1. INTRODUCTION

The intelligibility of the output of text-to-speech systems today is generally poorer than that of natural human speech. One way of improving the quality of synthetic speech is to insert speech pauses at selected positions in the utterances rather more frequently than the human reader would choose to do. Pause insertion has been reported to improve intelligibility in deaf-speech [1] as well as in speech synthesized from diphones [2]. In the studies cited here, the pauses were inserted at breaks in the syntactic structure of the spoken sentences, without explicit motivation for this particular pausing strategy. Although the choice is intuitively appealing, there may be other pausing strategies that are perceptually as adequate but easier to use in text-to-speech applications. The present research, therefore, aims to establish an optimal pause insertion strategy for low-quality speech synthesis.

There are at least two criteria that have to be considered when choosing a pausing strategy. Firstly, the pauses should convey as much useful information to the listener as possible; secondly, the positions where the pauses are to be inserted should be detectable by a simple algorithm that can easily be incorporated in a text-to-speech system. Generally, the more useful the information signalled by the speech pause, the harder it is to find its position automatically. In our experiments we systematically examined the effects of four pausing strategies, which will be discussed in our next sections in ascending order of complexity. The first strategy is one without pauses. The second strategy has pauses at word boundaries at regular intervals of six words. Here the listener gets information about some word beginnings in the sentence. In the third strategy we have marked word boundaries before content words so that information is given to the listener about beginnings of relatively important words in a sentence. In the last strategy we marked boundaries so as to reveal the prosodic structure of the sentence. The speech pauses in this version are located at the end of intonational or phonological phrases. We expect this last strategy to be the most helpful to improve the intelligibility of low quality speech.

In good quality speech, long words can generally be recognized before their final sounds have reached the listener (cf. [3,4]). For monosyllabic words, however, it is often impossible to determine on the spot whether they are indeed short words or the beginning of a longer word, e.g., cap may well be the beginning of the longer word captain. The ambiguity of word boundaries will be severely increased with low-quality speech, so that especially the recognition of monosyllabic words with little lexical redundancy will suffer. From this reasoning we predict that inserting speech pauses will improve
EFFECTS OF PAUSE INSERTION

Intelligibility in general; pauses after monosyllabic words will contribute more to intelligibility than those after longer words (cf. [2,5]).

In this paper we shall consider the following main question: what pausing strategy do we have to apply (e.g. in a reading machine) so that it facilitates recognition of artificial speech for a listener?

2. METHOD AND MATERIALS

The effect of speech pause insertion on intelligibility is not easy to measure. Using hifi quality speech we can measure no such effect: everything is intelligible as it is. Moreover, in short sentences the urge of pausing does not exist. Therefore we selected long sentences in low quality speech so as to maximize the effect of speech pause insertion.

We generated 36-word (68-syllable) sentences, all possible news bulletin messages. An example of such a sentence is:

(a) In de stukgoedsector in de Rotterdamse haven zijn gedwongen ontslagen niet meer te vermijden, omdat er volgens de werkgevers, en ook volgens de betrokken bond, op dit moment in die bedrijfstak minstens honderd personeelsleden overbodig zijn.

'In the general cargo sector in the port of Rotterdam forced lay-offs cannot be avoided, because according to the employers, and also according to the union involved, at this moment in this branch of trade there are at least one hundred employees redundant'.

There were 15 stimulus sentences preceded by 3 examples. Each sentence was synthesized from diphones at an information rate of 2 kBaud. Diphones are speech fragments that consist of the last half from one speech sound and the first half of the one following that sound thus containing the natural transition between the two sounds (cf. [6,7]). Intonation contours were generated as illustrated in figure 1, with a rise on each accented syllable and a gradual fall spanning the time interval between two consecutive rises; the final two accents were joined into a hat contour [8].

Example of a sentence with a synthetic 'sawtooth' contour.

Figure 1.
The 'sawtooth' contour, with the transcription from 't Hart & Collier [8].

In this way we generated utterances that have no segmental or prosodic boundary markers, which is necessary in the no-pause versions. In the other versions we inserted pauses of 200 ms on basis of the three different strategies, readjusting the pitch contours and lengthening the prepausal syllables by a factor 1.4 [9], as in figure 2:

262 7th FASE Symposium, Edinburgh
EFFECTS OF PAUSE INSERTION

We selected four speech pause insertion strategies:

(1) pauses inserted at regular intervals of 6 words. Although all pauses were inserted at word boundaries, most did not coincide with phrase boundaries. It has been found that speech pauses are likely to occur at what is called Intonational Phrase (IP) boundaries, and that IPs should have roughly uniform lengths within one sentence [10,11]. IPs are constituents or phrases within the sentence that are not only syntactically but also prosodically motivated. How long IPs should be in general is unknown; we chose an arbitrary fixed length of 6 words. The result was five pauses per sentence; this number was also inserted in each of the following versions of the sentences.

(2) pauses inserted immediately before selected content words. In order to obey the principle of uniform IP-length, we selected content words in the neighbourhood of the pause locations in version (1). This strategy mimics a speech style in human spontaneous speech sometimes used by professional talkers. It is unclear whether this is an appropriate strategy for reading text, but at least the positions are relatively easy to find automatically.

(3) pauses inserted at prosodically motivated locations. The pauses were all located at IP-boundaries, or – if no more IP-boundaries were available – at Phi-boundaries (boundaries between Phonological Phrases) in which case the similar length condition was applicable. An algorithm that derives these locations from orthographic text is complicated, and entails a syntactic analysis of some depth.

(0) a control condition with no pauses inserted. The extra time that listeners get in versions (1) through (3) when speech pauses were inserted, was compensated for in this version by a slight decrement of the speech tempo (about 6%), so that all versions were of the same length.

By way of illustration we have marked the pause positions as defined by strategies (1) through (3) in our sample sentence in (a'):

(a') In de stukgoedsector in de rotterdamse (1) haven (3) zijn (2) gedwongen ontslagen niet meer (1) te (2) vermijden, (3) omdat er volgens de (1) werkgevers, (3) en ook volgens de (2) betrokken (1) bond, (3) op dit (2) moment in die (1) bedrijfstak (3) minstens (2) honderd personeelsleden overbodig zijn.
EFFECTS OF PAUSE INSERTION

Each sentence was presented to 4 groups of 10 listeners in a fully blocked design. Subjects were instructed to listen to a sentence, to write down what they had heard with a blue pen, to listen again to the same sentence and to fill any gap with a red pen. To facilitate the writing task and to decrease possible memory problems we already printed the function words on the answer sheets.

3. RESULTS AND CONCLUSIONS

Let us first of all analyse the overall effect of the four pause strategies on the intelligibility of the synthetic speech used in our experiment:

Table I.
Scores (in rau) of correctly recognized morphemes (-inflection) in 4 pause strategies. In parentheses the total number of cases per cell.

<table>
<thead>
<tr>
<th></th>
<th>correct after 1st presentation</th>
<th>correct after 2nd presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) no pauses</td>
<td>50 (371)</td>
<td>81 (371)</td>
</tr>
<tr>
<td>(1) pauses after each six words</td>
<td>49 (371)</td>
<td>78 (371)</td>
</tr>
<tr>
<td>(2) pauses before content words</td>
<td>49 (371)</td>
<td>75 (371)</td>
</tr>
<tr>
<td>(3) pauses at prosodic boundaries</td>
<td>53 (371)</td>
<td>84 (371)</td>
</tr>
</tbody>
</table>

Taking the version without speech pauses as our base-line, we observe that a (modest but significant) improvement of intelligibility with 3 rau is obtained only when pauses are inserted at prosodic boundaries (condition 3). Conditions (2) and (3) show a small decrease of intelligibility relative to the base-line version, which is significant only for second presentation of the stimulus.

The difference between first and second presentation is roughly 30 rau, which effect is highly significant. This improvement is additive to all other effects with the exception noted above. Therefore the remainder of the data analysis is restricted to first presentations only.

The next question we ask is whether a better pausing strategy will be extra helpful for short words. In other words: is there an interaction between word length and pause strategy? Consider table II:
As predicted, the effect of speech pause insertion at prosodically motivated locations is very large for monosyllabic words. Not only compared with less obvious strategies like (1) and (2), but also in comparison with the no-pause version the improvement is substantial. We claim that this is the result of the fact that speech pause insertion after monosyllabic words in any case leads to considerable improvement of the recognition of such words and that this effect is optimized when speech pauses occur at prosodic boundaries in a sentence.

The next question we want to explore concerns the position of a word within a (phonological) phrase. Is the recognition of words at the end of a (phonological) phrase easier when the phrase boundary is overtly marked? Consider:
Phrase final words were recognized significantly better when followed by a speech pause than when they were not followed by a pause. When phrase-penult words were followed by a speech pause we found a negative effect of this pause. The same thing happened when we inserted pauses after non-phrase-final or non-phrase-penult words. We can conclude from this observation, that the intelligibility of the long sentences as we used them in this experiment has substantial benefit from speech pauses inserted after phrase-final words.

We already saw that monosyllabic words benefit from speech pauses inserted at prosodic boundaries. After concluding that pausing after phrase final words is the best strategy for improving intelligibility, we predict that recognition of short (monosyllabic) words will be easier when they are phrase final and followed by a pause. Consider the next table.

Table IV: Scores (in rau) after first presentation of correctly recognized morphemes (-inflection) from words followed or not followed by a pause, broken down by the position of the words within a phrase, for three different word lengths. In parentheses the total number of cases per cell.

<table>
<thead>
<tr>
<th></th>
<th>Monosyllabic words</th>
<th>Bisyllabic words</th>
<th>Polysyllabic words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not final/penult</td>
<td>phrase penult</td>
<td>phrase final</td>
</tr>
<tr>
<td>not followed by a pause</td>
<td>30 (49)</td>
<td>26 (37)</td>
<td>40 (63)</td>
</tr>
<tr>
<td>followed by a pause</td>
<td>9 (3)</td>
<td>38 (3)</td>
<td>75 (17)</td>
</tr>
<tr>
<td></td>
<td>-21</td>
<td>+12</td>
<td>+35</td>
</tr>
</tbody>
</table>

Pauses inserted after phrase-final words maximally improve the intelligibility of low quality speech. This effect is - as we predicted - stronger as the words are shorter. Importantly, phrase-penult monosyllabic words still benefit from a pause, whereas the longer words that are not phrase-final suffer severely. Short words are presumably very sensitive to pause insertion, and by marking the word-ending of those words the pause helps to recognize them, even when the pause does not coincide with a prosodic boundary.

There is one final question we wish to address: do phrase-penult (or even earlier) words benefit from a pause that is inserted after the phrase-final word? One could imagine that the positive effect of the pause at a prosodic boundary could stretch over more than just the phrase-final word. If phrase-penult words in the version with overtly marked prosodic boundaries should be recognized better than in the version without pauses, this has to be due to the speech pause after the phrase-final word. Consider table V:
**EFFECTS OF PAUSE INSERTION**

<table>
<thead>
<tr>
<th></th>
<th>not final or penult</th>
<th>phrase-penult</th>
<th>phrase-final</th>
</tr>
</thead>
<tbody>
<tr>
<td>no pauses</td>
<td>50 (92)</td>
<td>49 (92)</td>
<td>51 (187)</td>
</tr>
<tr>
<td>pauses at prosodic</td>
<td>50 (92)</td>
<td>50 (92)</td>
<td>56 (187)</td>
</tr>
<tr>
<td>boundaries</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Clearly, phrase-penult words do not benefit any more from the pause inserted after the following phrase-final word. We conclude that the effect of the speech pause does not stretch over more than one word.

4. DISCUSSION

The results of the experiments show that it is very helpful for listeners to insert speech pauses, but only if they are inserted at the right locations: the intelligibility of low quality speech improves when speech pauses are inserted at prosodic boundaries, but deteriorates when other locations are chosen. Therefore we suggest that when speech pauses are to be inserted we can best do this at prosodic boundaries. If the algorithm used to derive these locations in printed text should be too complicated for a given text-to-speech application, the only suitable alternative is to insert no pauses at all.

A speech pause has a positive effect on the recognition of only the word immediately preceding it, which in our pausing strategy should be a phrase-final word; there is no effect of this pause insertion on recognition of non-phrase-final words.

**Acknowledgement:**

This work is supported by the Linguistics Foundation, which is funded by the Netherlands Organisation of Scientific Research, NWO (grant 300-161-035), and by the Speech Technology Foundation, which is funded by the Netherlands Stimulation Project for the Information Sciences, SPIN.

**Notes:**

[1] Although the total experiment also contained material in comparable natural speech conditions, in this paper we will limit the presentation to the recognition of synthesized speech.

[2] Results are often expressed as proportions or percentages. A difference in percent however has a different likelihood of occurrence, by chance alone. In a different (especially the upper and lower) part of the
EFFECTS OF PAUSE INSERTION

scale; thus differences in percent are not comparable: the scale is not linear and additive. An arcsine transformation is often used to avoid this problem. A disadvantage here is that the arcsine values do not seem to have any obvious relation to the original percentages. Therefore we applied a simple linear transformation of the arcsine transform, that produces values that are numerically close to the original percentages. We will present all our results in rationalized arcsine units (rau); for details: cf. [12].

[3] Dutch words are very often morphologically complex, each containing several units that are separately stored in the mental lexicon. Therefore we expressed intelligibility in terms of correctly recognized morphemes excluding inflections.

5. REFERENCES