Early preparation of experimentally elicited minimal responses

Wesseling, W.; van Son, R.J.J.H.

Published in:
Proceedings of the 6th SIGdialWorkshop on Discourse and Dialogy

Link to publication

Citation for published version (APA):
Wesseling, W., & van Son, R. J. J. H. (2005). Early preparation of experimentally elicited minimal responses. In L. Dybkjær, & W. Minker (Eds.), Proceedings of the 6th SIGdialWorkshop on Discourse and Dialogy (pp. 11-18). Lisboa: SIGdial.

Early Preparation of Experimentally Elicited Minimal Responses

Wieneke Wesseling and R. J. J. H. van Son
Chair of Phonetic Sciences/ACLC,
University of Amsterdam
W.Wesseling@uva.nl and R.J.J.H.vanSon@uva.nl

Abstract

In both human-human and human-machine conversation, an important task for the participants is to identify the moment the other participant finishes speaking, giving them the possibility of taking over the turn in talk. In an RT experiment, consistent evidence was found for an intermediate stage in the planning and articulation of elicited minimal responses in the shape of early larynx and glottal movements in laryngograph recordings. Using a simple Response Time model, it is estimated that this intermediate stage occurs at approximately 2/3 of the integration-time needed for the articulation of a response. Impoverished *intonation only* stimuli were still adequate to elicit minimal responses, but a longer integration-time was required to initiate a response.

Keywords Minimal Responses, Response Times, Dialogs, Intonation, Spoken Language Processing, Random Walk

1 Introduction

In human-human as well as human-machine conversation, an important task for the participants is to identify the moment the other participant finishes speaking, giving them the possibility of taking over the turn in talk. The organization of turn-taking in interaction was described in a classical paper by Sacks et al. (1974), who introduced the notion of Transition Relevance Place (TRP), a point of possible completion of the current utterance. At this point a change of turn between speakers becomes relevant. This generally means that it is possible for the current speaker to select another speaker, or for another speaker to self-select and start talking. The latter can have the form of a full utterance or of a minimal response. Given the number of factors that are likely to be involved in this identification process of TRPs, one would expect this to be a difficult task. In human-machine interactions, smooth turn switches are still a largely unsolved problem. Nevertheless, transitions between human speakers are usually smooth, with little overlap and only small pauses. This implies that participants are able to predict, or project, end-of-turns fairly reliably before they actually take place, (see e.g. Liddicoat, 2004; Pickering and Garrod, 2004).

Information sources that are known to be used for the projection of TRPs include syntactic, semantic and pragmatic information, prosodic factors like pitch, loudness, tempo and duration and visual cues like gaze direction and gestures. In her experiments on the communicative function of (local) melodic elements in the Dutch turn-taking system, Caspers (2003) found that syntactic completion seems to be the main factor in the turn-taking mechanism, and that local melodic factors play a supporting as well as a constraining role in the process. At positions where pauses coincide with syntactic completion, *low* or *high* boundary tones are used. At positions where pauses and syntactic completion do not coincide, incompletion is signaled by the use of a *mid-register* tone. Caspers concludes that boundary tones can be used as a cue to TRP location, although they are generally subordinate to syntactic completion cues.

The goal of the present study is to provide *quantitative* data about the length of the speech interval over which TRP position is projected and the time-course over which information is used to project TRPs. Elsewhere, we study to what extent redundant subordinate cues for the projection of TRP location (i.e. intonation) can compensate for the loss of dominating cues (i.e. verbal or syntactic information and prosody)(see Wesseling and van Son, in press). This is done by measuring response-times (RTs) in an elicited minimal response task. Responses to annotated normal recordings are compared to responses to manipulated speech, containing nothing but the intonation and timing information of the original. The task involved subjects listening to recordings of natural dialogs and giving minimal responses (in our case, by saying ‘AH’) to both speakers in these dialogs. This task can be compared to conventional “press a button” RT tasks, but is better suited to analyzing long conversations due to the short latencies of spoken responses, which allow better attribution to putative TRPs. Minimal responses, or backchannels, are responses listeners give in conversations, signaling their role as listener. They indicate the speaker’s utterance is being heard and are here assumed to signal comprehension of at least part of the utterance’s structure and recognition of a possible end-of-turn (TRP). As such, understanding of the timing of minimal responses is crucial to understanding the dynamics of conversation.

In psychological research of Sigman and Dehaene (2005) involving response-times to investigate the mental decision-making process (c.f. Posner, 2005), this process was modeled as a noisy integrator that stochastically accumulates perceptual evidence from the sensory system in time. Three stages of processing can be identified: a perceptual component (P), a central decision making component (C), and a motor component (M). Sigman and Dehaene (2005) conducted a response-time experiment, in which they showed that the central component C was responsible for almost all of the variance in RTs. In a number-comparison task, subjects had to decide whether a presented digit was larger or smaller than 45. Three different factors were manipulated: number notation (Arabic digits or spelled words); numerical distance between the presented numbers and response complexity (tapping once or twice). These factors are assumed to be related to respectively the P, C and M components of the decision making process. The effects in RTs turned out to be additive for the three factors, but only the distance manipulation, associated with the C component, resulted in a significant increase of dispersion with the mean (see also Posner, 2005).

In the model used by Sigman and Dehaene (2005), 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 $\tau = \frac{1}{\alpha}$. In this model, the probability distribution of the RTs, $g(t)$, is derived from the probability of a random walk, crossing a threshold for the first time at time t, which can be written as:

$$g(t) = \frac{1}{\sigma \cdot \sqrt{2\pi \cdot (t - t_0)^3}} \cdot exp\left(-\frac{(1 - \alpha \cdot (t - t_0))^2}{2 \cdot \sigma^2 (t - t_0)}\right) \tag{1}$$

where the threshold is set at 1 without loss of generality. In this model the average RT becomes $\overline{RT} = t_0 + \tau$ and the variance $var(RT) = \frac{1}{2}\sigma^2\tau^3$ where σ is a task independent,mostly unknown, modeling parameter. 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:

$$\frac{\tau_i}{\tau_j} = \sqrt[3]{\frac{s_i^2}{s_j^2}} \tag{2}$$

Eq. 2 is independent of the difficult to estimate σ parameter (Sigman and Dehaene, 2005).

2 Materials and Methods

2.1 Speech Materials

All speech material used for this experiment was obtained from the Spoken Dutch Corpus (CGN, Oostdijk, 2000; Oostdijk et al., 2002) and consisted of informal and spontaneous Dutch dialogues in two settings: telephone switchboard dual channel speech recordings and volunteer face-to-face stereo home recordings. Telephone recordings in the CGN have been digitized at an 8 kHz sampling frequency and 8 bit precision. The two speakers in each telephone conversations were recorded on separate channels. Face-to-face conversations were recorded on Sony Minidisk and subsequently digitized at 16 kHz and 16 bit precision (c.f. van Son, 2005). The stereo signal allowed an auditory spatial separation of the speakers.

In a total of 61 informal and spontaneous Dutch dialogues from this corpus, 32 switchboard telephone conversations and 29 home recorded face-to-face dialogs, with a total duration of 588 minutes ($\approx 9\frac{1}{2}$ minutes/dialog), all change-of-speaker moments were categorized by a single annotator from SPEX as either a Minimal Response, a Question/Answer pair, or a General Turn switch. For each of the turn-switches, the audio quality of the adjacent utterances was also judged on a 4 point scale (0-3) from nearly incomprehensible to high-quality sound. For all 61 dialog recordings, hand-aligned utterances (“chunks”), word boundary segmentations, transliterations and phonetic transcriptions were available. In the context of the conversations used in this study, the hand-labeled utterances from the CGN can be interpreted as a very crude form of prosodic phrasing. About 75% of these utterances are followed by silent pauses. For the

Table 1: *Total number of utterances for each of the end-tone categories for the full set of conversations and for the present stimulus selection.*

material	low	mid	high	total
full set	5850	11198	5065	22113
stimulus set	1964	3354	1560	6878

present stimulus set, a subset of 7 switchboard and 10 home recordings with a total duration of 165 minutes was selected, based on high audio quality and coverage of the turn-switching categories.

Since boundary tones are an important cue to TRP projection (Caspers, 2003), their presence was noted in the current study. The end boundary tones of all utterances were automatically estimated as *low*, *mid* or *high* from the pitch contours. For each speaker in each dialog, the global standard deviation of the F_0 was calculated ($Sd(F_0)$) using the Praat pitch tracker at 5 ms increments (Boersma, 2001; Boersma and Weenink, 2004). For each utterance *i*, the mean ($\overline{F}_0^i$) and the end boundary pitch (F_{0end}^i) over the last 25 ms of voiced speech were measured. From this the relative boundary tone (Z_i) of utterance *i* was determined as:

$$Z_i = \frac{\overline{F}_0^i - F_{0end}^i}{Sd\left(F_0\right)} \tag{3}$$

The boundary tone of utterance *i* was considered *high* if $Z_i > 0.2$, *low* if $Z_i < -0.5$, and *mid-tone* otherwise. These values were determined heuristically. See table 1 for the distribution of intonation categories over utterances. Given their importance to TRP location, the three boundary tone classes were treated as independent categories in our statistical tests to obtain more uniform RT samples. For an evaluation of the influence of the individual boundary tones on RTs and integration-times, see Wesseling and van Son (in press).

2.2 Stimulus preparation

Two sets of stimuli were presented: a *full speech* set and an *intonation only* set. The 17 dialog recordings from the stimulus subset were each divided into two overlapping 6 minute stimuli, i.e. the first and last 6 minutes of each dialog. This is the *full speech* stimulus set (34 stimuli). The *intonation only* set of stimuli was created by converting the *full speech* stimuli to pitch contours with Praat (Boersma, 2001; Boersma and Weenink, 2004) and having them resynthesized as ”hummed” neutral-vowel speech, containing no loudness or spectral information, i.e. no verbal or syntactic information. The hummed speech contains nothing but the intonation and pause structure of the original speech. All stimuli were upsampled to 16 kHz where necessary.

2.3 Stimulus presentation

Stimuli were pseudo-randomized for presentation. Every subject heard a different subset of 4 *full speech* and 4 *intonation only* type dialog fragments of 6 minutes duration in alternating order, starting with a *full speech* stimulus. These first 8 dialog fragments (with a total duration of 48 minutes) were all selected 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 *intonation only* fragment from a dialog that was not in the stimulus set. Stereo stimuli were played directly from an Acer Travelmate 529 laptop running Knoppix (Linux 2.4.26) in console mode.

2.4 Response collection

Responses were registered with a laryngograph (Laryngograph Ltd, Lx proc) and recorded at a 16 kHz sampling rate on one channel concurrently on the same laptop used for stimulus presentation. A fed-back (summed) mono version of the stimulus was duplex recorded on the other stereo channel for alignment purposes (c.f. Bailly, 2001). 15 Naive subjects, between the age of twenty and seventy, 7 males, 8 females, all staff or students of the ACLC with no reported hearing problems, participated in the experiment. 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 were asked to act like they participated in the conversation they would hear. To get a well defined response onset timing, the subjects were asked to respond with ‘AH’, instead of more common responses like ’oh’, ’ok’, ‘hm’ or the Dutch ‘ja ’ (yes), as often as they could. After the practice set, none of the subjects had any problems with this task.

2.4.1 Voiced Responses

The laryngograph response recordings were automatically extracted and aligned with the original conversations using the re-recorded mono stimulus signal. The responses were automatically identified as the voiced parts of the laryngograph recordings. A Praat script located

Table 2: *Total number of articulated (voiced) and early responses to stimuli for each of the 3 end-tone categories and minimal responses for the total conversation set. The total number of responses including non-attributable responses is also given.*

response category	low	mid	high	total
full speech voiced	1860	2850	1374	6084
early	690	1144	515	2349
total				6565
intonation only voiced	1917	3205	1453	6575
early	663	1180	534	2377
total				7420
total set (voiced)	386	539	281	1206
total				1310

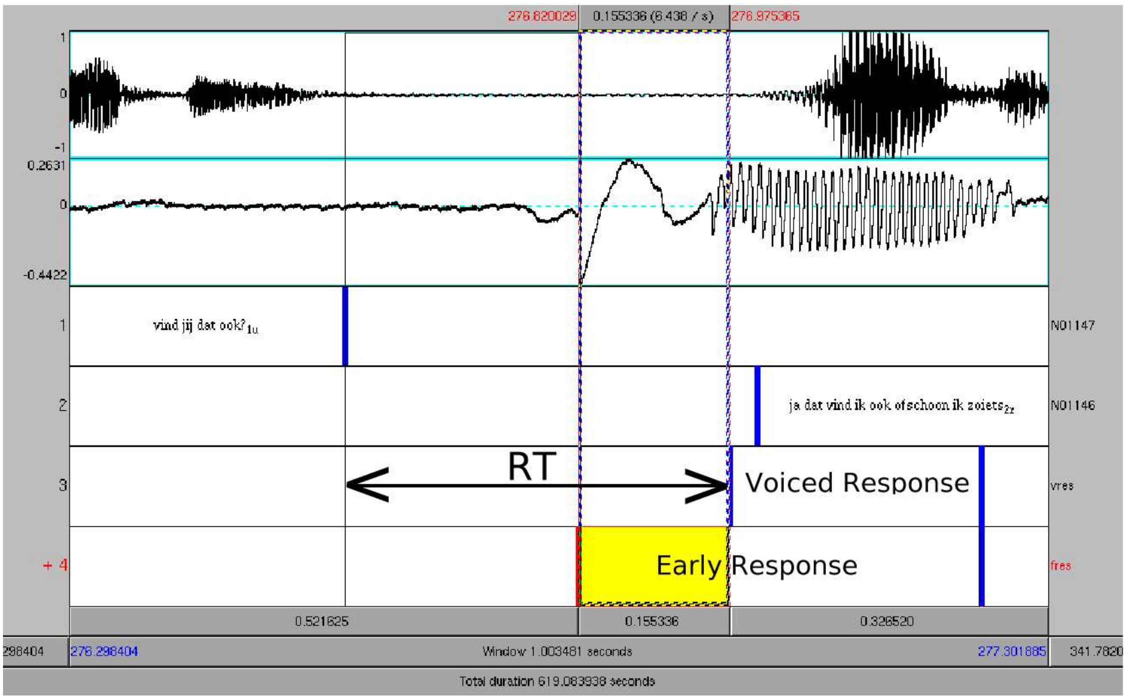

Figure 1: Example response waveform and segmentation. Top: Mono waveform of the stimulus, Center: laryngograph signal of a single response, Bottom: Annotation tiers for the transliterated utterances of the two speakers and the automatic segmentation of a *voiced* and *early response*. The two classes of response delays (and their difference, in color) are the intervals between the vertical lines.

and labeled these *voiced responses* in the recordings (see figure 1 and table 2).

It is assumed that each utterance end, defined as the end of the last (hand aligned) word in the hand labeled “chunk”, as given in the CGN, could function as a TRP. For each automatically determined response start, the distance to the closest utterance end (irrespective of the speaker), within a window of 1 second around the response start, was determined as the RT delay. To ensure that only causal responses were considered, the relevant utterance had to start at least 0.25 seconds before the start of the response. Furthermore, inspection of the laryngograph waverforms showed evidence of larynx movements that did not result in a noticeable voiced response but were sometimes still labeled as an extremely short voiced segment by Praat. Therefore, in this study, responses with a voicing duration shorter than 15 ms were discarded as spurious. Using the same criteria, minimal responses in the original (61) Spoken Dutch Corpus conversations were treated as responses to utterances of the other speaker in the dialog. These are presented here for comparison.

The distribution of responses with respect to the intonation boundary tones is given in table 2. Close to a thousand *voiced responses* were elicited for each of our experimental subjects (varying from 413 to 1374 *voiced responses* per subject), compared to less than a dozen “natural” minimal responses per participant in the original (61) conversations (122 speakers). Our subjects sometimes used more natural, and complex, responses than the prescribed ‘AH’, e.g. short utterances, laughing or giggling, or they corrected themselves. Utterances more complex than a simple syllable were often registered as multiple responses by the laryngograph. Therefore, any response starting less than 250 ms after the previous responses ended was discarded as spurious.

Each identified response was individually aligned with the corresponding part of the original conversation to compensate for small sample frequency differences between the original recordings and the response recording (c.f. Bailly, 2001). The sample ”drift” between these sounds was of the order of 90 ms for each 6 minute stimulus. The final alignment precision was 0.7 ms for the *full speech* stimuli and 2.1 ms for the *intonation only* stimuli which lacked almost all spectral information.

2.4.2 Early Responses

Quite often, a *voiced* minimal response is preceded by evidence of an early “preparation” of the larynx for the minimal response (in 2349 of 6084 *full speech* and 2377 of 6575 *intonation only* elicited *voiced responses*, see ta-

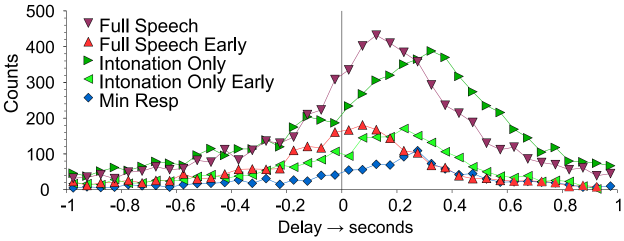

Figure 2: Distribution of reaction-time delays with respect to corresponding utterance-ends. Bin size is 50 ms. For total number of responses, see table 2

ble 2). These putative "preparation" responses will be referred to as *early responses*. These *early responses* are most likely caused by the laryngograph reacting to muscle movements and glottal closure well before the phonation starts (see figure 1). The total number of *early responses* varied widely per subject from 48 to 694 responses to the total 48 minute stimulus set. Since subjects were asked to say 'AH', this preparation could be a glottal closure for building up lung pressure. The large variation between subjects in number of detected *early responses* is possibly partly caused by the actual minimal response chosen by the subject which might not all induce measurable *early responses* (e.g. 'AH' versus 'hm' or 'ja') and partly by a laryngograph signal that sometimes was just too weak to allow the detection of *early responses*.

We were able to automatically label these *early responses* by segmenting the laryngograph signal around a *voiced response* at absolute (i.e. positive and negative) threshold crossings using Praat (with a threshold at 0.15 of the maximal amplitude). Threshold crossings should be no further apart than 200 ms. High amplitude low-frequency "waves" in the laryngograph signal were filtered out with a high-pass filter with a 4 Hz cutoff. The resulting segmentation proved to be quite consistent (see table 2). There is a sharp increase in the number of *early responses* just before the start of phonation. This suggests that these really short *early responses* are linked to the initiation of phonation itself, a phenomenon outside the scope of the current study. Therefore, *early responses* starting less than 40 ms before the start of a *voiced response* were ignored in the current study.

3 Results

RT measurements differ markedly between experimental subjects and were affected by the boundary tones (see Wesseling and van Son, in press). Therefore, all statistics were done on a subject-by-subject basis and end-tone categories (with a Bonferroni correction to $\alpha < 0.01$, two tailed). This was not really possible for the minimal response delays from the original conversations due to the huge number of speakers and low numbers of minimal responses per speaker. In total we recorded 6 hours of responses to each of the *full speech* and the *intonation only* stimulus set. These elicited 6565 and 7420 responses respectively (18.2 and 20.6 responses/minute). In the total set of 61 conversations, 1310 minimal responses were annotated (see table 2, 2.2 responses/minute). The differences between the number of responses to *full speech* and *intonation only* stimuli were not statistically significant ($p \geq 0.01$, Wilcoxon matched pairs signed ranks, WMPSR, test, on subjects). The distribution of the audible *voiced* and *early response* delays and the original minimal responses are presented in figure 2 and the distribution of the differences between the *voiced* and *early responses* in figure 3.

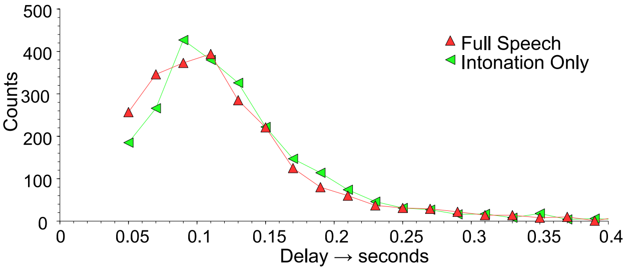

Figure 3: Distribution of the individual differences between the *voiced response* delay and the delay of the corresponding *early response*. Bin size is 20 ms, differences shorter than 40 ms are ignored. For total number of responses, see table 2

3.1 Voiced Responses

Figure 2 shows that response counts already start to increase before the end of the utterance, indicating that subjects were indeed able to predict upcoming utterance ends at least in some instances. The average response delays are 0.101 s ($Sd = 0.398$) for the *full speech* condition, 0.144 s ($Sd = 0.452$) for the *intonation only* condition and 0.127 s ($Sd = 0.414$) for the original minimal responses (see figures 4 and 5).

It is conceivable that the presence of an *early response* affects the delay of the actual *voiced response*, e.g. by delaying it even more. This was checked by comparing the RTs for *voiced responses* preceded by an *early response* to those *not* preceded by an *early response* for each subject, stimulus type, and boundary-tone class. Only for the *intonation only* stimuli was a 65 ms increase in delay found for voiced responses with an *early response* ($p < 0.01$, WMPSR test on subjects and boundary-tone classes). For *full speech* stimuli the increase was around 15 ms and not significant ($p \geq 0.01$, WMPSR test, id.). Also, we could not ascertain whether the effect of an *early response* on the *voiced response* differed between *full speech* and *intonation only* stimuli ($p \geq 0.01$, WMPSR test, idem). The presence of an *early response* had no effect whatsoever on the variance of the RT of the following *voiced response*. This means that whatever effect the presence of an *early response* has on the timing of the *voiced response*, it does not affect the integration-time in the C component, but more likely the M component (see Introduction). A possible explanation could be that *early response* are only initiated for specific types of *voiced responses* with an intrinsic longer t_0, e.g. only those responses that start with a glottal stop.

3.2 Early Responses

The average response delays for the *early responses* are -0.022 s ($Sd = 0.391$) for the *full speech* condition and 0.045 s ($Sd = 0.422$) for the *intonation only* condition. The differences between *voiced* and *early responses* are 0.130 s ($Sd = 0.165$) and 0.141 s ($Sd = 0.179$) for the *full speech* and *intonation only* condition respectively (see figures 4 and 5).

The differences between the mean delays and the standard deviations for *full speech* and *intonation only* stimuli are significant for both *voiced* and *early responses* ($p < 0.01$, WMPSR test on differences per subject and end-tone intonation) The mean, but *not* the standard deviations, of the differences between individual *voiced* and *early responses* differ between *full speech* and *intonation only* stimuli ($p < 0.01$, WMPSR test on differences per subject and end-tone)

The mean delays are, by construction, different for different response types (figure 4). For both stimulus types, the variance (standard deviation) of the differences between *voiced* and *early responses* was significantly lower than the variance of either of the *voiced* and textitearly responses itself (see figure 5 for variance of the differences; $p < 0.01$, WMPSR test on differences per subject and end-tone intonation). There was only a small difference between the variances for *voiced* and *early responses* ($p < 0.01$, WMPSR test on differences per subject and end-tone intonation, stimulus types pooled).

4 Discussion

Elicited minimal responses seem to be well suited to describe and analyze large conversational corpora. With around 6 hours of net listening time it was possible to get an average of 1 minimal response per utterance from 165 minutes of conversations (c.f. tables 1 and 2). This is only 2.2 times real time.

It is clear from the figures 4 and 5 that the *intonation only* stimuli induced both a longer RT and a larger variance, i.e. a larger effective integration-time. This was found for the audible *voiced responses* as well as the *early responses*. However, we did not find that the

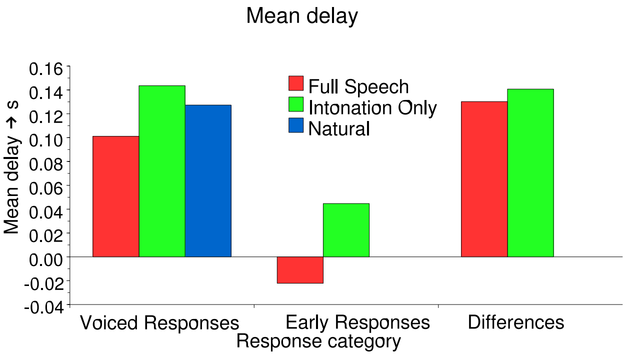

Figure 4: Mean delays for three types of response delays. For numbers, see table 2. See text for statistical results.

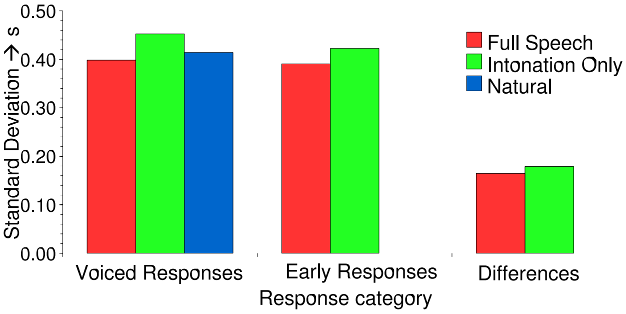

Figure 5: Mean standard deviations for three types of response delays. For numbers, see table 2. See text for statistical results.

variance of the differences between *voiced* and *early responses* was affected by the stimulus-type, although the mean delays were slightly different. This suggests that removing everything but intonation mostly affects the early integration-times, and much less the timing after the *early responses*, except that there seems to be an additional P, i.e. perceptual, type of delay. It is quite probable that the unnatural *intonation only* stimuli were more difficult to understand (c.f. Wesseling and van Son, in press). From these results it can be concluded that the intonation of speech in itself can be a sufficient, but impoverished, cue to project an upcoming TRP.

Using eq. 2 it is possible to determine the relative amounts of (integration) time, τ_{early}, it takes to decide to start an *early response* and then from there to subsequently start the actual *voiced response*, τ_{diff}. Using the variances corresponding to the *Early Response* and *Differences* columns of figure 5, the proportion shows to be:

$$\frac{\tau_{diff}}{\tau_{early}} \approx 0.55 \tag{4}$$

averaged over the individual τ estimates per speaker, stimulus type, i.e. *full speech* and *intonation only*, and boundary tone class. Note that the variances differed between the stimulus types and between *early* and *voiced responses* ($p < 0.01$, WMPSR test). This means that the integration-time τ_{early} to decide to start an *early response* is about twice as long as the integration-time τ_{diff} needed to decide to start the actual *voiced response*, after the *early response* has been initiated. If we use a simple model of Response Times, it can be assumed that the average difference times from figure 4 are already pure mean integration-times, τ_{diff}. That is:

$$\tau_{voiced} = \tau_{early} + \tau_{diff} \Leftrightarrow \qquad (5)$$
$$\tau_{diff} = RT_{voiced} - RT_{early}$$

Then from the average *difference* RT, 130 ms for *full speech* and 140 ms for *intonation only* (see figure 4), we can estimate that the integration-time, τ_{early}, for the *early responses* stimuli would be around 235 ms and 255 ms respectively (c.f. eq. 4). The total effective integration-times needed for the *voiced responses*, $\tau_{voiced} = \tau_{early} + \tau_{diff}$, might then be estimated to have been around 370 to 400 ms respectively.

The *early responses* we see in the laryngograph signals might be explained by the subjects preparing for the articulation of the minimal response (see section 2.4.2). From the results presented above it can be concluded that the generation of minimal responses involves at least one intermediate stage where the speaker starts preparing the intended utterance if needed. This preparation starts well before the actual articulation, on average, more than 100 ms before the actual start of phonation (see figure 4). At this point the subjects must have decided that a TRP is imminent, but they might still be unsure about its exact timing.

Intonation only stimuli showed to be quite capable of inducing the perception and projection of TRPs at utterance ends and elicit minimal responses (but see Wesseling and van Son, in press, about limitations). However, the increased integration-times for *intonation only* stimuli found under all circumstances is also evidence for the fact that subjects needed significantly more time to extract the information required to project the TRPs from the impoverished speech. The fact that not only the mean RTs, but also the dispersion, i.e. integration-time, increased, shows that it is really a question of less information being available in *intonation only* speech.

5 Conclusions

To summarize, we can conclude that the articulation of elicited minimal responses has at least one intermediate stage. An *early response* can often be observed in the laryngograph signal that suggests preparatory larynx and glottal movements. At this point, the subjects have obviously decided that a TRP is imminent and, possibly, what specific response they would articulate. Using the most simple model, a first estimate of the C or Central, component's integration-time, τ, of the three component PCM RT model, would be 235-255 ms up to the initiation of the preparatory movements and an additional 130-140 ms for the remaining time to initiate the actual phonation. The longer times are for the impoverished *intonation only* stimuli, which induced measurably longer integration-times than *full speech* stimuli, indicative for their lower information content. This shows, that while the (end-)intonation might be a sufficient cue to predict an upcoming TRP, it is measurably impoverished. With only intonation to go by, subjects definitely need more time to extract the information relevant to predict the utterance end.

From the average *early response* delays, from 20 ms before to 40 ms after actual end of the conversational utterances and the processing time needed to get there, over 235 ms, it is clear that our subjects used speech attributes from before the actual end of utterance to predict an upcoming TRP. Given the perceptual (P) and motor (M) delays involved in speech understanding and production (i.e. $\geq$50 ms under the most favorable circumstances, Bailly, 2001), we can tentatively conclude that planning (elicited) minimal responses starts more than 300 ms before the actual utterance end (TRP).

6 Acknowledgments

The authors would like to thank Louis ten Bosch and 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 and Louis Pols and Ingrid van Alphen for their valuable comments and suggestions. This project was made possible by grant 276-75-002 of the Netherlands Organization for Scientific Research (NWO).

References

Bailly, G. (2001). Close shadowing natural vs synthetic speech. In *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, SSW4*, page http://www.ssw4.org/.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345.

Boersma, P. and Weenink, D. (2004). Praat 4.2: doing phonetics by computer. Computer program: http://www.Praat.org/.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31(2):251–276.

Liddicoat, A. J. (2004). The projectability of turn constructional units and the role of prediction in listening. *Discourse Studies*, 6(4):449–469.

Oostdijk, N. (2000). The Spoken Dutch Corpus, overview and first evaluation. In Gravilidou, M., Carayannis, G., Markantonatou, S., Piperidi, S., and Stainhaouer, G., editors, *Proceedings of LREC-2000*, volume 2, pages 887–894.

Oostdijk, N., Goedertier, W., Eynde, F. V., Boves, L., Martens, J., Moortgat, M., and Baayen., H. (2002). Experiences from the Spoken Dutch Corpus project. In Rodriguez, M. and Araujo, C. S., editors, *Proceedings of LREC-2002*, volume 2, pages 340–347.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–226.

Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biology*, 3(2):e51.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, 50(4):696–735.

Sigman, M. and Dehaene, S. (2005). Parsing a cognitive task: A characterization of the mind's bottleneck. *PLoS Biology*, 3(2):e37.

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

van Son, R. J. J. H. (2005). A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. *Acta Acustica united with Acustica*, 91(4):771–778.

Wesseling, W. and van Son, R. J. J. H. (in press). Timing of experimentally elicited minimal responses as quantitative evidence for the use of intonation in projecting TRPs. In *Proceedings of Interspeech2005*, Lisbon.