# Timing of Experimentally Elicited Minimal Responses as Quantitative Evidence for the Use of Intonation in Projecting TRPs

*Wieneke Wesseling and R.J.J.H. van Son*

Chair of Phonetic Sciences/ACLC,
Department of Linguistics, University of Amsterdam, The Netherlands
W.Wesseling@uva.nl

## Abstract

In an RT experiment, subjects were asked to respond with minimal responses to prerecorded dialogs and a manipulated version of these dialogs that contained only intonation and pause information. Response delays and, especially, variances were higher to the impoverished, *intonation only*, stimuli than to the original recordings. It was also found that *intonation only* utterances ending in a mid-frequency pitch induced significantly longer response delays than utterances ending in a low pitch. These results are interpreted as evidence that just the intonation and pauses of a conversation already contain sufficient information to project end-of-utterance TRPs. However this information is measurably impoverished with respect to full speech to an extent that increases the "processing" time by 10%. Our subjects seemed to fall back to reacting to pauses when presented with *intonation only* utterances ending in a mid-frequency tone. This suggests that, in contrast to low or high end-tones, intonation contours that end in a mid-frequency tone might not contain any useful information for predicting end-of-utterance TRPs.

## 1. Introduction

An important task for participants in conversations is to identify the moment other participants finish speaking, allowing them to take the floor. Given the number of factors involved, the identification of possible Transition Relevance Places (TRPs) would be expected to be a difficult task. Nevertheless, transitions between 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 [1], [2]. Information sources that are known to be used for this projection include gaze direction, gestures, intonation, syntactic, and timing information (like speaking rate and pauses). In her experiments, Caspers [3] found that syntactic completion seems to be the main factor in the turn-taking mechanism. Boundary-tones tended to support the grammatical structure by signaling turn-incompleteness by *mid-register* tones at positions where pauses *did not* coincide with syntactic completion and by the use of *low* or *high* tones where pauses *did* coincide with a TRP.

The goal of the present study is to provide some *quantitative* data about the time-course with which information becomes available and about the way the various sources of information are combined in the comprehension of spoken language. This is done by measuring response times (RTs) in a simulated minimal response task, that involved subjects listening to recordings of natural dialogs and giving minimal responses (by saying 'AH') to both speakers in these dialogs. Minimal responses are the (natural) reaction of listeners and are here assumed to signal comprehension of at least part of the utterance's structure and a recognition of a possible end-of-turn (TRP).

Psychological research involving response times to investigate the mental decision-making process, has modeled this process as a noisy integrator that stochastically accumulates perceptual evidence from the sensory system in time [4], [5]. Three stages of processing can be identified: a perceptual component ($P$), a central decision making component ($C$), and a motor component ($M$). Experiments by Sigman and Dehaene [4] showed that the central component $C$ was responsible for almost all of the variance in RTs. In their model, 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 $\tau = \frac{1}{\alpha}$ [4]. The average RT becomes $\overline{RT} = t_0 + \tau$ and the variance $var(RT) = \frac{1}{2}\sigma^2\tau^3$ where $\sigma^2$ is a task independent (mostly unknown) modeling parameter. The proportion of the integration time constants ($\tau$) for two experimental conditions (e.g. $i$ and $j$) can be determined from their respective variances ($s_i^2$ and $s_j^2$) as:

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

Eq. 1 is independent of the difficult to estimate $\sigma^2$ parameter.

## 2. Materials and Methods

### 2.1. Speech Materials

All speech material was obtained from the Spoken Dutch Corpus (CGN) [6], [7]. In 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 [8] as either a Minimal Response, a Question/Answer pair, or a General Turn switch. For each turn-switch, 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

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 |

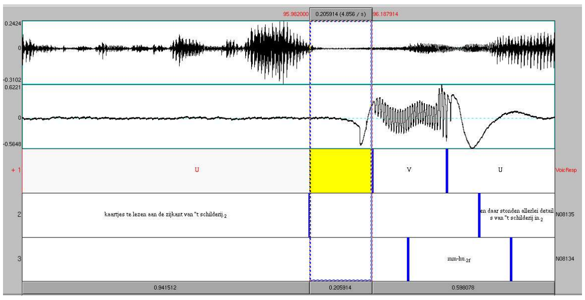

Figure 1: Example response waveform and segmentation. Top: Mono waveform of the stimulus, Center: laryngograph signal of a single response, Bottom: Annotation tiers for the automatic segmentation of the response and the transliterated utterances of the two speakers. The response delay is the interval between the vertical lines.

("chunks") and word boundary segmentations, transliterations and phonetic transcriptions were available. In the context of the conversations used in this study, the hand-labeled utterances can be interpreted as a very crude prosodic phrasing. About 75% of these utterances are followed by silent pauses. A subset of 7 switchboard and 10 home recordings with a total duration of 165 minutes were selected for the present stimulus set, based on high audio quality and coverage of the turn switching categories.

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 [9]. 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:

$$Z_i = \frac{\overline{F}_0^i - F_{0end}^i}{Sd\left(F_0\right)} \qquad (2)$$

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.

## 2.2. Stimulus preparation and presentation

Telephone switchboard stereo speech recordings in the CGN have been digitized at an 8 kHz sampling frequency and 8 bit precision. The two speakers in the telephone conversations were recorded on separate channels. Volunteer face-to-face stereo home recordings were digitized at 16 kHz and 16 bit precision. The stereo signal allowed an auditory spatial separation of the speakers.

Table 2: *Total number of responses to stimuli for each of the end-tone categories and minimal responses for the total conversation set. Between brackets the total number of responses including non-attributable responses is given.*

| material | low | mid | high | total |
|---|---|---|---|---|
| full speech | 1884 | 2888 | 1385 | 6157 (6565) |
| intonation only | 1970 | 3303 | 1480 | 6753 (7420) |
| total conversation set | 386 | 539 | 281 | 1206 (1310) |

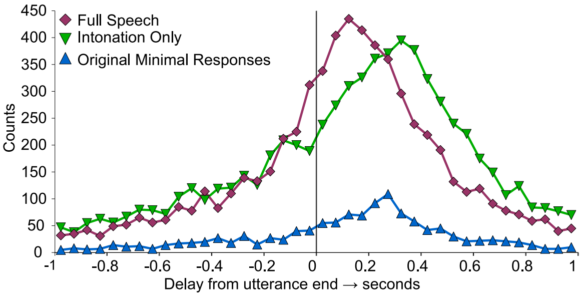

Figure 2: Distribution of reaction-time delays with respect to corresponding utterance-ends. Bin size is 50ms. For total number of responses, see table 2

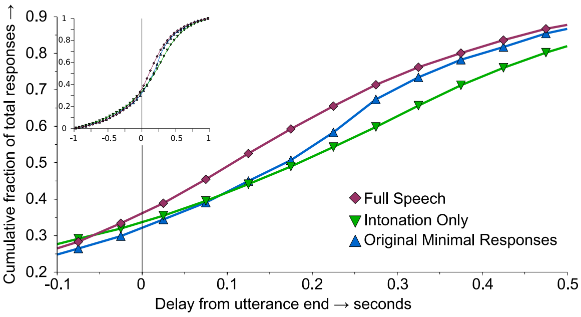

Figure 3: Cumulative distribution of reaction time delays of fig. 1. For total number of responses, see table 2

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). A second set of stimuli was created by converting these *full speech* stimuli to pitch contours with Praat and having them resynthesized as "hummed" neutral-vowel speech [9] (*intonation only* stimuli). This hummed speech contains nothing but the intonation and pause structure of the original speech, i.e. no loudness or spectral information from the original versions was present. All stimuli were upsampled to 16 kHz where necessary.

Stimuli were pseudo-randomized for presentation. Each of the 15 subjects heard a different subset and order 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 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 *intonation only* fragment from a dialog that was not in the stimulus set.

## 2.3. Response collection and processing

Stereo stimuli were played directly from an Acer Travelmate 529 laptop running Knoppix (Linux 2.4.26) in console mode. The same laptop recorded the laryngograph responses concurrently at a 16 kHz sampling rate on one channel. A fed-back (summed) mono version of the stimulus was duplex recorded

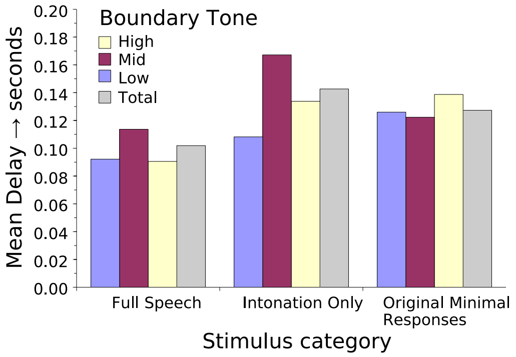

Figure 4: Mean delays for three categories of boundary tones. For numbers, see table 2. See text for statistical results.

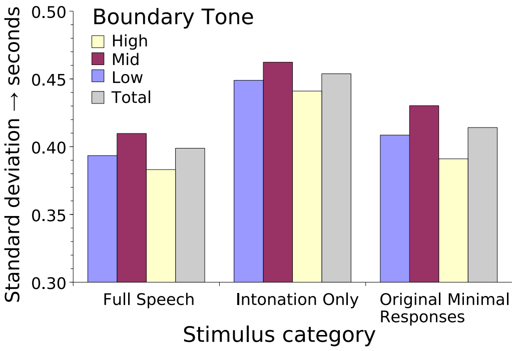

Figure 5: Standard deviation of delays for three categories of boundary tones. For number of responses, see table 2. See text for statistical results.

on the other stereo channel for alignment purposes [10]. 15 Naive subjects, 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 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 *intonation only* type.

Responses were recorded with a laryngograph (Laryngograph Ltd, Lx proc). The 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 [9] located and labeled these responses in the recordings, see fig. 1.

It is assumed that each utterance end, defined as the end of the last (hand aligned) word in the utterance, could function as a TRP. For each automatically determined response start, the distance to the closest utterance end was determined as the RT delay (irrespective of the speaker) within a window of 1 second around the response start. 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, responses with a duration shorter than 15ms 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. Our subjects sometimes used more natural, and complex, responses than the prescribed 'AH', e.g. short utterances, laughing or giggling, or corrected themselves. Complex responses were often registered as multiple responses by the laryngograph. At the current level of analysis, we did not filter out these complex responses. We have close to a thousand elicited responses for each of our experimental subjects, compared to less than a dozen "natural" minimal responses per participant from the original (61) conversations (122 speakers).

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 (cf, [10]). 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 (spectrally impoverished) *intonation only* stimuli.

## 3. Results

RT measurements differ markedly between experimental subjects. Therefore, all statistics were done on a subject-by-subject basis (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. 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 (2.2 responses/minute), see table 2. 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 response delays and the original minimal responses are presented in fig. 2 and fig. 3. Fig. 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. From fig. 3 it can be concluded that delays are shortest for *full speech* stimuli, while the delay for *intonation only* and the original minimal response delays are comparable. The average response delays are 0.102s ($Sd = 0.399$) for the *full speech* condition, 0.143s ($Sd = 0.454$) for the *intonation only* condition and 0.127s ($Sd = 0.414$) for the original minimal responses. The differences between the mean delays and the standard deviations for *full speech* and *intonation only* stimuli are both significant ($p < 0.005$, WMPSR test on differences per subject). None of the differences between the mean delays and the standard deviations for both experimental conditions and the minimal responses are significant ($p \geq 0.01$, Student-t test on means and F-test on variance, respectively).

Fig. 4 shows the average response delays for the three types of boundary tones. The difference between *full speech* and *intonation only* stimuli are only significant for the responses to the mid boundary tone utterances ($p < 0.001$, WMPSR test on subject mean delays). The relative RTs ordering and differences between the boundary tones are not significant in the *full speech* condition and the original minimal responses ($p \geq 0.01$), but the relative ordering is significant for the *intonation only* stimuli ($p < 0.001$, Friedman test $Q = 16.90$, per subject), mostly due to the difference between the mid and low boundary tone utterances ($p < 0.001$, WMPSR test per subject).

For the standard deviation (see fig. 5) none of the differences between boundary tones is significant ($p \geq 0.01$, Friedman test and WMPSR test, on subject differences). For all boundaries tones the difference in variances between responses to *full speech* and *intonation only* is significant ($p < 0.001$, WMPSR test on subject standard deviation).

## 4. Discussion and conclusions

Around 50% of all elicited and natural minimal responses were found within 25% of our 2s window (between -0.1s and 0.4s, fig. 2). The low number of natural minimal response delays prevented us from exploring the differences between experimental and natural minimal responses. This must be left to later studies. We do find that our subjects were responding at a high rate to the speech they heard. We recorded almost 20 responses per minute. Around 90% of the responses could be attributed to the ends of individual utterances (see table 2).

The main result of this study is that impoverished *intonation only* conversational speech elicited delayed and more variable responses (average delays statistically significant only for mid tone utterances) than the original *full speech* stimuli. ¿From these results it can be concluded that our subjects were able to point out TRPs at putative utterance ends at high response rates with low latencies in both *intonation only* speech and *full speech* stimuli. Response times to *full speech*, around 100ms, are only 30ms slower than those seen in the fastest spoken response latency experiments: close shadowing of a known text by trained subjects [10]. Responses to incomprehensible *intonation only* low boundary tone utterances were actually faster (i.e. 110ms) than most latencies for shadowing a synthesized version of a known text [10]. These rapid responses and the fact that the actual boundary tones *did* affect response delays in *intonation only* speech rule out that subjects reacted to the utterance ends themselves, instead of predicting the ends from the content or intonation of the stimuli. So it seems that the intonation into a high or low boundary tone is indeed sufficient for our subjects to estimate the position of an upcoming utterance end at least some of the time.

Both intonation (i.e. boundary tone) and verbal and prosodic information (in the *full speech* condition) help TRP projection in terms of reduced delays (fig. 4). From the variances of the delays per speaker (c.f. averages in fig. 5), the relative integration times (eq. 1) for the different experimental conditions can be estimated. Removing all acoustical information except for the intonation and pauses (i.e. *intonation only* stimuli), increases the integration time with around $10 \pm 1.3$ % (average proportion per speaker and tone). Relative differences between boundary tones are at most 4% in *full speech* and 3% in *intonation only* stimuli and never statistically significant. The significant difference in average delay between *intonation only* utterances with a mid and low tone (60ms, fig. 4) is too large to be explained by a modest 3% difference in integration time. We must assume that the deterministic $t_0$ is considerably higher for mid boundary tones in *intonation only* stimuli. A possible explanation is that subjects are unable to predict the end of an utterance if it has a mid boundary tone and have to wait for the end to be actually perceived before they can respond (i.e. they are responding to the pauses rather than the intonation). The same difference in end tone has less dramatic consequences in *full speech* stimuli (20ms, fig. 4) as the subjects can use the verbal information for end of utterance projection. For *full speech* stimuli, subjects have to wait less often for the speaker to stop speaking before they can initiate a response.

To summarize, the *intonation only* (+pauses) condition contains less information on upcoming (end-of-utterance) TRPs than the *full speech* condition, but is still sufficient for detecting TRPs (as end of utterances). On average, the integration (processing) time of the central, decision, component increases with approximately 10%. With a mid boundary tone the subjects might fall back to responding to the pause at the actual end of the utterance for lack of predictive information in the intonation, much more so for *intonation only* stimuli than for *full speech* stimuli.

## 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 and Prof. Louis C.W. Pols for his valuable comments and suggestions. This project was made possible by grant 276-75-002 of the Netherlands Organisation 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] Pickering M. J. and Garrod S.,“Toward a mechanistic psychology of dialogue”, Behavioral and Brain Sciences 27: 169-190, 2004.

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

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

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

[6] 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.

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

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

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

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