PERCEPTUAL GOALS OF SPEECH PRODUCTION

Sieb G.Nooteboom

Research Institute for Language and Speech, Utrecht University

ABSTRACT

In this paper predictions are made about the production of words embedded in fluent connected speech, based on a principle of cooperative behaviour combined with insights in perception. It is concluded that the effects of such cooperative behaviour, in as far they are not brought about by linguistic means, are real but relatively small.

1. INTRODUCTION

Sometimes speakers speak in order to be understood. On those occasions, they are well advised to adapt their pronunciation to the estimated needs of their audience. I will call this the principle of cooperative behaviour. It entails that a speaker spends more time and effort on the pronunciation of parts of his message that are essential to recognition and comprehension than on parts of his message that are to some degree redundant (output- oriented control when necessary versus system-oriented control when permitted [14]

In this paper I will predict regularities in speech production from insights in speech perception. In some cases these predictions are corroborated by experimental evidence, in other cases evidence is controversial, in other cases again it is still lacking, and has to be filled in by others who know more than me, or by future research. I will focus on embedded words and discuss lexical redundancy, word onsets, word boundaries, and contextual redundancy.

2. LEXICAL REDUNDANCY

A monosyllable like English PAT contains little redundancy. Each of its constituting phonemes is not fully predictable from the other ones. This is different for a polysyllabic word like HIPPOPOTAMUS, that is highly redundant, in the sense that the word remains recognizable even when a few of its constituting phonemes are missing out completely.

From a listener's point of view this means that the word HIPPOPOTAMUS remains recognizable when it is rapidly and sloppily pronounced, whereas the word PAT can only be recognized correctly when it is pronounced carefully, and therefore, I assume, more slowly, so that all three phonemes are identifiable.

In general, lexical redundancy increases with word length. Therefore the need for slow and careful articulation decreases with increasing word length. From this we may predict that speed of articulation increases with increasing word length. This seems a plausible explanation of the well known phenomenon of time compression in polysyllabic words. From hereon I will take for granted that, other things being equal, monosyllabic words are more vulnerable to communication noise than polysyllabic words, and therefore liable to be pronounced more slowly and more carefully.

3. WORD ONSETS

Lexically redundant words can be recognized on the basis of a fragmentary stimulus, for example when a considerable portion of either the end or the beginning of the word stimulus is missing. Some time ago I demonstrated that words are more easily and faster recognized from initial than from final word fragments, when initial and final fragments nominally contain equal amounts of lexical information [18]. This difference appears to be related to the fact that word onsets ensure fast and proper alignment of the stimulus with word candidates, whereas word endings do not [20]. I predict that cooperative speakers spend more time and effort on onsets than on endings of embedded words, especially when the words concerned are contextually little redundant.

I have no clear and direct evidence in favour of this prediction, and would be grateful to anyone who does. Indirect evidence would be the prevalence of regressive over progressive assimilation and coarticulation on word boundaries, as seems to be case in Germanic Languages. Whether this tendency is universal, I do not know.

The relative importance of word onsets to recognition should also make it profitable for a language to have word initial stress. For English it has been argued that a listener's strategy considering each stressed syllable as a potential word onset is profitable [6]. From this it has been predicted that special word onset markers are more to be expected when a word starts with an unstressed than when it starts with a stressed syllable. This expectation was to some extent corroborated in a production study [4].

In clear speech, for example in noisy environments, informationally important words are sometimes set off by speech pauses. From the relative importance of word onsets one would predict that such speech pauses are more liable to be made before than after important content words. This seems to occur in certain styles of reading aloud [8]. The evidence is easily confusing, however, because informationally important words are often at the end of phonological phrases, potentially followed by a phrase boundary marking speech pause.

4. WORD BOUNDARIES

Whatever the relative importance of word onsets and endings, word boundaries in connected speech are potentially important to word recognition. Knowing where a particular word ends, is knowing where the next word begins, and this saves an awful lot of trouble in lining up potential word candidates with the incoming signal, as is well known from problems in the machine recognition of connected speech. Real perceptual ambiguities such as **lettuce** versus **let us**, or **budget** versus **budge it** [5], however, attest to the fact that word boundaries need not (always) be clearly marked.

Yet, it seems reasonable to predict that cooperative speakers, also in fluent, perhaps even fast, connected speech, tend to make subtle, little time consuming phonetic word boundary markers that aid listeners in finding word boundaries. This was demonstrated to be correct for ambiguous Dutch two-word combinations of the type **known ocean** versus **no notion**. Excised from fast connected speech, such speech fragments were recognized correctly in a binary forced choice task well above chance (80%) [22]. 80% is not 100%, however. In the absence of disambiguating context, ambiguity remains.

Perceptual ambiguity can also arise due to assimilation and/or degemination on word boundaries, as in **hold back** being confusable with **whole back**. This type of perceptual confusion is an immediate function of global speech tempo. In rapid connected speech the **whole back**-stimulus leads as easily to a **hold back**-response as the **hold back**-stimulus leads to a **whole back**-response [16]. Apparently, optional assimilation and degemination on word boundaries is fully incorporated in the word recognition strategies employed by listeners. One expects that the actual occurrence of assimilation and degemination is a function of lexical and contextual redundancy. This has not yet been tested.

In normal conditions, long polysyllabic words often are recognized before the end of the stimulus word has come in [15], but short monosyllabic words, if not predictable from preceding context, can rarely be recognized before the end of the word, and, if embedded, are recognized only during or after the processing of the immediately following word [1]. A speech pause immediately following a monosyllabic content word would therefore be much more helpful to recognition than a speech pause following a lexically redundant polysyllabic word. This was experimentally shown to be correct [19]. From this I predict that cooperative speakers are more liable to insert speech pauses after monosyllables than after polysyllables. I do not know whether this is actually so.

5. CONTEXTUAL REDUNDANCY

An effect of contextual redundancy seems confirmed by Lieberman's famous example **a stitch in time saves nine**, used to show that highly predictable words are less clearly pronounced than unpredictable words [12]. Lieberman's findings were more recently confirmed and extended [10],[11]. However, none of these studies pulled apart the effects of contextual redundancy and accentuation.

It has been argued that accented words are processed faster by listeners than unaccented words [3],[7]. This is not always so. Terken and Nooteboom [21] showed that 'new' words are processed faster when (correctly) accented than when (incorrectly) unaccented, but 'given' words are processed faster when (correctly) unaccented than when (incorrectly) accented. From this it follows that cooperative speakers should take care to produce accents on words carrying new information, but not on words carrying given information.

The issue is somewhat more complicated, however, because of the phenomenon of 'unit accentuation' [2]. A word group like **french cheese** can be marked as informationally important or as carrying new information, by a single accent on **cheese**. The word **french** is then 'new' but not accented. The principle of cooperative behaviour predicts that it will nevertheless be more carefully and more slowly pronounced than when it is 'given'. This prediction is falsified by Eefting [9], who, taking advantage of the phenomenon of unit accentuation, varied contextual redundancy and context-induced accentuation independently in a production study with read aloud text. She found that, other things being equal, accented words are considerably longer than unaccented words, but unaccented 'new' words are not significantly longer than unaccented 'given' words. Apparently sometimes the consequences of cooperative behaviour on one level, in this case the level of the word, are suppressed by the consequences of cooperative behaviour at another level, in this case the level of accent patterns, the correct realization of which constrains temporal patterning.

6. CONCLUSION

In this contribution I have focussed on the level of the word, because of my belief that the struggle between cooperative and self-indulgent behaviour of speakers will be most apparent where form and meaning come together. In quite a few cases evidence was found for systematic variations in word production due to the alternation between output- and system-oriented control.

Yet, the most striking result of this enterprise is in my own judgement that these variations are often relatively small and unimportant, at least when they are not supported by or drawing on conventionalized rules or structures of the language. Of course, the effect of time compression in polysyllabic words is not a small effect, but does not seem to reflect spontaneous adaptive behaviour. It rather is canonized in the sense that it belongs to what native speakers know about their language: shrinking monosyllables and stretching polysyllables are perceived as incorrect [17]. Non-phonological acoustic-phonetic word boundary cues are there, but relatively small and not very effective in normal conditions. Phonotactic word boundary cues might be effective, but tend to be obliterated by regular conventionalized rules of assimilation and degemination. Contextual informativeness and redundancy, if not expressed by conventionalized and rule-governed accent patterns, have only marginal effects on pronunciation.

Summarizing, we can say that the effects discussed are real but not very impressive. It seems that for normal speech communication conditions, the tools a speaker finds at his disposal in the set of structures and rules that make up his language, are by and large sufficient for his purposes, and that he has some but relatively little need of adapting his speech in ways that are not rule-governed.

It is plausible, of course, that this state of affairs is the result of the adaptive nature of language [13]. Strong and regularly occurring adaptive behaviour of speakers is easily conventionalized and thereby becomes part of the language. If this is correct, the adaptive behaviour of speakers can be studied as a source of language change. Such adaptive behaviour should perhaps more than we have done so far be studied in less favourable but real communication conditions. But we should always be aware that the expected adaptations on one level can be severely constrained by the requirements of another level, as in the example of word durations being controlled by accent patterns.

7. REFERENCES

[1] Bard, E.G., Shillock, R.C., and Altmann, G.T.M. (1988), "The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context", *Perception & Psychophysics, 44, 5,* 395-408.

[2] Chafe, W. (1974), "Language and consciousness", *Language 50,* 111-133.

[3] Cutler, A. (1976), "Phoneme-monitoring reaction time as a function of preceding intonation contour", *Perception & Psychophysics, 20,* 55-60.

[4] Cutler, A. (1990), "Production and perception of word boundaries", paper presented at the ATR Workshop on Speech Perception and Production, held in Kyoto, november 15-16, 1990.

[5] Cutler, A. and Butterfield, S. (1990), "Syllabic lengthening as a word boundary cue", to be printed in the Proceedings of the SST-90, Melbourne.

[6] Cutler, A. and Carter, D.M. (1987), "The predominance of strong syllables in segmentation for lexical access", *Computer, Speech and Language, 2,* 133-142.

[7] Cutler, A. and Foss, D.J. (1977), "On the role of sentence stress in sentence processing", *Language and Speech, 20,* 55-60.

[8] De Rooij, J.J. (1979), *"Speech punctuation",* unpublished doctor's thesis, Utrecht.

[9] Eefting, W. (1991), "The effect of "information value" and "accentuation" on the duration of Dutch words, syllables and segments", *Journal of the acoustical Society of America, 89, 1,* 412-424.

[10] Fowler, C.A. and Housum, J. (1987), "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction", *Journal of Memory and Language, 26,* 489-504.

[11] Hunnicut, S. (1985), "Intelligibility versus redundancy-Conditions of dependency", *Language and Speech, 28,* 45-56.

[12] Lieberman, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech, 6,* 172-187.

[13] Lindblom, B. (1989), "Phonetic invariance and the adaptive nature of speech", *In: "Working Models of human Perception", edited by B.A.G.Elsendoorn and H.Bouma,* Academic Press, London, 139-173.

[14] Lindblom, B. (1990), "Explaining phonetic variation: a sketch of the H&H theory", *In: "Speech Production and Speech Modelling", edited by W.J.Hardcastle and A.Marchal,* Kluwer Academic Publishers, The Hague, 403-439.

[15] Marslen Wilson, W. (1980), "The temporal structure of spoken language understanding", *Cognition, 8,* 1-71.

[16] Menert, L. (1989), "Perceptual ambiguity as indicator of voice assimilation", *In: "OTS-Yearbook 1989", edited by P.Coopmans, B.Schouten, W.Zonneveld,* 63-73.

[17] Nooteboom, S.G. (1973), "The perceptual reality of some prosodic durations", *Journal of Phonetics 1,* 25-46.

[18] Nooteboom, S.G. (1981), "Lexical retrieval from fragments of spoken words: beginnings versus endings", *Journal of Phonetics, 9,* 407-424.

[19] Nooteboom, S.G., Scharpff, P. and Van Heuven, V.J. (1990), "Effects of several pausing strategies on the recognizability of words in synthetic speech", *Proceedings of the First International Congress on the Processing of Spoken Language,* Kobe, Acoustical Society of Japan, 385-387.

[20] Nooteboom, S.G. and Van der Vlugt, M.J. (1988), "A search for a word-beginning superiority effect", *Journal of the acoustical Society of America, 84, 6,* 2018-2032.

[21] Terken, J. and Nooteboom, S.G. (1987), "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information", *Language and Cognitive Processes, 2,* 145-163.

[22] Quené, H. (1991), "Acoustic-phonetic cues for word segmentation", submitted for publication.