CHAPTER 9

# Prosody in fragmentary grounding

The evaluation of the HIGGINS system presented in Chapter 7 showed that fragmentary grounding utterances often failed, in the sense that the users often did not seem to understand them and act as expected. As already noted, this may be explained partly by the fact that users do not expect such human-like behaviour from dialogue systems, partly because they were used in contexts where a human would not have used them, and partly because the prosodic model was very simplistic and not tested.

In this chapter, the effects of prosodic features on the interpretation of synthesised fragmentary grounding utterances in Swedish dialogue are studied. First, the users' interpretation of such utterances, depending on their prosodic realisation, will be explored in a perception experiment. In a second experiment, we will test the hypothesis that users of spoken dialogue systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behaviour accordingly in a human-computer dialogue setting.

The following scenario, taken from the pedestrian navigation domain used in previous chapters, was used in the first experiment presented in this chapter:

(65) U.1: Further ahead on the right I see a red building.
S.2: *Red (?)*

As discussed in 3.1.4, the evidence of understanding that the system provides in S.2 in this example may have different readings, depending whether we interpret it is as positive or negative evidence, and depending on what level of action the evidence concerns. Three possible readings of S.2 are shown in Table 9.1.

Table 9.1: Different readings of the fragmentary grounding utterance S.2 in example (65).

| Reading | Paraphrase | Evidence of understanding |
|---|---|---|
| ACCEPT | *Ok, red* | Display of understanding. Positive on all levels. |
| CLARIFYUND | *Do you really mean red?* | Clarification request. Positive perception, negative/uncertain understanding. |
| CLARIFYPERC | *Did you say red?* | Clarification request. Positive contact, uncertain perception. |

The reading "positive understanding, negative acceptance" (as discussed in 3.1.4) has not been included here. The reason for this is that it is hard to find examples which may be applied to spoken dialogue systems (at least in the studied domain) where reprise fragments may have such a reading.

## 9.1 Prosody in grounding and requests

Considerable research has been devoted to the study of question intonation in human-human dialogue. However, there has not been much study on the use of different types of interrogative intonation patterns in spoken dialogue systems. Not only does question intonation vary in different languages, but also different types of questions (e.g., wh and yes/no) can result in different intonation patterns (Ladd, 1996).

In very general terms, the most commonly described tonal characteristic for questions is high final pitch and overall higher pitch (Hirst & Cristo, 1998). In many languages, yes/no questions are reported to have a final rise, while wh-questions typically are associated with a final low. In Dutch, for example, van Heuven et al. (1999) have documented a relationship between incidence of final rise and question type, in which wh-questions, yes/no questions and declarative questions obtain an increasing number of final rises, in that order. Wh-questions can, moreover, often be associated with a large number of various contours. Bolinger (1989), for example, presents various contours and combinations of contours which he relates to different meanings in wh-questions in English. One of the meanings most relevant to the present study is what he terms the "reclamatory" question. This is often a wh-question in which the listener has not quite understood the utterance and asks for a repetition or an elaboration. This corresponds to the paraphrase, "What did you mean by red?"

In Swedish, interrogative mode is most often signalled by word order with the finite verb preceding the subject (yes/no questions) or by lexical means (e.g., wh-questions). Question intonation can also be used to convey interrogative mode when the question has declarative word order. This type of echo question is relatively common in Swedish especially in casual questions (Gårding, 1998). Question intonation of this type has been studied in scripted elicited questions and has been primarily described as marked by a raised topline and a widened $F_0$ range on the focal accent (Gårding, 1998).

In recent perception studies, however, House (2003) demonstrated that a raised fundamental frequency ($F_0$) combined with a rightwards focal peak displacement is an effective means of signalling question intonation in Swedish echo questions (declarative word order) when the focal accent is in final position. Furthermore, there was a trading relationship between peak height and peak displacement so that a raised $F_0$ had the same perceptual effect as a peak delay of 50 to 75 ms.

In a study of a corpus of German task-oriented human-human dialogue, Rodriguez & Schlangen (2004) found that the use of intonation seemed to disambiguate clarification types with rising boundary tones used more often to clarify acoustic problems than to clarify reference resolution.

## 9.2 Experiment I: Interpretations

In Experiment I, subjects were asked to listen to short dialogue fragments in Swedish, similar to example (65) above, where the computer is saying a fragmentary grounding utterance after a user turn, and to judge what was actually intended by the computer, based on prosodic features of the utterance.

### 9.2.1 Method

#### 9.2.1.1 Stimuli

Three test words comprising the three colours: blue, red and yellow (*blå, röd, gul*) were synthesized using an experimental version of LUKAS (Filipsson & Bruce, 1997) diphone Swedish male MBROLA voice (Dutoit et al., 1996) implemented as a plug-in to the WaveSurfer speech tool (Sjölander & Beskow, 2000).

For each of the three test words, the intonational contour (i.e., the $F_0$ curve) was manipulated by changing the following parameters: 1) $F_0$ peak POSITION, 2) $F_0$ peak HEIGHT, and 3) Vowel DURATION. Three peak positions were obtained by time-shifting the focal accent peaks in intervals of 100 ms comprising *early*, *mid* and *late* peaks. A *low* peak and a *high* peak set of stimuli were obtained by setting the accent peak at 130 Hz and 160 Hz respectively. Two sets of stimuli durations (*normal* and *long*) were obtained by lengthening the default vowel length by 100 ms. All combinations of three test words and the three parameters gave a total of 36 different stimuli. Six additional stimuli, making a total of 42, were created by using both the early and late peaks in the long duration stimuli which created a double peaked stimulus. A possible late-mid peak was not used in the long duration set since a late rise and fall in the vowel did not sound natural. The stimuli are presented schematically for the word "yellow" in Figure 9.1.

The first turn of the dialogue fragment in example (65) above was recorded for each colour word and concatenated with the synthesized test words, resulting in 42 different dialogue fragments similar to example (65).

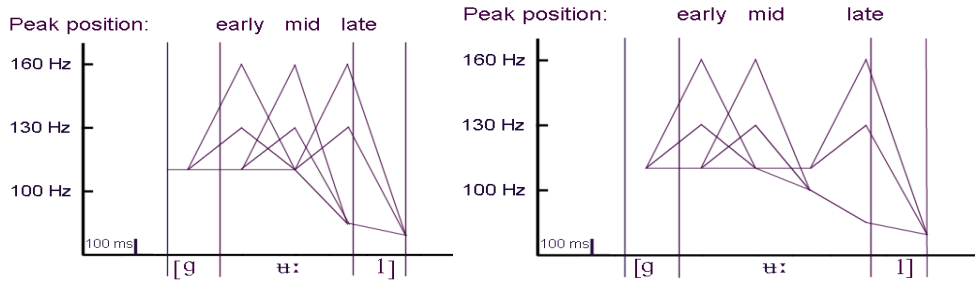


Figure 9.1: Stylized representations of the stimuli "gul" ("yellow"), showing the $F_0$ peak position. The left panel shows normal duration, the right lengthened duration.

#### 9.2.1.2 Experimental design and procedure

The subjects were 8 Swedish speakers in their 20s and 30s (2 women and 6 men, 2 second language speakers and 6 native speakers). All of the subjects had some knowledge of speech technology, although none of them worked with the issues addressed in the experiment.

The subjects were placed in front of a computer monitor in a quiet room. In order to give a sense of the kind of domain envisaged in the experiment, the subjects were shown a video demonstrating a typical dialogue between the HIGGINS spoken dialogue system and a user. The subjects were told that they would listen to 42 similar dialogue fragments containing a user utterance and a system utterance each, and that their task was to judge the meaning of the system utterance by choosing one of three alternatives and to rate their own confidence in that choice. They were also informed that they could only listen to each dialogue fragment once. After the instructions, the test was started and the subjects were left alone for the duration of the experiment.

During the experiment, the subjects were played each of the 42 stimuli once, in random order, on a loudspeaker. After each stimulus, they used the GUI shown in Figure 9.2 to pick a paraphrase for the system utterance and to judge their own confidence in that choice. The different paraphrases corresponded to the ones shown in Table 9.1 above. The subjects could not listen to the stimulus more than once, nor could they skip any stimuli. The total test time was around five to ten minutes per subject.

### 9.2.2 Results

There were no significant differences in the distribution of votes between the different colours ("red", "blue", and "yellow") ($\chi^2$=3.65, dF=4, p>0.05). There were not any significant differences for any of the eight subjects ($\chi^2$=19.00, dF=14, p>0.05), nor had the DURATION parameter any significant effect on the distribution of votes ($\chi^2$=5.72, dF=2, p>0.05).

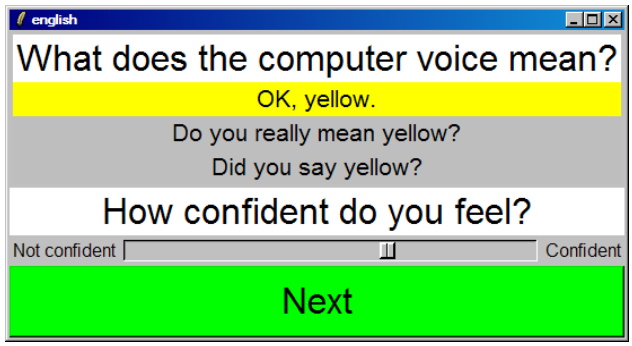


Figure 9.2: The test GUI (translated from Swedish).

Both POSITION and HEIGHT had significant effects on the distribution of votes, which is shown in Table 9.2 ($\chi^2$=70.22, dF=4, p<0.001 resp. $\chi^2$=59.40, dF=2, p<0.001). The interaction of the parameters POSITION and HEIGHT also gave rise to significant effects ($\chi^2$=121.12, dF=10, p<0.001), as shown in the bottom of Table 9.2. Figure 9.3 shows the distribution of votes for the three interpretations as a function of position for both high and low HEIGHT.

Table 9.2: Interpretations that were significantly overrepresented, given the values of the parameters POSITION and HEIGHT, and their interactions. The standardized residuals from the $\chi^2$-test are also shown.

| POSITION | Interpretation | Std. resid. |
|---|---|---|
| early | ACCEPT | 3.1 |
| mid | CLARIFYUND | 4.6 |
| late | CLARIFYPERC | 3.6 |

| HEIGHT | Interpretation | Std. resid. |
|---|---|---|
| high | CLARIFYUND | 3.2 |
| low | ACCEPT | 4.0 |

| POSITION* HEIGHT | Interpretation | Std. resid. |
|---|---|---|
| early*low | ACCEPT | 3.4 |
| mid*low | ACCEPT | 3.4 |
| mid*high | CLARIFYUND | 5.6 |
| late*high | CLARIFYPERC | 4.4 |

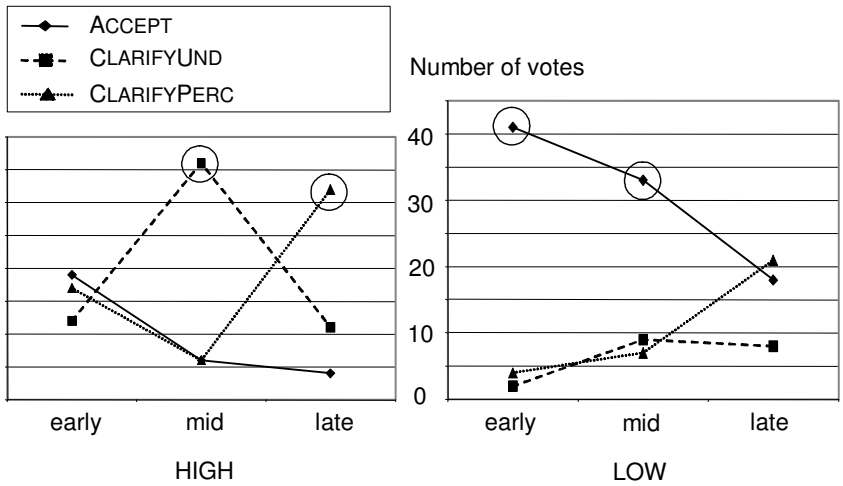


Figure 9.3: The distribution of votes for the three interpretations as a function of position: where HEIGHT is high on the left, and low on the right. The circles mark distributions that are significantly overrepresented.

Weighting the votes with the subjects' own confidence scores only seemed to strengthen the results, so they were not used for further analysis. Results from the double-peak stimuli were generally more complex and are not presented here.

In summary, this first experiment shows that three prototypical intonation patterns can be distinguished, corresponding to the different readings of the fragmentary grounding utterance: an early low $F_0$ peak corresponds to ACCEPT ("ok, red"), a mid high $F_0$ peak corresponds to CLARIFYUND ("do you really mean red?"), and a late high $F_0$ peak corresponds to CLARIFYPERC ("did you say red?").

## 9.3 Experiment II: User responses

In Experiment II, we wanted to test the hypothesis that users of spoken dialogue systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behaviour accordingly in a human-computer dialogue setting.

### 9.3.1 Method

To test our hypothesis, an experiment was designed in which subjects were given the task of classifying colours in a dialogue with a computer. They were told that the computer needed the subject's assistance to build a coherent model of the subject's perception of colours, and

that this was done by having the subject choose among pairs of the colours green, red, blue and yellow when shown various nuances of colours in-between (e.g., purple, turquoise, orange and chartreuse). An example classification task is shown in Figure 9.4.

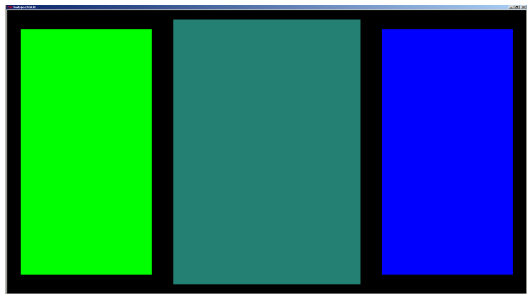

Figure 9.4: An example colour classification task. The computer asked the subject which of the two colours on the flanks was most similar to the one in the middle.

The subjects were also told that the computer may sometimes be confused by the chosen colour or disagree. The test configuration consisted of a computer monitor, loudspeakers, and an open microphone in a quiet room. An extra close-talking microphone was fitted to the subject's collar. An experiment conductor sat behind the subjects during the experiment, facing a different direction. The total test time was around ten minutes per subject.

The experiment used a Wizard-of-Oz set-up: a person sitting in another room – the Wizard – listened to the audio from the close talking microphone (a radio microphone). The Wizard fed the system the correct colours spoken by the subjects, as well as giving a go-ahead signal to the system whenever a system response was appropriate. The subjects were informed about the Wizard setup immediately after the experiment, but not before. Here is an example of a typical dialogue fragment (translated from Swedish):

(66)
S.1: *[presents turquoise flanked by green and blue] which colour is closest to the one in the middle?*
U.2: green
S.3: *green*
U.4: mm
S.5: *okay [presents orange flanked by red and yellow] and this?*
U.6: yellow perhaps

The Wizard had no control over what utterance the system would present next. Instead, this was chosen by the system depending on the context, just as it would be in a system without a Wizard. The grounding fragments (such as S.3 above) came in four flavours: a repetition of the colour with one of the three prototype intonations found in Experiment I (ACCEPT, CLARIFYUND or CLARIFYPERC) or a simple acknowledgement consisting of a synthesized /m/

or /a/ (ACKNOWLEDGE) (Wallers et al., 2006). The system picked these at random so that for every eight colours, each grounding fragment appeared twice.

All system utterances were synthesized using the same voice as the experiment stimuli used in Experiment I. The prosody of each utterance was hand-tuned before synthesis in order to raise the subjects' expectations of the computer's conversational capabilities as much as possible. As seen in the dialogue example above, the computer made heavy use of conversational phenomena such as backchannels and ellipses. There was also a rather high degree of variability in the exact rendition of the system responses. Each of the non-stimuli responses was available in a number of varieties, and the system picked from these at random. Due to the simplicity of the task and the Wizard-of-Oz setup, the system was very responsive, with virtually no delays caused by processing.

The subjects were 10 Swedish speakers between 20 and 65 years old (7 women and 3 men, 1 second language speaker and 9 native speakers). One of the subjects had some knowledge of speech technology, although he did not work with the issues addressed in the experiment.

### 9.3.2 Results

The recorded conversations were automatically segmented into utterances based on the logged timings of the system utterances. User utterances were then defined as the recorded audio segments in-between these. Out of ten subjects, two did not respond at all to any of the grounding utterances (i.e., didn't say anything similar to U.4 in the example above). For the other eight, responses were given in 243 out of 294 possible places. Since the object of our analysis was the subjects' responses, two subjects in their entirety and 51 silent responses distributed over the remaining eight subjects were automatically excluded from analysis.

In almost all cases, subjects simply acknowledged the system's grounding utterance with a brief "yes" or "mm" as the utterance U.4 in the example above. However, when listening to the dialogues, we got the impression that the response time differed. For example, the response time after a grounding fragment with the meaning "do you really mean red?" seemed to be longer than after a fragment meaning "did you say red?".

To test whether the response times were in fact affected by the type of preceding fragment, the time between the end of each system grounding fragment and the user response (in the cases there was a user response) was automatically determined using /nailon/, a software package for extraction of prosodic and other features from speech (Edlund & Heldner, 2006). Silence/speech detection in /nailon/ is based on a fairly simplistic threshold algorithm, and for our purposes, a preset threshold based on the average background noise in the room where the experiment took place was deemed sufficient. The results are shown in Table 9.3. The table shows that, just in line with our intuitions, ACCEPT fragments are followed by the shortest response times, CLARIFYUND the longest, and CLARIFYPERC between these. The differences are statistically significant (one-way within-subjects ANOVA; $F=7.558$; $dF=2$; $p<0.05$).

These response time differences are consistent with a cognitive load perspective that could be applied to the fragment meanings ACCEPT, CLARIFYPERC and CLARIFYUND. To simply acknowledge an acceptance should be the easiest, and it should be nearly as easy, but not quite,

for users to confirm what they have actually said. It should take more time to re-evaluate a decision and insist on the truth value of the utterance after CLARIFYUND. This relationship is nicely reflected in the data.

Table 9.3. Average of subjects' mean response times after grounding fragments.

| Grounding fragment | Response time |
|---|---|
| ACCEPT | 591 ms |
| CLARIFYUND | 976 ms |
| CLARIFYPERC | 634 ms |

## 9.4 Discussion

The results of these studies can be seen in terms of a tentative model for the intonation of fragmentary grounding utterances in Swedish. A low-early peak would function as an ACCEPT statement, a mid-high $F_0$ peak as a CLARIFYUND question, and a late high peak as a CLARIFYPERC question. This would hold for single-syllable accent I words. Accent II words and multi-word fragments are likely to be more complex.

For these single-word grounding utterances, the general division between statement (early, low peak) and question (late, high peak) is consistent with the results obtained for Swedish echo questions (House, 2003) and for German clarification requests (Rodriguez & Schlangen, 2004). However, the further clear division between the interrogative categories CLARIFYUND and CLARIFYPERC is especially noteworthy. This division is related to the timing of the high peak. The high peak is a prerequisite for perceived interrogative intonation in this study, and when the peak is late, resulting in a final rise in the vowel, the pattern signals CLARIFYPERC. This can also be seen as a yes/no question and is consistent with the observation that yes/no questions generally more often have final rising intonation than other types of questions. The high peak in mid position is also perceived as interrogative, but in this case it is the category CLARIFYUND which dominates as is clearly seen in the left panel of Figure 9.3. This category can also been seen as a type of wh-question similar to the "reclamatory" question discussed in Bolinger (1989). For example, the question "do you really mean red?" is similar to (and may have the same effect as) "what do you mean by red?"

Another interesting result is the evidence of an interaction between the parameters peak height and peak position when the peak position is mid. Here, the high-mid peak is perceived as the CLARIFYUND question, while the low-mid peak is perceived as the ACCEPT statement. A similar type of interaction is the trading relationship between peak height and peak displacement in House (2003), where a higher earlier peak has the same perceptual status as a lower later peak.

It is somewhat surprising that the longer duration was not perceived as more interrogative, as this was expected to be interpreted as hesitation and uncertainty. The fact that the majority of the stimuli ended in a very low $F_0$ may have precluded this interpretation.

Although we have not quantified other prosodic differences in the users' responses in Experiment II, we also got the impression that there were subtle differences in, for example, pitch range and intensity. These differences may function as signals of certainty following CLARIFYPERC and signals of insistence or uncertainty following CLARIFYUND. More neutral, unmarked prosody seemed to follow ACCEPT.

### 9.4.1 Future Work

When listening to the resulting dialogs from Experiment II as a whole, the impression is that of a natural dialogue flow with appropriate timing of responses, feedback and turn-taking. To be able to create spoken dialogue systems capable of this kind of dialogue flow, we must be able to both produce and recognise fragmentary grounding utterances and their responses. Further work using more complex fragments and more work on analysing the prosody of user responses is needed.

As grounding fragments become more complex, the interaction between focus and level of action must also be understood. Consider the following examples:

(67) U: I can see a blue brick building.
S: *A **red** brick building?*
U: No, blue

(68) U: I can see a red concrete building.
S: *A red **brick** building?*
U: No, concrete

By using different prosodic realisations in these examples, the system may signal more precisely where the uncertainty is located. This should in turn affect how a potential negation by the user should be integrated in the resulting semantic structure. However, it is also possible that the uncertainty may be associated with the different levels of action dealt with in this chapter. To understand the prosodic interplay between level of action and focus is an interesting challenge.

## 9.5 Summary

In this chapter, two experiments have been presented. In the first experiment, subjects were given the task of listening to short dialogue fragments containing synthesised fragmentary grounding utterances, and choosing the most likely paraphrase. The prosody of these utterances was systematically varied in order to study how the prosodic realisation affects the interpretation of them; whether they signalled acceptance or were interpreted as a clarification request, and which level of action was concerned. The results show that a low early $F_0$ peak is

interpreted as acceptance; a mid high $F_0$ peak is interpreted as a clarification of understanding; and a late high $F_0$ peak is interpreted as a clarification of perception.

The second experiment show that users of spoken dialogue systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behaviour accordingly in a human-computer dialogue setting. The results show that the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.