Prominent Words as Anchors for TRP Projection

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

Chair of Phonetic Sciences/ACLC,
Department of Linguistics, University of Amsterdam, The Netherlands
R.J.J.H.vanSon@uva.nl

Abstract

The effect of the position of the last accented word on the projection of TRPs was investigated with two RT experiments. Subjects were asked to respond with minimal responses to prerecorded dialogs and impoverished versions of these dialogs, containing either only intonation and pause information,*hummed* stimuli, or no periodic component at all, *whispered* stimuli. The distribution of these elicited response delays was comparable to that of natural turn switches. It is shown 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 that the presence of an prominent, informative, word starts the projection of a possible upcoming TRP. The availability of non-prominent, predictable, speech then allows listeners to improve their predictions of the exact timing of the TRP.

1. Introduction

In order to allow for smooth turn transitions in natural conversations, participants have to be able to predict the end of the previous speaker's turn [**?**]. Various information sources are known or suspected to help listeners in determining possible Transition Relevance Places (TRPs), like gaze direction, gestures, intonation, syntactic, and timing information (like speaking rate and pauses) [**?**, **?**, **?**, **?**]. The study of [**?**] concluded that only lexico-syntactic content was used for projection. However, other studies did find that intonation was used to project TRPs under experimental conditions [**?**, **?**, **?**].

Given that subjects are able to project TRPs reliably and are likely to use intonation, raises the question precisely what cues are used. Reaction Time (RT) paradigms are the most sensitive to information distribution and processing. However, projecting TRPs is not like the classical RT experiments (cf, [**?**, **?**]). Instead of starting at the start of the stimulus (information) presentation, the subject is asked to predict an end-point from an ongoing stimulus. In [**?**] it was argued that subjects started to integrate information over 500ms before the TRP. The RT to a TRP is then dependend on two types of information: Cues about the likelyhood of a TRP being prepared and cues about the exact location of the TRP. The boundary tone, or end-tone, and the coming end of the last word are strong cues about the exact location of the TRP. But these cues are often only available in the last syllable of the utterance [**?**].

The location of the end of the final word is often predictable if the last word itself is predictable. The same holds for the boundary tone, which can often be predicted from the end of the last pitch accent. These two cues merge on the last prominent word before the TRP. Given the normal *Information Structure* of prosody, pitch accents are placed on prominent words that generally are also informative, ie, unpredictable words. The last prominent word before a TRP will therefore likely carry the last pitch accent and also will be the last (highly) unpredictable word. The non-prominent words are following the last prominent one will be unaccented and predictable, as will be the remainder of the intonation contour. This leads to the prediction that the more words follow the last prominent word, the better subjects will be able to predict the upcoming TRP.

Subjects listened to original and manipulated versions of recordings of natural dialogs and were asked to give minimal responses by saying 'AH'. Their responses are assumed to signal comprehension of at least part of the utterance's structure and a recognition of a possible end-of-turn (TRP). A decision-making model by Sigman and Dehaene [**?**] is used to compare processing of the different stimuli (see fig. **??**). In this model, mental decision-making is modeled as a noisy integrator that stochastically accumulates perceptual evidence from the sensory system in time [**?**, **?**], through a perceptual (P), central decision-making (C) and 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 $\tau = \frac{1}{\alpha}$. Experiments by Sigman and Dehaene [**?**] showed that the central component C is responsible for almost all of the variance in response times. An important property of the model is that the proportion of the integration time constants (τ) for two experimental conditions (e.g. *i* and *j*) can be determined from their respective variances (s_i^2 and s_j^2) as:

$$\frac{\tau_i}{\tau_j} = \sqrt[3]{\frac{s_i^2}{s_j^2}} \qquad (1)$$

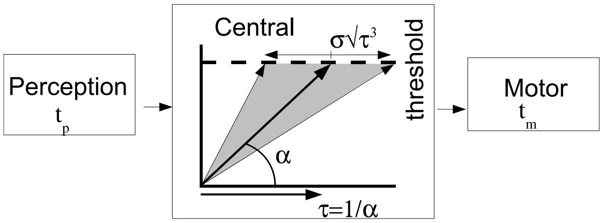

Figure 1: Perception-Central-Motor model of Reaction Times. $\tau = \frac{1}{\alpha}$ is the average central integration time. σ is an unknown noise term. The average reaction time $RT = t_p + t_m + \tau$. The variance is $var(RT) = \frac{1}{2}\sigma^2\tau^3$

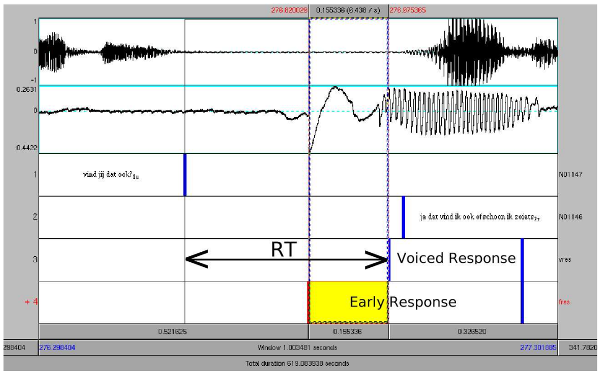

Figure 2: Example response waveform and segmentation. Top: Mono waveform of the stimulus, Center: laryngograph signal of a single response, Bottom: Annotation tiers for the automatic segmentation of the response and the transliterated utterances of the two speakers. The response delay is the interval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

2. Materials and Methods

2.1. Speech Materials

All speech materials were obtained from the Spoken Dutch Corpus (CGN) [**?**, **?**], making hand-aligned utterances ("chunks"), word boundary segmentations, transliterations, and phonetic transcriptions available. This study is based on the hand aligned word boundaries and the *pro1* and *pro2* prominence markings. In the CGN protocol, prominence was explicitly connected with the possibility of a pitch accent (cf, [**?**, **?**]). The last word that either of the two transcribers considered to be prominent, was marked as the last prominent word. We will refer to the prominent words as *accented* words. However, it must be kept in mind that the transcribers used a broader definition of prominence.

Based on audio quality and coverage of turn switching categories [**?**, **?**], 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.

2.2. Stimulus preparation and presentation

Stimulus selection and preparation was identical to [**?**, **?**]. The 17 dialog recordings were each divided into two overlapping 6 minute stimuli, i.e. the first and last 6 minutes of each dialog. This is the *original* stimulus set (34 stimuli). Two new stimulus sets were constructed. First, a set of *hummed* stimuli was created by converting the *original* stimuli to pitch contours with Praat [**?**] and having them resynthesized as neutral-vowel speech [**?**, **?**]. This *hummed* speech contains nothing but the intonation and pause structure of the *original* speech, i.e. no loudness or spectral information was present. Second, the *original* stimuli were resynthesized from an LPC analysis using white noise as the sound source. The LPC order was chosen as 8 poles for telephone speech and 16 poles for the home recordings. The amplitude was scaled to prevent clipping. These constitute *whispered* stimuli as they did not contain a periodic component. However, it must be remembered that both the *hummed* and *whispered* speech were artificial and sounded not like natural *humming* or *whispering*. The artificially *whispered* stimuli were still intelligible and did audibly contain non-periodic prosodic cues. All stimuli were upsampled to 16 kHz and 16 bit

Table 1: *Distribution of Voiced and Early responses over stimulus types by pitch accent positions. Only responses to utterances with at least 3 words are used. '-' indicates no accent in the last three words.*

Accent	position	1	2	3	-	total
Voiced	Orig. (32)	3545	1647	1011	944	7147
(subj.)	Hum. (21)	2425	1295	819	974	5513
	Whisp. (11)	1102	476	308	276	2162
Early	Orig. (32)	1446	647	411	369	2873
(subj.)	Hum. (21)	988	540	327	379	2234
	Whisp. (11)	541	228	132	110	1011
	Utterances	1480	766	491	534	3271

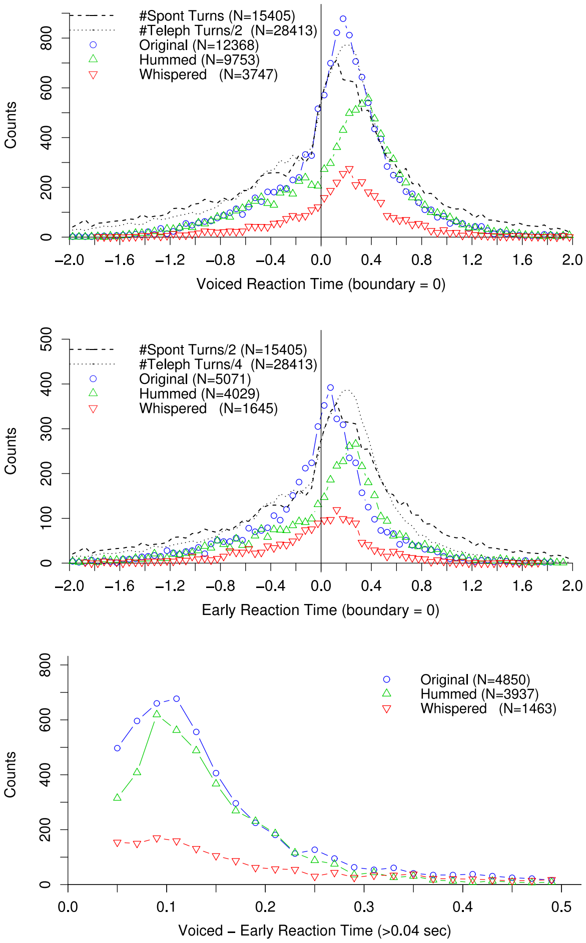

Figure 3: Distribution of reaction-time delays with respect to corresponding utterance-ends. Top: *Voiced* responses, Mid: *Early* responses, Bottom: Difference between *Voiced* and *Early* responses. Bin size is 40ms. *Early* responses must start more than 40ms before the *Voiced* response. (# responses)

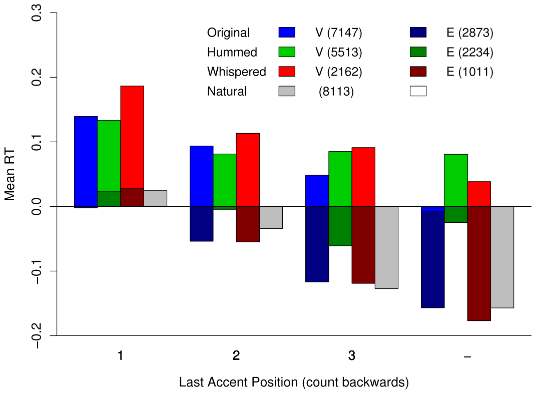

Figure 4: Mean delays for accent positions ('-': no accent in last three words). See text for statistical results (# responses). **V**: Voiced, **E**: Early responses.

where necessary.

Stimuli were pseudo-randomized and balanced for presentation. Each of the 32 subjects (with one exception due to an error) heard a different subset and order of 4 *original* and 4 manipulated dialog fragments of 6 minutes duration in alternating order, starting with an *original* stimulus. These first 8 dialog fragments were all from different full dialogs. These were followed by two repeat stimuli (ignored in the current study), the dialog complements of the first two stimuli. The whole 10 stimulus session contained two 2 minute breaks and was preceded by two 2 minute practice items, a *full speech* and *hummed* or *whispered* fragment from a dialog that was not in the stimulus set.

2.3. Response collection and processing

Stereo stimulus playback and response recording were done concurrently on a single laptop [**?**, **?**]. The laryngograph (Laryngograph Ltd, Lx proc) responses were recorded at a 16 kHz sampling rate on one channel, with the fed-back (summed) mono version of the stimulus on the other channel for alignment purposes [**?**, **?**]. 32 Naive, native Dutch subjects participated in the experiment. 21 Subjects heard the *original* and *hummed* stimuli and 11 subjects heard the *original* and *whispered* stimuli. Some subjects were paid. Only one subject had some knowledge of the aims of the experiment. Subjects were explained what Minimal Responses were (in layman's terms if necessary) and asked to act like they participated in the conversation they would hear. The subjects were asked to respond with 'AH' if possible, as often as they could. After the practice stimuli, none of the subjects had any problems with the tasks and all responded rather "naturally" to the stimuli, even to the *hummed* speech.

Responses were automatically extracted and individually aligned with the original conversations using the re-recorded mono stimulus signal [**?**, **?**, **?**]. These are the *Voiced* responses (see fig. **??**). About one third of all *Voiced* responses were preceded by a characteristic early larynchograph signal indicating muscle activity in the larynx. The start of this signal was automatically segmented and constitutes the *Early* response (see fig. **??**). A minimum difference of 40ms was used to ensure reliable identification.

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. The relevant utterance had to start at least 0.1 seconds before the start of the response. Furthermore, responses with a duration shorter than 15ms were discarded as spurious. Using the same criteria, Turn Transfer delays in the Spontaneous and Telephone dialogs of the hand aligned part of the Spoken Dutch Corpus were determined. The distribution of responses with respect to the intonation boundary tones is given in table **??**. At the current level of analysis, we did not distinguish between the prescribed 'AH' responses and other, more complex, responses [**?**, **?**].

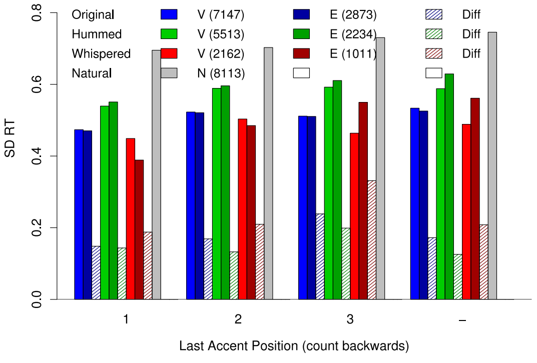

Figure 5: Standard deviation of delays for accent positions. As fig. **??**. **V**: Voiced, **E**: Early responses, **Diff**: Difference between V and E responses.

3. Results

In total, 25.6 hours of responses are used from 32 subjects, containing 14,822 responses to utterances of three words and longer (see table **??**). In fig. **??**, the distribution of *all* 25,868 response delays that could be attributed to specific utterances is compared to the natural turn start delays for home recordings and telephone speech in the CGN. The distributions of the Voiced responses corresponds quite well to that of the natural turn switch delays. This indicates that our elicited responses capture at least part of the natural conversational behavior. The distribution of the *Early* responses and the delay differences between *Voiced* and *Early* responses is as expected from [**?**] (note the 40ms lower cutoff in the latter).

In fig. **??**, the average RT delay to TRPs is presented against the position of the last accented word (1 is ultimate, etc.). A plot of the RT with respect to the start of the last accented word showed a large increase of over 200 ms in response delay going from ultimate to antepenultimate accent (not shown). So the TRP will be used as a reference point. A clear correlation between the average RT and the distance to the last accent is visible. For the natural turn switches and the *Voiced* and *Early* responses to the *original* and *whispered* stimuli, the relation between accent position and RT is statistically significant ($p < 0.001$, one-way ANOVA on Accent position). The *Voiced* responses to *hummed* utterances are affected by accent position ($p < 0.02$, one-way ANOVA). This can be attributed to the effect of the final accent (1) which differed from the rest combined ($p < 0.001$, t-test 1 vs. 2, 3,'-'). No difference was found for the *Early* responses to *hummed* stimuli. The effect of accent position on the delay difference between *Voiced* and *Early* responses is only significant for *whispered* stimuli ($p < 0.002$,

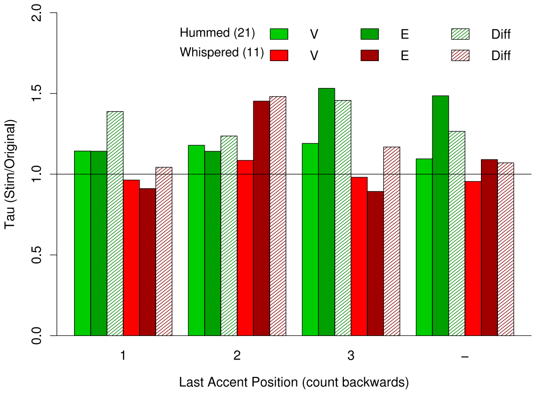

Figure 6: Relative "processing" time $\frac{\tau'}{\tau_{orig}}$ for accent positions and different stimulus types ('-': no accent in last three words). See text for statistical results (# subjects). **V**: Voiced, **E**: Early responses, **Diff**: Difference between V and E responses.

one-way ANOVA on Accent position for difference). There was a main effect of stimulus type for all data pooled for all response types ($p < 0.01$, ANOVA). There was also an effect of stimulus type on the *Voiced* responses to *whispered* stimuli ($p < 0.01$, ANOVA, by subject) and the *Voiced* minus *Early* difference for *hummed* stimuli ($p < 0.001$, ANOVA, by subject).

The variances of the RTs are related to the integration (decision) time [**?**]. Fig. **??** shows that accent position had little or no impact on the standard deviation of the RTs for any of the stimulus types ($p > 0.05$, one-way ANOVA). The natural turn switches had the highest variance, followed by the *Voiced* and *Early* responses of the *hummed* stimuli and both the *original* and *whispered* stimuli, which did not differ ($p < 0.001$, t-test on each pair of conditions). No differences were found for the delay between *Voiced* and *Early* responses.

Fig. **??** expresses the differences in variance in terms of the relative decision (integration) time, $\frac{\tau'}{\tau_{orig}}$, of [**?**] (see eq. **??** and fig. **??**). It is clear that there is no effect of accent position on the integration time, but only of stimulus type. *Hummed* stimuli take consistently more time to decide.

4. Discussion and conclusions

The delays of the elicited minimal responses had a distribution that was very close to those of natural turn switches (see fig. **??**). The larger variances of the natural turn switch delays can be explained from the fact that these were not (all) minimal responses and should be expected to require additional processing for formulation, and as a consequence, have a larger variance. This corroborates the use of elicited minimal responses as a probe into natural turn behavior.

It is clear from the results in fig. **??** that there is indeed a very strong effect of last accent position on response delays. This effect cannot be attributed to an increased integration time, as there was no effect found of accent position on the variance of the responses (see fig. **??**). This holds for both the audible *Voiced* responses as the inaudible *Early* responses.

These results suggest a model of TRP projection where the listener predicts the position of an upcoming TRP using the last, unpredictable, prominent word as a starting point. The more time the listener has to estimate the position of the upcoming TRP, the more exact, or earlier, she will respond. The hummed responses did show a slightly different effect. When the final word was (likely) accented, the response was delayed. But in all other cases the response was not affected. This suggests that a pitch movement on the final word disturbs the projection, but that there is no benefit of more than a short, word-length, part of the intonation contour.

5. Acknowledgments

The authors would like to thank Dr. Louis ten Bosch and Dr. Henk van den Heuvel of Radbout 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

[1] Liddicoat, A.J., "The projectability of turn constructional units and the role of prediction in listening", Discourse Studies 6: 449-469, 2004.

[2] Caspers, J., "Local speech melody as a limiting factor in the turn-taking system in Dutch", Journal of Phonetics 31: 139-278, 2003.

[3] Wesseling, W. and R. J. J. H. van Son, "Timing of Experimentally Elicited Minimal Responses as Quantitative Evidence for the Use of Intonation in Projecting TRPs", in *Proceedings of Interspeech2005*, Lisbon, 2005

[4] Wesseling, W. and van Son R.J.J.H. (2005), "*Early* Preparation of Experimentally Elicited Minimal Responses", in *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 2005

[5] De Ruiter, J.P., Mitterer, H., and Enfield, N.J., "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation", Language, In Press

[6] Caspers, J., Van Son, R.J.J.H., "Investigating the relationship between high, low and level boundary tones and punctuation symbols in Dutch", submitted to Interspeech2006.

[7] Sigman, M. and Dehaene, S., "Parsing a Cognitive Task: A Characterization of the Mind's Bottleneck", PLoS Biology 3, e37, 2005 (http://www.plos.org/)

[8] Posner, M.I., "Timing the Brain: Mental Chronometry as a Tool in Neuroscience", PLoS Biology 3, e51, 2005 (http://www.plos.org/)

[9] Oostdijk, N. et al., "Experiences from the Spoken Dutch Corpus Project.", eds M.G. Rodriguez and C.P. Surez Araujo, in *Proceedings of the third International Conference on Language Resources and Evaluation*: 340-347, 2002.

[10] Oostdijk N., "The Spoken Dutch Corpus, overview and first evaluation", in *Proceedings of LREC-2000*, Athens, Vol. 2: 887-894, 2000.

[11] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International 5: 341-345, 2001. (Praat is Free Software, http://www.Praat.org/)

[12] "Prosodic annotation", as part of the annotation documentation of the Spoken Dutch Corpus (http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/annot/prosody/info.htm)