Prominent Words as Anchors for TRP Projection

R.J.J.H. van Son, Wieneke Wesseling, and Louis C.W. Pols
Institute of Phonetic Sciences/ACLC, University of Amsterdam, The Netherlands
R.J.J.H.vanSon@uva.nl, W.Wesseling@uva.nl

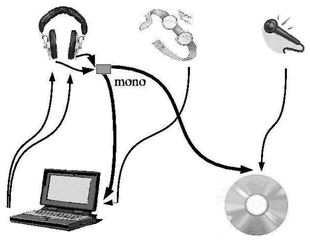
Introduction

We are interested in the relative importance of various sources of (prosodic) information, e.g. pitch*, pauses, stress, in the perception of speech. To reach this goal, we are comparing the recognition and projection of Transition Relevance Places, potential turn changes, in (natural) human conversation in 'normal' and manipulated versions.

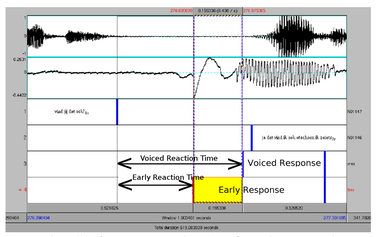
- How do final accents affect TRP projection?
- What do we know about the timing of TRP projection?

*See Wieneke Wesseling, R.J.J.H. van Son, and Louis C.W. Pols, 'On the Sufficiency and Redundancy of Pitch for TRP Projection', Interspeech 2006 session Thu1FoP,"Prosody", 10:30 Thursday

Reaction Time (RT) experiment



Recording setup with laryngograph and audio



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

Stimuli: 17 informal Dutch dialogs from Spoken Dutch Corpus (CGN), with basic annotation and hand aligned word boundaries (165 min. 7 switchboard and 10 home recordings). *Natural turn switches* are compared to Elicited RTs delays:

1. *Original* condition
2. *Hummed* condition (intonation / pause information)
3. *Whispered* condition (no periodic 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

Responses: recorded with a laryngograph and automatically labeled in PRAAT

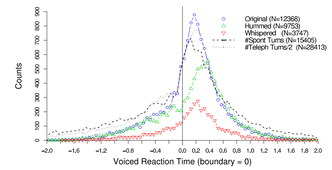
- *Voiced Reaction Time (RT):* Distance from the start of Voicing to the closest Utterance End (as defined in CGN) within a window of 1 second
- *Early Reaction Time (RT):* Distance from start of Laryngograph signal to the Utterance End

Subjects: 32 naive native Dutch speakers

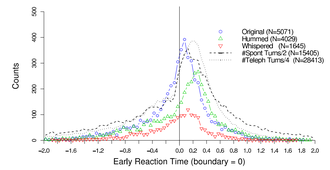
- Experiment 1, Original vs. Hummed, 21 subjects
- Experiment 2, Original vs. Whispered, 11 subjects

Pitch accents: for the last 3 words of each utterance, the last prominent word (as labeled in CGN) was marked

Results

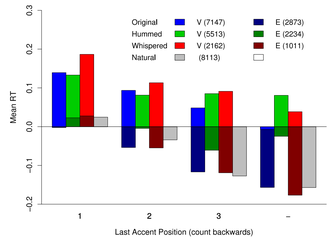


R1a Voiced RT distribution

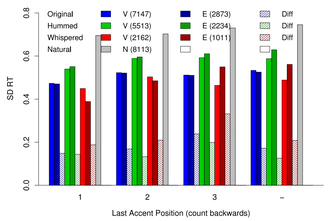


R1b Early RT distribution

R1 Response counts are already increasing before end of utterance → projection takes place in all experimental conditions as well as *natural turn switches*



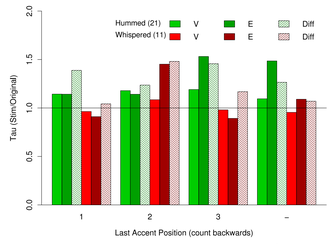
R2a Mean delays for accent positions
('-': no accent in last three words)



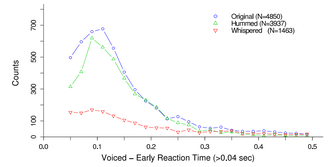
R2b Standard deviation of delays for accent positions

R2a Clear correlation between average RT and distance to last accent in all but *hummed* stimuli. Voiced responses to *hummed* utterances are only affected by the *final accent*

R2b Accent position has little or no impact on standard deviation



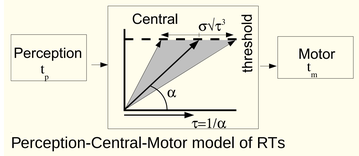
R3a Relative "processing" time $\frac{\tau'}{\tau_{orig}}$ for accent positions and different stimulus types
('-': no accent in last three words)



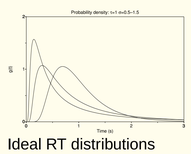
R3b Voiced-Early RT distribution (Early RT with a 40ms lower cut-off)

R3a Relative increase in "processing time" for *hummed* stimuli, no effect of accent position

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 $\frac{\tau'}{\tau_{orig}}$ can be determined from their respective variances.

Conclusions

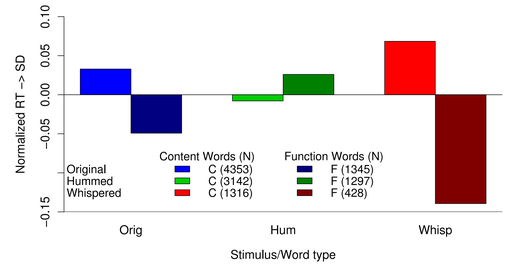
- Subjects can project TRPs with high reliability under all stimulus conditions used, even using only intonation
- *Original* and *Whispered* stimuli did not differ significantly
- RTs are strongly affected by the position of the last accent
- This accent effect cannot be attributed to increased integration time
- These results suggest that listeners predict TRPs using the last prominent word as a starting point

Discussion

- Our whispered stimuli might still contain intonational components (e.g. duration, loudness, reduction)
- Pitch movements on the final word might disturb projection
- An accent marks an unpredictable word, following words might be predictable

Current work

Predictability affects Reaction Times



R4 Reaction Time versus type of the utterance-final word

RT values are normalized for stimulus type, subject and accent position ($mean = 0$, $sd = 1$).
Every utterance ends either in a high frequency Function word, **F**, or in a Content word, **C**.
Differences are statistically significant for *Original* and *Whispered* stimuli ($p < 0.01$, t-test), but not for the *Hummed* stimuli ($p \geq 0.05$).

More to come

- Manipulated other modalities, eg. pauses, and loudness
- Add visual modality (video recordings)
- Integrate results with high level annotations (e.g., POS, syntax)

INTERSPEECH 2006 Pittsburg USA