EFFECTS OF CONTEXT AND LEXICAL REDUNDANCY ON CONTINUOUS WORD RECOGNITION

PETER J. SCHARPFF

Dept. of Linguistics/Phonetics Laboratory,
Leyden University, P.O. Box 9515,
2300 RA Leiden, The Netherlands

ABSTRACT

Word recognition research typically focusses on the recognition of isolated words. Yet in actual speech perception the correct or incorrect recognition of earlier words will be crucial to the recognition of later words in the sentence and vice versa. Using an ongoing gating technique, the effects of lexical redundancy (monosyllabic vs. polysyllabic words) and speech quality (synthetic speech, degraded natural speech, high quality natural speech) on word recognition were investigated.
The results reveal that sentences composed of short words are more difficult to understand than sentences with longer words, as can be predicted by e.g. the Cohort model of word recognition. Also, it appeared that when a word could not be recognized instantaneously (as often occurs in low quality speech), chances of a postponed recognition on the basis of following context abruptly decrease when more than 4 words (or 7 syllables) have elapsed. Such delayed recognition of earlier words typically occurs at constituent boundaries.

INTRODUCTION

When a listener hears a sequence of sounds like "Inabankmanagersoff..." he can't be sure yet whether this would be the beginning of the sentence

(1) In a bankmanager's office law and order must rule.

or

(2) In a bank, managers offer a lot of service to customers.

A decision as to how the incoming sounds should be divided into words can be made only when we have heard enough of the following context to solve the ambiguity. Such ambiguities pose problems to the listener, especially when the segmental quality of speech is poor, e.g. as a result of background noise or due to the fact that speech is produced by a machine.
The number of alternative interpretations that the listener must keep in mind during the process of recognition can be very large, and the listener will need relatively much of the following context to solve an ambiguity. These kinds of problems are caused by the fact that the listener does not know where to place word boundaries. When giving away those boundaries we will help the listener to solve ambiguities and to integrate the sounds he has already heard. This can be done by means of prosodic word boundary markers like a pitch rise at the end of a phrase, a non-final pitch fall between two rises or a speech pause (all three accompanied by lengthening of the preceding syllable).
In previous research (see [1] and [2]) it was shown that it is possible to reduce the negative effects of poor segmental intelligibility by placing a clear speech pause after, for instance, every related group of words. In this research the recognition percentage increased with 10 points as a result of pauses edited into the speech.
When prosodic boundary markers are to be edited in continuous speech, these have to be inserted at those places that help the listener recognize the speech as much as possible.
Not only does reduced speech quality affect the intelligibility but also word length can play an important role in the delay of word recognition. Long (polysyllabic) words will be recognized early relative to their word length as opposed to short (monosyllabic) words. This effect can be explained as a result of the inherent lexical redundancy of longer words. Such redundancy is generally absent in short words. When a listener hears the sound sequence "eleph..." he will undoubtedly recognize (under perfect listening conditions) the word "elephant" even if he has not heard the final syllable yet, because there is no other (monomorphematic) word in his vocabulary that begins with this sound sequence. The moment that a listener has heard enough of the sound material to determine which word it will be, is called the recognition point of that word. It will be clear that shorter words contain far less or even no lexically redundant material. The lack of redundancy in words results in a shift of the recognition point towards, or even beyond the word end. This tendency will even be increased by the effect of degraded speech quality. In such cases a listener will need more of the following context to solve his recognition problems.
In an experiment systematically varying word length and speech quality we have examined the following questions:

a. To what extent does word length (or lexical redundancy) influence the recognition of words in connected speech?
b. What is the maximal stretch of following context that a listener may use to facilitate the recognition of a word?

METHOD

When we want to establish the positions in a sentence where most of the recognition problems arise and how long such problems may persist for a listener, we must be able to trace responses from the listener from moment to moment. This is possible when we use a gating technique in presenting stimuli to subjects. The technique used in this experiment presents fragments of sentences to subjects that are lengthened on each following presentation, until eventually the listener has heard the whole sentence. The length of one increment used in this particular experiment is a speech fragment that begins in the middle of the vowel of a lexically stressed syllable and ends in the middle of the vowel of the next stressed syllable (roughly comparable to a 'foot'). The first fragment is of course from the sentence onset to the middle of the vowel from the first stressed syllable.
For each sentence three versions were constructed with different speech qualities: hi-fi natural speech, natural speech degraded by amplitude-modulated white noise, and diphone synthesis using a Philips MEA 8000 speech chip. The rationale behind including degraded natural speech was that we wished to check whether the same type of errors were obtained under poor speech quality irrespective of the precise type of degradation.

MATERIAL

Pairs of sentences were constructed in which we varied poly- and monosyllabic words in the same syntactical structure and with a similar meaning. For example:

(3) Een knecht vond het kind op de stoep van zijn huis.
(A servant found the child on the doorstep of his house.)

and

(4) Een agrarier ontdekte de vondeling in een weiland nabij zijn boerderij.
(An agrarian discovered the foundling in a field near his farm.)

Thirty subjects were asked to listen to the stimuli each time guessing what word the word fragment they heard last would be the beginning of. They had to type their responses into a computer, that was programmed to analyse the answers on what was correct and what was not. After having been informed what words had been correct, the subjects listened again to the sentence now lengthened with one 'foot' of context, corrected their earlier response when necessary and added what they had recognized of the newly heard sound sequence. All responses of the subjects throughout all stages of the experiment were stored in computer memory.

RESULTS AND CONCLUSIONS

Because in the material only content words were systematically varied with respect to word length, we analysed only the responses to those words.

Turning to the first question of the experiment, whether word recognition is more difficult in the versions with short words than in the versions with long words, we find that the longer words were indeed recognized better than the short words: 96% versus 92.5% correct. The difference is fairly small. However when we look at table I, we see that the difference in word recognition of long and short words is substantially larger for the synthetic speech quality:

	short words	long words
hifi	99.9%	99.8%
noise	95.5%	97.7%
synthetic	82.0%	90.4%
mean	92.5%	96.0%

Table I. Percentage correct recognized content words after final presentation. N [short words] = 2400; N [long words] = 2400.

There is no difference at all between the word recognition of long and short words under hifi speech quality. The versions with noise were still recognized better than the synthesized versions, because, as we analysed, we found that listeners get used to the noise; learning effects were much smaller for synthetic speech. In pilots the noise level masking the human speech was adjusted so as to make degraded human speech as (un)intelligible as the diphone synthesis. However, due to the much shorter exposure times in the pilots, no differences in learning effects were discovered before the main experiment.
The differences between the three speech qualities were all significant. This leads us to conclude that words are more difficult to recognize when speech quality gets worse. Moreover, it appears that recognition of short words suffers more from the negative effect of degraded speech quality than that of long words.

The next question to be answered concerns the maximal stretch of following context that a listener may use to facilitate the recognition of a word. Consider the next figure:

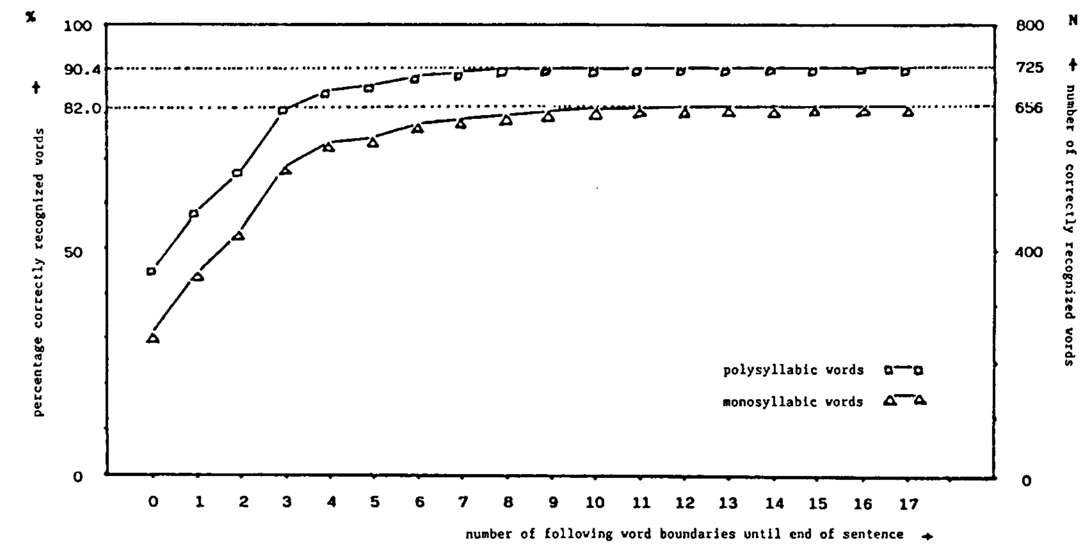

Figure 1. Word recognition of speech synthesized from DIPHONES as a function of the number of word boundaries following a target word. Zero boundaries means: subject heard only part of the target without any following context.

In this figure we have plotted % correctly recognized targets, for synthetic speech only, as a function of the length of the following speech context (expressed in number of words following the target in the audible fragment). Notice, first of all that words synthesized from diphones were recognized less then 40% correct when only their first part (up to and including half of the lexically stressed syllable) is made audible. Even when one foot is added (comprising the integral target as well as at least one other word), recognition is still at 50%. Recognition scores continue to rise as more of the following context is made audible, until 3 complete words have elapsed. The curve then quickly asymptotes when more than 3 words are added to the target.
Context further away than 3 words apparently does not help the listener in finding earlier words that he did not recognize. What has happened when the listener reaches the fourth word? Considering the structure of our stimulus sentences we find that most of the word groups (constituents) contain three words so that the next word is the onset of a new constituent. We argue that later words do not help the listener to recover an earlier unintelligible word across a constituent boundary. This is borne out by the following table which presents percentage content words recognized with or without later context, broken down by word position within the phrase (constituent).

	recognized at 1st partial presentation	recognized after adding one gate	
phrasefinal words 80% (1280)	39% (500)	73% (936)	34% (436)
non-phrasefinal words 20% (320)	35% (111)	79% (253)	44% (142)

Table II. Recognition of synthesized words at different positions in the constituent. N [diphone quality] = 1600. Increased recognition in the case of phrasefinal words is on basis of extra information from a following constituent, in the case of non-phrasefinal words on basis of added information from within the same constituent, $X^2(1)=7.28$ $(p<.01)$.

A phrase-penultimate word is recognized on the basis of later context significantly more often than a phrase-final word, $X^2(1)=7.28$ $(p<.01)$. We can explain this effect by assuming that transitional probabilities between words are much higher within constituents than across constituent boundaries.

DISCUSSION

Additional context within a constituent seems to enable listeners to recover non-recognized earlier words. We also found that non-phrasefinal words were recovered on the basis of following context more often than phrasefinal words. We take this to be an indication that listeners tend to recognize words in phrases. Therefore, if we are to help the listener recognize words in poor speech quality (synthesized speech), we shall have to mark phrase boundaries with effective prosodic markers.

Acknowledgement:

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for the advancement of pure research, ZWO.

References:

[1] B.A.G. Maassen, "Marking word boundaries to improve the intelligibility of deaf speech", in: Artificial corrections to deaf speech studies in intelligibility, Enschede, Holland, 1985.
[2] S.G. Nooteboom, "The temporal organisation of speech and the proces of spoken word recognition", IPO Annual Progress Report, Eindhoven, Holland, 1983.