BNAM 2018
Baltic-Nordic Acoustics Meeting
15-18 April 2018
Harpa, Reykjavík, Iceland

# The effect of speaking rate and vowel quantity on the perception of voice-onset-time in Icelandic

Jörgen L. Pind, Gyða Elín Björnsdóttir, Einar Þór Haraldsson, Ásdís Jónsdóttir, Emilie Anne Jóhannsdóttir Salvesen, Elísabet Ólöf Sigurðardóttir

Faculty of Psychology, University of Iceland, Reykjavík IS-101, jorgen@hi.is

Speech is a variable signal. In the temporal domain, speech has shown itself to be highly elastic. This elasticity is caused e.g. by differences in speaking rates. At faster rates, speech segments shorten, while at slower durations they lengthen. This durational variability presents a problem for perceivers. Consider thus a speech cue such as voice-onset-time (VOT). VOT is the time from the opening of the mouth for a stop to the onset of the voicing in the following vowel. In monosyllables such as 'ba' and 'pa', the VOT is short (20–30 ms) in the former, and long (50–60 ms) in the latter syllable. However, the duration of VOT is also influenced by the speaking rate, so that no absolute boundaries in terms of the duration of VOT can be found to distinguish between 'b' and 'p'. One theory holds that listeners "take account of" speaking rate in perception, and thus adjust for different speaking rates (exhibiting what is usually termed perceptual normalization). Before a short vowel, the VOT boundaries in perception between 'b' and 'p' are thus commonly found to lie at shorter VOT durations than before a long vowel. This has been found in many studies with speakers of English. In Icelandic, different vowel durations are ambiguous. On the one hand, different durations can signify differences in speaking rate, as in English. On the other hand, durational differences can signify a difference in quantity, i.e. between phonemically long and short vowels. This paper presents perceptual experiments which explore the effect of rate vs. quantity in the perception of VOT in Icelandic. The results show that with differences in the perceived rate of the following vowel, the listeners show normalization for rate in VOT perception, but not if comparable changes in vowel duration signify differences in quantity.

## 1 Introduction

A striking aspect of speech sounds is their context-dependent nature. One factor which has been shown to influence the manifestation of speech segments is speaking rate. Speech sounds compress and expand with changes in speaking rate. Such changes in the durations of speech sounds pose a potential problem for the listener, especially with regard to the perception of temporal speech cues, speech cues which are defined by their duration. How can the listener disentangle those durational properties of speech sounds which are phonemic, intrinsic to the phonetic message, from those which are due to extrinsic factors such as speaking rate? Previous research has shown that listeners are sensitive to the temporal structure of speech. In particular listeners' perceptions are often influenced by speaking rate, showing 'rate-dependent perception' [1].

One temporal speech cue that has been the focus of intensive research is that of voice-onset-time or VOT [2]. VOT is a cue for the voicing or aspiration of word-initial stops. Voiced or unaspirated stops, such as 'b', 'd' or 'g', have short VOTs (20-30 ms) whereas aspirated stops, 'p', 't' and 'k', have long VOTs (typically 50-60 ms or even longer). During the VOT period the voicing of the vowel is replaced by aspiration or noise.

Questions have arisen as the mechanism responsible for rate-dependent adjustments in speech perception. One theory holds that such adjustments reflect a process of "taking into account" analogous to that often posited for visual perception [3]. Here it is claimed that the perceptual system engages in a thought-like process to calibrate the perceptual boundaries, e.g. the VOT boundary separating unaspirated from aspirated stops. This has also been termed normalization [4]. While rate adjustments do occur that can be taken to imply a process of taking-into-account or normalization for speech rate, the adjustments in perception, e.g. for VOT, are often less than such a theory would predict [5, 6].

Another theory put forward by Diehl and Walsh [7] claims that rate adjustments in speech perception are in fact not properly interpreted as speech-specific adjustments, but rather reflect a "general auditory principle" of durational contrast. Focusing on the 'ba'–'wa' distinction (cued by syllable-initial vowel formant transition duration, short for 'ba' and long for 'wa') these authors claim that the perceptual boundary separating these two syllables should move to a longer transition if followed by a longer vowel, since the long vowel would tend to make any particular transition duration appear shorter than if it were followed by a short vowel. A similar line of argument could be put forward to account for the rate-adjustments observed in the perception of VOT.

The Icelandic language can provide an interesting test case for these two theories. Icelandic makes a distinction between phonemically long and short vowels and consonants. This distinction is cued by the duration of vowels and consonants, in many cases by the relationship of vowel and consonant durations, or vowel to rhyme duration. A terminological note: In a word such as *sál* (Icelandic for soul), the single syllable of the word consists of the syllable onset 's' followed by the rhyme 'ál'. Vowels are often abbreviated as V and consonants as C. Icelandic stressed syllables often show complementary duration of vowel and consonant durations, with syllables being either of the type V:C, a long vowel (denoted by the : after the symbol V) followed by a short consonant, or of the type VC: (short vowel followed by long consonant).

In one-syllable utterances (commonly employed in studies of rate-dependent perception) the correlation between the duration of the vowel and the perceived speech rate is high — the shorter the vowel, the faster the perceived speech rate presumably is. This relationship does not, however, necessarily hold when we consider disyllabic words. Consider thus the case of Icelandic where it is possible to use vowel duration to cue either for changes in speech rate or for changes in vowel quantity. Thus the word *kaka* [ka:ka] 'cake' has a phonemically long vowel followed by a phonemically short stop consonant, is of the type V:C. The word *kagga* [kak:a] 'fancy car, acc. sg.' has a phonemically short vowel followed by a phonemically long stop, VC:. In general, the phonemically long vowels have greater durations than the phonemically short ones, though speaking rate can influence these durations. Research has shown that a higher-order invariant, to borrow James J. Gibson's term [8], of vowel to rhyme duration, V/(V+C), will account for the perception of quantity in Icelandic in the face of extensive transformations of rate [9, 10].

Presumably, an auditory contrast theory would predict similar effects of vowel duration on the perception of VOT whether it signifies a difference in speaking rate or a difference in vowel quantity. This, does not hold, however, for the normalization theory which would only predict normalization effects with stimuli differing in perceived speaking-rate. The experiments presented here are intended to test this hypothesis using Icelandic stimulus words where the vowel duration in some conditions reflect a change of speaking rate (Experiments 1 and 2) in other conditions a change in vowel quantity (Experiment 3). All stimulus words begin with a stop with variable durations of VOT. The aim of the experiments is to test for the effect of the following vowel duration on the location of the VOT boundaries.

# 2 Method

## 2.1 Stimuli

Three perception experiments are reported in this paper. All made use of synthetic speech made with the Sensimetrics SenSyn synthesizer, a version of the Klatt cascade/parallel formant synthesizer [11, 12]. The synthesizer was run in the cascade configuration. Each experiment consisted of three stimulus continua with varying word-initial VOT ranging from 10 to 60 ms in 5 ms steps. Figure 1 shows schematic drawings of the stimulus continua in each experiment. The continua were comparable in that in all experiments the vowel of the first syllable was either 140, 200 or 260 ms long. Experiment 1 consisted of single CV-syllables where shortening the vowel is presumably perceived as a faster speaking rate. Experiments 2 and 3 consisted of disyllabic words of the type *gaka* [ga:ka] 'nonsense word', *gagga* [gak:a] 'to cackle', at the short end of the VOT continua. These would then change into the words *kaka* and *kagga* at the long end of the VOT continua. In Experiment 2 the overall durations of the stimuli shortened in concert with the shortening of the vowel-initial shortening, showing a clear effect of speaking rate. In Experiment 3 the overall duration of the stimuli was constant at 540 ms. The shortening of the vowel, identical to that of Experiments 1 and 2, now signifies a change of quantity, from a

V:C type word at top of Figure1 to a VC: type word at bottom, not a change of speaking rate. Note that in all experiments the three continua are defined by identical durations of the initial vowel, these being either 140, 200 or 260 ms long.

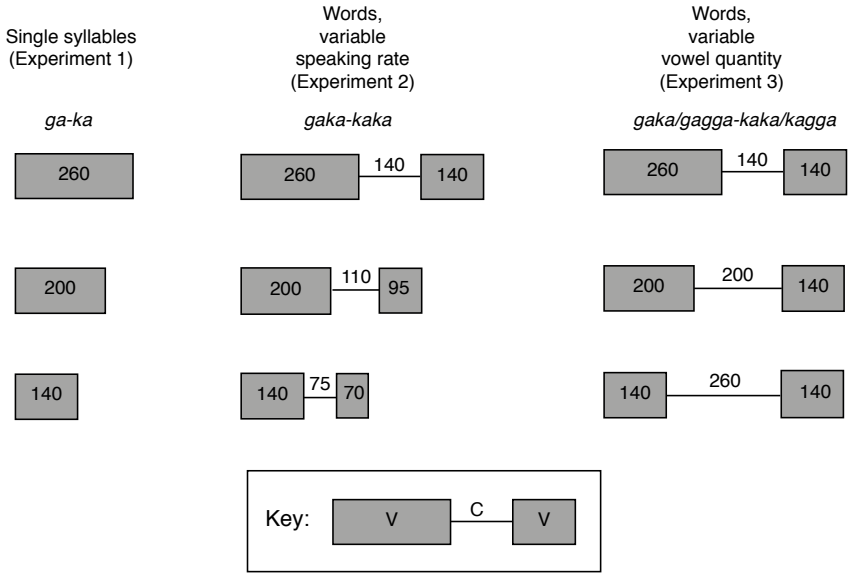


Figure 1: Schematic drawings of the stimuli used in the three experiments. The numbers denote the durations of the segments in ms. The syllable- or word-initial vowel had VOTs ranging from 10 to 60 ms in 5 ms steps for a total of 11 stimuli. All stimuli start with formant transitions appropriate for the stop consonant 'g' or 'k' and variable durations of VOTs. These are not specifically depicted in the figure but form part of the initial vowel.

The steady-state formants of the vowel 'a' had the following values: F1 = 750 Hz; F2 = 1280 Hz; and F3 = 2423 Hz. The transitions for the word-initial velar stop were made in the following manner: F1 started at 200 Hz and rose over 45 ms to the target value of 750 Hz; F2 started at 1800 Hz and fell over 55 ms to the target value of 1280 Hz; and F3 started at 2000 Hz and rose to the target value of 2423 Hz in 55 msec. The first 10 ms of the stimuli consisted of a noise burst centred at 1800 Hz, cueing the onset of the velar stops 'g' or 'k'.

The fundamental frequency of each stimulus was fixed at 100 Hz. The VOT continua were made by replacing the voiced excitation at the beginning of each stimulus with noise and aspiration 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 = 40 dB). The stimuli were synthesized with a sampling rate of 11.025 Hz on a Macintosh computer running OS X.

Each experiment consisted of 33 stimuli, 3 different initial vowel durations, each with 11 different values of word-initial VOTs ranging from 10 to 60 ms in 5 ms steps.

Figure 2 shows a spectrogram (made using the Praat program [13]) of one stimulus from Experiment 3. This particular stimulus has a 200 ms initial vowel with a 20 ms long VOT.

## 2.2 Participants

Eighteen participants took part in each experiment. The six authors of the paper took part in all experiments with a different group of additional twelve participants in each experiment. Most of the participants were students at the University of Iceland. All reported normal hearing and Icelandic as their mother tongue. The age of the participants ranged from 20 to 66 years.

## 2.3 Procedure

The perception experiment was run using the Praat software [13] on a Macintosh computer in a sound-insulated booth. Participants listened to the stimuli at a comfortable listening level over Sennheiser HD 201 headphones. The stimuli in each experiment were presented in 11 randomized blocks, each block consisting of the 33 stimuli of each experiment. The first block was considered a practice block and was discarded from the following data analysis which is thus based

on 10 blocks, 10 presentations of each stimulus. After each stimulus had been played the participants had to indicate whether it had started with 'ga' or 'ka' by pressing one of the keys G or K on the computer keyboard. Participants could take a break if they wanted between stimulus blocks.

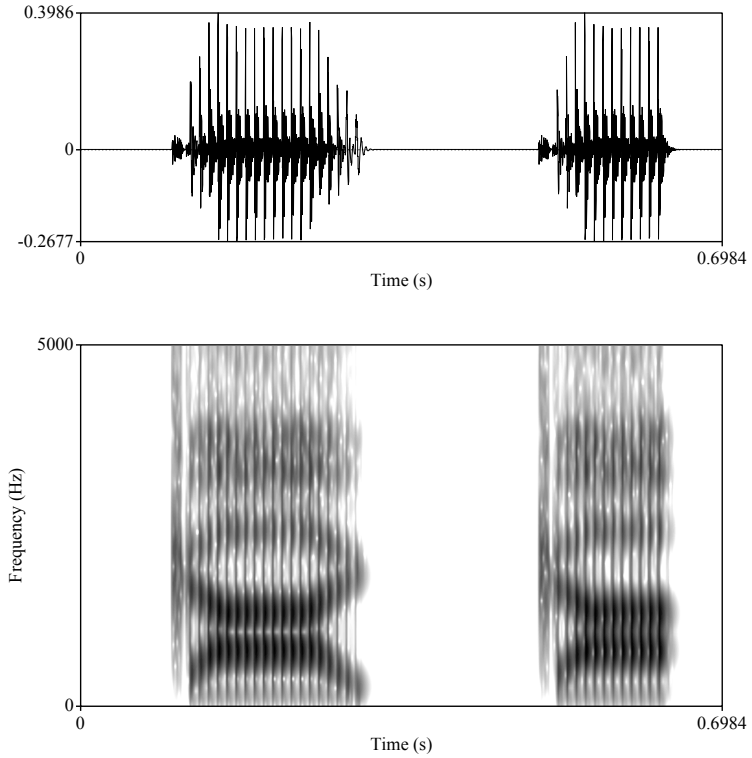


Figure 2. Waveform (above) and spectrogram (below) of one stimulus from Experiment 3. The initial vowel is here 200 ms long with a 20 ms long word-initial VOT (the aspirated part at the left border of the stimulus). The overall duration of the stimulus is 540 ms. Compare the schematic stimulus depicted in the middle row of the rightmost column of Figure 1.

# 3 Results

Pooled identifications curves for the responses of all participants in all three experiments are shown in Figure 3. Here the ordinate of each panel of the figure shows the percentage of *ka*-responses with the abscissa depicting the duration of VOT. Phoneme boundaries, the perceptual boundaries separating the two response categories 'ga' and 'ka', were calculated for each individual participant using the method of probits [14]. This method fits a cumulative normal distribution to the responses of each participant for each condition of the experiment and returns the value of the 50% *ka*-response from the probit curve. In speech perception research this is traditionally considered the perceptual boundary, or phoneme boundary, between the two response categories. The averages of the phoneme boundaries for each condition of the three experiments are shown in Table 1.

Table 1: Average phoneme boundaries (in ms of VOT) in the different conditions of the three experiments.

| Vowel duration | Experiment 1 | Experiment 2 | Experiment 3 |
|---|---|---|---|
| 140 ms | 37.6 ms | 31.9 ms | 32.3 ms |
| 200 ms | 38.3 ms | 33,5 ms | 32.6 ms |
| 260 ms | 40.0 ms | 35.2 ms | 32.9 ms |

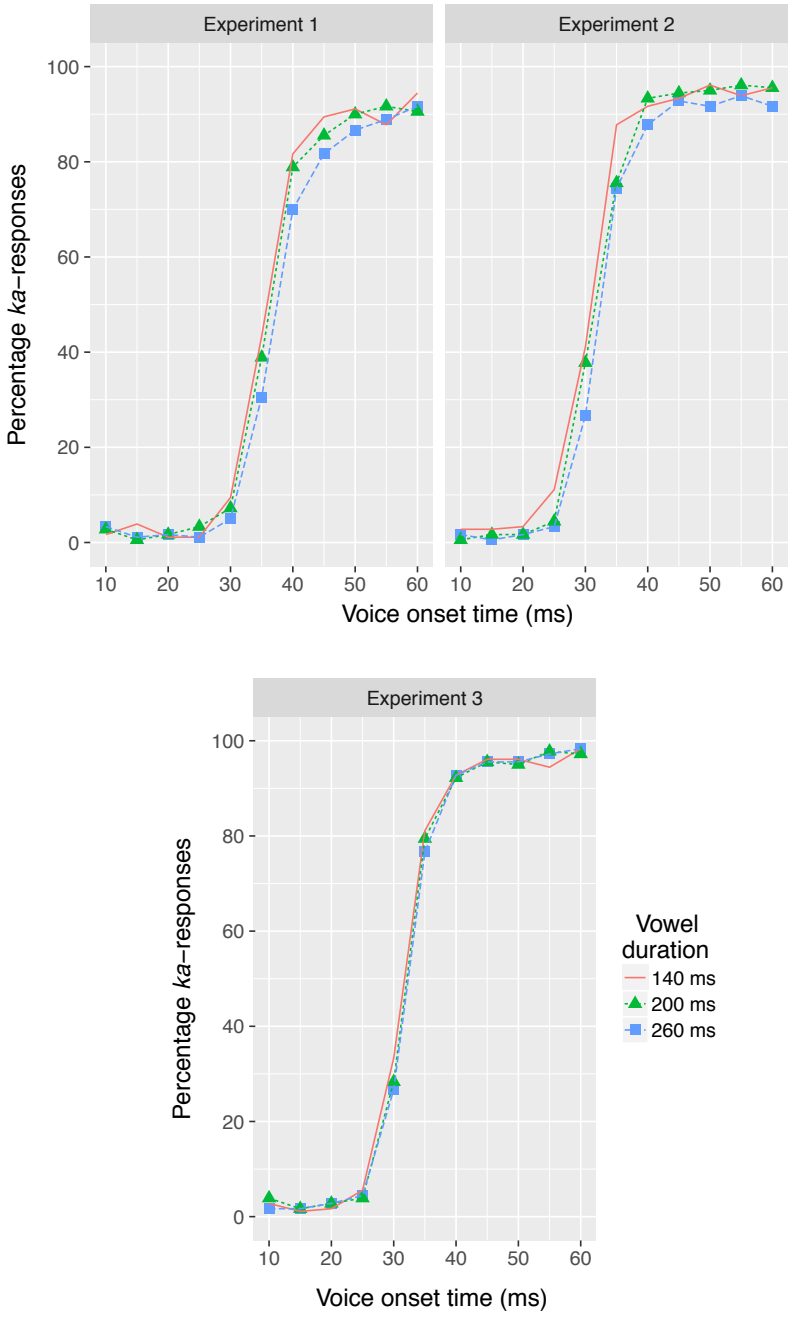


Figure 3. Pooled identification curves for 18 participants in each of the three experiments.

A cursory examination of the results shown in table 1 shows a lengthening of the phoneme boundaries in both Experiments 1 and 2 with increased vowel duration with only limited changes in Experiment 3. This is borne out by the statistical analyses.

Separate repeated measures ANOVAs were run on the results of each experiment, using the R statistical program [15]. For Experiment 1 the results show a significant effect of vowel duration on the location of the VOT boundaries in the three conditions, $F(2,34) = 6,776$, $p < 0.01$. Post hoc comparisons with paired t-tests show that the 140 and 260 ms conditions are significantly different, as well as the 200 and 260 ms conditions. The 140 and 200 ms conditions were not significantly different. Experiment 2 also showed an overall significant difference, $F(2,34) = 22,41$, $p < 0.01$, in the location of the phoneme boundaries. Post hoc comparisons used paired t-tests showed significant differences between all paired comparisons, 140 and 260 ms, 200 and 260 ms conditions as well as between the 140 and 200 ms conditions. Experiment 3 showed a non-significant effect of vowel duration on the location of the phoneme boundaries for VOT, $F(2,34) = 0.669$, $p = 0.519$.

# 4 Discussion

The results of the present experiment show that durational changes in a vowel following word-initial VOT do not all exert the same influence on the location of phoneme boundaries for VOT. If such changes in vowel duration can be traced to changes in perceived speaking rate, as in Experiments 1 and 2, then the VOT boundaries show the typical effect of rate-dependent normalization. Identical changes in vowel duration, when expressing changes in vowel quantity as in Experiment 3, do not lead to significant changes in the location of the phoneme boundaries for VOT. The results thus show that the effects of changes in vowel duration on the location of phoneme boundaries for VOT stimulus continua differ depending on the linguistic significance of the vowel duration change. If it reflects a change of speaking rate, the typical effects of rate normalization are seen. If identical vowel duration changes reflect a change in the quantity of the vowel then, naturally enough, no rate-dependent normalization is found. This can be taken to argue against the claims of Diehl and Walsh [7] that rate-dependent perception is based upon auditory contrast. The auditory contrast of VOT to vowel duration is presumably identical in all three experiments, yet the changes in the location of the VOT boundaries are different, a shifting of phoneme boundaries in Experiment 1 and 2, but not in Experiment 3. The present results can thus most easily be explained by assuming that the rate-dependent perception of VOT operates through a process of "taking into account of" speaking rate.

# References


[1] Miller, J.L., Rate-dependent processing in speech perception, in Progress in the psychology of language, A.W. Ellis, Editor. 1987, Lawrence Erlbaum Associates: London, England. pp. 119–157.

[2] Lisker, L. and A.S. Abramson, A cross-language study of voicing in initial stops: Acoustical measurements. Word, 20**,** 1964, 384–422.

[3] Epstein, W., The process of `taking-into-account' in visual perception. Perception, 2, 1973, 267–285.

[4] Naga, K. and K. de Jong, Perceptual rate normalization in naturally produced rate-varied speech. Journal of the Acoustical Society of America, 121, 2007, 2882–2898.

[5] Miller, J.L., K.P. Green, and A. Reeves, Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. Phonetica, 43, 1986, 106–115.

[6] Pind, J., Speaking rate, VOT and quantity: The search for higher-order invariants for two Icelandic speech cues. Perception & Psychophysics, 1995, 57, 291–304.

[7] Diehl, R.L. and M.A. Walsh, An auditory basis for the stimulus-length effect in the perception of stops and glides. Journal of the Acoustical Society of America, 85, 1989, 2154–2164.

[8] Gibson, J.J., The senses considered as perceptual systems. 1966, Boston, MA: Houghton Mifflin.

[9] Pind, J., The perception of quantity in Icelandic. Phonetica, 43, 1986, 116–139.

[10] Pind, J., Speech segment durations and quantity in Icelandic. Journal of the Acoustical Society of America, 106, 1999, 1045–1053.

[11] Klatt, D.H., Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 1980, 971–995.

[12] Klatt, D.H. and L.C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87, 1990, 820–857.

[13] Boersma, P. and D. Weenink, Praat: doing phonetics by computer (Version 6.0.36) [Computer program]. 2018. URL http://www.fon.hum.uva.nl/praat/.

[14] Finney, D.J., Probit analysis. 1971, Cambridge: Cambridge University Press.

[15] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.