# ACCEPTABILITY OF SEVERAL SPEECH PAUSING STRATEGIES IN LOW QUALITY SPEECH SYNTHESIS; INTERACTION WITH INTELLIGIBILITY

Vincent J. van Heuven & Peter J. Scharpff

Dept. Linguistics/Phonetics Laboratory,
Leyden University, The Netherlands

## ABSTRACT

We know from previous studies that inserting speech pauses at the end of coherent word groups, but not in any other positions, improves the intelligibility of low quality speech. The present study examines the effect of several pausing strategies on the acceptability, rather than intelligibility, of low quality speech for listeners who either did or did not know the verbal contents of the message beforehand.

## 1. INTRODUCTION

Our research started from the assumption that speech pauses may help the listener to decode the incoming message. In earlier reports [5,6] we studied the effects on word recognition in connected speech of four different speech pausing strategies applied to low quality diphone synthesis of Dutch sentences. Melodically and temporally well-formed pauses had been inserted in long sentences at more or less regular intervals. Subjects listened to each sentence twice (in order to reduce memory load), and were asked to fill in all the content words they had recognised on their anser sheets. In the answer sheets all the function words had been printed beforehand, interspersed with underlined blanks, one for each content word to be filled in. About 50% of the blanks were filled in correctly after listening to a sentence once, another 30% was added after hearing the sentence for the second time. At first sight, the effect of synthesizing pauses in the utterances facilitated word recognition only marginally: when taking the condition without any pauses at all as a base line, percent correctly filled in blanks was raised significantly, but no more than by 4 percent on average, as a result of inserting pauses at prosodically motivated boundaries (I and Phi-domain boundaries, cf. [1,3]). However, when pauses had been inserted before important content words (mimicking a speech pausing strategy of certain experienced broadcasters), no significant improvement was obtained relative our baseline condition. When speech pauses had been inserted by a fixed rule after every sixth word, irrespective of grammatical structure or of the communicative importance of the word, the subjects' performance was significantly poorer than in the baseline condition.

The differences between the conditions were larger (8 percent improvement re. baseline), however, when we considered the effects for monosyllabic words only. The results revealed that the recognition of monosyllabic words, but not of longer words, was facilitated by the insertion of grammatically motivated pauses (3 to 4 percent improvement re. baseline).

This interaction between pausing and word length is predictable from what we know about word recognition in connected speech (cf. [4]). Longer words are



usually recognized before the final sounds belonging to the word's acoustic make-up have reached the listener. For instance, the word elephant can be recognized when only the sound sequence corresponding to eleph has been heard: there is no other word in the English lexicon than elephant (and its derivations) that begins with this sound sequence. The final portion of longer words is lexically redundant. This means that, in connected speech, the listener can predict exactly where a new word will start, thus reducing the number of competing recognition hypotheses. Short, monosyllabic words, however, cannot be recognized with certainty until at least some of the following context has been heard: monosyllables can very often be the first syllablle of a longer word (cf. cap - captain; cat - caterpillar, etc.). Therefore the number of competing parses for a sequence of monosyllabic words is typically greater than for sequences of polysyllabic words, with better word recognition performance for the latter type [5]. The difference between monosyllabic and polysyllabic words will increase when the average number of competing parses is raised, as happens in poor quality speech. In such cases, word segmentation ambiguity can be reduced by inserting speech pauses, which always occur at word boundaries. Moreover, if the pauses are inserted at grammatically motivated positions, they contain not only information on word boundary location, but also reveal part of the grammatical structure of the input sentence.

In the present experiment we wished to study the influence of the various pausing strategies on acceptability, rather than on word recognition. Conceivably, frequent interruption of the utterance by conspicuous and time consuming pauses - even if conducive to better intelligibility - may be disruptive and annoying to the listener.

Furthermore, we reasoned that one may expect different acceptability results when the listener knows the text beforehand, than when the text is new to him. If in the latter case the pauses do indeed help the listener to resolve ambiguous word boundaries and recognise the grammatical structure of the sentence, he will gladly pay the price of having to put up with the time delay. However, when the listener is familiar with the message, pauses are not needed, and will sooner be felt as a nuisance. We therefore predict that frequent pauses, especially when they do not contribute to word recognition, will be negatively valued by the listener. However, in novel utterances, which are difficult to understand, pauses that increase intelligibilty will be positively valued.

## 2. METHOD

Seven Dutch sentences, each 36 words and 68 syllables long, were selected from the stimulus material used in the earlier intelligibility test [6]. These sentences had been concatenated from severely quantized diphones with a resulting speech quality that was equal to that of the Philips MEA8000 formant synthesis chip, and were given appropriate intonation contours. Pauses were 200 ms long, marked by a pitch fall B (cf. [2]), and preceded by a 40% lengthened syllable. This means that sentences with pauses lasted longer (by some 250 ms for each pause) than sentences without any pauses. We took the precaution of creating an extra stimulus condition without pauses with a slower speaking rate so that the overall duration here was equal to that of a sentence with pauses inserted. Suspecting that a melodically marked boundary could be counterproductive in the middle of a coherent phrase, we added a sixth condition in which speech pauses before important content words (as

in condition 4 below) were not accompanied by the boundary marking pitch movement. The six different boundary marking conditions are listed below:

1. No pauses, no adaptation of speaking rate.
2. As condition 1, but with speaking rate slowed down so as to make the duration of the utterance equal to that of versions with pauses.
3. Six pauses inserted at fixed intervals (after every sixth word), disregarding any structural considerations.
4. Six pauses inserted at more or less regular intervals, but always immediately preceding important content words; these pauses did never occur at the end of an intonation domain (I) or of a phonological phrase (Phi).
5. As condition 4, but with pauses marked temporally only (no boundary marking pitch movements were executed).
6. Six pauses interserted more or less regularly, but always at the end of an I or Phi domain.

The full set of 7 (lexically different sentences) * 6 (pausing conditions) = 42 stimuli, preceded by 6 practice stimuli, were presented to two groups of 60 listeners. The first group had taken part in the intelligibility test described in section 1, immediately prior the present test. Each of these listeners had heard the sentences twice before; also, they had printed versions of the stimuli before them. The stimulus material should therefore be perfectly intelligible to this group of prepared listeners.

The material presented to a second group of 60 unprepared listeners who had never heard the sentences before. In this group each listener heard each of the 7 lexically different sentences only once, with maximal variation of pausing conditions within subjects.

All listeners heard the stimuli over headphones, and rated each sentence along a 7-point acceptability scale, where 1 stood for 'very unpleasant to listen to' and 7 for 'very pleasant to listen to'.

## 3. RESULTS

The results are presented in Table I.

Table I: mean acceptability score broken down by type of listener (prepared vs. unprepared) and pausing condition (1 through 6, see text); in parentheses the number of reponses.

| | LISTENER TYPE | |
|---|---|---|
| PAUSING CONDITION | prep. | unpr. |
| 1. no pauses | 4.7 (120) | 4.4 (10) |
| 2. as 1, but slowed down | 4.6 (120) | 4.2 (10) |
| 3. pauses after every 6th word | 2.7 (120) | 3.5 (10) |
| 4. pauses before imp.cont.words | 2.6 (120) | 3.3 (10) |
| 5. as 4, but no pitch movement | 3.1 (120) | 3.7 (10) |
| 6. pauses at word group boundary | 4.1 (120) | 4.3 (10) |

When the listener knows the text beforehand (prepared), the condition with no pauses at all, no matter whether speaking rate is slowed down (cond. 2) or not (cond. 1), is rated most favorably. Pauses at the end of word groups (cond. 6), though still above the middle of the scale, are rated less favorably. Considerably lower ratings are obtained for the three remaining conditions.

When the listener is unfamiliar with the message and intelligibility is therefore poor (unprepared) the results are rather different. The differences between the six pausing conditions are less extreme, although the relative ordering of the six conditions is hardly changed. Crucially however, the condition with pauses at grammatically motivated locations (cond. 6) is now rated in between the two conditions with no pauses at all. Moreover, condition 6 is rated more favorably in an absolute sense by unprepared listeners than by prepared listeners. Since condition 6 was already rated above the middle of the scale by the prepared listeners, the improvement runs counter to the general tendency of unprepared listeners to regress towards the middle of the rating scale.

A classical ANOVA with listener type and pausing condition as fixed factors shows significance for pausing condition and for the pausing*listener type interaction. Newman-Keuls tests for contrasts ($p < .05$) show that conditions 1 and 2 do not differ from each other with prepared listeners; conditions 1, 2 and 6 do not differ from one another with unprepared listeners, as do conditions 3, 4 and 5.

## 4. CONCLUSION

Listeners evaluate the presence of speech pauses differently depending on the intelligibility of the stimulus. When they do not need the speech pauses in order to decode the message, all pauses, whether placed appropriately or not, are considered a nuisance. However, when the listener is not familiar with the text, and therefore needs the speech pauses in order to decode the message, one type of pausing is evaluated as positively as not pausing at all.

We now know that in the normal situation when the listener is unfamiliar with the message, e.g., when hearing a news broadcast, pauses inserted at the end of coherent word groups, and only these, help word recognition in continuous speech of low quality. Moreover, listeners do not judge the presence of such pauses unpleasant, even though the input speech is interrupted quite frequently. We therefore generally recommend pausing at grammatical boundaries (but nowhere else) as a means of improving the intelligibity of low quality synthesis of continous speech.

NOTE

We thank S.G. Nooteboom for comments and discussion. This research was partly supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organisation for Research, NWO, under grant # 300-161-035.

## 5. REFERENCES

[1] GEE, J.P, GROSJEAN, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal, Cogn. Psych. 15, 411-458.

[2] HART, J. 'T., COLLIER, R., COHEN, A. (1990), A perceptual study of intonation, an experimental phonetic approach to speech melody, Cambridge: Cambridge UP.

[3] NESPOR, M., VOGEL, I. Prosodic phonology, Dordrecht: Foris.

[4] NOOTEBOOM, S.G. (1985). A functional view of prosodic timing in speech, in J.A. Michon, J.L. Jackson (eds.): Time, mind and behavior, Berlin: Springer, 242-252.

[5] SCHARPFF, P.J. (1987). Effect of context and lexical redundancy on continuous word recognition, Proc. 11th ICPhSc, vol. 5, 43-47.

[6] SCHARPFF, P.J., HEUVEN, V.J. VAN (1988). Effects of pause insertion on the intelligibility of low quality speech, Proc. 7th FASE/SPEECH-88, 261-268.