RATE-DEPENDENT PERCEPTION OF VOT: AUDITORY CONTRAST OR RATE NORMALISATION

Jörgen Pind
Faculty of Social Sciences, University of Iceland, Reykjavík, Iceland

ABSTRACT

Listeners are sensitivity to the temporal variability of speech. The mechanisms underlying this sensitivity are, however, unclear. One theory holds that listeners perceive temporal speech cues by "taking into account" the speech context. Another theory holds that rate-dependent perception is based on auditory contrast. An experiment on the perception of VOT in Icelandic indicates that auditory contrast is not a sufficient explanation for rate-dependent speech perception.

INTRODUCTION

A striking aspect of speech sounds is their context-dependent nature. The realisation of individual sounds is heavily dependent on other neighbouring sounds. One factor which has been shown to influence the behaviour and 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 as regards 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?

Research over the past decades has shown that listeners are sensitive to the temporal structure of speech. In particular, experiments on rate-dependent perception, show that listeners' percepts are often influenced by speaking rate [1].

One unresolved issue is the nature of the underlying mechanism responsible for these rate-dependent adjustments. One theory holds that these reflect a process of "taking into account" analogous to that often posited for visual perception [2], where the perceptual system engages in a thought-like process to establish the perceptual boundaries e.g. the VOT boundary separating /g/ from /k/. In fact, it turns out that the rate-adjustments seen e.g. for VOT are typically less than such a model would imply [3,4].

Another theory put forward by Diehl and Walsh [5] claims that rate adjustments in perception are in fact not properly interpreted as speech-specific adjustments, but rather in terms of a "general auditory principle" of durational contrast. Focusing on the /ba-wa/ distinction (cued by vowel transition duration) 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. Experiments using non-speech analogs have been taken to support to this theory.

If, indeed, rate-dependent speech perception is primarily a process involving auditory contrast the question arises as to what the domain of the contrast-inducing speech segment is. If it is to be possible to predict the extent of rate-dependent adjustment in perception it is necessary to establish how far the contrast-inducing elements extends beyond the segment of interest, e.g. word-initial VOT.

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 speech rate. This relationship does not, however, necessarily hold when we consider whole words. Consider thus a language like Icelandic which 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. Research has shown that a higher-order invariant of vowel to rhyme duration will account for the perception of quantity in the face of extensive transformations of rate [4,6].

If the major contrast-inducing factor is that of the following vowel it is possible, using suitably chosen Icelandic words, to arrange for changes in vowel duration to *either* cue vowel quantity *or* speaking rate. Using such stimuli it should be possible to distinguish between the two theories of rate normalisation and auditory contrast. If the theory of rate normalisation ("taking into account") is correct only the vowel durations signifying a rate change should lead to a shift in the phoneme boundaries for VOT. If the auditory contrast theory is correct both manipulations, whether rate-specific or quantity-based, should lead to comparable changes in VOT boundaries since both involve the same contrast of VOT to the following vowel segment.

METHOD

Stimuli

This experiment made use of synthetic speech made with the Sensimetrics SenSyn™ synthesiser, a version of the Klatt cascade/parallel formant synthesiser [7]. The synthesiser was run in the cascade configuration. Six VOT stimulus continua were made containing three different vowel durations. In three of the continua this vowel duration was a cue for speaking rate, in the other three the identical vowel durations were a cue for different vowel quantities (see Figure 1). The words synthesised were tokens of the words 'gaka' [ka:ka], (nonsense word) 'gagga' [kak:a] (to cackle), 'kaka' [k^ha:ka] (cake) and 'kagga' [k^hak:a] (car, acc. sg.).

The different transformations of rate and quantity were accomplished in the manner shown in Figure 1. In the base stimulus the vowel was 260 ms long (including any word-initial VOT) and the closure was 140 ms long. The initial syllable was thus 400 ms long. In this syllable the vowel to rhyme ratio is thus 260/400 = 0.65 which is appropriate for the perception of a word with a long vowel followed by a short consonant. The initial syllable was followed by a 140 ms long [a] for the second syllable. Depending on the duration of the VOT this word would either be perceived as 'gaka' or 'kaka'.

In the Quantity series the duration of the initial syllable was kept constant at 400 ms while the vowel was shortened, first to 200 ms (in this case the closure duration was increased to 200 ms) and then to 140 ms (closure duration 260 ms). The latter stimulus has a vowel to rhyme ratio of 0.35, appropriate for words with a short vowel and following long consonant, i.e. the words 'gagga' and 'kagga'.

In the rate series the same vowel durations were used as in the Quantity series but, contrary to the Quantity series, other segments were also shortened to the same extent as the initial vowel. In this series the vowel to rhyme ratio is therefore kept constant at 0.65 in all three stimulus continua. In the two stimulus series, Quantity and Rate, the very same manipulations of vowel duration can in one case be traced to a change in phonemic make-up, in the other to a change of speaking rate.

The steady state formants of the vowel [a] had the following values. F1 was 750 Hz, F2 1280 Hz and F3 2425 Hz. Appropriate transitions for a velar place of articulation were synthesised at the beginning of both vowels and also at the end of the first vowel. The fundamental frequency of each stimulus was fixed at 100 Hz.

All six continua had variable VOT for the word-initial velar stop ranging in 5 ms steps from 15 ms to 70 ms, made by replacing the voiced excitation with aspiration and noise excitation in the region of F2 and F3 and by increasing the bandwidth of F1 from 90 Hz (the default) to 200 Hz. The shortest VOT of 15 ms consisted of a 10 ms word-initial burst followed by 5 ms of silence. The total number of stimuli in the experiment thus amounted to 3 (vowel durations) × 2 (stimulus series, quantity or rate) × 12 (VOT steps) = 72. Notice that one stimulus series, at the very top of Figure 1, is the same in both the Quantity and the Rate series.

The stimuli were recorded on two tapes with an inter-stimulus interval of 2.5 seconds. One tape contained the three Quantity series the other the three Rate series in randomised order.

Subjects

Twelve subjects took part in the experiment, two members of staff at the University of Iceland and ten undergraduate students of Psychology. All subjects reported normal hearing.

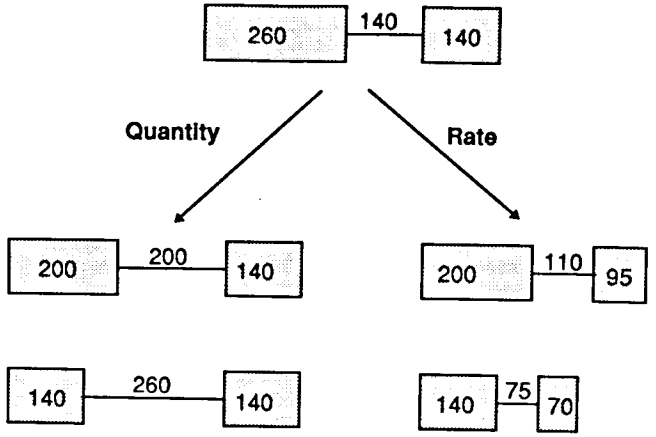

Figure 1. Schematic diagrams illustrating the structure of the stimulus continua used in the present experiment. The shaded boxes denote vowels, the lines connecting them the closures. Numbers refer to the durations of the segments in ms. Note that identical transformations of vowel durations (260 → 200 → 140 ms) signify either a change of quantity or a change of rate.

Procedure

Six subjects listened to the Quantity tape followed by the Rate tape. For the other six subjects the order was reversed. The testing took place in a quiet room. Subjects listened to the stimuli, which were played at a comfortable listening level, over Sennheiser HD-530-II circumaural headphones, and indicated their responses in forced-choice tests by marking appropriate fields on answer sheets.

RESULTS AND DISCUSSION

Pooled identification curves for 11 subjects (one being left out, showing highly irregular response patterns) are presented in Figure 2. The left-hand figure shows the responses for the Quantity series, the right-hand figure the responses for the Rate series. The figures show the percentage of /ka-/ responses as a function of the duration of VOT in the different continua. Phoneme boundaries for individual subjects were calculated using the method of probits. Table 1 shows the average location of the boundaries for 11 subjects.

A two-way repeated measures ANOVA (series × vowel duration) shows that the effect of series is not significant $F(1,10) < 1$ while that of vowel duration is, $F(2, 20) = 8.565$, $p < 0.01$. The interaction is not significant, $F(12,20) = 1.014$, $p = 0.38$. Pairwise comparisons, in all cases using the Bonferroni correction, reveal that none of the means in the Quantity series differ significantly from each other. In the Rate series the VOT boundaries are significantly different in the 260 and 140 ms continua, $F(1,10) = 14.934$, $p = 0.018$, not significantly different in the 200 and 140 ms continua, and just misses significance in the 260 and 200 ms continua, $F(1,10) = 10.317$, $p = 0.054$. Though both series, Quantity and Rate, show a shift towards shorter VOT boundaries with shorter vowel duration only in the Rate conditions do these shifts achieve statistical significance.

Table 1. Average VOT phoneme boundaries (in ms) for eleven subjects.

	Vowel duration		
Series	260 ms	200 ms	140 ms
Quantity	39.27	37.57	36.56
Rate	39.25	36.83	34.25

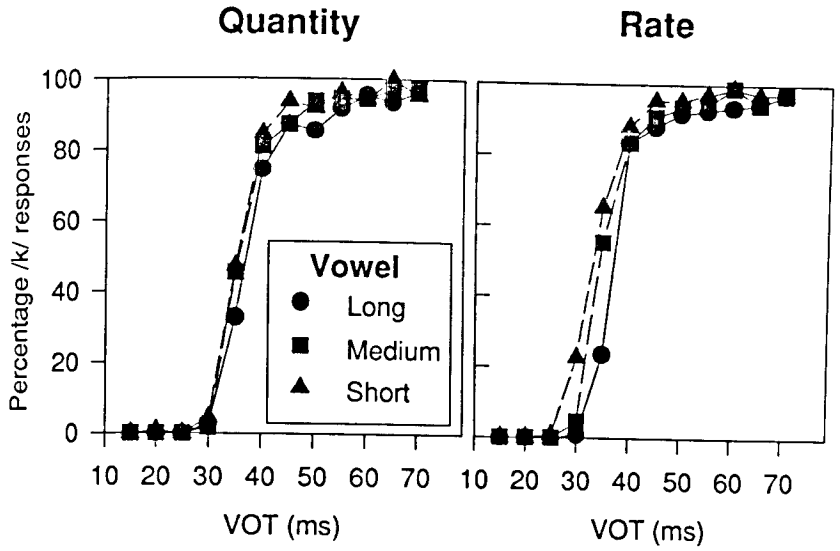

Figure 2. Pooled identification curves (11 subjects) in the present experiment. The figure on the left shows the results for the Quantity series, the figure on the right the results for the Rate series.

Table 1 reveals a shift of 2.7 ms in the location of the VOT phoneme boundary (from 39.27 to 36.56 ms of VOT) and of 5 ms in the Rate series (from 39.25 to 34.25 ms).

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 the VOT boundaries. If such changes in vowel duration can be traced to changes in speaking rate then the VOT boundaries show the typical effect of rate-dependent normalisation. The same changes in vowel duration, when expressing changes in vowel quantity, do not lead to significant rate-normalisation. The present results can most easily be explained by assuming that the rate-dependent perception of VOT operates by a process of "taking account of" the speaking rate.

ACKNOWLEDGEMENT

This research was supported by the Icelandic Science Foundation and the Research Fund of the University of Iceland.

REFERENCES

[1] Miller, J. L. (1987), "Rate-dependent processing in speech perception". In A. W. Ellis (Ed.), *Progress in the Psychology of Language*, Vol. III (pp. 119–157). Hove: Lawrence Erlbaum Associates.

[2] Epstein, W. (1973), "The process of 'taking-into-account' in visual perception", *Perception*, vol. 2, pp. 267–285.

[3] 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*, vol. 43, pp. 106–115.

[4] Pind, J. (1995), "Speaking rate, voice-onset time and quantity: The search for higher-order invariants for two Icelandic speech cues", *Perception & Psychophysics*, vol. 57(3), in press.

[5] 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*, vol. 85, pp. 2154–2164.

[6] Pind, J. (1986), "The perception of quantity in Icelandic", *Phonetica*, vol. 43, pp. 116–139.

[7] 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*, vol. 87, pp. 820–857.