User Responses to Prosodic Variation in Fragmentary Grounding Utterances in Dialog

Gabriel Skantze, David House & Jens Edlund

Computer Science and Communication, Department of Speech, Music and Hearing
KTH, Stockholm, Sweden
[gabriel,davidh,edlund]@speech.kth.se

Abstract

In a previous study we demonstrated that subjects could use prosodic features (primarily peak height and alignment) to make different interpretations of synthesized fragmentary grounding utterances. In the present study we test the hypothesis that subjects also change their behavior accordingly in a human-computer dialog setting. We report on an experiment in which subjects participate in a color-naming task in a Wizard-of-Oz controlled human-computer dialog in Swedish. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

Index terms: dialog systems, prosody, error handling

1. Introduction

Detecting and recovering from errors is an important issue for spoken dialog systems. A common means for verifying the system's hypothesis of what the user says is *explicit* and *implicit* verification: the system makes a clarification request or repeats what it has understood in order to ground its hypothesis, possibly based on the confidence score of the whole user utterance. Unfortunately, these error handling techniques are often perceived as tedious and unnatural. One of the reasons for this is that they are, in most cases, constructed as full propositions verifying the complete user utterance. In contrast, humans often use fragmentary, elliptical constructions when grounding and clarifying what has been said. If dialog systems considered confidence scores on smaller units than whole utterances, elliptical clarifications could be utilized to focus on problematic fragments and thereby make the dialog more efficient [1]. However, the interpretation of elliptical constructions is often dependent on both context and prosody.

In a previous experiment, the effects of prosodic features on the interpretation of fragmentary grounding utterances were investigated [2]. Using a listener test paradigm, subjects were asked to listen to short dialog fragments in Swedish where the computer replies after a user turn, and to judge what was actually intended by the computer. The results indicated that listeners could use prosodic features (primarily peak height and alignment) to make different interpretations of the elliptical constructions. In the present paper, we look at actual user responses to fragmentary grounding utterances in a human-computer dialog setting.

Both studies are part of the research in the HIGGINS spoken dialog system [3], and will be used in further dialog studies. The primary domain of HIGGINS is pedestrian navigation. In this domain, the system does not have access to the user's position when guiding the user, but has to rely on the user's descriptions of the environment. Table 1 exemplifies how fragmentary grounding can make the dialog more natural and efficient in this domain, compared to explicit or implicit verification of whole utterances [1]. This is also the scenario used in the previous experiment [2].

Table 1. *Example use of fragmentary grounding (translated from Swedish)*

User	Further ahead on the right I see a red building
System	Red (?)

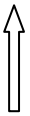
2. Grounding and prosody

Grounding is the process by which speakers establish information as part of common ground [4]. They do this by giving positive and negative evidence of their understanding of what their interlocutor says. Positive evidence can be given in different ways, for example by asserting understanding with an acknowledgement ("okey", "mhm"), by following up with a relevant next contribution, or by displaying a fragment of what has been understood ("red").

A clarification request is a means to give both positive and negative evidence – showing what has been understood as well as what is needed for complete understanding. Clarification requests may have different forms and different functions. In a study of the British National Corpus (BNC), Purver et al. [5] studied the form and function of clarification requests. 45% of the clarification requests were fragmentary or elliptical. The form of clarification requests studied in this paper, as shown in Table 1, are called reprise fragments in their study. Such reprise fragments may have the same lexical form as a fragmentary display of understanding ("red").

We will use a distinction made by both Clark [4] and Allwood et al. [6] in order to classify possible readings of such fragmentary grounding utterances. They suggest four levels of action that take place when speaker S is trying to say something to hearer H:

- Acceptance: H accepts what S says.
- Understanding: H understands what S means.
- Perception: H hears what S says.
- Contact: H hears that S speaks.



For successful communication to take place, communication must succeed on all these levels. The order of the levels is important; to succeed on one level, all the other levels be-

low it must be completed. Also, if positive evidence is given on one level, all the other levels below it are presumed to have succeeded. When making a clarification request, the speaker is signaling failure or uncertainty on one level and success on the levels below it. When displaying understanding, success on all levels may be signaled.

According to the scheme given above, the fragmentary grounding utterance in Table 1 may have three different readings:

- Ok, red. (positive on all levels)
- Do you really mean red? What do you mean by red? (positive perception, negative/uncertain understanding)
- Did you say red? (positive contact, uncertain perception)

In terms of prosody, we would like to find a relationship between prosodic realization and the three different readings. The first reading would normally be categorized as a prosodic statement, while the second and third readings would be prosodic questions. In the context of this work we are interested not only in differentiating between question and statement, but also between two types of questions.

In very general terms, the most commonly described prosodic characteristic for questions is high final pitch and overall higher pitch [7]. 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. [8] 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 [9], 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?" Kohler [10] suggests that questions with rising intonation generally have a wider interpretation, while questions with falling intonation are associated with routine replies.

In Swedish, interrogative mode is most often signaled 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 [11]. 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 [11]. In recent perception studies, House [12] also demonstrated that a raised fundamental frequency (F_0) combined with a rightwards focal peak displacement is an effective means of signaling question intonation in Swedish echo questions (declarative word order) when the focal accent is in final position.

In a study of a corpus of German task-oriented human-human dialog, Rodriguez & Schlangen [13] 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.

3. Previous experiment

In the previous experiment [2], three test words comprising the three colors: blue, red and yellow (*blå, röd, gul*) were synthesized using an experimental version of LUKAS diphone Swedish male MBROLA voice [14], implemented as a plug-in to the WaveSurfer speech tool [15]. 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 duration were used differing by 100 ms. The combinations peak position, peak height, duration and color resulted in 36 different stimuli. Subjects were asked to listen to short human-computer dialog fragments in Swedish (as shown in Table 1) in which a synthetic voice was making a fragmentary grounding after a user turn. The subjects were asked to judge what was actually intended by the computer by choosing between the paraphrases shown in Table 2. The results show that an early, low F_0 peak signals acceptance (display of understanding), that a late, high peak is perceived as a request for clarification of what was said, and that a mid, high peak is perceived as a request for clarification of the meaning of what was said. Duration had no significant effect. The results are summarized in Table 2 and demonstrate the relationship between prosodic realization and the three different readings.

Table 2. *Prototype stimuli found in the previous experiment.*

Position	**Height**	**Paraphrase**	**Class**
Early	Low	Ok, red	ACCEPT
Mid	High	Do you really mean red?	CLARIFYUND
Late	High	Did you say red?	CLARIFYPERC

In the present study, we want to test the hypothesis that users of spoken dialog systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behavior accordingly in a human-computer dialog setting.

4. Method

To test our hypothesis, an experiment was designed in which subjects were given the task of classifying colors in a dialog 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 colors, and that this was done by having the subject choose among pairs of the colors green, red, blue and yellow when shown various nuances of colors in-between (e.g. purple, turquoise, orange and chartreuse). They were also told that the computer may sometimes be confused by the chosen color 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 colors 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. A typical dialog is shown in Table 3.

Table 3. *A typical dialog fragment from the experiment (translated from Swedish).*

S1-1a	[presents purple flanked by red and blue]
S1-1b	what color is this
U1-1	red
S1-2	red (ACCEPT/CLARIFYUND/CLARIFYPERC) *or* mm (ACKNOWLEDGE)
U1-2	mm
S1-3	okay
S2-1a	[presents orange flanked by red and yellow]
S2-1b	and this
U2-1	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 (S1-2 in Table 3) came in four flavors: a repetition of the color with one of the three intonations described in Table 3 (ACCEPT, CLARIFYUND or CLARIFYPERC) or a simple acknowledgement consisting of a synthesized /m/ or /a/ (ACKNOWLEDGE) [16]. The system picked these at random so that for every eight colors, each grounding fragment appeared twice.

All system utterances were synthesized using the same voice as the experiment stimuli [14]. Their prosody 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 dialog example, 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. In general, 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.

5. 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 gaps in-between these. Out of ten subjects, two did not respond at all to any of the grounding utterances. 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

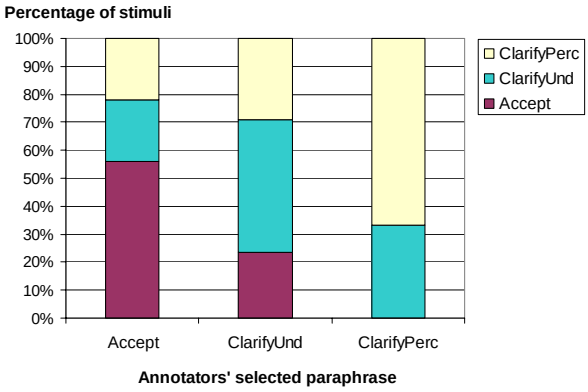


Figure 1. *The percentage of preceding system utterance types for the classifications on which the annotators agreed.*

over the remaining eight subjects were automatically excluded from analysis.

User responses to fragmentary grounding utterances from the system were annotated with one of the labels ACKNOWLEDGE, ACCEPT, CLARIFYUND or CLARIFYPERC, reflecting the preceding utterance type.

In almost all cases subjects simply acknowledged the system utterance with a brief "yes" or "mm" as the example U1-2 in Table 3. However, we felt that there were some differences in the way these responses were realized. To find out whether these differences were dependent on the preceding system utterance type, the user responses were cut out and labeled by two annotators. To aid the annotation, three full paraphrases of the preceding system utterance, according to Table 2, were recorded. The annotators could listen to each of the user responses concatenated with the paraphrases, and select the resulting dialog fragment that sounded most plausible, or decide that it was impossible to choose one of them. The result is a categorization showing what system utterance the annotators found to be the most plausible to precede the annotated subject response. The task is inherently difficult – sometimes the necessary information simply isn't present in the subjects' responses – and the annotators only agreed on a most plausible response in about 50% of the cases. The percentage of preceding system utterance types for the classifications on which the annotators agreed is shown in Figure 1.

The figure shows that responses to ACCEPT fragments are significantly more common in the group of stimuli for which the annotators had agreed on the ACCEPT paraphrase. In the same way, CLARIFYUND, and CLARIFYPERC responses are significantly overrepresented in their respective classification groups (χ^2=19.51; dF=4; p<0.001). This shows that the users' responses are somehow affected by the prosody of the preceding fragmentary grounding utterance, in line with our hypothesis.

The annotators felt that the most important cue for their classifications was the user response time after the paraphrase. For example, a long pause after the question "did you say red?" sounds implausible, but not after "do you really mean 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 automati-

cally determined using /nailon/ [17], a software package for extraction of prosodic and other features from speech. 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 4.

Table 4. *Average of subjects' mean response times after grounding fragments.*

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

The table shows that, just in line with the annotators' 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; $F=7.558$; $dF=2$; $p<0.05$).

6. Conclusions and discussion

In the present study, we have shown that users of spoken dialog systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behavior accordingly in a human-computer dialog setting. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

The response time differences found in the data 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 reevaluate a decision and insist on the truth value of the utterance after CLARIFIYUND. This relationship is nicely reflected in the data.

Although we have not quantified other prosodic differences in the users' responses, the annotators felt that there were subtle differences in e.g. pitch range and intensity which may function as signals of certainty following CLARIFYPERC and signals of insistence or uncertainty following CLARIFYUND. More neutral, unmarked prosody seemed to follow ACCEPT. When listening to the resulting dialogs as a whole, the impression is that of a natural dialog flow with appropriate timing of responses, feedback and turntaking. To be able to create spoken dialog systems capable of this kind of dialog flow, we must be able to both produce and recognize fragmentary grounding utterances and their responses. Further work using more complex fragments and more work on analyzing the prosody of user responses is needed.

7. Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations and was also supported by the EU project CHIL (IP506909).

8. References

[1] Skantze, G. (2005). Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proceedings of SIGdial*. Lisbon, Portugal.

[2] Edlund, J., House, D., & Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech 2005*. Lisbon, Portugal.

[3] Edlund, J., Skantze, G., & Carlson, R. (2004). Higgins - a spoken dialogue system for investigating error handling techniques. In *Proceedings of ICSLP*, 229-231.

[4] Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

[5] Purver, M., Ginzburg, J., & Healey, P. (2001). On the means for clarification in dialogue. In *Proceedings of SIGDial*.

[6] Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9, 1-26.

[7] Hirst, D. & Di Cristo, A. (1998). A survey of intonation systems. In D. Hirst and A. Di Cristo (eds.) *Intonation Systems*. Cambridge: Cambridge University Press, 1-45.

[8] Heuven, V. J. van, Hann, J., & Kirsner, R. S. (1999). Phonetic correlates of sentence type in Dutch: Statement, question and command. In *Proceedings of ESCA International Workshop on Dialogue and Prosody*, 35-40.

[9] Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. London: Edward Arnold.

[10] Kohler, K. J. (2004). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In G. Fant, H. Fujisaki, J. Cao, & Y. Xu (Eds.) *From traditional phonology to modern speech processing*, 205-214. Beijing: Foreign Language Teaching and Research Press.

[11] Gårding, E. (1998). Intonation in Swedish, In D. Hirst and A. Di Cristo (eds.) *Intonation Systems*. Cambridge: Cambridge University Press, 112-130.

[12] House, D. (2003). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. In *Proc 15th ICPhS*, Barcelona, 755-758

[13] Rodriguez, K. J. & Schlangen, D. (2004). Form, intonation and function of clarification requests in German task oriented spoken dialogues. In *Proceedings of Catalog '04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04)*, Barcelona, Spain.

[14] Filipsson, M. & Bruce, G. (1997). LUKAS - a preliminary report on a new Swedish speech synthesis. *Working Papers 46*, Department of Linguistics and Phonetics, Lund University.

[15] Sjölander, K. & Beskow, J. (2000). WaveSurfer - a public domain speech tool, In *Proceedings of ICSLP 2000*, 4, 464-467, Beijing, China.

[16] Wallers, Å., Edlund, J., & Skantze, G. (2006). The effect of prosodic features on the interpretation of synthesised backchannels. In *Proceeding of Perception and Interactive Technologies*. Kloster Irsee, Germany.

[17] Edlund, J. & Heldner, M. (2005). Exploring Prosody in Interaction Control. *Phonetica*, 62(2-4), 215-226.