On the sufficiency and redundancy of pitch for TRP projection

Wesseling, W.; van Son, R.J.J.H.; Pols, L.C.W.

Published in:
Interspeech

Citation for published version (APA):
On the Sufficiency and Redundancy of Pitch for TRP Projection

Wieneke Wesseling, R.J.J.H. van Son, and Louis C.W. Pols

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

Abstract

In two RT experiments, subjects were asked to respond with minimal responses to prerecorded dialogs and impoverished versions of these dialogs, containing either only intonation and pause information, hummed stimuli, or no periodic component at all, whispered stimuli. For the hummed, stimuli, response delays and, especially, variances were higher than the original recordings. Responses to mid-frequency pitch utterance-ends were significantly longer than responses to low pitch utterance-ends, suggesting that our subjects fell back to reacting to pauses when presented with hummed 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. We conclude that just the intonation and pauses of a conversation contain sufficient information for projection of TRPs. However this information is measurably impoverished with respect to original to an extent that increases the “processing” time by 10%. No difference was found between whispered and original speech. This lack of any effect of removing all periodic sound components from the speech signal indicates that in natural speech the pitch signal itself might be redundant for predicting TRPs.

1. Introduction

In order to allow for smooth turn transitions in natural conversations, participants have to be able to predict the end of the previous speaker’s turn [1]. Various information sources are known or suspected to help listeners in determining possible Transition Relevance Places (TRPs), like gaze direction, gestures, intonation, syntactic, and timing information (like speaking rate and pauses). Syntactic completion seems to be the main factor in the turn-taking mechanism. Caspers [2] found that boundary-tones tend to support the grammatical structure. Where pauses coincide with a TRP, low or high tones are used, where pauses do not coincide with syntactic completion, turn-incompleteness is signaled by mid-register tones. Wesseling and Van Son [3] also found boundary tones to help TRP projection.

The present study is a continuation of earlier research and tries to collect evidence about the sufficiency and necessity of pitch in the projection of TRPs using an RT paradigm. Subjects listened to original and manipulated versions of recordings of natural dialogs and were asked to give minimal responses by saying ‘AH’. Their responses are assumed to signal comprehension of at least part of the utterance’s structure and a recognition of a possible end-of-turn (TRP).

To compare processing of the original and manipulated stimuli, a decision-making model by Sigman and Dehaene [4] is used (see fig. 1). In this model, mental decision-making is modeled as a noisy integrator that stochastically accumulates perceptual evidence from the sensory system in time [4, 5], through a perceptual (P), central decision-making (C) and a motor component (M). RTs are the sum of a P + M related deterministic response time, t0, and a C related random walk to a decision threshold, fully determined by an integration time \( \tau = \frac{1}{\alpha} \). Experiments by Sigman and Dehaene [4] showed that the central component \( C \) is responsible for almost all of the variance in response times (RTs). An important property of the model is that 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:

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

\[ (1) \]

2. Materials and Methods

2.1. Speech Materials

All speech materials were obtained from the Spoken Dutch Corpus (CGN) [6, 7], making hand-aligned utterances (“chunks”), word boundary segmentations, transliterations, and phonetic transcriptions available. Based on audio quality and coverage of turn switching categories [3, 9], a stimulus set of 7 switchboard (8 kHz, dual channel telephone recordings) and 10 volunteer home recordings (16 kHz, stereo face-to-face) of 10 minutes each (total duration 165 min.) was selected. The end boundary tones of all utterances were automatically estimated as low, mid or high from the pitch contours [3, 9]. These automatic estimations were then verified by a human lister at SPEX [8].
2.2. Stimulus preparation and presentation

Stimulus selection and preparation was identical to [3, 9]. The 17 dialog recordings were each divided into two overlapping 6 minute stimuli, i.e. the first and last 6 minutes of each dialog. This is the original stimulus set (34 stimuli). Two new stimulus sets were constructed. First, a set of hummed stimuli was created by converting the original stimuli to pitch contours with Praat [10] and having them resynthesized as neutral-vowel speech [3, 9]. This hummed speech contains nothing but the intonation and pause structure of the original speech, i.e. no loudness or spectral information was present. Second, the original stimuli were resynthesized from an LPC analysis using white noise as the sound source. The LPC order was chosen as 8 poles for telephone speech and 16 poles for the home recordings. The amplitude was scaled to prevent clipping. These constitute whispered stimuli as they did not contain a periodic component. However, it must be remembered that both the hummed and whispered speech were artificial and sounded not like natural humming or whispering. The artificially whispered stimuli were still intelligible and did audibly contain non-periodic prosodic cues. All stimuli were upsampled to 16 kHz and 16 bit where necessary.

Stimuli were pseudo-randomized and balanced for presentation. Each of the 32 subjects (with one exception due to an error) heard a different subset and order of 4 original and 4 manipulated dialog fragments of 6 minutes duration in alternating order, starting with an original 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 hummed or whispered fragment from a dialog that was not in the stimulus set.

Table 1: Distribution of Voiced and Early responses over stimulus types by end-tone categories.

<table>
<thead>
<tr>
<th>end-tone</th>
<th>low</th>
<th>mid</th>
<th>high</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced (subjects) Orig. (32)</td>
<td>5240</td>
<td>3652</td>
<td>3476</td>
<td>12368</td>
</tr>
<tr>
<td>Hum. (21)</td>
<td>3926</td>
<td>3164</td>
<td>2663</td>
<td>9753</td>
</tr>
<tr>
<td>Whisp. (11)</td>
<td>1435</td>
<td>1242</td>
<td>1070</td>
<td>3747</td>
</tr>
<tr>
<td>Early (subjects) Orig. (32)</td>
<td>2143</td>
<td>1488</td>
<td>1440</td>
<td>5071</td>
</tr>
<tr>
<td>Hum. (21)</td>
<td>1630</td>
<td>1274</td>
<td>1125</td>
<td>4029</td>
</tr>
<tr>
<td>Whisp. (11)</td>
<td>649</td>
<td>517</td>
<td>479</td>
<td>1645</td>
</tr>
<tr>
<td>Utterances</td>
<td>2430</td>
<td>2543</td>
<td>1697</td>
<td>6670</td>
</tr>
</tbody>
</table>

2.3. Response collection and processing

Stereo stimulus playback and response recording were done on a single laptop [3, 9]. The laryngograph (Laryngograph Ltd, Lx proc) responses were recorded at a 16 kHz sampling rate on one channel, with the fed-back (summed) mono version of the stimulus on the other channel for alignment purposes [3, 9]. 32 Naive, native Dutch subjects participated in the experiment. 21 Subjects heard the original and hummed stimuli and 11 subjects heard the
original and whispered stimuli. Some subjects were paid, only one 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 hummed speech.

Responses were automatically extracted and individually aligned with the original conversations using the re-recorded mono stimulus signal [10, 3, 9]. These are the Voiced responses (see fig. 2). About one third of all Voiced responses were preceded by a characteristic early larynchograph signal indicating muscle activity in the larynx. The start of this signal was automatically segmented and constitutes the Early response (see fig. 2). A minimum difference of 40ms was used to ensure reliable identification.

The RT delay was defined as the time between the start of the Voiced response and the closest utterance end (irrespective of the speaker) within a window of 2 seconds. The relevant utterance had to start at least 0.1 seconds before the start of the response. Furthermore, responses with a duration shorter than 15ms were discarded as spurious. For comparison, Turn Transfer delays in the Spontaneous and Telephone dialogs of the hand aligned part of the Spoken Dutch Corpus were determined, using the same criteria (see fig. 3). The distribution of responses with respect to the intonation boundary tones is given in table 1. At the current level of analysis, we did not distinguish between the prescribed ‘AH’ responses and other, more complex, responses [3, 9].

3. Results

In total, 25.6 hours of responses are used from 32 subjects, containing 25,868 responses that could be attributed to specific utterances in the dialogs (see table 1). In fig. 3, the distribution of response delays is compared to the natural turn start delays for home recordings and telephone speech in the CGN. The distribution of the Early responses and the delay differences between Voiced and Early responses is as expected from [4] (note the 40ms lower cutoff in latter).

The effect of stimulus type and end-tone on RT delays is clearly visible in fig. 4. In general, hummed stimuli induced longer RTs in all types of responses (hummed versus original by subject: $p < 0.001$, ANOVA). However, this stimulus effect was not significant for the low end-tone ($p > 0.1$, t-test) and limited to the mid and high end tones ($p < 0.001$, t-test). There was a difference for Voiced RTs between Whispered and Original stimuli when tested on pooled data ($p < 0.01$, t-test). However, this effect was not corroborated for any of the response types when subject was taken into account (whispered versus original by subject, $p > 0.1$, ANOVA). The RTs were different by end-tone for the hummed Voiced and Early responses (end-tone for hummed by subject: $p < 0.001$, ANOVA) and the Voiced responses to the original stimuli (end-tone for original stimuli by subject $p < 0.001$, ANOVA) and maybe for the Early responses (id., $p < 0.02$, ANOVA). In all these cases, the mid end-tone was different from both the low ($p < 0.001$, t-test) and the high ($p < 0.01$, t-test) end-tones. For the whispered stimuli, there might be an effect of end-tone on the RT difference (end-tone for whispered stimuli by subject, $p < 0.02$, ANOVA). No other effects of end-tone were found (id., $p > 0.1$, ANOVA). So, the presence of a mid end-tone increased the RT in hummed and original stimuli with respect to the other end-tones. No such effect was found for whispered stimuli. Note that neither stimulus type nor end-tone had a statistically significant effect on the interval between Voiced and Early response.

Stimulus type had a strong effect on all response types for hummed versus original stimuli (stimulus type by subject, $p < 0.001$, ANOVA). No effect was found for whispered versus original stimuli (stimulus type, $p > 0.1$, ANOVA).

In all cases, there was a strong effect of subject identity which was expected (subject main effect, $p < 0.001$, ANOVA). There were interactions between stimulus type and end-tone for all responses pooled (stimulus:end-tone, $p < 0.001$, ANOVA) for Voiced responses and for Early responses (stimulus type:end-tone, $p < 0.01$, ANOVA) but not for RT differences (stimulus type:end-tone, $p > 0.1$, ANOVA). There may be such an interaction for the Voiced responses to hummed with respect to the original stimuli (by subject, $p < 0.02$, ANOVA). No such interaction was found for the other responses nor for any responses to whispered stimuli (stimulus type:end-tone by subject, $p > 0.1$, ANOVA).

An important aspect of RT delays is their variance [4], (see fig. 4). The time intervals between the Early and Voiced responses are
clearly less variable than these responses themselves. This shows that these two types of responses are (causally) related. Most likely, the Early response is some kind of preparatory phase of the audible response. For each condition (stimulus type and end-tone), the variance was calculated on a per subject basis. These variances were then entered in a Wilcoxon matched pairs signed ranks (WMPSR) test for main effects and in ANOVA calculations directly as separate measurements to allow estimations of interactions. Although variances are not exactly normally distributed, the large number of subjects (32) gives some assurance for relevance.

The results are rather simple. There is a strong effect of stimulus type with the hummed stimuli inducing a larger variance in both Voiced and Early responses (stimulus type, \( p < 0.001 \), ANOVA; id. WMPSR by subject). There may be an effect of end-tone on the variance of Early responses to hummed stimuli (end-tone, \( p < 0.02 \), ANOVA). No other main effects of stimulus type or intonation nor interaction effects could be found (\( p > 0.05 \), ANOVA). Plainly said, hummed stimuli increase the variance of Voiced and Early responses with respect to original and whispered stimuli. No other factor has any effect.

Using equation 1, it is possible to determine the relative increase in decision time (the C component in fig. 1) due to the manipulations. These relative decision times are plotted in fig. 6. The statistics of these data are the same as those of the variances. It is obvious that hummed stimuli induced increased decision times with respect to the original stimuli while the whispered stimuli either did not differ or might have slightly faster decision times.

4. Discussion and conclusions

The main result of this study is that impoverished hummed conversational speech elicited delayed and more variable responses than the original stimuli. However, our subjects were still able to project TRPs with high reliability using only intonation, without other prosodic or lexical information (see fig. 4). So intonation is clearly a sufficient but impoverished cue for TRP projection when the end-tone is high or low.

No systematic effect could be found for the whispered stimuli. Informal listening to the whispered stimuli showed that they were reasonably intelligible and the prosody and some aspects of intonation were still audible. It is quite possible that the first LPC formant in the resynthesis has often followed the \( F_0 \) which might lead to a pitch perception. Still, it is rather remarkable that so heavily modified stimuli with no periodic component and a decreased intelligibility did not affect the RT in measurable ways. This suggests that the TRP projection cues are very robust with many redundant components.

Contrary to [11], we conclude that intonation is a sufficient cue to project TRPs when the utterance end-tone is low or high, but not when an utterance ends in mid-tone. However, there is no evidence found that pitch is not a completely redundant cue to TRP projection in normal speech.

5. Acknowledgments

The authors would like to thank Dr. Louis ten Bosch and Dr. 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. This project was made possible by grant 276-75-002 of the Netherlands Organization of Scientific Research.

6. References

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