On the Sufficiency and Redundancy of Pitch for TRP Projection

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

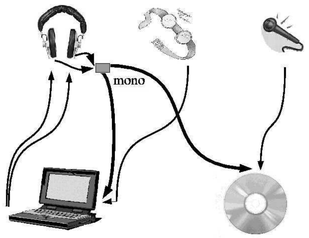
Introduction

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're comparing the recognition and projection of Transition Relevance Places, or potential turn changes in (natural) human conversation in 'normal' and manipulated versions. Claims:

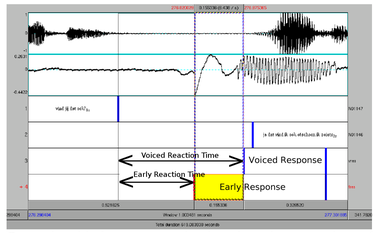
- Intonation is sufficient for TRP projection, perhaps redundant.
- End intonation (low - mid - high) affects TRP projection.
- TRP projection can start early in the utterance.

* See R.J.J.H. van Son, Wieneke Wesseling, and Louis C.W. Pols, 'Prominent Words as Anchors for TRP Projection', Interspeech 2006 session Mon2FoP, "Spoken Dialog Systems I", 14:00 Monday

Reaction Time (RT) experiment



Recording setup with laryngograph and audio



Speech with laryngograph signal and annotation of Speech, Voiced/Early RT and their difference (yellow)

Stimulus set: 17 informal Dutch dialogs from Spoken Dutch Corpus, with basic annotation and hand aligned word boundaries (165 min., 6670 utterances, 7 switchboard and 10 home recordings).

1. *Original* condition
2. *Hummed* condition - only intonation, resynthesized neutral vowel speech
3. *Whispered* condition - no intonation, resynthesized from LPC analysis using white noise as sound source

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, automatically labeled in PRAAT

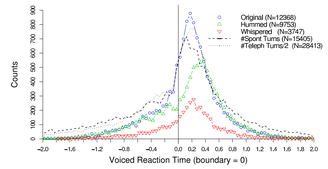
- *Voiced Reaction Time (RT):* Distance from the start of Voicing to the closest Utterance End (as defined in CGN) within a window of 1 second.
- *Early Reaction Time (RT):* Distance from start of Laryngograph signal to the Utterance End.

Subjects: 32 naive native Dutch speakers

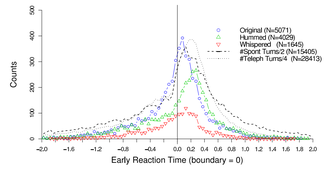
- Experiment 1, Original vs. Hummed, 21 subjects
- Experiment 2, Original vs. Whispered, 11 subjects

Intonation: for each utterance, the *end intonation* was automatically marked and hand-checked by human labelers, as *low*, *mid* or *high*.

Results

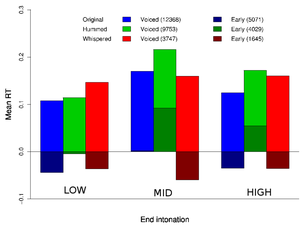


R1a Voiced RT distribution

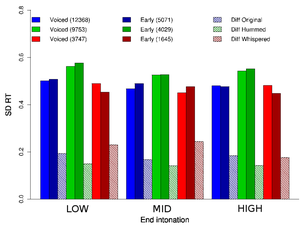


R1b Early RT distribution

R1: Response counts are already increasing before end of utterance → projection takes place in all conditions.



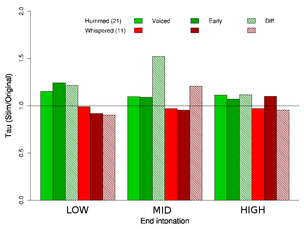
R2a Mean delays for three categories of boundary tones (low, mid, high).



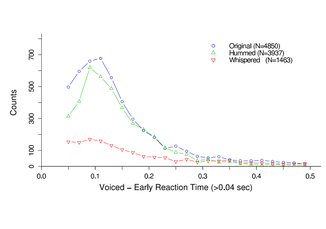
R2b Standard deviation of delays for the three boundary tones (low, mid, high).

R2a: Longer RTs for *hummed* stimuli are only significant for *mid* and *high* boundary tones.

R2b: Larger variances in both Voiced and Early responses for *hummed* stimuli.



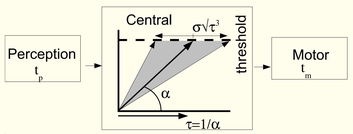
R3a Relative "processing" time $\frac{\tau'}{\tau_{orig}}$ for three categories of boundary tones and different stimulus types



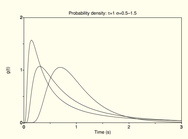
R3b Voiced-Early RT distribution

R3a: There is a relative increase in "processing time" for *hummed* stimuli. There is no effect or only slightly faster decision times for *whispered* stimuli.

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 $\tau = \frac{1}{\alpha}$.
- The proportion of integration times $\frac{\tau'}{\tau_{orig}}$ can be determined from their respective variances.

Conclusions

- Impoverished *hummed* speech elicits *delayed* and *more variable* responses than original stimuli, but subjects are still able to project TRPs with high reliability using only intonation.
- Subjects might react to mid-tone *hummed* speech by waiting for the pause. *Whispered* stimuli did not differ/have slightly faster decision times.

Discussion

- Intonation is a sufficient cue to project TRPs when the utterance ends in high or low pitch.
- We found no evidence that pitch is not a redundant cue for TRP projection in normal speech.
- However, our whispered stimuli might still contain intonational components (e.g. duration, loudness, spectro-temporal properties)

Latest work: Pragmatic annotation

All utterances were labeled in terms of their discourse function such as:

beginning:	new subject or change of subject
continuation:	current subject is continued
elicitation:	elicits a respons from the other speaker
reaction:	respons to an elicitation
formula:	grounding act of formulaic utterance
repetition:	elements from earlier utterances are repeated, without new information
interruption:	jokes and utterances about subject matter that don't add new information
interjection:	not subject-related, but informative remark
hesitation:	no information, but marks that speaker wants to say more

Probability and distribution of resp. over pragmatic categories

function	*stim type*	*N.*	*resp.*	*prob. of resp.*	*avg. delay*	*stddev. delay*
beginning	*Orig.*	*1102*	*540*	*.49*	*.130*	*.60*
	Hum.	*764*	*469*	*.61*	*.061*	*.65*
	Whisp.	*349*	*115*	*.33*	*.102*	*.55*
continuation	*Orig.*	*13152*	*7141*	*.54*	*.104*	*.50*
	Hum.	*8585*	*4768*	*.56*	*.105*	*.57*
	Whisp.	*4513*	*2072*	*.46*	*.157*	*.49*
elicitation	*Orig.*	*2846*	*1190*	*.42*	*.130*	*.47*
	Hum.	*1957*	*1050*	*.54*	*.129*	*.54*
	Whisp.	*1031*	*372*	*.36*	*.145*	*.51*
reaction	*Orig.*	*2175*	*977*	*.45*	*.170*	*.42*
	Hum.	*1484*	*734*	*.49*	*.204*	*.49*
	Whisp.	*784*	*313*	*.40*	*.160*	*.43*
formula	*Orig.*	*8780*	*2418*	*.28*	*.104*	*.40*
	Hum.	*6052*	*2110*	*.35*	*.344*	*.41*
	Whisp.	*3154*	*797*	*.25*	*.197*	*.39*
repetition	*Orig.*	*536*	*256*	*.48*	*.064*	*.40*
	Hum.	*358*	*171*	*.48*	*.146*	*.47*
	Whisp.	*179*	*63*	*.35*	*.135*	*.47*

Notes:

- NB: No content recognisable in *Hummed* condition, so any effects here can't be attributed to function type.
- Slow responses to formulaic utterances in *Hummed* condition may be caused by shortness of the signal.
- Tentative result: *repetitions* elicit significantly faster responses.

Future work

- Manipulated other modalities, eg. pauses, and loudness.
- Add visual modality (video recordings).

INTERSPEECH 2006 Pittsburgh