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

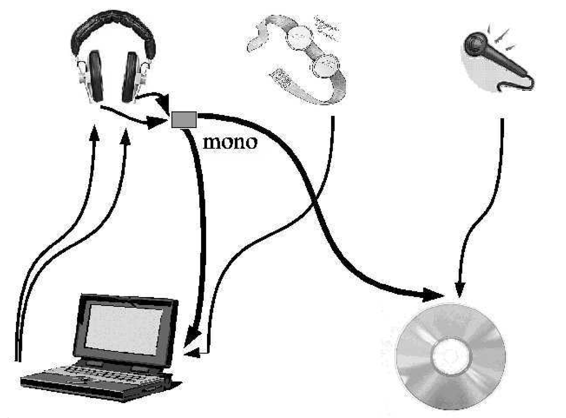
Wieneke Wesseling and R.J.J.H. van Son, ACLC/Chair of Phonetic Sciences, Department of Linguistics, University of Amsterdam, The Netherlands `W.Wesseling@uva.nl, R.J.J.H.vanSon@uva.nl`

Introduction

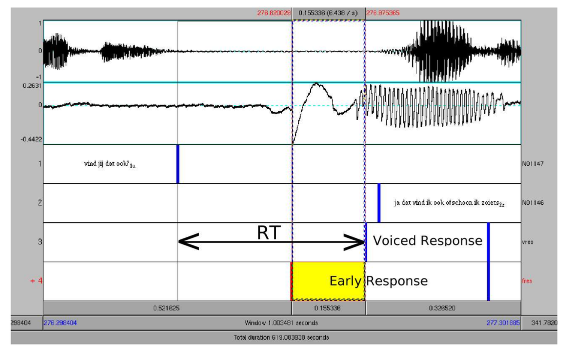
Our interest is the relative importance of various sources of information in understanding language, in particular in the recognition and projection of Transition Relevance Places (TRPs), or potential turn changes in (natural) human conversation.

- Is intonation enough for TRP projection?
- How is the use of intonation integrated with other sources of information?
- What do we know about the timing of TRP projection?

Reaction Time (RT) experiment



Recording setup with laryngograph and audio



Speech with laryngograph signal and annotation of Speech, RTs and their difference

Stimuli: Dialogs from Spoken Dutch Corpus (CGN):

1. *Full Speech* condition
2. *Intonation Only* condition (intonation and pause information)

Task: Recognition of end-of-turns; Respond with 'minimal responses' (*AH*) to prerecorded dialogs. The assumption is that at this point there is recognition of (at least part of) the utterance.

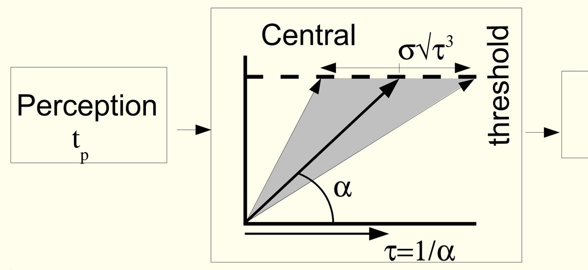
Voiced Reaction Time (RT): Voicing Start - Utterance End: the distance from the onset of voicing to the closest utterance-end (as defined in CGN) within a window of 1 second (0.25s refractory period between responses).

Early Reaction Time (RT): Start of Laryngograph signal - Utterance End: As Voiced RT but with a 40ms lower cut-off.

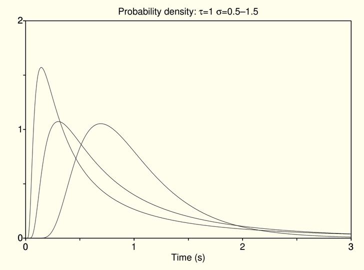
Boundary Tones: for each utterance, the *end intonation* Z_i was established (see materials)

Responses were recorded with a laryngograph and automatically labeled in Praat

Perception-Central-Motor model of Reaction Times



Perception-Central-Motor model of RTs



Ideal RT distributions

- Three stages of processing: a perceptual component (P) and a motor component (M), with a deterministic response-time t_0 and a central **decision making component** (C), characterized by a random walk to a decision threshold, determined by an integration-time $\tau = \frac{1}{\alpha}$.
- From this model, the proportion of integration times can be determined from their respective variances (see Appendix for formulas)
- The difference between the *Voiced* and the *Early* part of a response behaves like an RT, in a first order approximation (i.e., $\tau_{diff} = \tau_{voiced} - \tau_{early}$ with identical t_0).

Materials

Full set: 61 informal Dutch dialogs with basic annotation (588 min.), 32 switchboard telephone, 29 home recorded face-to-face dialogs

- Basic Utterances
- Minimal Responses

Stimulus set: 17 dialogs with hand aligned word boundaries (165 min.), 7 switchboard and 10 home recordings

Subjects: 18 naive native Dutch speakers

Boundary tones: for each utterance end, the end intonation Z_i was established as:

$$Z_i = \frac{\overline{F_0^i} - F_{0end}^i}{Sd(F_0)}$$

High: $Z_i > 0.2$
Mid: $0.2 \geq Z_i \geq -0.5$
Low: $Z_i < -0.5$

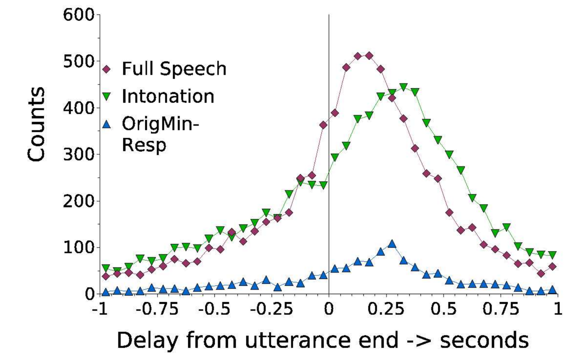
Total number of utterances for each of the end-tone categories for all conversations and for the stimuli

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

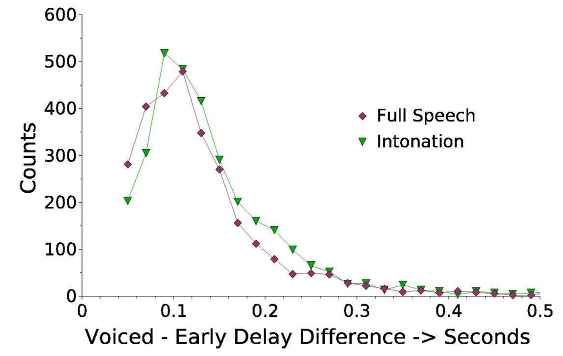
Total number of (minimal) responses to stimuli and full set for the end-tone categories

response category	*low*	*mid*	*high*	*total*
full speech	*2294*	*3410*	*1700*	*7404*
intonation only	*2316*	*3893*	*1778*	*7987*
full set (min resp)	*386*	*539*	*281*	*1206*

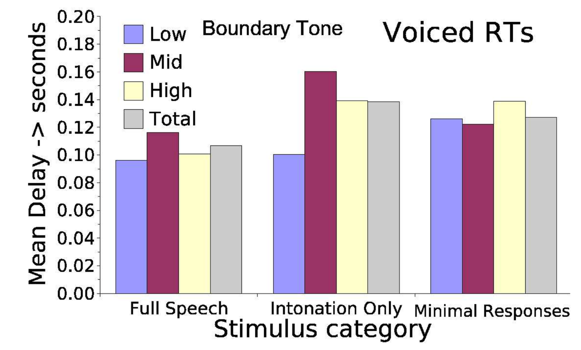
Results



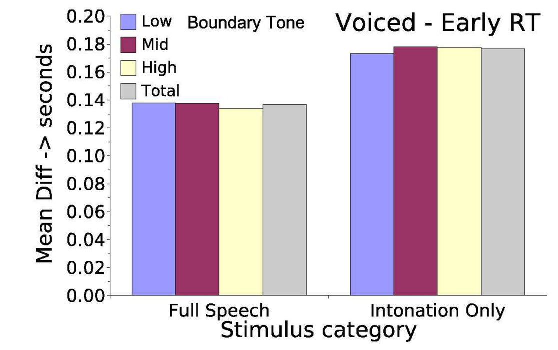
R1a Voiced RT distribution



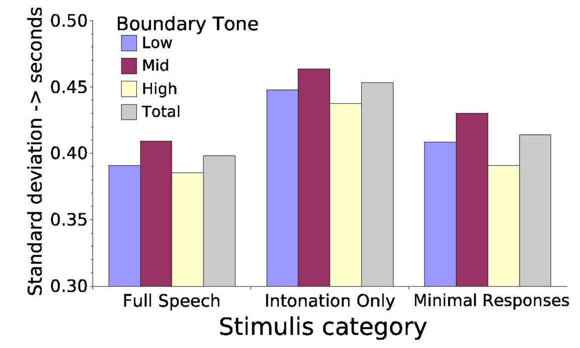
R1b Voiced-Early RT distribution



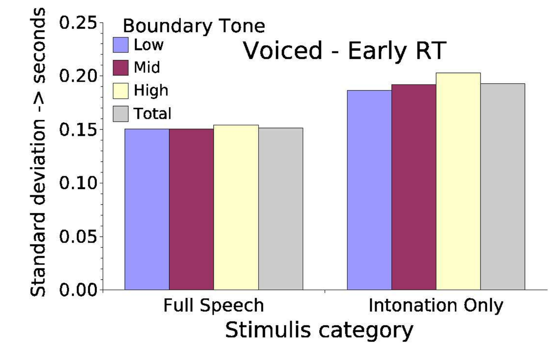
R2a Mean Voiced delays for three categories of boundary tones



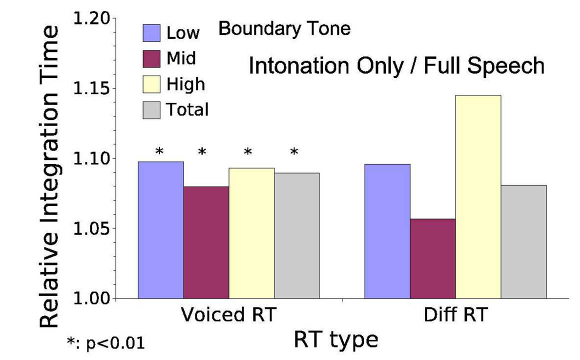
R2b Mean differences between Voiced and Early responses



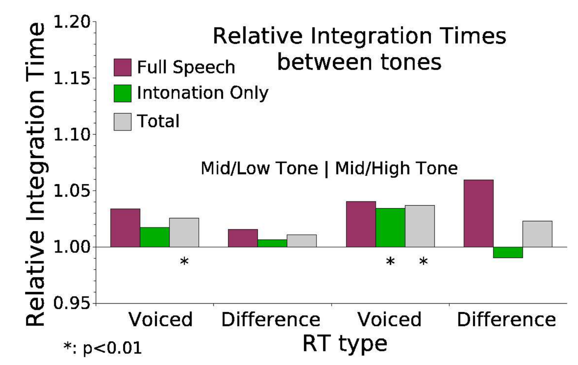
R3a Standard deviation of delays for three categories of boundary tones



R3b Standard deviations for differences between Voiced and Early responses



R4a Proportion of τ_{io} and τ_{fs} for Voiced RTs and Difference RTs (weighted by response)



R4b Mid-tone versus Low (left) and High (right) tones (*id.*)

- R1 Response counts are already increasing before end of utterance → projection takes place in both conditions.
- R1 Delays are shorter for *Full Speech* stimuli.
- R2 Difference between *full speech* and *intonation only* is only significant for *mid* boundary tones.
- R2 Relative ordering is significant only for *intonation only* stimuli (mostly between mid and low boundary tones)
- R3 None of the differences between boundary tones is significant
- R3 For all boundary tones the difference in variances between responses to *full speech* and *intonation only* is significant

Conclusions

- Impoverished *intonation only* speech *increases* the *Reaction Times*
- It *increases* integration times by 10 ± 1.0 % (unweighted average of τ per subject)
- Mid-tone *intonation only* speech has longer *plain* RTs (by 60ms)
- But *Standard Deviations* and *Integration Times* are *not* increased
- ⇒ Mid-tone *intonation only* speech induces a higher t_0, but not a higher τ
- Subjects might react to mid-tone *intonation only* speech by waiting for the pause

Discussion

- 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 *mid* boundary tones, 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.

Future work

- Use manipulated pauses, intonation and loudness;
- Use manipulated visual speech;
- Integrate results with high level annotations (e.g., syntax).

Appendix

Reaction time distribution $g(t)$:

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

Define integration time $\tau = \frac{1}{\alpha}$

Average Reaction Time: $\overline{RT} = t_0 + \tau$

Variance: $var(RT) = \frac{1}{2}\sigma^2\tau^3$ (with σ as a modeling parameter)

Proportion of integration times τ_i and τ_j: $\frac{\tau_i}{\tau_j} = \sqrt[3]{\frac{s_i^2}{s_j^2}}$