Influences of segmental content on the perception of word duration: a first approach towards a new perceptual model of speech rhythm

Dellwo, Volker ; Hagmann, Lea

INFLUENCES OF SEGMENTAL CONTENT ON THE PERCEPTION OF WORD DURATION: A FIRST APPROACH TOWARDS A NEW PERCEPTUAL MODEL OF SPEECH RHYTHM

Volker Dellwo & Lea Hagmann

Phonetisches Laboratorium, Universitaet Zurich, Switzerland
volker.dellwo@uzh.ch; lea.hagmann@uzh.ch

ABSTRACT

The present research tested the segmental influences on the perception of monosyllabic word durations. 12 listeners of Swiss-German heard pairs of speech and non-speech sounds (monosyllabic words and rectangular-gated sinusoids). They were asked to change the duration of the second sound so that it would match the duration of the first one. Results showed that the Weber Fraction (ΔT/T) for tones is normally distributed around 0 (perfect alignment) while the Weber Fraction for different monosyllabic word pairs can vary significantly from this. In particular quantitative vowel length in contrastive position was found to have an effect on the perception of word duration. Possible implications for models of speech rhythm are discussed.

Keywords: prosody, speech rhythm, speech timing, duration, speech perception

1. INTRODUCTION

Speech rhythm is a complex phenomenon and to date it is unclear what its precise acoustic correlates are. Most models have the durational variability of different types of speech intervals at their center. It had been assumed, for example, that speech rhythm was influenced by the durational variability of syllables or feet. The auditory impression of syllable-timing (percept of regularly timed syllables), for example, was thought to be the result of syllables of roughly the same duration [1, 7]. Such a view implies that listeners can judge the measurable durations of syllables objectively, even if they consist of varying segmental content. To judge, for example, whether the syllables /maː/ and /θɪŋ/ are isochronous we must be able to compare their durations objectively despite the fact that they consist of varying numbers of segments with varying qualities.

Some more recent models of speech rhythm have moved to consonantal and vocalic interval durations as rhythmic units arguing that these units are less variable in more regularly timed languages [4, 8]. Other models assume that rhythmic events of different magnitude (syllabic beats) reoccur in speech in oscillating fashion (coupled oscillator model; [2, 6]). All models of rhythm are based on measurements of absolute durations and do not take into account that their rhythmic units consist of highly variable phonetic content which may contribute to our perception of the duration of these units and thus to speech rhythm.

There is perceptual evidence that durational characteristics derived from vocalic intervals (for example the percentage over which speech is vocalic, %V) are acoustic correlates for rhythmic variability between languages [8]. Such measures, however, capture evidence for very global durational variability between languages. It remains unclear to what degree %V accounts, for example, for the durational variability of vocalic intervals within a sentence. But our percept of speech rhythm must inevitably be influenced by the perceived durational variability of rhythmic units (e.g. syllables, consonantal or vocalic intervals, inter-beat intervals, etc) and it appears improbable that the perceived duration of rhythmic units is solely dependent on the physical unit duration. The psychoacoustic literature revealed that simple qualitative characteristics like the type of signal (e.g. periodic or aperiodic, filled or empty) or ramping and damping can warp listeners' perception of durations to a high degree [9]. Speech is characterized by such phenomena and by now there is an extensive body of literature suggesting that perceived speech segment duration is influenced by the segment quality (cf. [3, 5] for reviews of the literature). However, the influence of varying qualities in segment clusters on the percept of duration of higher level intervals like syllables or consonantal and vocalic intervals is widely unknown.

The present paper is an approach to study segmental influences on the perception of rhythmic units by studying the influence of segmental content on duration perception in monosyllabic words. In a pilot experiment we asked listeners to adjust the duration of the second word in monosyllabic word pairs of varying segment quality but equal segment quantity (three segment words only) until the words would appear of the same duration to them. The performance was compared to the adjustment of rectangular-gated sinusoid tone pairs.

2. EXPERIMENT

2.1. Method

2.1.1. Subjects

12 speakers of Swiss German (7 female, 5 male; all students at Zurich University) between 20 and 29 years of age (mean: 23.8, stdev: 3.2) without reported hearing difficulties took part in the listening experiment. One female speaker of Swiss German (not part of the listener group) was recruited as a speaker. Listeners and the speaker were paid for their participation.

2.1.2. Stimulus and apparatus

The female Swiss German speaker was recorded reading 100 monosyllabic words in the carrier sentence "Ich han nomal *word* gseit" (literally: 'I have again *word* said'). The position of the target word prompted speakers to produce an intonational accent and short prosodic pauses before and after the target word. To avoid that the speaker would fall into a monotonous reading mode she read a random sentence from the BKB corpus (translated into Swiss) after every carrier-word phrase. From the 100 monosyllabic words 10 word-pairs were selected. Table 1 contains the word pairs with phonological transcriptions.

The pairs were selected to differ in:

- (a) quantitative vowel length only (pairs 1 & 2)
- (b) manner of articulation of the final consonant only (pairs 3 & 4)
- (c) voice and manner of the final consonant (pairs 5 & 6)
- (d) quantitative vowel length as well as voice and manner of the final consonant (pairs 7 & 8)
- (e) vowel quantity and quality and at least in manner of all consonants and sometimes also in consonant voicing (pairs 9 & 10).

For each word pair two stimulus pairs were built, in which the second word was either 200 ms longer or shorter than the first one. All durational adjustments of words were performed with the overlap-add methods (as provided in Praat; www.praat.org). The first word was adjusted to 488 ms (average duration of monosyllabic words in the list). Each signal period in each word was set to 5 ms (overlap-add method), i.e. F0 was 200 Hz throughout the entire stimulus set. The average intensity of each word was normalized to 70 dB(SPL).

Table 1: Word pairs used for building the stimuli.

pair number	word 1	word 2
1	Haas /haːs/	Hass /has/
2	Wil /viːl/	will /vɪl/
3	sum /zʊm/	surr /zʊr/
4	Wahn /vaːn/	Wahl /vaːl/
5	Ball /bal/	Bach /baχ/
6	Bill /bɪl/	Biss /bɪs/
7	Wahl /vaːl/	wach /vaχ/
8	lahm /laːm/	lach /laχ/
9	Maass /maːs/	will /wɪl/
10	Schal /ʃaːl/	Riff /rɪf/

For the sinusoid pairs two rectangular-gated 500 Hz sinusoid tones were paired of which the first tone was 500 ms in duration and the second either 200 ms longer or shorter. Sinusoids were not adjusted using the overlap-add method as this introduced audible artifacts; instead they were generated to the respective durations. In all stimulus pairs there was a 500 ms silent gap between the two sound events.

A duration adjustment interface was programmed using Praat scripting language (see Figure 1). With the play button in the center of the screen listeners could play the stimulus pair (words or sinusoids) at any time. With the '+' and '-' labeled buttons listeners could lengthen and shorten the duration of the second sound event in a stimuls pair (single labeling: change by 10 ms; double labeling: change by 50 ms; triple labeling: change by 100 ms). Listeners had to do at least three adjustments to the stimulus pair until they could proceed to the next pair. The number of rounds were counted down in the field below the play button (see figure). When the countdown went below 1 the label 'beenden moeglich' (finish possible) would appear in this field. The instructions in the upper half of the screen translate as 'Change the duration of the second sound until both sounds are equally long'. For the word pairs the word 'Klang' (sound) was replaced with 'Wort' (word).

Figure 1: The duration adjustment interface.

Note that each time an adjustment was made to the second word in a stimulus pair this word would be resynthesized from the original word and not from a previously adjusted version. This method minimized the introduction of audible noise artifacts by the overlap-add method.

Stimuli were played via a PC using high quality headphones in a sound treated room.

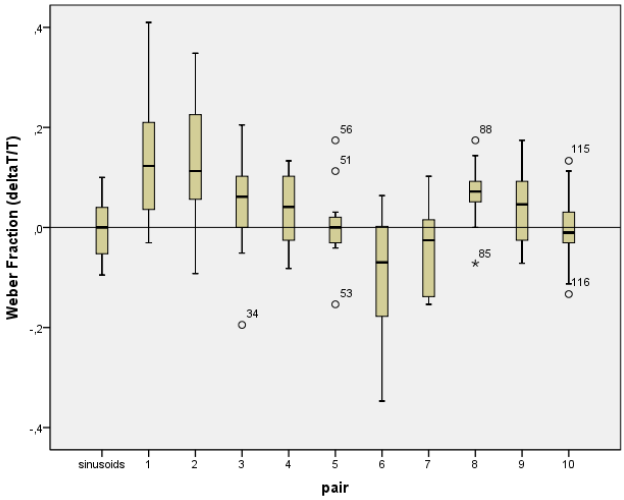
2.1.3. Procedure

Listeners first went through a demonstration session (results were not recorded) in which they adjusted one sinusoid pair and one word pair to familiarize themselves with the task and the interface. After that listeners judged the two sinusoid pairs (second one longer and second one shorter) twice, making for four adjustments in total. Finally listeners adjusted 20 word stimuli pairs (10 second one longer and 10 second one shorter; order randomized).

2.2. Results & discussion

For each final response to a stimulus pair we processed ΔT (difference in duration between the first and the second sound event, i.e. sinusoid or monosyllabic word). From this we calculated the Weber Fraction (ΔT/T; T = duration of the first sound event). Figure 2 contains a box-plot showing the distribution of the Weber Fraction for the sinusoids (first plot) and the 10 word pairs (see Table 1 for identification of the pair). Results for the sinusoid pairs were averaged for each listener across the four stimulus pairs (2x second one longer, 2x second one shorter) and for the word pairs across the two presentations (second one longer, second one shorter). A Weber Fraction of 0 means a 100% correct adjustment (488:488 ms for words, 500:500 ms for sinusoids). The horizontal line at 0 facilitates a visual judgment of the deviation of the distributions from 0. For values that fall above the zero line the second word was adjusted at a longer duration compared to the first one; for values below the zero line the second word was shorter.

Figure 2: Distribution of Weber Fraction values (ΔT/T) for sinusoid and monosyllabic pairs (word pairs can be obtained from numbers in Table 1).



On a descriptive level the results show that the Weber Fraction distributions for different word pairs deviate more or less strongly from 0 (apart from pairs 5 and 10 which are roughly normally distributed around 0) which indicates that the segmental content of the word pairs had an influence on listeners' duration adjustment ability. We performed two types of inferential tests to check which of the effects are significant: (a) a one-sample t-test for each pair to test for the effect of a difference from 0 and (b) a univariate ANOVA with a Tukey's post-hoc test to analyse the effects between each of the word pairs and the sinusoid pair. For test (a) highly significant effects were obtained for stimulus pairs 1 ($t[11]=3.9$; $p=.003$), 2 ($t[11]=3.5$; $p=.005$) and 8 ($t[11]=3.8$; $p=.003$) and a significant effect was obtained for pair 6 ($t[11]=-2.5$; $p=.03$). The sinusoid pair reveals no effect ($t[11]=0$; $p=1$), as is the case with all other pairs (3, 4, 5, 9, 10) at probabilities between .083 and .911. This shows that in particular the quantitative vowel length difference in a contrastive environment (pairs 1 and 2) has a significant effect on the perception of the duration of monosyllables. It means that listeners tend to attribute higher durations to words containing the short vowel in a contrastive pair. It is interesting,

however, that quantitative vowel length did not have this influence in non-contrastive position (7, 9 and 10). Whether quantitative vowel length played a role in stimulus 8 thus remains unclear. The significant effect in word pair 6 suggests that final fricative voicing can influence the perception of duration, however, pair 5 shows that this is not always true.

The results for the sinusoidal variability is well in line with previous results [9] showing that the Weber Fraction varies up to +-.2 for rectangular gated sound events (tones and noise). In the present experiment the variability of the Weber Fraction for the sinusoid pairs is distributed equally around 0 at a range of about +-.1.

For test (b) the results of the ANOVA revealed highly significant effects between the stimulus pairs ($f[11, 130]=6.5$; $p<.001$). Comparing the different word pairs to the sinusoid pair in the post-hoc analysis the data revealed that only pairs 1 and 2 vary significantly from the sinusoid pair ($p=.025$ and .035 respectively). This means that the strongest influence on duration perception is contrastive vowel length while performance for all other word pairs is at the level of rectangular-gated tones.

2.3. Discussion & conclusion

This paper has demonstrated that in particular contrastive vowel length can have an influence on listeners' ability to adjust monosyllabic word durations. It would now be interesting to perform this same test with listeners that do not have phonologic durational vowel contrasts in their language (e.g. Spanish) to find whether general acoustic or language specific phonological factors are responsible for this result.

Apart from contrastive vowel length a number of other segmental influences show influences on a descriptive level but inferential evidence is still weak. Larger numbers of stimuli might lead to stronger effects here in the future. It is particularly interesting to see that there is no tendency observable for complex segmental differences between words (pairs 9 & 10) to have an influence on listeners' duration adjustment performance. This is an interesting and rather unexpected finding of this study. It might imply that complex acoustic differences between monosyllabic words actually aid listeners to pay attention to overall word duration differences.

So what do these results tell us about models of speech rhythm? The results of this study might be interpreted in the way that listeners are actually well able to differentiate durational variability between syllables, independent of the segmental content. Only cases concerning contrastive vowel length might have an influence (which might be language specific). We feel, however, that our results are preliminary and further studies need to be carried out to support this view. Speech rhythm is not only formed by three segment mono-syllables but typically by a complex variability in the number of segments in rhythmic units (although, again, languages may vary strongly regarding this point [4, 8]). Additionally, we feel that the word duration needs to be varied in future experiments as results may vary with absolute word duration. Also artifacts of the overlap-add manipulations on the durational adjustments of words need to be taken into account in future studies. It further seems conceivable that prosodic acoustic parameters like F0 and amplitude may have an effect on perceived speech interval durations. With this study we have now laid out a method to study such effects in the future.

3. ACKNOWLEDGEMENTS

The authors wish to thank two anonymous reviewers for very valuable and in depth comments.

4. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: University Press.

[2] Barbosa, P. 2002. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. *Proceedings of Speech Prosody* Aix-en-Provence, 163-166.

[3] Bochner, J.H., Sneli, K.B., MacKenzie, D.J. 1988. Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing-impaired listeners. *J. Acoust. Soc. Am*. 84(2), 493-500.

[4] Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology* 7. Berlin: Mouton de. Gruyter, 515-546

[5] Lehiste, I. 1970. *Suprasegmentals.* Cambridge/MS: MIT.

[6] O'Dell, M., Nieminen, T. 1999. Coupled oscillatro model of speech rhythm. *Proceedings of ICPhS* San Francisco, 1075-1078

[7] Pike, K. 1945. *The Intonation of American English.* Ann Arbor: University of Michigan Press.

[8] Ramus, F., Mehler, J., Nespor, M. 1999. Acoustic correlates of linguistic rhythm. *Cognition* 73, 265-292.

[9] Schlauch, R.S., Ries, D.T., DiGiovanni, J.J. 2001. Duration discrimination and subjective duration for ramped and damped sounds. *J. Acoust. Soc. Am*. 109(6), 2880-2887.