HEADS AND TAILS IN WORD PERCEPTION: EVIDENCE FOR 'EARLY-TO-LATE' PROCESSING IN LISTENING AND READING

Sieb G. Nooteboom[1] *and Meinou van Dijk*[2]

[1]Utrecht Institute of Linguistics OTS, [2]Former student at the Phonetics department, Utrecht University

ABSTRACT

Sequential models of word perception assigned a very special role to word onsets. This accounted in a natural way for evidence that lexical access is easier from word beginnings than from word endings, a property speech perception shares with reading. Sequential models badly failed on other scores, however. More recent competition models seem to give equal weight to stimulus information, independent of position within the word. The present word recognition experiment aimed at testing the hypothesis that, other things being equal, mismatches are more damaging to word perception at onsets than at offsets of embedded words, both in speech perception and in reading. Results show that word recognition is quite good in all conditions, even when word onsets are mutilated, and mis-timed, thus lending support to competition models. Yet, the results also show that lexical access is modulated by some early-to-late or left-to-right component, as if human word perception displays a mixture of sequential and competition processing.

1. INTRODUCTION

Recent models of lexical access from speech have abandoned the earlier idea that words are recognized sequentially. Sequential recognition of words was a basic feature of earlier models (e.g. Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978), and implies that perceived word boundaries result from word recognition. The idea was that most words in connected speech are recognized before the acoustic offset is heard, and that so the listener knows where to start recognition of the next word. Hearing *cardin..* leads to recognition of CARDINAL, and thus processing for the next word starts at or after the *l*. In such models lexical forms compete only with other forms having the same input phoneme (or same other input unit) as a possible onset. It has been argued convincingly that such models cannot account for the perception of speech, mainly because of the frequency of "words within words" (such as FOR and OR in FORM), and because it can be shown that in connected speech very often word tokens cannot be recognized without later coming information (Norris, 1994; McQueen, Cutler, Briscoe, & Norris, 1995; Bard, Shillcock, & Altmann, 1988).

In more recent competition models lexical hypotheses potentially compete as soon as there is some temporal overlap in the way they fit the input string. Given the string *mycardinal..*, at some point in time there may be simultaneous competition between MY, MICA, ICON, CAR, CARD, CARDIAC, ARE, ARDENT. etc. Note that for example CARD can only be dismissed after it becomes evident that CARD + *inal* does not give a meaningful string. Such aspects of word perception are best modeled by competition models, as the localist SHORTLIST model proposed by Norris (1994), taking a phoneme string as its input, or the distributed network model proposed by Gaskell and Marslen-Wilson (1997), taking distinctive features as input units. Typically such models are much more robust than sequential models in that they efficiently use all the available input information, and make the process of word perception not depend on the availability in the stimulus of the word onset. A corollary of this is that, other things being equal, such models give equal weight to input units independent of position in the (hypothesized) word. Below we will mainly focus on the SHORTLIST model, because this gives the most precise predictions for our purposes.

The SHORTLIST model among many other things nicely accounts for the phenomenon of phonemic word restoration (Warren, 1972). It allows for an input symbol standing for noise that neither matches nor mismatches any phoneme. This ensures that *#igaret,* where # stands for input noise, is recognized as CIGARETTE (Norris, 1994). It is, of course, reasonable to assume that the noise should have been such that it could have masked the missing phoneme (Warren, Obusek, & Ackroff, 1972; Bashford, Riener, & Warren, 1992). If, for example, the noise is incompatible in duration with the missing phoneme or phoneme string, one expects lexical access to be more difficult than when the duration fits the missing phoneme or phoneme string. Assuming that noise of a duration that is much too long for a single phoneme would translate into a sequence of phoneme-sized input symbols (suggested by Norris, personal communication), and following a recent extension of SHORTLIST (Norris, McQueen, Cutler, Butterfield, 1997), this could be modeled by the Possible-Word Constraint. This constraint employs potential word boundaries in the input, derived from silences, strong syllable onsets, and phonotactic constraints, and penalizes lexical hypotheses that leave non-vocalic material dangling between two potential word boundaries. The important point here is that the penalty is independent of whether the constraint is violated at the onset or the offset of the hypothesized lexical unit. It is precisely this point the present experiment is taking up, by testing the opposite hypothesis, viz. that an acoustic-phonetic mismatch of this kind is more damaging to word perception at word onset than at word offset. We assume that noise of fitting duration does not and noise of inappropriate duration does constitute a mismatch (Cf. Nooteboom and Van der Vlugt, 1988)

Below we will describe an experiment testing this hypothesis both for speech perception and for reading, where a word beginning superiority effect is also well documented (Cf. Nooteboom, 1981; Nooteboom and Van der Vlugt, 1988). This allowed us to use, instead of continuous noise, countable discrete symbols (as the noise input symbols in SHORTLIST). Visual stimuli also allowed us to measure reaction times. This is difficult in speech when the targets are utterance-embedded words.

2. METHOD

2.1. Stimulus materials

We employed 80 Dutch three-syllable monomorphemic words comparable to English ELEPHANT or CANNIBAL. For each word we determined a visual and auditory forward and backward uniqueness point by going through the word form from the onset onwards or the offset backwards until the string of phonemes or characters uniquely determined the intended word. For each word we constructed a separate sentence that did not semantically constrain the target word, of the type:

I did NOT dream of an *elephant* last NIGHT

For the speech stimuli, the resulting 80 sentences were spoken by an experienced male speaker, each sentence with pitch accents on the capitalized content words. The target word always remained unaccented, and thus durationally maximally constrained by the surrounding utterance. The utterances were stored on disk with a sample frequency of 22,050 Hz. In the experiment four versions of each utterance were used, obtained by manipulating the target word, here exemplified by *elephant*, with #'s instead of noise: 1) onset audible; offset replaced with noise of fitting duration: *eleph###*; 2) onset audible; offset replaced with overly long noise: *eleph################*; 3) offset audible; onset replaced with noise of fitting duration: *##ephant*; 4) offset audible; onset replaced with overly long noise: *##############ephant*. Audible parts of the target words always corresponded to lexically unique parts of the phonological form. This was done to keep the stimulus information under strict control. Removing of the complements was done under auditory and visual control, taking care that the abutting phoneme remained clearly identifiable, but that all traces of coarticulation with the initial or final phoneme of the complement were removed. Noise was inserted during the experiment by a computer programme for experimental control. Noise amplitude was chosen such that the noise could just have masked the removed parts but was not disagreeably loud. Noise duration either corresponded exactly to the duration of the removed fragment, or was fixed at 800 ms, depending on the condition. As all 80 words were to be used in all four conditions, we had 320 different speech stimuli. Organization of the visual stimuli was similar. These were presented on a computer screen in a large font (Courier) that was nonproportional. Each sentence was first presented with an empty space for the mutilated target and then the mutilated target appeared in the empty space (see *Procedure*). Of course there were also four conditions, corresponding to those for the speech stimuli, and 320 different visual stimuli. In conditions 2) and 4) the sequence of #'s, playing the same role as noise in speech, had a length of 14.

2.2. Design

Independent variables followed a 2 × 2 × 2 matrix: auditory versus visual; fitting versus fixed; onsets replaced versus offsets replaced. This gave 8 cells. We used a fully blocked design, with 8 groups of 10 target words and 8 groups of 10 subjects.

2.3. Subjects

Subjects were 80 students having Dutch as their native language and with no known hearing or seeing deficiencies. Ages varied from 18 to 29. Subjects were paid for their participation.

2.4. Procedure

The experiment was run under computer control. For each group of 10 subjects there were four blocks of stimuli, viz. auditory with fitting noise, auditory with fixed noise, visual with fitting string of #'s, visual with fixed string of #'s. Within each block the order of replaced onsets and offsets was random. Order of blocks was systematically varied over subject groups. Subjects performed their task individually, in a sound treated booth, sitting at a table. They used high quality headphones for hearing the speech stimuli and a high resolution computer color screen for watching the visual stimuli. Each block of stimuli was preceded by 6 exercise items. Between each speech stimulus and a short tone announcing that the next stimulus would come after 350 ms, the subject had 5 seconds to write down the perceived word on an answering form, with stimulus numbers preprinted. The current stimulus number was displayed on the computer screen. Each visual frame sentence was displayed on the screen for 3 seconds with an empty slot for the stimulus. Then the target stimulus appeared in the empty slot for 350 ms. The subjects were instructed to pay attention to the frame sentence while it was visible, and to react as fast as possible to the target stimulus by speaking the perceived word. Reaction times were measured automatically with a voice key from the onset of visual target stimulus presentation to the detectable acoustic onset of the response. Responses were recorded for later analysis. After a visual block the subject was confronted with a list of 14 sentences and asked to indicate which of these had occurred in the experiment.

3. RESULTS

Some of our stimuli were responded to with more than a single existing word. We removed these words from the data set, together with some randomly selected other words, such that there remained an equal number of 72 (instead of 80) stimulus words in each of the cells of the matrix. The main hypothesis to be tested in the present data set is that inappropriate duration of an extraneous stimulus replacing part of a word form is more damaging when the replaced part of the word is a word onset than when it is a word offset. The hypothesis includes that this is so both in auditory and in visual presentation. Table 1 provides a first breakdown of the raw percentages unsuccessful recognition:

	AUDITORY			VISUAL	
	onsets	offsets		onsets	offsets
fitting	18	6		12	6
fixed	26	9		23	13

Table 1. Percentages incorrect recognition separately for auditory and visual recognition, for replaced onsets and replaced offsets, and for fitting and fixed noise durations/lengths of strings of #'s.

From the data in Table 1 we see that overall our subjects did quite well. This, of course, is related to the fact that each stimulus was meant to give enough information for the intended word to be recognized. We predict from our hypothesis that the values for fitting onsets, fitting offsets and fixed offsets would be basically equal, and the percentage of errors for fixed onsets would be significantly higher, because of uncertainty of word onset timing.This is not exactly what we find. There. appears to be an unpredicted main effect of onsets versus offsets. On closer inspection this effect was found to be caused for a large part by the fact that phoneme strings (and less so character strings) were much more often incorrectly identified with removed onsets (and thus perceptible offsets) than with removed offsets (perceptible onsets). Whatever the causes of this effect, this is not what we are interested in. Our interest is in responses to stimuli that are correctly perceived as a string of phonemes or characters, and that uniquely determine a lexical item. We therefore limited our analysis to those cases of unsuccessful recognition where we had no evidence that the perceptible string of phonemes or letters was misperceived. Those were the cases were subjects either gave a nonsense word containing the correct string, most often identical to the string of phonemes or letters in the stimulus, or gave no response at all. All responses reflecting misperception of the input string were removed. The new breakdown of the data is given in Table 2:

	AUDITORY			VISUAL	
	onsets	offsets		onsets	offsets
fitting	12.5	5		6.3	2.6
fixed	16.9	6.8		13.8	6.4

Table 2. Percentages incorrect recognition separately for auditory and visual, for replaced onsets and replaced offsets, and for fitting and fixed noise durations/lengths of strings of #'s. 100% is the number of responses where there was no evidence for misperception of phoneme or character string.

The data in Table 2 show, as predicted, the highest percentages of errors for "fixed onsets", both for auditory and for visual presentation, but otherwise the pattern of data is not as predicted. We ran two analyses of variance, with "auditory versus visual", "fitting versus fixed" and "onsets versus offsets" as fixed factors, one with listeners and one with words as replications. In both analyses two main effects, and in one analysis all three main effects were, against prediction, highly significant. The predicted interaction between "fitting versus fixed" and "onset versus offset" did not reach reliable significance (Listeners as replicas: auditory vs visual $F(1,576)=13$, $p<0.001$; fixed vs fitting $F(1,576)=11$, $p<0.001$; onset vs offset $F(1,576)=25$, $p<0.001$; fixed vs fitting × onset vs offset $F(1,576)= 1.9$, $p<0.162$; Words as replicas: fixed vs fitting $F(1,568)=23$, $p<0.001$; onset vs offset $F(1,568)=53$, $p<0.001$; fixed vs fitting × onset vs offset $F(1,568)=3.3$, $p<0.07$).

The errors of which percentages are given in Table 2 fall in two categories, viz. nonsense words agreeing with the input and zero responses. If replacing a word onset with some stimulus of inappropriate duration or length would be more upsetting to the perceiver than replacing a word offset with such a stimulus, one would certainly expect the subject more often to fail to give a response at all in the first than in the second condition. Table 3 presents the relevant breakdown of the data:

	AUDITORY			VISUAL	
	onsets	offsets		onsets	offsets
fitting	1.4	.14		3.3	3.33
fixed	5.7	.28		7.5	2.5

Table 3. Percentages zero responses separately for auditory and visual, for replaced onsets and replaced offsets, and for fitting versus fixed noise durations/lengths of strings of #'s. 100% is the number of responses where there was no evidence for misperception of phoneme or character string.

Here again we see that, as predicted, percentages of zero responses are highest for "fixed onsets", both for auditory and visual presentation. We ran again two analyses of variance, one with listeners and one with words as replicas. Against prediction these showed three significant main effects. This time the predicted interaction was also significant in both analyses. (Listeners as replicas: auditory vs visual $F(1,576)=17$, $p<0.001$; fixed vs fitting $F(1,576)=24$, $p<0.001$; onset vs offset $F(1,576)=45$, $p<0.001$, fixed vs fitting × onset vs offset $F(1,576)=12$, $p<0.001$; Words as replicas: auditory vs visual $F(1,568)=10$, $p<0.001$; fixed vs fitting $F(1,568)=15$, $p<0.001$; onset vs offset $F(1,568)=51$, $p<0.001$; fixed vs fitting × onset vs offset $F(1,568)=12.5$, $p<0.001$).

In the visual part of the experiment we also measured reaction times. Here the prediction would be that reaction times for "fixed onsets" are longest. Table 4 gives the relevant breakdown:

	VISUAL	
	onsets	offsets
fitting	930	880
fixed	1110	995

Table 4. Reaction times in ms to visual stimuli in recognition of embedded words, separately for replaced onsets and replaced offsets of stimulus words, and for fitting and fixed noise durations/strings of #'s.

As predicted, we find the longest reaction times for "fixed onsets". An analysis of variance with "onsets versus offsets" and "fitting versus fixed" as fixed factors reveals once again significant main effects for both factors. We also find the predicted interaction between those factors, again lending support to our initial hypothesis (onsets vs offsets $F(1,2408)=58$, $p<0.001$; fitting vs fixed $F(1,2408)=91$, $p<0.001$; onsets vs offsets × fitting vs fixed $F(1,2408)=5$, $p<0.05$).

4. DISCUSSION

In the present experiment we have replaced either initial or final fragments of embedded word forms with some extraneous stimulus, broadband noise in the case of speech and a string of #'s in the case of printed words. The duration or length of the extraneous stimulus was either the same as that of the missing word fragment or overly long. The remaining perceptible part of the stimulus word was meant to provide just enough information to retrieve the intended word. From older sequential models of word recognition (e.g. Cole & Jakimik, 1980, Marslen-Wilson & Welsh, 1978) one would predict that obliterating the moment

of onset of a word would severely damage lexical access, because in these models temporal alignment of lexical hypotheses with the incoming stimulus proceeded from the word onset onward. In our data all effects of replacing word fragments with an extraneous stimulus were minor, certainly if we limit analysis to those cases where we have no evidence that the perceptible parts of target stimulus words were misperceived. Obviously, in the great majority of cases, obliterating word fragments, either onsets or offsets, or even obliterating the timing/position of virtual word onsets, does not prevent subjects from correctly retrieving the intended words from the nonredundant stimuli. This in itself strongly supports competition models such as proposed by Norris (1994) and more recently by Gaskell and Marslen-Wilson (1979), where there is no special status of word onsets. Our data do show some other effects, however.

First of all we found that word recognition from print is slightly easier than word recognition from speech. This is not amazing. Far more amazing is that otherwise both sets of data follow exactly the same pattern, as if they are to the same extent and in the same way controlled by the early-to-late processing of speech. Then we found that lexical access is easier from word onsets than from word offsets, both in speech and in print. This replicates many experimental findings in the older literature, for lexical access from speech, from print, and in speech production (Cf. Nooteboom, 1981) It has been shown, though, that in speech perception this "word beginning superiority effect" does not depend on correctly perceived phonemic structure of the word onset. Correct duration of sufficiently speech-like unidentifiable sound seems to be enough (Nooteboom & Van der Vlugt, 1988). In competition models lexical hypotheses are activated by input units corresponding to any parts of the lexical forms. Yet, given the structure of our data, we have to assume that activation is strongest from input units corresponding to onsets of lexical forms, as if some shadow of the old Cohort model is still modulating lexical access.

We also found that inappropriate duration/length of the extraneous stimulus is more damaging than appropriate duration/length, both for onsets and offsets. This suggests that the duration or length of the extraneous stimulus, for all the subject knows masking uninterrupted speech or print, is actively used in accepting or rejecting lexical hypotheses. One way to look at this is as follows. From a lexical hypothesis for a partly unidentifiable stimulus the perceiver estimates the location of the potential word boundaries. If one of the potential word boundaries leaves an extensive part of the extraneous noise unaccounted for, this hypothesis is penalized, as predicted by the Possible-Word Constraint invoked by Norris et al. (1997).

Finally, and in agreement with our initial hypothesis, we found that inappropriate duration or length of an extraneous stimulus is more damaging when this stimulus replaces a word onset than when it replaces a word offset. There evidently is some asymmetry here, as if, in terms of the Possible Word Constraint, the penalty for unaccounted material is stronger at word onset than at word offset, supplying a weak ghost of the older sequential models of word perception.

Our findings, in agreement with earlier research on visual and auditory word perception, suggest that, although the word onset does not play the very special role it was assigned in the older sequential models, lexical access appears to be modulated by some early-to-late or left-to-right component favouring stimulus information corresponding to word onsets. It is as if each lexical hypothesis takes stimulus information more serious as its activation level is still low than when it is high.

5. REFERENCES

1. Bard, E.G. , Shillcock, R.C., and Altmann, G.T.M. "The recognition of words after acoustic offsets in spontaneous speech: Effects of subsequent context", *Perception and Psychophysics, Vol. 44, 395-408, 1988.*

2. Bashford, J.A., Riener, K.R. & Warren, R.M. "Increasing the intelligibility of speech through multiple phonemic restorations", *Perception & Psychophysics, Vol. 51 (3), 211-217, 1992.*

3. Cole, R.A. and Jakimik, J. "A model of speech perception", In: R.A. Cole (Ed.) *Perception and Production of fluent speech*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1980, pp 133-163.

4. Gaskell, M. G. and Marslen-Wilson, W.D. "Integrating form and meaning: a distributed model of speech perception", *Language and Cognitive processes, Vol. 12, 613-656, 1997.*

5. Marslen-Wilson, W.D. and Welsh, A.. "Processing interactions and lexical access during word recognition in continuous speech", *Cognitive Psychology, Vol. 10, 29-63, 1978.*

6. McQueen, J.M., Cutler, A., Briscoe, T., and Norris, D. "Models of continuous speech recognition and the contents of the vocabulary", *Language and Cognitive Processes, Vol. 10 (3/4), 309-331, 1995.*

7. Nooteboom, S.G. "Lexical retrieval from fragments of spoken words: beginnings versus endings," *J. of Phonet. Vol. 9, 407-424, 1981.*

8. Nooteboom, S.G. and Van der Vlugt, M.J. "A search for a word-beginning superiority effect," *J. Acoust. Soc. Amer. Vol. 84 (6), 2018-2032, 1988.*

9. Norris, D. "Shortlist: a connectionist model of continuous speech recognition", *Cognition Vol. 52, 189-234, 1994.*

10. Norris, D., McQueen, J.M., Cutler, A., Butterfield, S. "The possible-word constraint in the segmentation of speech", *Cognitive Psychology, Vol. 34, 191-243 1997.*

11. Warren, R.M. "Perceptual restoration of missing speech sounds", *Science, Vol. 167, 392-393, 1970.*