Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues

JÖRGEN PIND

University of Iceland, Reykjavík, Iceland

The temporal structure of speech has been shown to be highly variable. Speaking rate, stress, and other factors influence the duration of individual speech sounds. The highly elastic nature of speech would seem to pose a problem for the listener, especially with respect to the perception of temporal speech cues such as voice-onset time (VOT) and quantity: How does the listener disentangle those temporal changes which are linguistically significant from those which are extrinsic to the linguistic message? This paper reports data on the behavior of two Icelandic speech cues at different speaking rates. The results show that manipulations of rate have the effect of slightly blurring the distinction between unaspirated and aspirated stops. Despite great changes in the absolute durations of vowels and consonants, the two categories of syllables—V:C and VC—are nonetheless kept totally distinct. In two perceptual experiments, it is shown that while the ratio of vowel to rhyme duration is the primary cue to quantity and remains invariant at different rates, no such ratio can be defined for VOT. These results imply that quantity is the only one of these two speech cues that is self-normalizing for rate. Models of rate-dependent speech processing need to address this difference.

Many speech cues are primarily temporal in nature, being defined either by the temporal relationships between acoustic events or by the duration of individual segments (Lehiste, 1970; Lisker, 1974). A noticeable feature of normal speech is the highly variable utterance rate. The fact that listeners do not appear to be overly bothered by such speaking-rate variations poses a challenge for perceptual theory, not least concerning the perception of temporal speech cues. How is the listener able to disentangle those temporal aspects which properly are extrinsic to the phonemic message (i.e., changes of tempo) from those which define the relevant phonemic contrasts, given that both are carried simultaneously in the speech stream?

Numerous studies have shown convincingly that rate-dependent speech processing does in fact occur and is found to affect numerous speech cues. Miller and Liberman (1979) found that with a longer overall duration of the syllable, listeners required longer formant transition durations to hear /wa/ than to hear /ba/. Summerfield (1981) found that the longer the following vowel, the longer the voice-onset times (VOTs) that were needed

to cue aspirated stops, thus showing a pattern quite similar to that found by Miller and Liberman (1979). At first glance, these results would seem to show some kind of rate adjustment in speech perception. Rate dependent effects are by no means limited to durational variation in the relevant syllables, though these effects tend to be larger than those connected with extrinsic factors such as precursor sentences (Pind, 1986; Summerfield, 1981).

Researchers have been divided over questions concerning the precise nature of the mechanism involved in such rate-dependent processing, including that of whether it could possibly be explained in purely auditory terms. Thus, Diehl and Walsh (1989), following upon earlier work by Pisoni, Carrell, and Gans (1983), have claimed that the rate adjustments seen for the [ba-wa] distinction are in fact properly interpreted not as speech-specific adjustments, but rather in terms of a "general auditory principle" of durational contrast. Their claim is that a perceptual boundary—for example, one separating /ba/ and /wa/—would move to a longer transition duration if followed by a long vowel, since the long vowel would tend to make any particular transition duration appear shorter than it would if it were followed by a short vowel. Using nonspeech analogs of Miller and Liberman's (1979) original synthetic-speech stimuli, Diehl and Walsh were able to replicate most of the findings of the original study, though their interpretation of the significance of these findings has not gone unchallenged (Fowler, 1990).

Diehl and Walsh's (1989) auditory theory is put forward as an alternative to the theory which assumes a

---

This research was supported by the Icelandic Science Foundation and by the Research Foundation of the University of Iceland. I am grateful to Sigurður Pálsson for carrying out the measurements reported in Experiment 1. The paper has benefited considerably from discussions with Harlan Lane, Joanne Miller, Kenneth Stevens, and from comments on an earlier draft by Joanne Miller, Robert Port, and two anonymous reviewers. Correspondence should be addressed to Jørgen Pind, Institute of Lexicography, University of Iceland, 127 Reykjavík, Iceland (e-mail: jorgen@lexis.hi.is).

process of temporal normalization, whereby the listener may be assumed to set an interior "clock" according to the perceived speech rate, and to adjust the boundary values for temporal phonetic cues accordingly (Nooteboom, 1979; Summerfield, 1981). As an example, consider the fact that VOT values tend to increase with slower speaking rates (Miller, Green, & Reeves, 1986), a finding whose perceptual analog lies in boundary shifts seen in perception. According to a normalization view, the perceptual adjustment could be explained by assuming that with a slower speaking rate, the listener internally moves his or her perceptual boundaries to a longer VOT value. One unresolved issue that has plagued this view concerns the distinction between cue and context (Pind, 1986). Judgments of rate would seem to depend on the judgment of phonetic identity, which in turn depends on judgments of rate, and so on. The theory of rate normalization is crucially dependent upon the possibility of independently extracting rate and using it to drive the normalization process. So far, this issue has not received the attention it deserves.

Theories of speech perception have commonly been divided on the issue of the extent to which it is necessary to assume active processes on the part of the listener—essentially to make up for what are considered ambiguous speech cues—and the extent to which invariants can be found in the speech signal. This line of division also holds for the study of rate-dependent processing, in which a number of investigators have attempted to locate higher-order invariants which would preserve "the essential acoustical qualities of phonetic identity ... despite variations in rate of articulation" (Summerfield, 1981), and thus eliminate the need for a process of normalization (Pind, 1986; Port & Dalby, 1982; Summerfield, 1981).

In the case of temporal speech cues, the interest in higher-order invariants has primarily focused on durational ratios. Most studies of rate-dependent processing have been conducted on English, focusing on the behavior of temporal cues at syllabic onsets, with VOT, as a cue for voicing, and transition duration, as a cue for the /b–w/ distinction, receiving the most attention (Miller, 1981, 1987). The possibility needs to be kept in mind, though, that there is no particular reason to expect all temporal speech cues to behave identically under transformations of rate or, indeed, to expect that "the essential acoustic qualities of phonetic identity" will be comparable in all cases. Although durational ratios are sometimes mentioned as possible higher-order invariants for the perception of syllable-initial contrasts (e.g., of VOT [Green, Stevens, & Kuhl, 1994]), the results so far argue against such invariant durational ratios for VOT.

This by no means implies that durational ratios play no role in the perception of temporal speech cues. In particular, such ratios do seem to play a role in the perception of segments making up the syllable rhyme. Thus, arguments have been made for the importance of durational ratio as a plausible invariant in the perception of stop-consonant voicing in word-medial position in Eng-

lish (Port & Dalby, 1982; but see Massaro & Cohen, 1983, for a different interpretation). Pind (1986, in press) has argued for the role of durational ratios in Icelandic in the perception both of quantity (a contrast involving the durations of vowels and consonants) and of preaspiration (the phoneme [h] inserted between a vowel and closure). The study of quantity (Pind, 1986) showed that the ratio of the perceived-vowel boundary to the duration of the whole-syllable rhyme, $V/(V+C)$, was highly resistant to transformations of rate and came close to the ideal of an invariant cue.¹ It was not absolutely invariant, since extrinsic manipulations (e.g., of precursor rate) did cause small shifts in the phoneme boundaries. These shifts, however, were of an order of magnitude smaller than those seen to accompany intrinsic changes (i.e., changes of the syllable-rhyme duration) which accompany changes of speaking rate. But in those cases, the boundaries stayed close to invariance when expressed in terms of their ratio to the rhyme duration. On the basis of these results, Pind (1986; see also Summerfield, 1981) hypothesized that two mechanisms were involved in rate normalization of quantity—namely, (1) a perception of the intrinsic durational ratios, the defining characteristic of the speech cue; and (2) an extrinsic type of normalization involving, perhaps, an adjustment to a perceptual "clock." The latter effect shows up as minor perturbations in the perceptual ratios defined by the former mechanism. No previous studies have been reported on rate-dependent processing of VOT in Icelandic (or of other temporal speech cues associated with the syllable onset), but it is to be expected that these would show a pattern of results similar to those found in English, since the pattern of VOT for Icelandic is similar to that seen in English. The rate adjustments for VOT in English would typically seem to be of the same order of magnitude as the extrinsic effects seen in the rate adjustments for quantity.

The previous discussion has involved the terms *rate-dependent processing* and *rate normalization*. These concepts need to be kept apart. Rate-dependent processing is the more general of the two, focusing as it does on the established fact that speech perception often shows effects of speech rate, and it is theoretically neutral as to the actual mechanism involved. The concept of normalization has traditionally been used in the context of a particular view of the mechanism involved in rate-dependent processing (i.e., as an active adjustment by the listener to the perceived speech rate). However, this traditional usage of the term can be further sharpened by drawing a distinction between extrinsic and intrinsic normalization, following Summerfield (1981). Extrinsic normalization is synonymous with the traditional view of normalization, which assumes an active listener adjusting his or her perceptual criteria depending on the perceived speech rate. Intrinsic normalization, on the other hand, is exemplified by the kind of relational perception explored in this paper—namely, that involving durational ratios.

In the experiments reported here, two temporal cues—one forming part of the syllable onset (VOT) and the

other making up the syllable rhyme (quantity)—were directly compared for rate-dependent processing. It was hypothesized that the nature of the rate-dependent processing would differ and, in particular, that a ratio-like higher-order invariant would be seen for quantity (relational processing) but not for VOT. To make a meaningful comparison of the extent of normalization for these two speech cues, it was necessary to closely compare production and perception data.

Production studies of rate effects have commonly shown that with slower speech, the contrast between temporally contrasting phonemes increases. Miller and Baer (1983) showed this for transition durations as a cue for the /b–w/ distinction, and Miller et al. (1986) reported a similar finding for VOT. In both cases, the greatest change was seen in the segment cued by the longer transitions /w/ or the longer VOTs (aspirated stops). Thus, in the case of VOT, slower utterance rates will lengthen the VOT for aspirated stops more than the VOT for unaspirated stops, increasing the separation of the two categories.

Miller and Baer (1983) specifically compared their production measurements with the earlier results of Miller and Liberman's (1979) perception experiment, remarking that the "simplest prediction" of the fit between the perceptual and production data would be for the perceptual boundary to be "equidistant from the /b/ and /w/ regression lines." In fact, the shifts seen in the perceptual boundary tend to taper off as syllable durations increase beyond a particular point.

In their study of the behavior of VOTs under transformations of rate, Miller et al. (1986) introduced the notion of a *variable optimal boundary* which is similar to, but slightly different from, the "equidistance" prediction of Miller and Baer (1983). Miller et al. (1986) calculated this variable optimal boundary by sorting their VOTs measured at various utterance rates into 50-msec bins and calculating hypothetical phoneme boundaries that would optimally separate the two categories of unaspirated and aspirated stops for each bin. Miller et al. showed that this optimal boundary increased as utterance rates became slower (as judged by the overall duration of the syllables), from 14 msec of VOT in 100–149-msec long syllables to 52.5 msec of VOT in 650–700-msec long syllables. Typically, the perceptual shifts are considerably more limited.

This notion of a variable optimal boundary is a useful yardstick against which to measure the actual rate-dependent processing seen in perception (as long as the perceptual stimuli are modeled on those underlying the production data), and it will be used in the experiments reported in this paper.

The present paper reports the results of three experiments, the first of which gathered production data on the effects of rate on VOT and on vowel and consonant duration in two-syllable Icelandic words. From these data, variable optimal boundaries were calculated for both the VOT and the quantity contrast. The other two experi-

ments, which were concerned with gathering perceptual data using synthetic stimuli modeled on those used in Experiment 1, tested the hypothesis that rate normalization for VOT is quite different from that for quantity, with only the latter being accomplished through durational ratios.

EXPERIMENT 1

This experiment gathered production data on VOTs, vowel durations, and consonant durations in words of the type used in the perception experiments (Experiments 2 and 3). Before describing it, the relevant phonetic facts about Icelandic will be briefly discussed.

Word-initial Icelandic stops are traditionally described, on the one hand, as being voiceless unaspirated [p, t, c, k], and on the other, as being voiceless aspirated [p^h, t^h, c^h, k^h] (e.g., *bera* [pe:ra], to carry, vs. *pera* [p^he:ra], pear; *galdur* [kaltʏr], magic, vs. *kaldur* [k^haltʏr], cold).

In Icelandic, lexical stress always falls on the first syllable of a word. This syllable is also the only one to show oppositions of quantity in noncompound words. The rules for quantity are rather simple: A vowel is long if followed by one or no consonant; otherwise, the vowel is short. An exception to this rule is that a vowel is long if it precedes one of the following: /p, t, k, s/ + /v, j, r/ (e.g., *risa* [rɪ:sa], giant, accusative singular, vs. *rissa* [rɪs:a], to sketch; *betra* [pe:tra], better, vs. *bedda* [pet:a], bed, accusative singular).

In some cases, the quantity opposition presents itself as a *complementary* opposition of a long vowel followed by a short consonant, and vice versa (e.g., *risa/rissa*). The complementary nature of this opposition was brought out quite clearly in the earliest experimental phonetic studies of Icelandic. Einarsson (1927), for example, found that in two-syllable words of this type, long vowels had an average duration of 190 msec, with the following short stops being on average 151 msec long, whereas for short vowels followed by long stops, the values were 91 and 278 msec, respectively. In this case, the syllable-rhyme duration averaged 341 msec in words of the type V:C (long vowel followed by short consonant) and 369 msec in words of the type VC: (short vowel followed by long consonant). Similar results were reported by Pind (1982), who measured the durations of long vowels, short vowels, and stop consonants in one-, two-, and three-syllable words of the type [pa:k-] and [pak:-]. The words were spoken either singly or in a carrier sentence which was spoken at either a fast or a slow rate. The results showed that the duration of the segments was highly variable and depended both on the number of syllables in the word and on the utterance rate. The mean duration of the long vowels was 161 msec averaged over all conditions of the experiment, while the long consonants had an average duration of 158 msec. The short vowels were on average 105 msec long, and the short consonants averaged 97 msec. The rhyme durations thus totaled 258 msec for V:C syllables, but 263 msec for

VC: syllables. Again, these results quite clearly point to the complementary nature of the quantity contrast in these word types.

Experiment 1 gathered production data for use in the specification of synthetic-speech stimuli used for perceptual tests in Experiments 2 and 3, to enable a closer comparison of speech production and perception to be made.

Method

Stimuli. The stimulus words chosen for this experiment were: *gala* [ka:la], to yell; *galla* [ka:la], overalls, accusative singular; *kala*, [kʰa:la], suffer frostbite; and *Kalla* [kʰa:la], familiar form of the name Karl, accusative.² In these words, a stop with either a short VOT [k] or a long VOT [kʰ] is followed by either a V:C-syllable rhyme (long vowel, short consonant) or a VC:-syllable rhyme (short vowel, long consonant).

Subjects. Four Icelandic subjects—the author and three language students at the University of Iceland—took part in the experiment.

Procedure. To enable the effects of utterance rate on the durations of VOTs, vowels, and consonants to be investigated, the subjects were asked to read the words at five different tempos: normal, faster than normal, as fast as possible, slower than normal, and as slow as possible. A similar magnitude production technique had earlier been used by Miller et al. (1986) to investigate the effects of rate on VOT.

For the reading, a list was drawn up containing a randomized sample of the words, each repeated ten times. The words were printed in columns, one word to a line. The intention was to measure five tokens of each word, although ten tokens of each word were recorded in case misreadings, hesitations, pauses, and other similar phenomena would interfere with the reading.

All of the subjects started by reading the word list at the normal tempo. One subject then read at the two fast rates and then at the two slow rates, while the other three subjects read at the slow rates before they read at the fast rates.

The readings were recorded in a quiet room using a Sennheiser ME 40 microphone connected to a Marantz SD 315 cassette tape deck. The recordings were made with the Dolby noise-reduction system turned off and then transferred to computer disk using the Sensimetrics SpeechStation and its accompanying software. The sampling rate was set at 16 kHz.³ Five good tokens of each word at each rate for each subject were put into separate files, resulting in a total of 400 sound files, each containing one word [2 stops (/g/, /k/) × 2 syllable types (V:C, VC:) × 5 rates × 5 repetitions × 4 subjects].

Measurements were carried out on the spectrograms presented on the screen by the SpeechStation (aided, especially in the case of VOT, by the accompanying waveform displays). For measurement purposes, the following points in each spectrogram were marked: (1) the beginning and end of each word; and (2) within each word, the beginning of the voicing for the vowel. The interval (measured in msec) from the beginning of a word to the onset of voicing gives a measure of the VOT. The boundaries between [a], [l], and [a] were in each case defined as lying in the middle of the formant transitions connecting the steady-state portions of these segments. In the following sections, the terms *vowels* and *consonants* will refer to these acoustic definitions. In particular, the vowel is considered to extend from the beginning of voicing to the middle of the /a/-/l/ transition, while the consonant is seen as extending from that point onward to the middle of the /l/-/a/ transition.

Results

The durations of the individual words ranged from 304 to 1,002 msec. The average word duration was 531 msec

Table 1
Average Word Durations (in Milliseconds) for Each Subject
at Each Speaking Rate in Experiment 1

Rate	Subjects									
	AK		HH		JP		SP		M	SD
	M	SD	M	SD	M	SD	M	SD		
1	524	33	594	25	728	50	883	76	682	146
2	484	34	528	32	604	37	585	52	550	62
3	422	20	507	26	580	44	464	27	493	66
4	453	28	464	19	480	25	460	36	464	29
5	365	23	388	23	395	21	357	31	377	29
M	450	61	496	73	558	120	550	188		

Note—M, mean; SD, standard deviation. Speaking rates are numbered from slowest (1) to fastest (5).

(SD = 128 msec). The subjects were mostly able to pronounce the words at the five different rates (see Table 1). The ratio of the slowest rate to the fastest ranged from 1.44 to 2.47 for the individual subjects. The word durations at the individual rates (numbered 1–5 in Table 1) show an ordered sequence for every subject except Subject AK, for whom the average word duration at Rate 3 is slightly shorter than that at Rate 4.

Measurements of VOT. The average VOT for /g/, collapsed over all conditions and subjects, was 31 msec (SD = 10 msec). For /k/, the average duration of VOT was 80 msec (SD = 22 msec). These figures accord well with the traditional description of these consonants as being voiceless and distinguished by aspiration.

A two-way analysis of variance (ANOVA), with the factors being aspiration (/g/ vs. /k/) and rate (1–5), shows both main effects to be significant [aspiration: $F(1,3) = 203.45, p < 0.001$; rate: $F(4,12) = 9.72, p < 0.001$]. The interaction of aspiration and rate was also significant [$F(34,12) = 7.013, p = 0.004$]. For a consideration of the interaction, it is more profitable to view the VOT durations as a function of actual syllable durations rather than of planned speech rate, since the subjects showed some variability in whether they achieved consistently different rates.

Figure 1 plots the VOT durations against the overall durations of the stressed syllables, measured from the beginning of the stressed vowel to the end of the /l/ (i.e., to the middle of the /l/-to-final-/a/ transition). It is evident that while the stop categories are quite clearly separated at long syllable durations (> 450 msec), at shorter durations, there is some overlap of the categories. The interaction between rate and stop category also emerges quite clearly in this figure, in which rate is seen to have a much greater effect on the VOT of the aspirated stops than on that of the unaspirated stops. This finding is well known from other studies (Miller et al., 1986; Volaitis & Miller, 1992).

Regression lines were fitted to both stop series. There was found to be a correlation between VOT and syllable duration in the aspirated series [$r = 0.81, F(1,198) = 387.19, p < 0.001$]; the slope of the regression line was 0.15. No significant relationship was found for the unaspirated series [$r = 0.08, F(1,198) = 1.4, p = 0.24$].

Figure 1. A scatter diagram showing individual measurements of VOTs for /g/ and /k/. The VOTs are clearly separated at slow utterance rates (with long syllable durations), but there is some overlap at faster rates.

Estimation of optimal VOT boundaries. In the preceding data set, it is possible to estimate a single optimal boundary separating /g/ from /k/ (optimal in the sense of categorizing correctly [as /g/ or /k/] the highest possible number of individual tokens). It turns out that a boundary value of 52 msec gives the lowest number of misses, classifying correctly all but 21 of the 400 tokens, the hit rate being 94.8%. This may be compared with the results for English presented by Miller et al. (1986), who found that a similarly estimated boundary at 23 msec of VOT yielded the highest hit rate (89.8%) for their data set. It is unclear why this difference should exist between the hit rates of the two experiments. Language differences or the lexical status of the words in the present experiment (compared with the one-syllable nonsense items in Miller et al.'s experiment) could conceivably contribute to this difference. At this point, it is probably not worth attaching too much significance to this difference.

In their paper, Miller et al. (1986) also propose a *variable optimal VOT boundary* for their English data by sorting the syllables according to their durations into 50-msec-long bins and then calculating the optimal VOT boundary for each bin. Using this method, Miller et al. found that a variable VOT boundary ranging from 14 msec for the shortest syllables (100–149 msec in duration) to 52.5 msec for the longest syllables (650–700 msec in duration) increased the hit rate to 97.6%, or by almost 10%.

The results of a similar analysis of the present data set are shown in Table 2, in which it can be seen that when a variable-criterion VOT was employed (ranging from 36.9-msec VOT for words having stressed syllables in the 200–249-msec range to 80.9-msec VOT for the longest syllables [800–849 msec]), the number of incorrect clas-

sifications fell to 10. The hit rate thus became 97.5%—an increase of almost 3%.

An interesting question that arises in connection with the present experiment concerns whether the following vowel's phonemic identity (i.e., as long or short) has any effect on the duration of the VOT. It turns out that this effect is negligible: /g/ showed an average VOT of 31 msec ($SD = 10$) before a short vowel, and a VOT of 31 msec ($SD = 10$) before a long vowel, and the corresponding values for /k/ were 79 msec ($SD = 21$) before the short vowel and 80 msec ($SD = 23$ msec) before the long vowel.

Durations of vowels and consonants. In addition to VOT, Experiment 1 also looked at another temporal speech cue—that of duration as a cue to quantity. As mentioned in the introduction, in one type of Icelandic syllable, complementary quantity exists, a long vowel being followed by a short consonant, and vice versa.

The average duration of [a:] was 224 msec ($SD = 87$ msec), whereas [a] had an average duration of 91 msec ($SD = 29$ msec). The average duration of [i:] was 233 msec ($SD = 87$ msec), while that for [i] was 118 msec ($SD = 27$ msec). From these numbers, it is easily calculated that the average rhyme duration (V + C) for the V:C-type words was 342 msec, while that for the VC:-type words was 324 msec. Although there is a significant 18-msec difference in the average rhyme durations [$t(199) = 6.37, p < 0.001$], these data must be considered to fit well with the description of the quantity contrast as being one of complementary opposition.

Individual measurements of vowel and consonant durations have been plotted on a two-dimensional V-C plane in Figure 2, which reveals that there is absolutely *no* overlap between the two categories of syllables, V:C and VC:. The effects of lengthening also emerge quite clearly from this figure, in which it is seen that lengthening a syllable with a long vowel leads to considerable stretching of the vowel but to limited lengthening of the following short consonant. Thus, the V/C ratio changes quite dramatically from 1.31 at the shortest vowel durations to 2.92 at the longer durations (in both cases, the

Table 2
Calculation of a Variable Optimal VOT Boundary

Syllable Duration (msec)	N	N _e	Variable Optimal VOT Boundary	N _{eo}
200–249	18	3	36.9	0
250–299	63	7	41.6	2
300–349	96	6	50.2	5
350–399	85	3	53.9	3
400–449	52	1	46.9	
450–499	36		54.4	
500–549	15		69.1	
550–599	8		64.0	
600–649	8		79.4	
650–699	5		71.7	
700–749	6	1	80.8	
750–799	6		76.1	
800–849	2		80.9	

Note—N_e = number of erroneous classifications (as /g/ or /k/) with a fixed VOT boundary of 52 msec in each bin; N_{eo} = corresponding value using the variable optimal VOT boundary.

Figure 2. The distribution of vowel and consonant durations in V:C- and VC:-type words as a function of utterance rate (1 being the slowest and 5 the fastest), showing the complementary nature of the quantity opposition in Icelandic. Long vowels are followed by short consonants (open symbols) and short vowels are followed by long consonants (closed symbols). The phonemically long member of each pair (V:C or VC:) shows greater stretchability than the short member.

ratios represent the average ratios of the 10 syllables containing the shortest or longest vowels). Note that the results also reveal that the short consonants show *some*, albeit limited, stretchability. For the VC:-type syllables, it is the long consonants that stretch quite dramatically, with the short vowels showing only limited stretchability. Again, the V/C ratio is variable, ranging from 0.44 for the shortest consonant durations to 0.27 for the longest durations (averaged over 10 syllables at each end). It also emerges from Figure 2 that a truly optimal boundary separating the two syllable categories would consist of a line drawn between the two clusters having a slope close to 1 (corresponding to a V/(V+C) ratio of 0.5). I will return to this issue in the General Discussion section at the end of the paper.

One interesting feature is the difference in vowel durations depending on the aspiration of the preceding consonant. The average duration of vowels (both long and short) following /g/ was 170 msec, whereas the average duration of the vowels following /k/ was 145 msec. The difference in vowel durations depending on the previous consonant suggests the hypothesis that the aspiration would, to some extent, be considered part of the vowel when judging quantity. If all of the aspiration contributed to the perception of quantity in the following vowel, one would expect the duration of the first syllable to be approximately the same, whether it starts with /g/ or with /k/. In fact, though, the /ga/ syllables averaged 201 msec, whereas in the /ka/ syllables, the average was 225 msec. However, it should be noted that the /ga-/

words were somewhat shorter (504 msec on average) than the /ka-/ words (522 msec on average). The durational ratio of the initial /ga/ syllable to the whole word is, therefore, 0.40, while that for /ka/ is 0.43. While there is thus some variability, the figures do suggest that the aspiration does, to a considerable extent, influence the perception of duration.

Discussion

In this experiment, it was found that the duration of temporal speech cues is highly susceptible to the effects of rate. However, another interesting finding was that under these transformations of speech rate, some aspects of the speech cues remain relatively invariant; the variability seen in VOT as a function of speech rate, for example, is primarily to be found in the aspirated member [k^h] of a cognate stop pair, not in the unaspirated member [k]. Indeed, this would seem to indicate that the VOT region occupied by the aspirated member puts severe limits on the stretchability of the VOT of the unaspirated member, since otherwise there would be considerable overlap between the two categories—overlap, that is, which would exceed the rate normalization typically seen in studies of VOT perception. For the aspirated member, however, no such upper limiting factor exists, thus allowing its VOT to increase with slower rates of speech.

Very similar behavior can be seen in duration as a speech cue for quantity. In this case, it is the phonemically short member that is highly restrained, and the phonemically long member that has considerable latitude in its stretchability. In Icelandic, it can be seen in the present data that, with syllables showing complementary duration, *no* overlap occurs between the syllables V:C and VC:.

While the VOT and duration speech cues showed some interesting parallels in this experiment, it is also worth noting the differences between them. At least in the present context of Icelandic V:C and VC: syllables, duration as a cue for quantity is clearly relational, a long vowel being followed by a short consonant, and vice versa. In this case, there is not much reason to doubt that the listener is readily able to use this relational property in perception (Pind, 1986). In the case of VOT, however, the use of a relational cue—of VOT to the following vowel, for example—seems unlikely.

EXPERIMENT 2

Experiment 1 showed that VOTs for velar stops are variable in words spoken at different rates, and especially so for aspirated stops, where VOT was shown to be stretchable with differences in rate.

These results might be taken as evidence of a variable optimal boundary between the two cognate stop categories, as argued for by Miller and her associates (Miller et al., 1986; Miller & Volaitis, 1989). Miller and Volaitis (1989) found that the phoneme boundary, separating /b/ from /p/ in perceptual experiments, does change with

changes in the overall syllable duration. With a 125-msec-long /bi-pi/ continuum, the boundary was found to lie at 35.61 msec, but in 325-msec-long syllables, it was found to be at 43.89 msec. Notice that these shifts in the location of the phoneme boundaries, although in the predicted direction, are appreciably smaller than those predicted by Miller et al.'s (1986) calculation of a variable optimal boundary that moves from 14 msec of VOT (in 100–149-msec-long syllables) to 27.5 msec (in 300–349-msec-long syllables).⁴ This discrepancy is noticeable and should cast considerable doubt on the theory that listeners are using a strategy of optimal-boundary calculation in their perception of VOT. At the very least, it shows that the boundary shifts are constrained—though why they should be so constrained has not so far been the focus of great interest.

While the data presented in Experiment 1 show changes in the location of a similarly calculated optimal boundary, they also show that such a change does not greatly increase the number of correct classifications made by the imaginary listener, since they only rise from 94.8% to 97.5%. In Experiment 2, listeners were presented with a continuum of synthetic-speech stimuli having variable VOTs. These stimuli were modeled on the stimuli measured in Experiment 1, having three different overall durations (corresponding to three different speaking rates). Additionally, their phonemic makeup was varied, so that stimuli were either of the V:C or the VC: type.

Specifically, Experiments 2 and 3 tested the hypothesis that in Icelandic, both VOT and quantity show rate-dependent processing, though to different extents. On the basis of previous studies of rate-dependent processing of VOT in English, it was further hypothesized that in Icelandic, normalization seen for VOT would be limited, in the sense that it would fall appreciably short of the optimal boundaries previously calculated, and thus that it would not show any evidence of relational processing. This prediction was tested in Experiment 2, while Experiment 3 tested the hypothesis that for quantity, the normalization would be much more extensive—so extensive, in fact, that it might best be described as relational processing.

Method

Stimuli. Experiment 2 used synthetic speech made with the Sennheiser SenSyn synthesizer, a version of the Klatt cascade/parallel formant synthesizer (Klatt, 1980; Klatt & Klatt, 1990). The synthesizer was run in the cascade configuration. Six stimulus continua were made by independently varying two parameters, one of which was the overall duration of the stimuli, which was set to either 605 msec (slow rate), 460 msec (normal rate), or 360 msec (fast rate). The other parameter was that of phonemic makeup, with three stimulus continua (one at each rate) of the type V:C (long vowel followed by short consonant) and three of the type VC: (short vowel followed by long consonant).

The stimuli were modeled on the words measured in Experiment 1, being tokens of the words *gala* and *galla*. All continua had variable VOT for the word-initial velar stop, ranging, in 5-msec steps, from 10 msec to 60 msec. The VOT continua contained 11 steps. The total number of stimuli in the experiment thus amounted to 66 (i.e., 3 [word durations] $\times$ 2 [syllable types] $\times$ 11 [VOT steps]).

The different syllable types were synthesized by manipulating the lengths of the [a] and the [l] segments in the first syllable. In the same way as in Experiment 1, the vowel was equated with the voiced portion, and thus started at time = 10 msec in the syllable with the shortest VOT. The ratio of the steady state of the [a] to the steady state of the [l] was kept approximately constant, at 0.62 for the V:C-type words, and at 0.39 for the VC:-type words. With this criterion in mind, the different series had the following durations for [a] and [l]: In the V:C syllables, ranging from the slowest rate to the fastest, 340 + 185 msec, 250 + 130 msec, and 190 + 90 msec (the rhymes [V + C] thus ranged from 525 msec at the slowest rate through 380 msec at the normal rate to 280 msec at the fastest rate), and in the VC: syllables, again from the slowest to the fastest rate, 225 + 300 msec, 170 + 210 msec, and 130 + 150 msec (the rhyme durations were, of course, identical to those in the VC:-type words).

The steady-state formants of the vowel [a] had the following values: F1 = 750 Hz; F2 = 1280 Hz; and F3 = 2450 Hz. For [l], the steady-state values were: F1 = 440 Hz, F2 = 1150 Hz, and F3 = 2550 Hz. The values of F4 and F5 were held steady throughout the utterances, at 3250 and 3700 Hz, respectively. The transitions for the word-initial velar stop were made in the following manner: F1 started at 200 Hz and rose over 45 msec to the target value of 750 Hz; F2 started at 1800 Hz and fell over 55 msec to the target value of 1280 Hz; and F3 started at 2000 Hz and rose to the target value of 2450 Hz in 65 msec. The first 10 msec of the stimuli consisted of a noise burst centered at 1800 Hz.

The transitions from [a] to [l] were 20 msec long, with a linear fall of 750–440 Hz and 1280–1150 Hz for F1 and F2, respectively, and a linear rise of 2450–2550 Hz for F3. The transitions from [l] to [a] were an exact mirror image of those transitions from [a] to [l]. Additionally, the overall amplitude of the voiced waveform was 5 dB lower in the [l] than it was in the preceding vowel. The second vowel of each stimulus was 70 msec long (counting from the middle of the [l]–[a] transition), was identical in all series, and had formants identical to those of the first [a].

The fundamental frequency of each stimulus was fixed at 100 Hz, declining linearly to 95 Hz over the last 20 msec of each stimulus. The VOT continua were made by replacing the voiced excitation with noise excitation and by increasing the bandwidth of F1 from 90 Hz (the default) to 200 Hz. The utterances were made with a slightly breathy sound (synthesizer parameter AH = 20 dB). Figure 3 shows schematic diagrams of three of the stimuli used in this experiment.

The stimuli were synthesized at 11025 Hz with 14-bit quantization on a PC-compatible computer. The file format was then changed to Microsoft Multimedia-defined wav-files for playback on a Turtle Beach Multisound card.

The stimuli were recorded on two tapes, one of which contained the V:C-type stimuli (*gala-kala*), and the other the VC:-type stimuli (*galla-Kalla*). At the beginning of the tape, the 33 stimuli were each played once in randomized order as a practice test. This was followed by 5 blocks of 2 $\times$ 33 stimuli in randomized order. The interstimulus interval was 3 sec. A short tone of 1000 Hz was inserted after every 22 trials to make it easier for the subjects to keep track on the response sheets with which they were provided.

Subjects. The subjects were 10 Icelandic students at the University of Iceland. Three of them had earlier taken part in Experiment 1. All reported normal hearing. They were paid for their participation in the experiment.

Procedure. Five subjects listened to the *gala-kala* tape followed by the *galla-Kalla* tape; the order was reversed for the other five subjects. The testing took place in a quiet room. The subjects listened to the stimuli, which were played at a comfortable listening level over Sennheiser HD-530-II circumaural headphones.

The subjects were provided with response sheets which contained two fields for each stimulus presented. In the V:C test, one field was marked with the word *gala* and the other with the word *kala*. In the VC: test, one field was marked with the word *galla* and

Figure 3. Schematic drawings illustrating three of the stimuli used in Experiment 2. The top diagram shows a fast word of the V:C type with a 10-msec-long VOT; the middle diagram shows a VC:-type word of normal duration having a 35-msec VOT; and the bottom diagram shows a slow word of type V:C with a 60-msec VOT. In each case, the lines illustrate the values for the first three formants. All stimuli start at time = 100 msec. The top diagram also shows the boundaries of [a] and [l] in the stimuli.

the other with *Kalla*, using the normal Icelandic orthography. The subjects were instructed to mark the appropriate box for each stimulus presented and to guess if they were not sure which response was appropriate. They did not report any difficulty in carrying out this task. As mentioned above, the first 33 trials of each run were used for familiarization purposes, so for each stimulus, a total of 10 responses was tabulated for each subject.

Results

Pooled identification curves for all 10 subjects are presented in Figure 4, in which it can be seen that the curves for both V:C- and VC:-type words are very similar.

Phoneme boundaries for individual subjects were calculated using the method of probits. These results are shown in Table 3, in which it can be seen that the phoneme

boundaries shifted in the expected direction with changes in rate. On average, the phoneme boundaries are located at 37.83 msec for the stimuli at the slow rate, at 37.31 msec for the stimuli at the normal rate, and at 36.41 msec for the stimuli at the fast rate. A two-way repeated measures ANOVA (word type $\times$ rate) shows significant effects of both rate [$F(2,18) = 7.346, p = 0.005$] and word type [$F(1,9) = 5.696, p < 0.05$]. The interaction of these factors was not significant [$F(2,18) = 0.151$].

The main effects, though significant, are small. The different rates (with syllable durations ranging from 280 to 525 msec or over a 1:1.9 range) only shift the phoneme boundaries for VOT by 1.4 msec. As for the effect of the phonemic makeup of the syllable, the average VOT boundary is located at 37.80 msec in the V:C series, and at 36.56 msec in the VC:-series. Again, the shift is of comparable magnitude, at 1.24 msec.

Figure 4. Pooled identification curves for Experiment 2 showing the percentage of /k/ responses as a function of VOT. The top graph shows the results for V:C-type words with different overall durations, while the bottom graph shows comparable results from VC:-type words.

Table 3
Location of Average Phoneme Boundaries for the
Perception of the Unaspirated/Aspirated Contrast in
Icelandic (in Milliseconds of VOT) in Experiment 2

Rate	V:C Words	VC: Words	<i>M</i>
Slow	38.35	37.31	37.83
Normal	38.09	36.52	37.31
Fast	36.97	35.84	36.41
<i>M</i>	37.80	36.56	

Discussion

The results of Experiment 2 show that overall syllable duration, as well as the quantity of the following vowel, have significant, though quite small, effects on the perception of VOT in Icelandic. The first of these findings replicates a well-known finding in studies of the perception of VOT in English (Miller & Volaitis, 1989; Summerfield, 1981; Volaitis & Miller, 1992).

The effects found in the present experiment are quite small, and considerably smaller than those found by, for example, Miller and Volaitis (1989) or Volaitis and Miller (1992). A plausible reason for this may lie in the difference between the overall durations of stimuli used in those studies and those used in the present experiment; Miller and Volaitis (1989) used shorter (one-syllable) stimuli, 125 and 325 msec long, whereas the durations of the stressed syllables in the present experiment ranged from 280 to 525 msec. Summerfield (1981) has shown that increasing the length of stimulus words leads gradually to a diminishing shift in the phoneme boundaries. The fact that the perceptual normalization for rate tapers off in this manner, contrary to the behavior of VOTs in production, is of considerable interest, since it speaks against a simple "clock" type of normalization. Presumably, such a process would a priori be predicted to set its clock in a manner illustrated above by the optimal-boundary calculations. These optimal boundaries do not in fact do a particularly good job of predicting the shifts actually observed in the VOT boundaries. This is considered further in the General Discussion section of this paper.

The effect of the nature of the word (i.e., V:C or VC:) on the location of the phoneme boundaries also merits consideration. It shows that longer values of VOT are needed for the perception of an aspirated stop in a word starting with a phonemically long vowel. Experiment 1, though, showed no such effects in production.

EXPERIMENT 3

In Experiment 1, very considerable variability was found in the duration of vowels and consonants in stressed syllables with changes in speaking rate. While phonemically short segments underwent some rate-dependent changes in duration, those changes were much more limited than those seen in the phonemically long segments, in which this variability was especially noticeable.

Yet it was also found in Experiment 1 that this variability of individual segments did not blur the contrast

between the two syllable types, V:C and VC:, which were kept distinct through the durational ratios of the vowels and consonants. The perceptual reality of such a ratio was tested in Experiment 3. It was hypothesized that the rate normalization for quantity would follow invariant ratios of vowel to rhyme duration (see Pind, 1986). Note that, in order for the ideal of relational processing to prevail, this would involve quite extensive shifts in the phoneme boundaries, when expressed in terms of vowel duration at the boundaries. This contrasts with Experiment 2, in which limited boundary shifts were both predicted and found.

Method

Stimuli. The stimuli were based upon the stimuli used in Experiment 2 and, like them, were made with the SenSyn version of the Klatt synthesizer. In Experiment 3, six stimulus continua were made by independently varying two parameters, one of which was the overall duration of the stimuli, which, as in Experiment 2, was set to either 605 msec (slow rate), 460 msec (normal rate), or 360 msec (fast rate). The other parameter, that of the word-initial VOT, was varied by one series of stimuli beginning with /g/ (VOT = 10 msec), and the other with /k/ (VOT = 60 msec).

The stimuli were again modeled on the words measured in Experiment 1, being tokens of either the word *gala* or the word *kala*. The variable of interest in this experiment was the perception of the quantity opposition. To investigate this, the durations of [a] and [l] were systematically varied. The duration of the syllable rhymes in the *gala* series, counting from the offset of the word-initial aspiration to the middle of the [l-a] transition at the beginning of the second syllable, was 525 msec at the slow rate, 380 msec at the normal rate, and 280 msec at the fast rate. These values are identical to those in Experiment 2.

The stimulus sets consisted of 12 stimuli at the slow rate, 11 at the normal rate, and 9 at the fast rate. At the slow rate in the *gala-galla* series, the duration of the [a] and [l] ranged from 325 + 200 msec to 215 + 310 msec in 10-msec steps. For the *kala-Kalla* series, the range was from 275 + 200 msec to 165 + 310 msec (the voiced vowel was 50 msec shorter in these stimuli than in the *gala-galla* series). At the middle rate, the durations for [a] and [l] ranged from 260 + 120 msec to 160 + 220 msec (*gala-galla* series) and from 210 + 120 msec to 110 + 220 msec (*kala-Kalla* series). Finally, at the fast rate, the durations of the [a] and [l] ranged from 200 + 80 msec to 120 + 160 msec (*gala-galla* series) and from 150 + 80 msec to 70 + 160 msec (*kala-Kalla* series). In all of these series, the duration manipulation consisted of shortening the [a] in 10-msec steps and lengthening the [l] by the equivalent amount for each step of the continuum. Figure 5 shows more clearly the location of the stimuli with a 10-msec VOT in a two-dimensional V-C plane.

The steady-state formants, the [a]-[l] and [l]-[a] transitions, and the word-initial transitions for the velar stop were synthesized in the manner described for Experiment 2. The fundamental frequency contour was also the same. Again, the stimuli were synthesized with a slightly breathy voice. The second vowel of each stimulus was 70 msec long (counting from the middle of the [l]-[a] transition), and was identical in all of the series.

The stimuli were synthesized at 11025 Hz with 14-bit quantization on a PC-compatible computer and played through a Turtle Beach Multisound card onto two tapes, one containing the *gala-galla* series, and the other the *kala-Kalla* series. At the beginning of each tape, the 32 stimuli were each played once in randomized order as a practice test. This was followed by 5 blocks of 2 × 32 stimuli in randomized order. The interstimulus interval was 3 sec. A short tone of 1000 Hz was inserted after every 20 trials to make it easier for the subjects to keep track on the response sheets with which they were provided.

Figure 5. The durations of vowels and consonants in the three stimulus series used in Experiment 3. Each dot corresponds to vowel and consonant durations in a single stimulus word. Compare this figure with Figure 2, showing the measured durations. The stimuli were chosen such that they would span the interval separating the two types of words (V:C and VC:).

Subjects. The subjects were the same as those who had participated in Experiment 2. They were paid for their participation in the experiment.

Procedure. Five of the subjects listened to the *gala-galla* series followed by the *kala-Kalla* series; the order was reversed for the other five subjects. The testing took place in a quiet room. The subjects listened to the stimuli, which were played at a comfortable listening level, through Sennheiser HD-530-II circumaural headphones.

The subjects were provided with response sheets which contained two fields for each stimulus presented. In the *gala-galla* series, one field was marked with the word *gala* and the other with *galla*, and in the *kala-Kalla* series, one field was marked with *kala* and the other with *Kalla*. The subjects were instructed to mark the appropriate box for each stimulus presented and to guess if they were not sure which response was appropriate. Some subjects mentioned that they had some difficulty distinguishing between the two series at the slow rate. Since the first 32 trials of each run were used for familiarization purposes, a total of 10 responses was tabulated for each subject.

Results

Pooled identification curves for all 10 subjects are presented in Figure 6, in which the percentage of V:C-type responses is plotted as a function of the *voiced-vowel duration*, calculated from the offset of VOT to the onset of [l].

Both graphs in Figure 6 reveal that the identification curves produced by the stimuli at the slow rate were not as sharp as those yielded by the other two rates; one subject in particular had some difficulty hearing these stimuli presented at the slow rate as V:C-type words.

Phoneme boundaries for individual subjects were calculated using the method of probits. These results are shown in Table 4, in which it can be seen that the pho-

neme boundaries, expressed in terms of voiced-vowel durations, show massive shifts, in the expected direction. Both main effects were found to be significant in a two-way repeated measures ANOVA (word-initial VOT $\times$ rate).

The effect of word-initial VOT is highly significant [$F(1,9) = 268.56, p < 0.001$]. The phoneme boundaries are located on average at 206.54 msec in the words starting with /g/, but at 163.05 msec in the words starting with /k/. The effect of rate is also highly significant [$F(2,18) = 143.08, p < 0.001$], and can be seen in the movement of the phoneme boundaries from 249 msec of vowel duration at the slow rate to 128.21 msec at the fast rate. The interaction of word-initial VOT and rate was not significant [$F(2,18) = 1.097, p = 0.355$]. The fact that word-initial VOT led to differences in the location of the vowel quantity boundaries is of interest; note that this differ-

Figure 6. Pooled identification curves for Experiment 3 showing the percentage of V:C responses (*gala* or *kala*) as a function of vowel duration. The top graph shows the results for the *gala-galla* series, while the lower graph shows the results for the *kala-Kalla* series. Notice that in both cases, the vowel is equated with the voiced portion of the stimuli, measured from the end of VOT.

ence averages 43.5 msec over all conditions of the experiment. The difference in word-initial aspiration in the two series—*gala-galla* and *kala-Kalla*—is 50 msec, showing that most of the aspiration is considered as part of the vowel when judgments of quantity are called for—in fact, all but 6.5 msec of the aspiration seems to be ascribed to the vowel in judgments of quantity.

While the phoneme boundaries for vowel quantity shift considerably as a function of syllable-rhyme duration (i.e., utterance rate), this is to be expected. If these phoneme boundaries are expressed instead in terms of a $V/(V+C)$ ratio (of phoneme boundary to syllable-rhyme duration), these effects are diminished. In Table 5, which shows average values for these ratios, it can be seen that the average of the ratios for the two series (/g/ vs. /k/) is 0.50, regardless of the speaking rate. If the individual series are examined, however, with the faster rate, the ratio is seen to rise for the *gala* series, but to drop for the *kala* series. A two-way repeated measures ANOVA ($VOT \times$ rate) shows that the effect of VOT on the ratios is significant [$F(1,9) = 58.713, p < 0.001$], but that the effect of rate is nonsignificant [$F(2,18) < 1$]. The interaction of VOT and rate is significant [$F(2,18) = 18.821, p < 0.001$].

These ratios, however, depend crucially on the way they are calculated. We have seen earlier that aspiration contributes significantly to the judgment of quantity. Another way of calculating the ratios would take that effect into account by shifting the phoneme boundaries in the aspirated series up by 43.5 msec, the value which was calculated above as the contribution of the aspiration to the quantity judgments. The average ratios calculated with these “shifted” values are shown in Table 6, in which a much closer correspondence between the values can be seen. These values were submitted to a two-way repeated measures ANOVA ($VOT \times$ rate), in which neither main effects nor interactions were significant [VOT,

Table 6
Average Ratios of Shifted Phoneme Boundaries to Syllable Rhyme Durations in Experiment 3

Rate	Series		<i>M</i>
	<i>gala</i>	<i>kala</i>	
Slow	0.51	0.52	0.52
Normal	0.53	0.52	0.52
Fast	0.54	0.53	0.54
<i>M</i>	0.53	0.53	

$F(1,9) < 1$; rate, $F(2,18) < 1$; interaction, $F(2,18) = 1.465, p = 0.257$].

Discussion

The results of Experiment 3 show that the phoneme boundaries for duration as a cue to quantity are very much dependent on the overall duration of the syllable rhyme and thus on the utterance rate. They also show quite clearly that the durational ratio of vowel to syllable-rhyme duration remains constant, or nearly so, under quite severe rate transformations, thus confirming the earlier findings of Pind (1986).

A secondary finding of some interest concerns the contribution made by aspiration to the perception of quantity of the following vowel. In connection with the findings of Experiment 1, the question was raised as to whether words starting with /ga/ and /ka/ would show the same gestural patterns with the ratios of the initial /ga/ or /ka/ syllables to the overall word durations being constant. It turned out that although the ratios were not quite constant, they were approximately so. The same finding emerged from Experiment 3, in which the aspirated portion of the aspirated stops contributed significantly to the quantity judgment.

GENERAL DISCUSSION

The present paper has looked at the behavior of two temporal speech cues in Icelandic and at the effects of differences in speech rate on their production and perception. In a production experiment, it was shown that both VOT, as a cue for aspiration in word-initial stops, and segment duration, as a cue for quantity, are influenced by speaking rate. The production experiment revealed interesting commonalities, as well as differences, in the behavior of these two speech cues. In both cases, it is possible to identify one member of the binary contrast as having much more freedom than the other in terms of durational variation. In the case of aspiration, this is the aspirated member of a two-valued contrast (non-aspirated vs. aspirated, [k] vs. [k^h]). A considerable change in the speaking rate reveals little variation in the VOTs for the unaspirated member of the pair, [k], whereas VOTs for [k^h] increase in duration with slower speaking rate. In the case of segment duration as a cue to quantity in Icelandic, to some extent the findings are comparable, the phonemically long segments showing a much greater effect of rate than do the phonemically short members.⁵

Table 4
Location of Average Phoneme Boundaries Separating Short and Long Vowels (in Milliseconds of Voiced-Vowel Duration) in Experiment 3

Rate	Series		<i>M</i>
	<i>gala</i>	<i>kala</i>	
Slow	269.27	228.73	249.00
Normal	199.84	154.54	177.19
Fast	150.52	105.90	128.21
<i>M</i>	206.54	163.05	

Table 5
Average Ratios of Phoneme Boundaries to Syllable-Rhyme Duration in Experiment 3

Rate	Series		<i>M</i>
	<i>gala</i>	<i>kala</i>	
Slow	0.51	0.48	0.50
Normal	0.53	0.47	0.50
Fast	0.54	0.46	0.50
<i>M</i>	0.53	0.47	

Note—These ratios have been calculated using the phoneme boundaries expressed in terms of voiced-vowel duration (see Table 4).

In this case, however, it needs to be borne in mind that the durational contrast involves simultaneous contrast of both vowel *and* consonant duration, since the listener always hears both segments.

In both types of speech cue, it is possible to calculate an optimal boundary between the two categories, following the example of Miller et al. (1986). This was done for the VOT contrast in Table 2. The values from Table 2 have been plotted in Figure 7 (top), and a regression line fitted through the points. Similarly estimated optimal-boundary values for the quantity contrast have been plotted in Figure 7 (bottom), and a line fitted to the data. These curves can be seen to be quite similar. In addition, the actual phoneme boundaries, as found in Experiments 2 and 3, have been plotted, and here a striking difference in the behavior of these two temporal speech cues emerges. The phoneme boundaries for VOT evidently do not adhere to any significant degree to the calculated optimal boundaries, whereas those for the perception of quantity do. It must therefore be concluded that different mechanisms are involved in the perception of the two cues. In particular, it is clear that the perception of the quantity contrast is relational and based on the ratio of vowel and consonant duration. While some have speculated as to the possibility of such relational cues for voicing (Green et al., 1994; Summerfield, 1981), the specifications of such cues have so far proven elusive, and the present results definitely argue against any role for a VOT-to-syllable ratio in perception (as suggested by Green et al., 1994).⁶ In fact, even though the present data show a rate-dependent effect on VOT, it is so small as to be confined to the region of uncertainty separating the two phonetic categories, and is thus similar to the effects of extrinsic normalization reported by Pind (1986) for the perception of quantity.

The very marked rate adaptation of vowel and consonant duration might at first sight be considered to support the case for rate normalization in the perception of duration—how else will the speaker know whether a particular vowel of average duration is a phonemically short vowel spoken at a slow rate or a phonemically long vowel spoken at a fast rate? In fact, however, it is not at all necessary in this case to appeal to the notion of rate normalization—at least not in the original extrinsic sense of a listener adjusting for rate by setting a “clock” based on speaking rate and using that to readjust his or her boundaries. The speech cue of quantity is defined by the durational relationship of two segments; this relationship is therefore intrinsic to the speech cue (to borrow Summerfield’s [1981] terminology), and automatically encompasses the major rate effects (i.e., the changes in syllable-rhyme duration). In other words, rate adjustments for this particular speech cue do not entail any mechanisms other than those used in the perception of the speech cue itself, and the cue for quantity is thus *self-normalizing* with respect to rate. The case of rate adjustments for VOT is clearly different, which implies that at least two mechanisms are operative in rate-dependent speech processing.

Figure 7. Summary of the results of the perceptual experiments. The top graph contrasts the results of Experiment 2 with the optimal boundaries calculated to separate the two categories of unaspirated and aspirated word-initial stops. The actual results are averaged over both conditions of the experiment, since there was very little difference between them. The results for the perceptual experiment are noticeably different from the calculated optimal boundaries. Quite a different pattern of results is shown in the bottom graph, in which data from Experiment 3 are compared with an optimal boundary separating the two quantity categories, V:C and VC:, depicted in Figure 2. Here, the slopes of the phoneme boundaries in both conditions of the experiment are quite similar to the slope of the calculated optimal boundary.

At this stage, it is not possible to give a definitive answer as to why these speech cues should show such different patterns of rate-dependent processing. One obvious difference between them lies in their specifications. In Icelandic, quantity is a speech cue defined over the syllable rhyme, and is a prosodic cue involving two segments. Since it is the durational relationship between these two segments which defines the quantity of the syllable, relational processing of the speech cues is quite simply the optimal strategy that the listener can employ. The results of Experiment 3 show that this is indeed the strategy employed by listeners. Although rate effects in the per-

ception of VOT have been extensively investigated in English, no consensus has so far been reached on whether the mechanism employed by listeners is speech specific (Miller, 1987; Summerfield, 1981) or auditory (Diehl & Walsh, 1989). One possible explanation for the differences seen in the present study in the rate-dependent perception of VOT and quantity is that they involve two different levels of processing, an auditory level and a linguistic level.

Such a theory would also, however, need to explain why, in Icelandic, the perception of voice offset time (VOFT, an [h]-like sound inserted, as a cue for preaspiration, between a stressed vowel and the following closure) is also, to a large extent, relational, as shown by a relatively constant h/V ratio at the perceptual boundaries in utterances of different durations (Pind, in press). With voicing giving way to aspiration before the closure, VOFFT is a mirror image of VOT and would thus, at first sight, be expected to behave similarly to VOT with respect to rate-dependent processing. But in fact, VOFFT shows extensive self-normalization, in a similar way to quantity, even though it is not a prosodic cue. Considerations of quantity have some effect on the perception of preaspiration in Icelandic, insofar as syllables with long vowels need more VOFFT for perception of preaspiration (which, linguistically, is always preceded by a short vowel) than would be predicted by a strictly relational view (Pind, in press). Accounting for this effect still leaves the rate-dependent processing of VOFFT more similar in nature to that of quantity than to that of VOT. This is of interest since, as speech cues, VOFFT and quantity both belong to the syllable rhyme, whereas VOT belongs to the onset. The syllable rhyme shows much more durational variability than the onset (see Figures 1 and 2), and it is thus perhaps not surprising that rate-dependent processing for speech cues belonging to the rhyme should in fact differ from those belonging to the onset. It is, nonetheless, of considerable interest that an invariant for quantity is both suggested by the production data and borne out by the results of the perceptual experiments. So far, the specification for such an invariant for VOT remains elusive, suggesting that should any invariant for VOT be found, it would not be relational.

Obviously, further studies are needed to shed more light on these issues. Currently, we are left with results showing markedly different rate-dependent processing of the two speech cues of VOT and quantity. Theories of rate-dependent speech perception need to address this discrepancy.

REFERENCES

- BANNERT, R. (1979). The effect of sentence accent on quantity. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the Ninth International Congress of Phonetic Sciences* (Vol. 2, pp. 253-259). Copenhagen: Institute of Phonetics.
- DIEHL, R. L., & WALSH, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.
- EINARSSON, S. (1927). *Beiträge zur Phonetik der Isländischen Sprache* [Contributions to the phonetics of the Icelandic language]. Oslo: A. W. Brøgger.
- FOWLER, C. A. (1990). Sound-producing objects as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, **88**, 1236-1249.
- GANONG, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 110-125.
- GREEN, K. P., STEVENS, E. B., & KUHLM, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, **55**, 249-260.
- KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.
- KLATT, D. H., & KLATT, L. C. (1990). Analysis, synthesis, and perception of voice quality among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.
- LEHISTE, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- LISKER, L. (1974). On time and timing in speech. In T. A. Sebeok (Ed.), *Current trends in linguistics* (Vol. 12, pp. 2387-2418). The Hague: Mouton.
- MAGEN, H. S., & BLUMSTEIN, S. E. (1993). Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics*, **21**, 387-409.
- MASSARO, D. W., & COHEN, M. M. (1983). Consonant/vowel ratio: An improbable cue in speech. *Perception & Psychophysics*, **33**, 501-505.
- MILLER, J. L. (1981). Effects of speaking rates on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.
- MILLER, J. L. (1987). Rate-dependent processing in speech perception. In A. W. Ellis (Ed.), *Progress in the psychology of language* (Vol. 3, pp. 119-157). Hove, U.K.: Erlbaum.
- MILLER, J. L., & BAER, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America*, **73**, 1751-1755.
- MILLER, J. L., GREEN, K. P., & REEVES, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, **43**, 106-115.
- MILLER, J. L., & LIBERMAN, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.
- MILLER, J. L., & VOLAITIS, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.
- NOOTBOOM, S. G. (1979). Complex control of simple decisions in the perception of vowel length. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the Ninth International Congress of Phonetic Sciences* (Vol. 2, pp. 298-304). Copenhagen: Institute of Phonetics.
- PIND, J. (1982). *The perception of quantity and preaspiration in Icelandic*. Unpublished doctoral thesis, University of Sussex, Brighton.
- PIND, J. (1986). The perception of quantity in Icelandic. *Phonetica*, **43**, 116-139.
- PIND, J. (in press). Constancy and normalization in the perception of voice offset time as a cue for preaspiration. *Acta Psychologica*.
- PISONI, D. B., CARRELL, T. D., & GANS, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.
- PORT, R. F., & DALBY, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, **32**, 141-152.
- SUMMERFIELD, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.
- VOLAITIS, L. E., & MILLER, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

NOTES

1. The rhyme is defined as the syllable less any initial consonants. Thus, in an English word such as "bracket," the initial rhyme is "ack."

The notion of a V/(V+C) ratio is from Bannert (1979), who used it in phonetic investigations of languages having quantity systems similar to that of Icelandic, with long vowels followed by short consonants and vice versa. This particular ratio has no magical properties (e.g., for each V/(V+C) ratio, it would also be possible to calculate a V/C ratio).

2. Two considerations determined the choice of stimulus words for the present series of experiments. First, to eliminate any possible effect of lexical status on perception (see Ganong, 1980), artificial words were eschewed. Second, since previous studies of the perception of quantity in Icelandic had used stops and fricatives (Pind, 1986), liquids were chosen here to further extend the class of consonants tested. The prediction was that consonant class would not in fact make a difference to the results.

3. The SpeechStation card includes its own anti-aliasing filter which is automatically set for the sampling frequency. The same is true for the Multisound card used in Experiments 2 and 3.

4. Miller et al. (1986) in fact calculate a "variable optimal VOT boundary" for syllables up to 700 msec long (at which upper limit the variable boundary is found to lie at 52.5 msec). In the perceptual studies (Miller & Volaitis, 1989; Volaitis & Miller, 1992), however, only

syllables with relatively short durations were used. These typically show the greatest rate-dependent effects.

5. This may be contrasted with the results of Magen and Blumstein (1993), who found phonemically short vowels in Korean to be affected similarly to phonemically long vowels by changes of speaking rate (i.e., with slower speaking rate, the short vowels were lengthened to approximately the same degree as their long counterparts).

6. In the present experiments, which used word-initial VOTs and isolated words as stimuli, any ratio effects observed would derive from the part of the stimuli following the word-initial VOT. In natural, connected speech, however, other contextual factors come into play (e.g., the duration of the closure preceding the VOT), whose effects might to some extent be mediated relationally. See Summerfield (1981) for further discussion, and Pind (in press) for an example of the role of speech-segment ratios in the perception of Voice Offset Time (VOftT) as a cue for preaspiration.

(Manuscript received January 21, 1994;
revision accepted for publication September 15, 1994.)