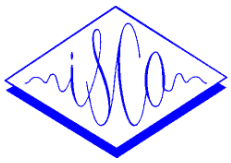

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

RHYTHMIC CONSTRAINTS IN DURATIONAL CONTROL

Cynthia Grover and Jacques Terken

Institute for Perception Research
Postbus 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

Two potential factors in durational control are addressed. First, we investigate whether lengthening a syllable implies lengthening all of its constituent phonemes in a regular way. Analysis of a small corpus of syllables shows that this is not the case. Second, we investigate the influence of rhythm by inspecting the pattern of compression of unstressed or unaccented syllables as a function of the number and position of syllables in the stress or accent group. We find no evidence of compression. We conclude that speakers do not control the timing of phonemes within a syllable very precisely, and that speakers do not compress or stretch unstressed syllables to produce more rhythmic speech.

I. INTRODUCTION

The control of duration in speech synthesis is a complicated affair. There are many potentially relevant factors, and there may be interactions between them. As a result, the extraction of duration rules from speech databases runs into considerable problems: one doesn't know all the relevant factors, or how they combine. But even if all the factors were known, the number of observations for a given combination of factors is often so small that the statistical evaluation of different models is problematic ([11]). The use of experimental data is one way to overcome this problem, but in experiments only a few factor combinations can be studied.

A problem of a different kind concerns the segmentation into units of description, usually phonemes. Since speech provides a continuously varying signal and transitions from one sound to another extend over time, segmentation necessarily is to some extent arbitrary. This imposes problems of reliability, both within and between segmenters. Also, some phonemes may be elided, but acoustic evidence for their underlying presence may show up in the colouring of adjacent phonemes. This has convinced some of those working on segmentation that one really needs an alternative way of looking at the dynamics of acoustic information ([2, 5, 7]). Since these alternative ways are not yet easily applicable to the problem of durational control in speech, we have taken a more traditional approach to looking at phoneme durations.

A further problem with phoneme-based approaches to duration is the adequacy of the duration rules. A given set of rules may generate too long or too short durations for particular phonemes in particular contexts. Now, if we take the rule-based durations for the phonemes making up a synthesized syllable as predictions of the durations in the same syllable taken from natural speech, we will certainly observe deviations between the two sets of durations, which will count as prediction errors. In one case, the errors may all go in the same direction; thus, all phoneme durations might be too long, or too short, resulting in an overall syllable that is also too long or too short. In the other case, the errors might take different directions: some phonemes may be too short, and others too long. In this case, the errors may compensate for one another such that the overall syllable duration is the same as that of the natural one. Such errors may lead to a disruption of the speech rhythm in the first case, but not in the second case. There is evidence that such compensation occurs in natural speech both in production and perception ([6, 9]). The explanation given for such compensation is that speakers aim to maintain a certain rhythmicity and listeners expect speech to have this property.

Rhythmic properties by themselves have received little attention in speech synthesis, which may be one of the causes of the mechanical impression that often arises with synthetic speech. In this contribution, we investigate properties of speech which should reflect tendencies toward rhythmicity. In doing so, we hope to achieve two aims. In the first place, if such tendencies exist, the segmentation problem may be less critical, since rules for the duration of syllables would constrain the variability of phoneme durations. That is, the particular phoneme durations would not matter very much as long as the overall syllable duration was appropriate. Secondly, by appropriate modelling of such tendencies towards rhythmicity, we might be able to make synthetic speech sound less mechanical.

Below, we present two experiments addressing these issues. The first experiment concerns the question of variability: Do we find evidence for variability at the level of phonemes which is constrained at the level of syllables? The second experiment addresses more directly the manifestation of potential rhythmic tendencies in speech. If speakers aim to maintain a certain rhythm, this should show up in the pattern of compression of syllables in different contexts. One may predict that a given syllable will be less compressed if it is the only syllable between two accented syllables than when it is one of two unaccented syllables.

II. PHONEME VARIABILITY

If the durations of individual phonemes have little importance and durational constraints apply mainly at the level of the syllable, we might expect less variability to be associated with replications of a given syllable in a fixed context than with the replications of the constituent phonemes. On the other hand, the elasticity hypothesis maintains that lengthening a syllable can be achieved by stretching its constituent phonemes (either linearly or non-linearly, cf. [4, 12]). In order to evaluate these possibilities, we measured phoneme and syllable durations using a traditional method of segmentation in a corpus of German radio announcers' weather reports.

2.1 Method

The elasticity hypothesis states that the phoneme durations all increase (or decrease) as the syllable duration increases (or decreases). Statistically this means that the standard deviations (SD) for the component phonemes' durations should sum to the SD of the syllable durations. The alternative hypothesis is that there is syllable-internal compensation, or a trade-off between the durations of the phonemes. Some phonemes would increase while their neighbours decrease, even though the syllable duration need not change much. Or some phonemes would decrease even when the duration of the syllable as a whole increases. In this case the sum of the SDs of the component phonemes will exceed the SD of the syllable.

Materials and Procedure. Weather forecasts from German radio broadcastings were recorded at different hours on the same day, from four different speakers. From these recordings we selected sentence fragments which were linguistically identical across realizations. In this way we obtained a corpus of 44 syllables: 22 read by 3 speakers, and 22 by 4 speakers. Twenty-three syllables contained 2 phonemes, 15 had 3, 5 had 4, and 1 had 5 phonemes. All syllables were manually segmented into phonemes.

Analysis. Standard deviations (SD) were computed across the replications of each syllable, and for the separate phonemes within each syllable (cf. Table 1 for an illustrative example). The phoneme SDs were summed. This gave 44 pairs, the first member in each pair being the sum of the SDs for the individual phonemes and the second member being the SD for the syllable.

2.2 Results and discussion

In 36 pairs out of 44 the sum of the phoneme SDs exceeded the syllable SD by more than 10%. The median difference was 12.4 ms. For illustration, Table 1 gives the data for a typical case near the median. The sum of the phoneme SDs is 19.7, the SD for the syllable durations is 6.5, which gives a difference of 13.2. Looking at the particular phoneme durations, one may see that even when the duration of the syllable as a whole increases, the duration of the phoneme /I/ may decrease (cf. speaker 3). The data suggested no systematic differences between speakers.

Thus, the differences in most of the 44 pairs cannot be attributed to small inaccuracies (in terms of a few ms only) in the timing of phonemes by the speaker or to imprecision in the segmentation. Instead, it means that in a considerable number of cases, syllable tokens with constant durations show such a trade-off relation, and that syllable tokens with longer durations are associated with shorter durations of some of the constituent phonemes, relative to tokens with shoter durations.

Table 1. Phoneme and syllable durations (in ms) for the syllable /sI/ from /mEsIg@/ ("modest") for four speakers, with means and associated standard deviations.

speaker	s	I	syll. dur.
1	76	58	134
2	79	58	137
3	99	41	140
4	87	62	149
mean	85	55	140
sd	10.3	9.4	6.5

From these findings we conclude that the boundaries between phonemes, and by implication the phoneme durations, are not precisely programmed by the speaker. Instead, we adopt the view that the mere presence of acoustic information is primary, and the precise timing of the transition from one state to another is of minor importance.

III. RHYTHMIC INFLUENCES ON SYLLABLE DURATION

Speech rhythm is a much debated issue, and we will not go into this debate. For the present, we define speech rhythm as the occurrence of speech events at more or less regular intervals. For the Germanic languages, two levels of rhythmic organization have been proposed: the stress group and the accent group. The stress group consists of a stressed syllable and all following unstressed syllables up to the next stressed syllable. Several definitions of syllable stress are available: one may count only primary stresses, or both primary and secondary, or all heavy syllables. The division of a sentence into stress groups is dependent on the definition one adopts. If one wants to find out which factors contribute to durational variation, all the different definitions have to be taken into account. The accent group consists of an accented syllable and the following unaccented syllables. We use accent mainly in the sense of 'associated with a pitch accent'.

We assume that if speech rhythm is an independent factor in the control of syllable duration, it will show up as a tendency to compress the subordinate units of a stress or accent group as the number of unstressed or unaccented syllables in the group increases. Thus we let the stressed or accented syllables serve as the events occurring at more or less regular intervals in our application of the definition of speech rhythm.

Furthermore, the interpretation of syllable duration patterns and the conclusions that one draws about compression depend on the assumptions one makes about the shape of the compression function. If there is compression of syllables, the pattern may be flat or curved. If flat, all compressed syllables in a stress or accent group will show the same amount of compression. If curved, syllables closer to the stressed or accented syllable will

show less compression than syllables farther away, as it takes some time to switch between the relatively slow output rate associated with stressed or accented syllables to the relatively fast output rate associated with unstressed syllables.

A third level of rhythmic organization has been proposed by [3] for Swedish, implying a different pattern of compression. He argues that unstressed (light) syllables in even-numbered positions relative to an upcoming stressed (heavy) syllable last longer than unstressed syllables in odd-numbered positions relative to the stressed syllable (counting backwards from the stressed syllable).

Factors to control. Confounding influences upon syllable duration may include the position of a syllable relative to the word boundary and sentence end (see [1, 8]). The materials were designed to allow us to control experimentally for these factors. Lastly, a repeated measures design allowed us to control statistically for individual variation among subjects.

3.1 Method

Materials. We constructed sentences in which the position of the unstressed syllable [g@] varied with respect to the number of syllables from sentence end (2 to 14), and with respect to position within the word (word-initial, word-internal, or word-final). In certain sentences the syllables surrounding [g@] were either all light (C*V) or combinations of light and heavy (C*V: or C*VC*). This opposition enabled us to test whether Bruce's effect holds for German.

In order to test the hypothesis of compression, we contrasted materials where [g@] (where @ stands for schwa) was the only syllable, one of two, or one of four syllables between stresses. To investigate the pattern of compression, we placed [g@] (when it was one of four syllables between stresses) either directly before the upcoming stress (i.e. it was the fourth unstressed syllable), or 3 syllables before the upcoming stress (i.e. it was the second unstressed syllable). In order to test Bruce's effect, we placed [g@] (when it was one of two syllables between stresses) in either ultimate or penultimate position before the upcoming stress.

Subjects and Procedure. Eight native speakers of German from Philips Research Laboratory in Aachen, Germany were recorded onto digital audio tape as they read 186 German sentences aloud. The order of sentences was varied across subjects. Sentences were printed on sepaprate pieces of paper to prevent list effects. Filler sentences were included at the beginning and end, and at various positions in-between. Each sentence was read at least four times by each speaker. The data of 4 male speakers are analyzed here.

Segmentation. Time-amplitude graphs and spectrograms were obtained of each sentence, the [g@] targets were extracted and their durations measured. Acoustic evidence of closure for the [g] (energy only at low frequencies, or silence) was taken as the start of the syllable, and the syllable end was taken to be the disappearance of voicing and/or the appearance of acoustic features of the phoneme following the [@]. Checks by ear were conducted as necessary. This choice of syllable definition was motivated by a desire for compatibility with relevant work, such as [6].

Statistical Analysis Mean [g@] durations were calculated across replications of each sentence for each speaker. Between 3 and 8 tokens contributed to each mean. An analysis of variance and repeated measures multiple regression were conducted on these means.

3.2 Results and Discussion

Preliminary tests showed the duration of the syllables to be normally distributed. The first potentially confounding factor, distance from sentence end, failed to account for a statistically significant proportion of the variance in duration, and so it was disregarded in further tests. The other factors which we wished to control were statistically significant; for the position of the syllable relative to the word boundary, $F(2,144) = 3.92$, $p < .05$, and for the variance due to individual variation across subjects, $F(3,144) = 8.10$, $p < .05$. Syllables at the beginning of a word were longest, and those in medial position shortest. The variance due to these two factors was then removed from the analysis prior to every test given below. The mean duration of the unstressed syllable, [g@], was 101 ms, with a standard deviation of 22 ms over the sample size of 150.

Contrary to our prediction, unstressed syllables do not compress as the number of syllables between stresses increases (see Table 2); in fact, when [g@] is one of a group of four unstressed syllables, it is longer than when it occurs as the only syllable between stresses ($F(1,46) = 5.87$, $p < .05$). Accordingly, it is more appropriate to talk of expansion of unstressed syllables, rather than compression. However, our data do concur with [8] in that they show that slightly shorter syllable durations arise when a stressed syllable is adjacent. In all rows of Table 2 except the 4th from the top, [g@] was next to a stressed syllable. The 4th row shows the longest duration, although the effect is not large.

Despite the failure to find compression, we wished to investigate the pattern of duration further. There was no statistically significant effect of position of the unstressed syllable within the group; those in second position did not last significantly longer (only 4 ms) than those in fourth position, $F(1,69) = 3.65$, $p > .05$. We may conclude that the duration of unstressed syllables does not diminish in any important way with distance from a stressed syllable. Syllable duration may well be a flat function, but we will not know until we measure the duration of syllables in all inter-stress positions.

Furthermore, it appears that this pattern is not modulated by rhythmic effects such as suggested by [3] for Swedish. In the data presented by Bruce, unstressed syllables in odd-numbered positions, counting backwards from the upcoming stressed syllable, were shorter than those in even-numbered positions. IN the relevant conditions in the present data (when there were two light syllables between stresses) this effect held, $F(1,45) = 4.84$, $p < .05$, but the difference between the duration of syllables in ultimate and penultimate position before a stress was minimal, being less than 1 ms; [g@] in ultimate position lasted on average 97.8 ms, while in penultimate position it lasted 98.3 ms. This result, while statistically significant, is not meaningful, and will not be further considered in the data analysis.

Table 2. Mean durations (ms), associated standard deviations (SD), and number of observations of [g@] as a function of the number of syllables between stresses and position of target syllable. Each data point in the sample is the mean across at least three repetitions.

Number of [−*stress*] syllables	Syllable Position	Mean Dur	SD	Sample size
1	only	92.3	24.3	24
2	1st	98.3	21.6	24
	2nd	97.8	17.7	27
4	2nd	108.7	21.5	32
	4th	104.2	21.1	43

It may be that the heavy/light distinction and the syllable's position relative to an upcoming stress are nonetheless important, but over longer series of unstressed syllables. It was not possible to test the hypothesis of rhythmic variation on groups of four unstressed syllables because of the difficulty of creating appropriate sentences.

IV. CONCLUSION

Our main findings are that there is usually a trade-off between the durations of the component phonemes of an unstressed syllable (cf. section II), and that the duration of the unstressed syllable [g@] is consistent, both across different positions within the stress group, and across stress groups containing different numbers of syllables: there is no compression of unstressed syllables (cf. section III).

These findings suggest that speakers do not precisely control the durations of phonemes making up a syllable. This implies that the duration of a syllable with a particular phoneme make-up cannot be explained in terms of the duration of its constituent phonemes, and that explanations for timing control have to be looked for elsewhere, either in terms of the dynamic nature of speech production ([10]), or of rhythmic influences. However, we have not been able to demonstrate the influence of rhythmic constraints on syllable duration in German.

Our failure to find compression of unstressed syllables supports work by others such as [6], who assume that unstressed elements maintain a fairly constant duration, and deny that the stress group is a unit of organization. Of course, we do not want to argue against the possibility of rhythmic influences in speech. For instance, the realization of stressed syllables as accented or unaccented may have to do with rhythmical considerations.

The duration of syllables could be organized as a very simple step function; consistently shorter syllables for unstressed (or unaccented) syllables, and longer syllables for stressed (or accented) syllables. This consistency of unstressed syllable duration should help to simplify speech synthesis algorithms.

Taken together, the results argue against the view that speakers accurately control the timing of individual phonemes, and against the contribution of compression effects to the timing of syllables. Alternative accounts for phoneme and syllable timing need to be looked for, e.g. in terms of the dynamic aspects of speech production.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Allen, G.D., "Speech rhythm: its relation to performance universals and articulatory timing", Journal of Phonetics, Vo. 3, 1972, pp. 75-86.

[2] Browman, C.P. and Goldstein, L.M., "Towards an Articulatory Phonology", Phonology Yearbook, Vo. 3, 1986, pp. 219-252.

[3] Bruce, G., "On the Phonology and Phonetics of Rhythm: Evidence from Swedish", Phonologica 1984, Proceedings of the Fifth International Phonology Meeting, Eisenstadt, 25-28 June 1984, (eds. W.U. Dressler et al.) London: Cambridge University Press, 1987, pp. 21-31.

[4] Campbell, W.N., "Multi-level timing in speech", Ph.D. Thesis, University of Sussex, 1992.

[5] Coleman, J., "The Phonetic Interpretation of Headed Phonological Structures Containing Overlapping Constituents", Phonology, Vo. 9, pp. 1-44, 1992.

[6] Crystal, T.H. and House, A.S., "Articulation rate and the duration of syllables and stress groups in connected speech", Journal of the Acoustical Society of America, Vo. 88, 1990, pp. 101-112.

[7] Hertz, S.J., "Streams, Phones and transitions: Toward a New Phonological and Phonetic Model of Formant Timing", Journal of Phonetics, Vo. 19, 1991, pp. 91-101.

[8] Hoequist, C., "Syllable duration in stress-, syllable- and mora-timed languages", Phonetica, Vo. 40, 1983, pp. 203-237.

[9] Huggins, A.W.F., "On the Perception of Temporal Phenomena in Speech", J. Acoust. Soc. Am., Vo. 51, 1972, pp. 1279-1290.

[10] Munhall, K, Fowler, C., Hawkins, S., and Saltzman, E., "Compensatory Shortening in Monosyllables of Spoken English", Journal of Phonetics, Vo. 20, 1992, 225-239.

[11] van Santen, J.P.H., "Assignment of Segmental Duration in Text-to-speech Synthesis, Computer Speech and Language, Vo. 8, 1994, pp. 95-128.

[12] Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P., "Segmental durations in the vicinity of prosodic phrase boundaries", J. Acoust. Soc. Am., 1992, Vo. 91, pp. 1707-1717.