOT. Accurate identification of segments with aberrant VOT values suggests listeners use signal attributes in addition to VOT to differentiate the contrast. The general outcome of this perceptual tuning is that listeners’ identifications of the voicing category are extremely good, even in the face of extreme variability. Listeners effectively deal with rate-induced variation in categorization tasks. However, in extremely fast rates, there are persistent mismatches between production and perception, which suggest that the perceptual system has some sort of generalization capacity which mis-characterizes the production variability in these margins. We suspect that examining such cases of production-perception mismatch provides an important insight into the nature of perceptual generalization in speech, and has important and relatively poorly characterized implications for perceptual modelling as well as the function of speech communication systems. Of course, to do so requires us to examine speech perception more directly in light of actual speech production.
AKNOWLEDGEMENTS

This research was supported by NIH Grant No. DC04095 from the National Institute on Deafness and Other Communication Disorders and by Grants numbered BCS9910701 and BCS 04406540 from the National Science Foundation. We thank Byung-jin Lim for his assistance in the perception experiments. We also thank the three anonymous reviewers for their helpful comments on improving this paper.

1 When specific values are not available in the text, approximate values were read from figures. Mean values in Kissinger and Blumstein (1997) were estimated from Figure 4, and their range values were estimated from Fig. 6. Syllable structures and the vowel following an initial consonant vary. Data on Kessinger and Blumstein (1998) was taken from the results of words ‘peak’ and ‘peep’. Allen and Miller (1999)’s data was taken from the results of words ‘big’ and ‘pig’. For data from Lisker and Abramson (1964), values for pre-voiced cases are not reported here. Newman (2003) used the syllable /pa/.

2 Summerfield (1981)’s data was taken from the Series B results in Experiment 6, and results from the Series /bi-pi/ results in Experiment 6. Miller, Dexter, and Pickard (1984a)’s data was based on the results of Experiment 4a. Repp and Lin (1989)’s data was based on the result of the isolated stimuli in Experiment III and the values reported in Repp and Lin (1991). In Repp and Lin (1989, 1991), initial periodic segments of natural production of /b/ were replaced by a periodic segment of equal duration of natural production of /p/. As for Repp and Lin (1991), data was taken from the result of the isolated stimuli. For Newman and Sawusch (1996), data was taken from the results of /blos/-/plos/ series in Experiment 5.
Data for syllable durations longer than 300 ms were not plotted in Fig. 1 because there was only one sample.

REFERENCES


Table I. Mean and range of production VOT values (in msec) in the word initial /p/ and /b/ in English.

<table>
<thead>
<tr>
<th>Source</th>
<th>Task</th>
<th>Speech rate</th>
<th>Mean VOT (Range)</th>
<th>VOT at boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>/b/</td>
<td>/p/</td>
</tr>
<tr>
<td>Fast rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller, Green, &amp; Reeves</td>
<td>Magnitude&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Fast</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(1986)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volaitis &amp; Miller (1992)</td>
<td>Magnitude&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Fast</td>
<td>9.6</td>
<td>45.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kessinger &amp; Blumstein</td>
<td>Sentence&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Fast</td>
<td>13 (0-39)</td>
<td>63 (20-119)</td>
</tr>
<tr>
<td>(1997)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kessinger &amp; Blumstein</td>
<td>Sentence&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Fast</td>
<td>-</td>
<td>66 (30-110)</td>
</tr>
<tr>
<td>(1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allen &amp; Miller (1999)</td>
<td>Isolated words&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Fast</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allen &amp; Miller (1999)</td>
<td>Isolated words&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Fast</td>
<td>39</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal / Medium rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lisker &amp; Abramson</td>
<td>Sentence&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Normal</td>
<td>7 (0-15)</td>
<td>28 (10-45)</td>
</tr>
<tr>
<td>(1964)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lisker &amp; Abramson</td>
<td>Isolated words&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Normal</td>
<td>1 (0-5)</td>
<td>58 (20-120)</td>
</tr>
<tr>
<td>(1964)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller, Green, &amp; Reeves</td>
<td>Magnitude&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Medium</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(1986)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volaitis &amp; Miller (1992)</td>
<td>Magnitude&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Medium</td>
<td>13.1</td>
<td>80.2</td>
</tr>
</tbody>
</table>

50
<table>
<thead>
<tr>
<th>Study</th>
<th>Condition</th>
<th>Type</th>
<th>NA</th>
<th>Slow</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kessinger &amp; Blumstein (1997)</td>
<td>Isolated</td>
<td>words</td>
<td>NA</td>
<td>13</td>
<td>88 20-39</td>
</tr>
<tr>
<td>Newman (2003)</td>
<td>Imitation</td>
<td>NA</td>
<td></td>
<td>- 73 (51-125)</td>
<td>-</td>
</tr>
<tr>
<td>Miller, Green, &amp; Reeves (1986)</td>
<td>Magnitude</td>
<td>Slow</td>
<td>-</td>
<td>- 29-63</td>
<td>(SDur =500-799)</td>
</tr>
<tr>
<td>Volaitis &amp; Miller (1992)</td>
<td>Magnitude</td>
<td>Slow</td>
<td>13.8</td>
<td>103.1</td>
<td>- (SDur =500-799)</td>
</tr>
<tr>
<td>Kessinger &amp; Blumstein (1997)</td>
<td>Sentence</td>
<td>Slow</td>
<td>15 (0-39)</td>
<td>95 (40-149)</td>
<td>30-39</td>
</tr>
<tr>
<td>Kessinger &amp; Blumstein (1998)</td>
<td>Sentence</td>
<td>Slow</td>
<td></td>
<td>91(70-130)</td>
<td>(VDur = 95)</td>
</tr>
<tr>
<td>Allen &amp; Miller (1999)</td>
<td>Isolated</td>
<td>words</td>
<td>6</td>
<td></td>
<td>(VDur = 201)</td>
</tr>
<tr>
<td>Allen &amp; Miller (1999)</td>
<td>Isolated</td>
<td>words</td>
<td>61</td>
<td></td>
<td>(VDur = 173)</td>
</tr>
</tbody>
</table>

Note: VDur and SDur stand for mean vowel duration and syllable duration in msec.

a Syllables elicited by a magnitude production technique. b Words produced in a sentence.

c Words in an isolated condition. d Nonsense syllables in an isolated condition. The model stimuli were presented before each speaker produce the test syllables.
Table II. VOT values at the perceptual boundary (PB) for English initial bilabial stops.

<table>
<thead>
<tr>
<th>Sources</th>
<th>Task</th>
<th>Stimuli type</th>
<th>Stimuli duration</th>
<th>VOT at PB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short (Fast rate) stimuli</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lisker &amp; Abramson (1970)</td>
<td>Consonant ID</td>
<td>Synthetic speech(^a)</td>
<td>-</td>
<td>About 23 ms</td>
</tr>
<tr>
<td>Abramson &amp; Lisker (1985)</td>
<td>Consonant ID</td>
<td>Synthetic speech(^b)</td>
<td>-</td>
<td>Around 15-25 ms</td>
</tr>
<tr>
<td>Summerfield (1981)</td>
<td>Consonant ID</td>
<td>Synthetic speech(^c)</td>
<td>95 -140 ms</td>
<td>28.1 ms</td>
</tr>
<tr>
<td>Miller &amp; Vollaitis (1989)</td>
<td>Consonant ID</td>
<td>Synthetic speech(^d)</td>
<td>125 ms</td>
<td>35.61 ms</td>
</tr>
<tr>
<td>Vollaitis &amp; Miller (1992)</td>
<td>Consonant ID</td>
<td>Synthetic speech(^e)</td>
<td>125 ms</td>
<td>39.18 ms</td>
</tr>
<tr>
<td>Repp &amp; Lin (1991)</td>
<td>Consonant ID for “bin”-“pin” series</td>
<td>Edited speech by splicing</td>
<td>VOT (0-61ms) + /^|/</td>
<td>35.2 ms</td>
</tr>
<tr>
<td>(Data of Exp 3 in Repp &amp; Lin, 1989)</td>
<td>Consonant ID for “bin”-“pin” series</td>
<td>Edited speech by splicing</td>
<td>VOT (10-53ms) + /^|/</td>
<td>23-52 ms</td>
</tr>
<tr>
<td>Summerfield (1981)</td>
<td>Consonant ID in Exp 6</td>
<td>Synthetic speech(^e)</td>
<td>167.5 - 212.5ms</td>
<td>23.0 ms</td>
</tr>
<tr>
<td>Summerfield (1981)</td>
<td>Consonant ID in Exp 6</td>
<td>Synthetic speech(^e)</td>
<td>170 - 215ms</td>
<td>29.5 ms</td>
</tr>
<tr>
<td>Miller, Dexter, &amp; Pickard (1984)</td>
<td>Consonant ID for “beef”-“peef” series</td>
<td>Edited speech by splicing</td>
<td>210 ms</td>
<td>About 20 ms</td>
</tr>
<tr>
<td>Newman &amp; Sawusch (1996)</td>
<td>Goodness rating with a 6-point scale for /blos/-/plos/ short series in Exp5</td>
<td>Edited speech by splicing</td>
<td>187-224 ms (varying /l/)</td>
<td>34.9 ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>136-173 ms (varying /o/)</td>
<td>29.6 ms</td>
</tr>
<tr>
<td>Study</td>
<td>Task Description</td>
<td>Methodology</td>
<td>Duration</td>
<td>VOT Range</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------------------------</td>
<td>-------------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>Newman &amp; Sawusch (1996)</td>
<td>Goodness rating with a 6-point scale for /blos/-/plos/ long series in Exp 5</td>
<td>Edited speech by splicing (varying /l/ dur)</td>
<td>299-336 ms</td>
<td>39.6 ms</td>
</tr>
<tr>
<td>Newman (2003)</td>
<td>Goodness rating with a 10-point scale in Exp 1</td>
<td>Edited speech by splicing (=VOT) + /ə/</td>
<td>8.25- 291 ms</td>
<td>44.7 ms</td>
</tr>
</tbody>
</table>

### Long (slower rate) stimuli

<table>
<thead>
<tr>
<th>Study</th>
<th>Task Description</th>
<th>Methodology</th>
<th>Duration</th>
<th>VOT Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pisoni &amp; Tash (1974)</td>
<td>Consonant ID Synthetic speech</td>
<td>300 ms</td>
<td>About 30 ms</td>
<td></td>
</tr>
<tr>
<td>Miller &amp; Vollaitis (1989)</td>
<td>Consonant ID</td>
<td>Synthetic speech</td>
<td>325 ms</td>
<td>43.89 ms</td>
</tr>
<tr>
<td>Vollaitis &amp; Miller (1992)</td>
<td>Consonant ID</td>
<td>Synthetic speech</td>
<td>325 ms</td>
<td>45.22 ms</td>
</tr>
<tr>
<td>Miller, Green, &amp; Schermer (1984)</td>
<td>Consonant ID for “bath”-“path” series embedded in a sentence</td>
<td>Edited speech by splicing</td>
<td>425 – 491 ms</td>
<td>34.8 -36.5 ms</td>
</tr>
</tbody>
</table>

Note. The duration of stimuli used in Lisker & Abramson (1970) and Abramson & Lisker (1985) are not available, but grouped into the studies used short stimuli based on the short VOT at the perceptual boundary. Only the durations of VOT are available in Repp & Linn (1991) and Newman (2003). Most of the studies listed below varied VOT duration and the
fixed duration for the rest of syllable except Miller & Volaitis (1989), Volaitis & Miller (1992), Miller, Dexter, & Pickard (1984), and Pisoni & Tash (1974). Newman & Sawusch (1996) varied the duration of adjacent phonemes /l/ and /o/ separately, which is indicated as “varying /l/ dur” and “varying /o/ dur” in the table.

aSynthesis by a parallel resonance synthesizer. bSynthesis by the Haskins Lab. formant synthesizer. The version of synthesizer was not reported. cSynthesis by a serial resonance synthesizer. dThis is not equivalent to the /p/-/b/ boundary because the value was taken from the VOT at the lowest end of the VOT continuum for the /p/ category.
Table III. Mean goodness ratings for correctly identified /b/- and /p/-stimuli in three rates (1= “Terrible” to 10= “Excellent”).

<table>
<thead>
<tr>
<th>Rate</th>
<th>/b/-stimuli</th>
<th>/p/-stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast</td>
<td>5.41</td>
<td>6.67</td>
</tr>
<tr>
<td>Mid</td>
<td>7.20</td>
<td>8.00</td>
</tr>
<tr>
<td>Slow</td>
<td>7.67</td>
<td>8.82</td>
</tr>
</tbody>
</table>
FIGURE CAPTIONS

Figure 1. The filled diamonds and the filled squares represent the short and long stimulus used in the perception experiments in Volaitis & Miller (1992). The unfilled triangles interpolated with a thin line represents the rate-dependent optimal VOT to yield 90% or high categorization performance for /bi/- and /pi/- productions within successive 50-ms syllable duration interval reported in Miller et al (1986). The thick line represents VOT at the perceptual /b/-/p/ category boundaries reported in Volaitis & Miller (1992).

Figure 2. Mean VOT values of /bi/ and /pi/ tokens within successive 50-ms intervals of syllable duration. The filled circles represent /bi/ tokens, and the unfilled triangles /pi/ tokens.

Figure 3. VOT values for /bi/ and /pi/ tokens plotted as a function of syllable duration (/pi/: r = 0.68, /bi/: r = −0.40). The filled circles represent /bi/ tokens, and the unfilled triangles /pi/ tokens. The straight line represents estimated /b/-/p/ boundary locations, and the dotted line the /b/-/p/ boundary locations reported in Miller, Green, and Reeves (1986).

Figure 4. Stimuli used in the perception experiments.

Figure 5. VOT values for /bi/ and /pi/ responses plotted as a function of syllable duration (/pi/: r = 0.71, /bi/: r = −0.36). The filled circles represent /bi/ percepts, and the unfilled triangles /pi/ percepts. The size of the markers indicates mean percent correct identifications of the consonant by 18 listeners in Experiment 1 (The size decreases as mean percent identifications decreases). The straight line represents estimated /b/-/p/ boundary locations, and the dotted line the /b/-/p/ boundary locations reported in Volaitis & Miller (1992).

Figure 6. Mean goodness ratings as a function of VOT for correctly identified /bi/- and /pi/- stimuli. Judgments were made on 10-point scales (1 = Terrible, 10 = Excellent).
Figure 7. Mean goodness ratings for correctly identified /bi/-stimuli (top left panel), /pi/-stimuli (bottom left panel), misidentified /pi/-stimuli (top right panel), and misidentified /bi/-stimuli (bottom right panel) as a function of VOT at the three rates (Fast, Mid, and Slow). Three rates are defined by the average syllable durations (SDur) of the second and the third syllables in each stimulus: Fast (SDur ≤ 125msec), Mid (125msec < SDur ≤ 200msec), and Slow (SDur > 200msec).

Figure 8. The VOT values at the estimated category boundaries (CB) of /b/-/p/ productions in Experiment I, the VOT values at the perceptual CB estimated in Experiment III, the average VOT values for the /b/ and /p/ tokens at three rates (Fast, Mid, and Slow), and average VOT values for the best /b/ and /p/ stimuli at the three rates as a function of syllable duration. Three rates are defined by the average syllable durations (SDur) of the second and the third syllables in each stimulus: Fast (SDur ≤ 125msec), Mid (125msec < SDur ≤ 200msec), and Slow (SDur > 200msec).
- Short stimuli in V & M (1992)
- Optimal VOT values at the /b/-/p/ boundaries in Miller et al. (1986)
- VOT values at the perceptual /b/-/p/ boundaries in V & M (1992)