SOME NOTES ON THE PERCEPTION OF TEMPORAL PATTERNS IN SPEECH

Rolf Carlson*, Björn Granström*, and Dennis H. Klatt, Mass. Inst. of Tech., Cambridge, MA 02139 USA. [*Also Dept. of Speech Communication, KTH, S-10044 Stockholm, Sweden.]

Introduction. Prosodic factors in speech have recently attracted a remarkable amount of linguistic and phonetic research. A prevalent point of view is that prosody is of paramount importance, both for naturalness and intelligibility of speech. As a result of this belief, a change can now be seen in the methods adopted in speech training for hard-of-hearing and foreign language students. The increased focus on suprasegmental compared to segmental articulation is possibly advantageous. From a scientific point of view, however, very little evidence is yet available on the quantitative importance of prosody. This is especially true of the relative importance of different aspects of the prosodic pattern.

From a study employing synthetic speech (Huggins, 1976), we know that really deviant durations and fundamental frequency contour decreases intelligibility. Prosodic parameters have also been shown to be effective in disambiguating sentences (Lehiste et al., 1976). Our concern, however, has more to do with what information an explicit description of prosody has to supply and the precision with which it is supplied.

Descriptive models for segmental duration and fundamental frequency have been designed for a number of languages. Typically these models are based on material read repeatedly by a single speaker in a neutral, non-emphatic way. Subjects can perform remarkably consistently within such a recording session, but an examination of spontaneous speech reveals great variability in the prosodic realizations of a given sentence.

Thus it is not clear how precise the specification of duration is in the speech code common to speaker and listener. We also know that perception imposes certain restrictions on how prosodic effects could be appreciated (Klatt and Cooper, 1975). From previous studies (Carlson and Granström, 1975; Fujisaki, 1975), we know that the sensitivity to durational changes is greater in vowels than in consonants. The durational balance between syllable nuclei, as well as the interval between onsets of stressed vowels (a measure related to the foot concept) have been shown to be perceptually important (Carlson and Granström, 1975; Huggins, 1972; Lehiste, 1977).

This leads to the questions that we wish to address: given a primary interest in the functional properties of a model of prosody, what demands should we put on it? What aspects of the description are most important? Will different models be ranked in the same order if different criteria such as naturalness and intelligibility are used?

In our present study we have evaluated both the naturalness and intelligibility of sentences with several different durational structures. As a starting point we have used a version of Klatt's durational rules for American English (Klatt, 1979) that we use in the MIT text-to-speech project (Allen, 1976).

Test Material. An algorithmically complete rule system is meant to generate a first order approximation to the durational structure of any spoken English sentence. In order to evaluate such a system of rules, a variety of syntactic and phonological structures ought to be tested. The test material in our experiments, presented below, could include only a small sample of such structures. These include the active, passive, question, simple, compound, and complex embedded sentence types. Both short and long noun phrases are represented. In Sentence 8, the ")n" after "seafood" is specially used to indicate that the following prepositional phrase is a sentential modifier, rather than modifying the "icy seafood" noun phrase.

Test Sentence	measured dur. (msec)	synth.dur. (msec)
1. Someone at the table)n ordered hot and sour soup.	2365	2615
2. Going to school)n was an adventure.	1625	1860
3. He who eats too much)c will become fat.	2105	2295
4. If Kate)n goes, Bill)n will eat her orange.	2430	2385
5. Old eggs)n often spoil french bread.	2200	2330
6. The fat brown turkey)n was chased by everyone.	2495	2415
7. Do you think that it will rain?	1370	1450
8. Pete)n ate icy seafood)n on the veranda.	2195	2155
9. Frank)n saw pretty streetcars)n in San Francisco.	2755	2845

where: Noun Phrase Boundary =)n, Clause Boundary =)c

These sentences were recorded several times, the most natural sounding recording was selected, and the duration of each segment was measured. Since the rules are intended to match the speech of a particular speaker (DHK), the same subject was employed in the recording session. Nine different versions of each sentence were synthesized and put on language master cards. The synthesis algorithm is discussed in Klatt (1979). The versions listed below include three (3-5) that might be expected to be preferred over version Rule (since the Rule durations are adjusted in part toward Ref durations) and four versions (6-9) expected to be worse than Rule (since various rules contained in Rule have been deleted).

1 Ref — Synthesis by rule using the measured durations from natural speech, but normalized linearly over the whole sentence to get the same total duration as Rule. This adjustment of the speech rate was rather small, averaging 6 percent.

2 Rule — Synthesis using the rule system.

3 Vowel — Synthesis using the vowel durations from Ref and the consonant durations from Rule.

4 Cons — Synthesis using the consonant durations from Ref and the vowel durations from Rule.

5 StressVO — Synthesis using the durations from Rule but linearly normalized between stressed vowel onsets to get the same durations between onsets of stressed vowels as in Ref.

6 NoParse — Synthesis using the rule system, but disregarding the syntactic boundaries marked by ")n" and ")c".

7 SimpleFL — Same as Simple (below) but with clause-final lengthening at punctuation marks. Each segment after and including the last stressed vowel is assigned increased duration by a factor of 1.65.

8 Simple — Synthesis using a very simple rule system:
Stressed vowel : Dur= .80 * inherent duration
Unstressed vowel : Dur= .60 * inherent duration
Stressed consonant : Dur= .90 * inherent duration
Unstressed consonant : Dur= .65 * inherent duration

9 Random — Synthesis using the reference duration, but randomly multiplied or divided by a factor determined by the deviation between Rule and Ref. This condition was included as a clear example of a bad system.

Experiment I: Naturalness. Nine phonetically trained subjects (native speakers of American English, working at RLE,MIT) were asked to sort the nine versions of each sentence according to naturalness of the durational structure. The subjects used a language master and headphones. After the order for a particular sentence type was settled, the subject assigned a number corresponding to subjective naturalness (from 0 - 100) to each version. Most of the subjects finished the task within two hours. In some cases, the task required several sessions. Since the subjects used different scales in the rating task, the data from each subject were normalized to produce a mean of 0 and a standard deviation of 100. Mean ratings across sentences (Table 1, Column labeled "mean") indicate that Ref, Rule, Vowel, and StressVO are judged to be significantly more natural than the others.

The reproducibility of the naturalness rating for a subject was estimated from sentence seven, which had no syntactic markers in the input representation, and was thus identical in versions Rule and NoParse. The mean normalized distance across subjects for this pair was 26, which compares favorably to a typical standard deviation of 60 for the observations underlying an element in the matrix (Table 1). This suggests that subjects were quite consistent in their ratings compared to the intersubject variability. The estimated standard deviation of each element in the matrix is about 20 (60 divided by the square root of 9). The standard deviation of the mean across sentences is given in the table for each version.

Table 1. Naturalness ratings from Experiment I, as averaged across nine subjects. Column A indicates the number of errors for 6 versions used in an intelligibility test described in Experiment II. Versions 3,4 and 9 were not included in Experiment II.

version	sentence 1	2	3	4	5	6	7	8	9	mean	st.d.	A
1	78	58	17	33	21	68	-18	51	78	43	8	2
2	18	33	47	32	70	70	86	17	23	44	7	12
3	86	60	28	0	52	104	44	81	40	55	7	-
4	0	-2	6	73	38	71	-25	-23	61	22	8	-
5	99	28	-2	81	35	108	-25	50	92	52	8	9
6	14	66	3	30	-101	33	70	11	54	20	8	12
7	-36	4	22	21	-114	4	43	6	-80	-14	9	15
8	-63	-66	3	-169	-146	-30	-125	13	-127	-79	11	24
9	-116	-168	-169	-132	-163	-148	-150	-190	-80	-146	9	-

Experiment II: Intelligibility. Some of the versions used in Experiment I were included in an intelligibility test that was presented individually to 18 MIT students. These subjects were phonetically naive, native speakers of American English, and unfamiliar with synthetic speech. Before the test was run, the subject listened to a short passage of synthetic speech (75 sec)

to get acquainted to the speech quality. This familiarization process has been shown to be very rapid (Carlson et al., 1976). The number of word errors out of 122 possible words (excluding articles) is shown in Column A of Table 1, and is plotted against the naturalness rating data in Fig.1.

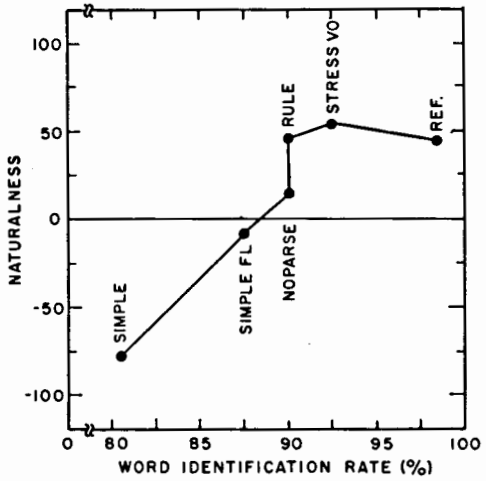

Figure 1. Mean naturalness ratings for six versions are plotted against the word identification rate from Table 1. The two measures are positively correlated, but the improvement in naturalness from NoParse to Rule is not accompanied by an improvement in intelligibility. It should be emphasized that the intelligibility figures are based on a small amount of data and should be interpreted with some caution.

Discussion. It is clear from ratings and comments given on the answer sheets, that subjects have different preferences. For example, Ref is not considered best by all subjects for all sentences. This might be a question of dialectal preference or idiosyncratic differences. Another possibility is that durations from natural speech, imposed on synthetic speech with a somewhat different realization of F0 and segmental content could constitute an incompatible combination. There is no way of controlling for this in the present study. A parallel study using LPC-coded natural speech might shed some light on this issue.

Ref and Rule have about the same mean naturalness score in Table 1, indicating that the durational rules produce as natural a durational structure as our reference speaker <POINT 1>. However, it should be noted that the test material consist of rather short sentences without e.g. the semantic relations between sentences that exist in paragraph-length material and that the intelligibility of Rule was somewhat lower than that of Ref.

We wanted to examine how the intermediate versions between Ref and Rule (Vowel, Cons, and StressVO) are ordered. This could not be done if we are not sure that the relation between Ref and Rule is the same for all subjects. Therefore, in Table 2, the results are presented after discarding the data on a sentence for each subject who rated Rule higher than Ref. (This will, of course, reduce the naturalness score for Rule relative to all other versions.)

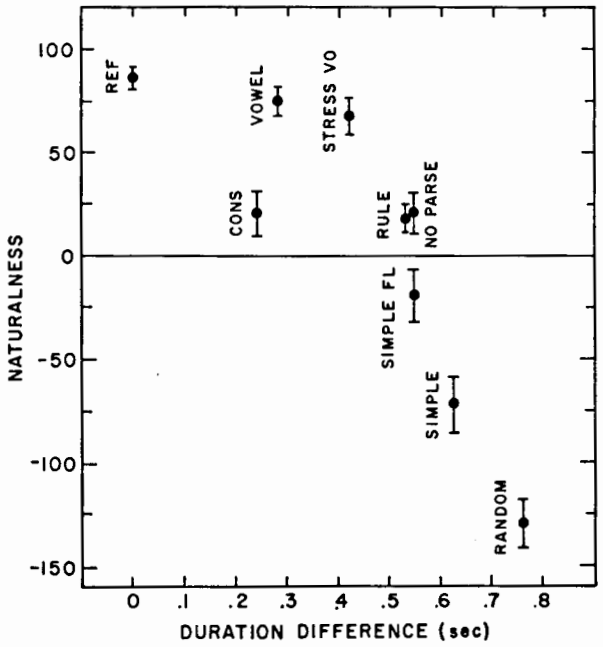

Figure 2. Mean naturalness ratings from Table 2 are plotted against one measure of the physical durational distance to Ref (city block), i.e., the sum, over all segments, of the absolute difference in duration between the version and Ref (average per sentence). There is a general correlation between naturalness ratings and physical difference in duration, but Rule, StressVO and Vowel are rated more natural than one might expect given the durational differences involved.

Table 2. Naturalness ratings after excluding Ss who rated Rule better than Ref. The number of subjects for each sentence is marked in the last line.

version	sentence 1	2	3	4	5	6	7	8	9	mean	st.d.
1	102	85	66	55	73	102	93	96	83	85	6
2	13	12	10	4	59	42	44	14	12	18	8
3	110	85	78	6	73	116	72	82	40	74	8
4	2	-20	-42	92	90	91	25	-43	52	20	11
5	105	27	27	53	42	116	41	56	90	67	9
6	-4	43	-30	52	-125	37	44	19	56	20	10
7	-40	-10	36	26	-159	-11	17	13	-82	-20	13
8	-42	-68	-27	-176	-208	-30	-71	-4	-119	-73	14
9	-125	-133	-150	-121	-122	-150	-116	-180	-75	-130	12
	7	6	5	5	2	5	2	7	8	(# subjects)	

The most striking result seen in Figure 2 is that both Vowel and StressVO are significantly more natural than Cons, despite their greater durational distance from Ref. This corroborates earlier observations that these two durational units i.e. vowel duration and interval between onset of stressed vowels are of great perceptual importance <POINT 2>. Cons, which has all consonant durations right but vowel durations done by rule, does not score significantly better than Rule, reinforcing this interpretation. Furthermore it is obvious that physical distance is clearly not a reliable predictor of perceptual distance.

Isochrony, i.e. the tendency toward equal durations between certain units, has been discussed in the literature. It might be suspected that the high scores for Ref and StressVO are because they preserve the isochrony of real speech. We compared Rule and Ref to see which has a greater tendency toward equal distances between stressed vowel onsets. If anything, Rule is more isochronous than Ref, suggesting that the amount of isochrony implemented in the rules via, e.g., cluster shortening and unstressed segment shortening is probably sufficient, and no "isochrony rule" per se need to be added <POINT 3>.

Versions Rule and NoParse have the same naturalness score in Figure 2. However, it must be remembered that an editing has taken place which selectively lowers the score of Rule. In Table 1, these two versions are significantly different. For some sentences, however, the score for NoParse is higher than that for Rule. Even if these differences are not highly significant, they indicate that in these instances, NoParse is regarded as close in quality to Rule. One possible reason could be that the rules dealing with phrase final lengthening overexaggerate the lengthening effect. An analysis yielded no support for this interpretation in our data. Another possibility which seems more reasonable is that the phrase-final lengthening rule is applied too frequently <POINT 4>. A simple-minded cure might be to ensure that short phrases containing only one content word are not affected by phrase final lengthening although recent work by Cooper et al (1978) indicates the likelihood of a more complex relation between surface structure and lengthening.

Comparing Simple and Rule, we can conclude that rules modifying the duration of a segment as a function of syntax and segmental context are of significant importance for both naturalness and intelligibility. Approximately half of the difference between the two versions seems to be explained by the extremely simple clause final lengthening rule used for SimpleFL <POINT 5>.

The intelligibility results shown in Figure 1 indicate a clear correlation between intelligibility and naturalness. Correct durations result in significantly better intelligibility and naturalness. This confirms in part the current belief in the importance of prosody to sentence perception. <POINT 6>.

References

Allen, J. (1976), "Synthesis of Speech from Unrestricted Text", Proc. IEEE 64, 433-442.

Carlson, R. and Granström, B. (1975), "Perception of Segmental Duration", in Structure and Process in Speech Perception, A. Cohen and S.G. Nooteboom (Eds.), Springer-Verlag, Berlin, 90-104.

Carlson, R., Granström, B., and Larsson, K. (1976), "Evaluation of a Text-to-Speech System as a Reading Machine for the Blind", STL QPSR 2-3/1976, 9-13.

Cooper, W.E., Paccia, J.M., and Lapointe, S.G. (1978), "Hierarchical Coding in Speech Timing", Cognitive Psychology 10, 154-177.

Fujisaki, H., Nakamura, K., and Imoto, T. (1975), "Auditory Perception of Duration of Speech and Non-speech Stimuli" in Auditory analysis and perception of speech , G. Fant and M. Tatham (Eds.), Academic Press, London.

Huggins, A.W.F. (1972), "On the Perception of Temporal Phenomena in Speech", J. Acoust. Soc. Am. 51, 1279-1290.

Huggins, A.W.F. (1976), "Speech Timing and Intelligibility", Proc. Attention and Performance 7, J. Reguin (Ed.).

Klatt, D.H. (1979), "Synthesis of Segmental Durations in English Sentences", 9th International Congress of Phonetic Sciences, Copenhagen.

Klatt, D.H. and Cooper, W.A. (1975), "Perception of Segment Duration in Sentence Contexts", in Structure and Process in Speech Perception, A. Cohen and S.G. Nooteboom (Eds.), Springer-Verlag: Heidelberg.

Lehiste, I. (1977), "Isochrony Reconsidered", J. Phonetics 5, 253-263.

Lehiste, I., Olive, J.P., Streeter, L.A. (1976), "The Role of Duration in Disambiguating Syntactically Ambiguous Sentences", J.Acoust. Soc. Am. 60, 1199-1202.