The Influence of masking words on the prediction of TRPs in a shadowed dialog

Wesseling, W.; van Son, R.J.J.H.; Pols, L.C.W.

Published in:
Interspeech

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
The influence of masking words on the prediction of TRPs in a shadowed dialog

Wieneke Wesseling, R.J.J.H. van Son, and Louis C.W. Pols

IFA/ACL, University of Amsterdam, The Netherlands
W.Wesseling@uva.nl

Abstract

It is well known that listeners can ignore disturbances in speech and rely on context to interpolate the message. This fact is used to determine the importance of individual words for projecting Transition Relevance Places, TRPs. Subjects were asked to shadow manipulated pre-recorded dialogs with minimal responses, saying ‘ah’ when they feel it is appropriate. In these dialogs, at random, of each utterance, either one of the last four words was replaced by white noise (masked condition), or no word was replaced (non masked condition). The reaction times were analyzed for effects of masked words. The presence of masked words, even prominent words, did not affect the response times of our subjects unless the very last word of the utterance was masked. This indicates that listeners are able to seamlessly interpolate the missing words and only need the identity of the last word to determine the exact position of the TRP.

Index Terms: turn taking, masking, prominence

1. Introduction

Various studies show that listeners are able to reliably predict -or project - the end of the present speaker’s turn (Transition Relevance Places, TRPs), helping them to achieve smooth transitions of speaker turns in natural conversations [3, 2]. To be able to determine if utterances are coming to an end, listeners can use a variety of information sources. Our previous experiments showed that subjects could predict TRPs reliably in an ‘intonation only’ condition [14], proving that in the absence of syntactic and lexical information, a rising or falling end intonation alone is a sufficient - although impoverished - cue for TRP projection. However, intonation might not be used to predict TRPs in normal speech [7, 14]. Another factor that seems to be an important cue for TRP detection, is the position of prominent words in the utterance. In [10] we showed that the presence of non-prominent words before a TRP reduces the delays of elicited and natural responses alike, even in impoverished speech. This suggests a model of TRP projection where the upcoming TRP can be predicted by the listener, using the last - unpredictable - prominent word as a starting point.

The information needed to predict an upcoming TRP can be split into global information about the number and type of words to expect before the TRP, and the precise end point of the last word. It can be expected that disturbing the last word will directly interfere with the timing of a TRP response. Disturbing words preceding the last word can be expected to interfere with predicting the relative position of the last word and, possibly, preparing for a response. Both of these effects depend on the predictability of the disturbed word. Prominent words are generally considered to be less predictable, and more important for understanding, than non-prominent words [10]. Therefore, it can be expected that disturbing prominent words will interfere more with the prediction of an upcoming TRP than disturbing a non-prominent word.

To test this, a reaction time paradigm was used, where subjects listened to recordings of natural dialogs, in which either one of the last four words of each utterance, or no word, was replaced by a Mask of white noise. They were asked to shadow the dialog and respond with minimal responses, saying ‘ah’, as if they were participants. The exact timing of the resulting responses is a sensitive probe into the processing of the available cues.

2. Materials and Methods

To compare processing of the masked and non-masked stimuli, a decision-making model by Sigman and Dehaene [8] is used (see fig. 1). In this model, mental decision-making is modeled as a noisy integrator that stochastically accumulates perceptual evidence from the sensory system in time [8], through a perceptual ($P$), central decision-making ($C$) and a motor component ($M$). RTs are the sum of a ($P + M$)-related deterministic response time, $t_0$, and a $C$-related random walk to a decision threshold, fully determined by an integration time $\tau = \frac{1}{\alpha}$, a measure of processing effort. Experiments by Sigman and Dehaene [8] showed that the central component $C$ is responsible for almost all of the variance in response times (RTs). An important property of the model is that the proportion of the integration time constants ($\tau$) for two experimental conditions (e.g., $i$ and $j$) can be determined from their respective variances ($s^2_i$ and $s^2_j$) as:

$$\frac{\tau_i}{\tau_j} = \sqrt{\frac{s^2_i}{s^2_j}}$$  \hspace{1cm} (1)
2.1. Speech Materials

All speech materials were obtained from the Spoken Dutch Corpus (CGN) [5, 4], making hand-aligned utterances (“chunks”), word boundary segmentations, transcriptions, and phonetic transcriptions available. Based on audio quality and on coverage of turn switching categories [12, 13], a stimulus set of 7 switchboard (8 kHz, dual channel telephone recordings) and 10 volunteer home recordings (16 kHz, stereo face-to-face) of 10 minutes each (total duration 165 min.) was selected. Each utterance was labeled by two judges on its discourse value.

2.2. Stimulus preparation and presentation

Stimulus selection and preparation was identical as described by [12, 13, 14, 10]. The 17 dialog recordings were each divided into two overlapping 6 minute stimuli, i.e. the first and last 6 minutes of each dialog. From these 34 dialog fragments, the Stimuli Set was created by replacing one or none of the last four words of each utterance by a Mask of white noise, at a comfortably convincing level. For each utterance in the stimuli, a number between -1 and 3 was randomly selected, with -1 representing No Mask, 0 representing a Mask on the last word, 1 on the penultimate word and so on. If the selected number was higher than the number of words in the utterance, there was no masking. Note that this procedure created a bias for no masking of the position of the masked word. Bin size is 40ms. (# responses, all data) NM/2, unmasked utterances, counts are divided by 2 for scale. 0 is last word etc. #Spont/20 and #Tel/20 are the corresponding distributions of turn-switch delays for spontaneous speech and telephone conversations from the Spoken Dutch Corpus ([10], counts divided by 20 for scale).

2.3. Response collection and processing

Stereo stimulus playback and response recording were done on a single laptop [12, 13]. Responses were recorded with a laryngograph (Laryngograph Ltd, Lx proc) at a 16 kHz sampling rate on one channel, with the feedback (summed) mono version of the stimulus on the other channel for alignment purposes [12, 13]. 24 Naive, native Dutch subjects participated in the experiment. Subjects were paid, and had no prior knowledge about the aims of the experiment. To subjects was explained what Minimal Responses were (in layman’s terms if necessary) and asked to act as if they were participating in the conversations they would hear. The subjects were asked to respond with ‘AH’ if possible, as often as they could. After two minutes of practice stimuli, none of the subjects had any problems with the tasks and all responded rather “naturally” to the stimuli.

Responses were automatically extracted as the voiced parts of the laryngograph recordings and individually aligned with the original conversations using the re-recorded mono stimulus signal. A Praat script [1, 12, 13] located and labeled these responses in the recordings, see fig. 2.

The RT delay was defined as the time between the start of the Voiced response and the closest utterance end (irrespective of the speaker) within a window of 2 seconds. Responses with a duration shorter than 15ms were discarded as spurious. The relevant utterance had to start at least 0.1 seconds before the start of the response. Furthermore, to make sure only utterances adding content to the conversation were regarded, all ’functional’ utterances that were labeled as minimal responses, grounding acts or fixed, ’formulaic’ expressions (e.g. ’listen’, ‘functional’ utterances that were labeled as minimal responses, grounding acts or fixed, ’formulaic’ expressions (e.g. ‘listen’, ‘functional’ utterances that were labeled as minimal responses, grounding acts or fixed, ’formulaic’ expressions (e.g. ‘listen’).

![Figure 2: Example response waveform and segmentation. Top: Mono waveform of the stimulus, Center: laryngograph signal of a single response, Bottom: Annotation tiers for the automatic segmentation of the response and the transcribed utterances of the two speakers. The response delay is the interval between the vertical lines.](image)

![Figure 3: Distribution of reaction-time delays with respect to the position of the masked word. Bin size is 40ms. (# responses, all data) NM/2, unmasked utterances, counts are divided by 2 for scale. 0 is last word etc. #Spont/20 and #Tel/20 are the corresponding distributions of turn-switch delays for spontaneous speech and telephone conversations from the Spoken Dutch Corpus ([10], counts divided by 20 for scale).](image)

Table 1: Probability of response, given the position of the Mask (all data)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>last</td>
<td>13,066</td>
<td>8,872</td>
<td>4,194</td>
<td>0.321</td>
</tr>
<tr>
<td>last-1</td>
<td>9,538</td>
<td>5,944</td>
<td>3,594</td>
<td>0.377</td>
</tr>
<tr>
<td>last-2</td>
<td>8,558</td>
<td>5,200</td>
<td>3,358</td>
<td>0.392</td>
</tr>
<tr>
<td>last-3</td>
<td>7,650</td>
<td>4,666</td>
<td>2,984</td>
<td>0.390</td>
</tr>
<tr>
<td>No Mask</td>
<td>2,5627</td>
<td>1,8411</td>
<td>7,216</td>
<td>0.282</td>
</tr>
<tr>
<td>Total</td>
<td>64,439</td>
<td>43,093</td>
<td>21,346</td>
<td>0.331</td>
</tr>
</tbody>
</table>

Table 2: Probability of response by position of Mask and length of utterance (all data)

<table>
<thead>
<tr>
<th>Position</th>
<th>Length of utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>of Mask</td>
<td>1</td>
</tr>
<tr>
<td>last</td>
<td>0.195</td>
</tr>
<tr>
<td>last-1</td>
<td>-</td>
</tr>
<tr>
<td>last-2</td>
<td>-</td>
</tr>
<tr>
<td>last-3</td>
<td>-</td>
</tr>
<tr>
<td>No Mask</td>
<td>0.225</td>
</tr>
</tbody>
</table>
Figure 4: Mean Reaction-time delays for different Mask positions with respect to utterance length.

Figure 5: Mean relative reaction-time delays for Length and Mask position.

'hold on', 'no really?') were disregarded (except where indicated that all data were used), as well as interjections, hesitations, and all utterances labeled with 'x', like coughs and unintelligible speech.

3. Results

In total, 24 hours of recordings containing 64,439 utterances were presented to the 24 subjects, which elicited 21,436 responses (see table 1). The distribution and probability of responses with respect to the position of the Mask is given in table 1. At the current level of analysis, we did not distinguish between the prescribed ‘AH’ responses and other, more complex, responses [12, 13].

In table 2 the probability of a response is given as a function of the utterance length and the position of the masked word. Subjects were more likely to respond to unmasked utterances ($p < 0.001$, $\chi^2$ test) and longer utterances ($p < 0.001$, $\chi^2$ test). The distribution of response times is displayed in figure 3. The distribution of reaction times in our experiments are comparable to the delays we found in natural turn-switches as exemplified by the spontaneous speech and telephone recordings.

Figure 4 shows the average reaction-time delays for the different Mask positions with respect to utterance length. A clear effect can be seen for utterance length for all utterances (length for all data, $p < 0.001$, ANOVA), specifically for unmasked utterances, (length for unmasked data, $p < 0.001$, ANOVA) and all utterances with masked words pooled (length for all masked data, $p < 0.01$, ANOVA but individual Mask positions did not differ significantly $p > 0.05$).

Figure 5 shows the relative RT for the Mask position and length, compared to unmasked utterances of the same length from the same subject. If the last word of the utterance is masked (in figure 5 number of words following Mask = 0), an effect of masking can be found on the reaction times compared to the unmasked condition ($p < 0.01$, Wilcoxon Matched Pairs Signed Ranks test). An exception may be the single word utterances, where responses might be faster in the (whole-utterance) masked condition than in the unmasked condition. However, this could not be resolved statistically in this data set. No statistically significant effect of utterance length on reaction times is found for masked words before the last word.

Figure 6 shows the standard deviations for the different Mask positions with respect to utterance length. For utterances with a length of 1 or 2 words, masked utterances have a larger standard deviation, and thus a larger integration time (see fig.1) than unmasked utterances (Length=1, $p < 0.001$, Length=2, $p < 0.01$, F-test). For longer utterances, this trend continues, but is no longer significant. No effect for length on standard deviation of delays was found.

Fig 7 shows the mean RTs for the position of the last prominent word and Mask position (utterance length > 2).
prominent word (Mask before, on or after the last prominent word). No effect of the position of the prominent word is found. Only if the penultimate word is prominent and the ultimate word was masked, responses were significantly slower than in the unmasked condition ($p < 0.001$, WMPSR test).

Figure 8 shows the standard deviation of the reaction times for the position of the last prominent word and the position of the Mask, relative to the last prominent word. All masked utterances pooled have longer standard deviations (and thus longer integration/decision times) compared to the unmasked utterances ($p < 0.001$, F-test). This effect is found for all mask positions (after, on or before the prominent word).

4. Discussion and conclusions

The number and delays of our subjects’ responses were primarily determined by the length of the utterance. Whether or not a word in the utterance was masked, had no effects on the reaction times, unless it was the last word in the utterance that was masked. Masking did have an effect on processing efforts. There was a systematic increase in standard deviation due to the presence of masked words.

Contrary to expectations, whether or not the last prominent word was masked had no effect on reaction times. Only when the penultimate word was prominent and the last word masked did we find a statistically significant delay in response time.

The presence of masked words, even masked prominent words, did not affect the RTs of our subjects unless the very last word of the utterance was masked. This indicates that listeners are able to seamlessly interpolate the missing words, just like they are able to restore masked phonemes in words [11]. Only to determine the exact position of the TRP, is the identity of the last word needed.

These results support the assumption that the identity of the last word before a TRP is used to predict the timing of the response. Masking other words had no measurable effects on the last word before a TRP is used to predict the timing of the response. Masking other words had no measurable effects on the last word before a TRP is used to predict the timing of the response. All masked utterances pooled have longer standard deviations (and thus longer integration/decision times) compared to the unmasked utterances ($p < 0.001$, F-test). This effect is found for all mask positions (after, on or before the prominent word).

5. Acknowledgements

The authors would like to thank Dr. Louis ten Bosch and Dr. Henk van den Heuvel of Radboud University Nijmegen for selecting and annotating the dialogs. We also want to thank Ton Wempe for his technical assistance. This project was made possible by grant 276-75-002 of the Netherlands Organization of Scientific Research.

6. References

[9] Speech Processing Expertise Centre (SPEX), Radboud University Nijmegen, the Netherlands, (http://www.spex.nl/).