The Effects of Prosodic Features on the Interpretation of Clarification Ellipses

Jens Edlund, David House and Gabriel Skantze[*]

Centre for Speech Technology, KTH, Sweden
{edlund, davidh, gabriel}@speech.kth.se

Abstract

In this paper, the effects of prosodic features on the interpretation of elliptical clarification requests in dialogue are studied. An experiment is presented where subjects were asked to listen to short human-computer dialogue fragments in Swedish, where a synthetic voice was making an elliptical clarification after a user turn. The prosodic features of the synthetic voice were systematically varied, and the subjects were asked to judge what was actually intended by the computer. The results show that an early low F_0 peak signals acceptance, 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. The study can be seen as the beginnings of a tentative model for intonation of clarification ellipses in Swedish, which can be implemented and tested in spoken dialogue systems.

1. Introduction

Detecting and recovering from errors is an important issue for spoken dialogue 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, 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 clarifying what has been said. As shown by Purver et al. [1], 45% of the clarification requests in the British National Corpus (BNC) were elliptical. If dialogue systems considered confidence scores on smaller units than whole utterances, elliptical clarifications could be utilized to focus on problematic fragments and thereby make the dialogue more efficient [2]. However, the interpretation of elliptical constructions is often highly dependent on both context and prosody, and the prosody of clarification requests has not been studied to a great extent.

In this paper, the effects of prosodic features on the interpretation of elliptical clarification requests in dialogue are studied. An experiment is presented where subjects were asked to listen to short dialogue fragments in Swedish where the computer is making an elliptical clarification after a user turn, and to judge what was actually intended by the computer, based on prosodic features of the clarification. The study is part of the research in the HIGGINS spoken dialogue system [3], and will be used in further dialogue studies. The primary domain of HIGGINS is pedestrian navigation, and in the example scenario shown in Table 1, we see that the system does not have access to the user's position, but has to rely on the user's descriptions of the environment.

Table 1: Example scenario in the HIGGINS domain (translated from Swedish)

User	I want to go to an ATM.
System	OK, where are you?
User	I'm standing between an orange building and a brick building.
System	OK, is the brick building a three storey building?
User	Yes.

Clarification ellipsis could be very useful in this domain. Table 2 shows the scenario that is used in the experiment presented in this paper.

Table 2: Example use of clarification ellipsis (translated from Swedish)

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

1.1. Clarification

Clarification is part of a process called grounding [4] or interactive communication management [5]. In this process, speakers give positive and negative evidence or feedback of their understanding of what the interlocutor says. A clarification may often give both positive and negative evidence – showing what has been understood as well as what is needed for complete understanding.

Clarification requests may have both different forms and different readings (i.e. functions). In a study of the BNC, Purver et al. [1] studied the form and function of clarification requests. According to their scheme, the form of clarification ellipses studied in this paper, as exemplified in Table 2, is called reprise fragments.

We will use a distinction made by both Clark [4] and Allwood et al. [5] in order to classify possible readings of reprise fragments. 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-

[*] Names in alphabetical order

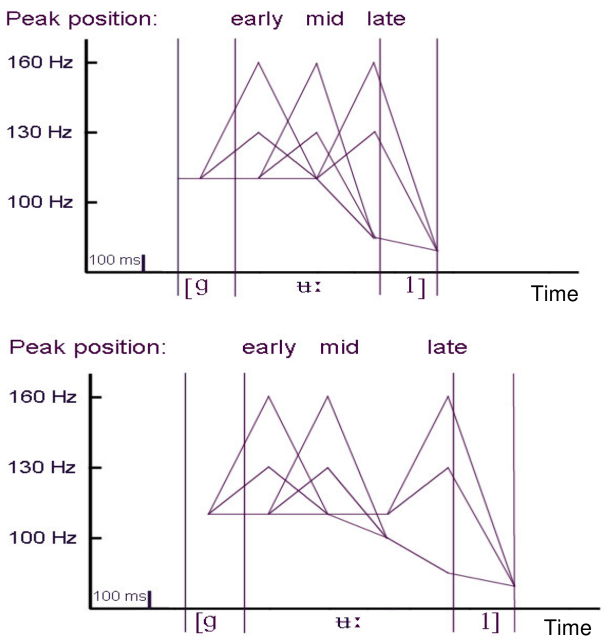


Figure 1. Stylized representations of the stimuli "gul" ("yellow"), showing the F_0 peak position. The top panel shows normal duration, the bottom lengthened duration.

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.

Other classifications of clarification readings have been made. In [6] a more fine-grained analysis of the understanding level is given. In [7], a distinction is made between what is called the "clausal reading" and the "constituent reading" of clarification ellipsis. Using the scheme above, the clausal reading could be described as a signal of positive contact and negative perception, and the constituent reading as a signal of positive perception and negative understanding.

According to the scheme given above, the reprise fragment in Table 2 may have three different readings:

- Ok, red. (No clarification request; 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)

The reading "positive understanding, negative acceptance" 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, where reprise fragments may have such a reading.

1.2. Prosody

In spite of the fact that considerable research has been devoted to the study of question intonation, the use of different types of interrogative intonation patterns has not been routinely represented 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 kinds of question intonation [8].

In very general terms, the most commonly described tonal characteristic for questions is high final pitch and overall higher pitch [9]. 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. [10] 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 [11], 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 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 [12]. 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 [12].

In recent perception studies, however, House [13], 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. 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 [14] 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.

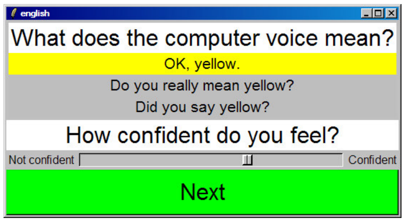
2. Method

In the following experiment, we explore the relationship between prosodic features and the interpretation of single-word clarification ellipses.

2.1. Stimuli

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 [15], implemented as a plug-in to the WaveSurfer speech tool [16].

For each of the three test words the following prosodic parameters were manipulated: 1) Peak POSITION, 2) 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 dura-



english

What does the computer voice mean?

OK, yellow.

Do you really mean yellow?

Did you say yellow?

How confident do you feel?

Not confident | Confident

Next

Figure 2: The test GUI (translated from Swedish)

tion stimuli which created a double peaked stimuli. 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 1.

The first turn of the dialogue fragment in Table 2 was recorded for each color word and concatenated with the synthesized test words, resulting in 42 different dialogue fragments similar to the one in Table 2.

2.2. 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 have some knowledge of speech technology, although none of them work 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. A transcription of part of the dialogue shown in the video is presented in Table 1. 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 Fostex loudspeaker. After each stimulus, they used the GUI shown in Figure 2 to pick a paraphrase for the system utterance and to judge their own confidence in that choice. The different paraphrases were (where X was the color used in the fragment):

- ACCEPT: Ok, X
- CLARIFYUNDERSTANDING: Do you really mean X?
- CLARIFYPERCEPTION: Did you say X?

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.

3. Results

There were no significant differences in the distribution of votes between the different colors (“red”, “blue”, and “yellow”) (χ^2=3.65, dF=4, p>0.05), nor were there any significant differences for any of the eight subjects (χ^2=19.00, dF=14, p>0.05). Neither had the DURATION parameter any significant effect on the distribution of votes (χ^2=5.72, dF=2, p>0.05).

Table 3: Interpretations that were significantly over-represented, given the values of the parameters POSITION and HEIGHT, and their interactions. The standardized residuals from the χ^2-test are also shown.

POSITION	**Interpretation**	**Std. resid.**
Early	ACCEPT	3.1
Mid	CLARIFYUNDERSTANDING	4.6
Late	CLARIFYPERCEPTION	3.6
HEIGHT	**Interpretation**	**Std. resid.**
High	CLARIFYUNDERSTANDING	3.2
Low	ACCEPT	4.0
POSITION* HEIGHT	**Interpretation**	**Std. resid.**
Early*Low	ACCEPT	3.4
Mid*Low	ACCEPT	3.4
Mid*High	CLARIFYUNDERSTANDING	5.6
Late*High	CLARIFYPERCEPTION	4.4

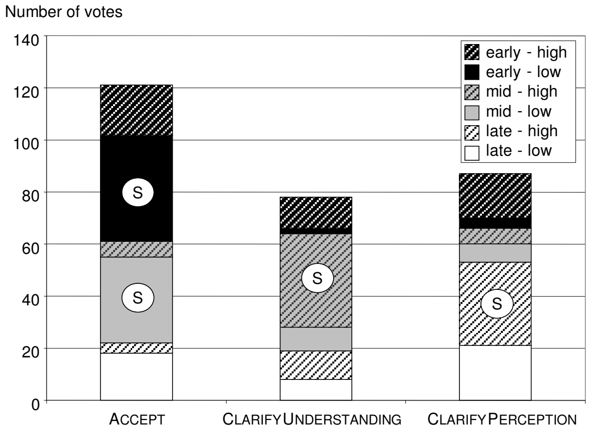


Figure 3: The distribution of votes for all combinations of position and height, split over interpretation. “S” mark distributions that are significantly overrepresented.

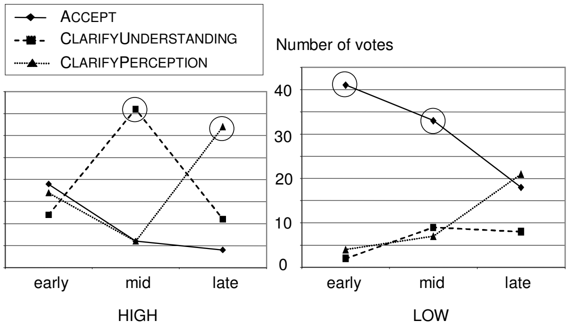


Figure 4: The distribution of votes for the three interpretations as a function of position: where HEIGHT is “high” on the left, and “low” on the right. The circles mark distributions that are significantly overrepresented.

Both POSITION and HEIGHT had significant effects on the distribution of votes, which is shown in Table 3 (χ^2=70.22, dF=4, p<0.001 resp. χ^2=59.40, dF=2, p<0.001). The interaction of the parameters POSITION and HEIGHT also gave rise to significant effects (χ^2=121.12, dF=10, p<0.001), as shown in the bottom of Table 3. Figure 3 shows the distribution of votes for all combinations of position and height, split over interpretation. Figure 4 shows the distribution of votes for the three interpretations as a function of position for both high and low HEIGHT.

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.

4. Discussion

The most interesting result in this experiment from both a spoken dialogue system perspective and a prosody modeling framework concerns the strong relationship between intonational form and meaning. For these single-word utterances used as clarification ellipses, the general division between statement (early, low peak) and question (late, high peak) is consistent with the results obtained for Swedish echo questions in [13] and for German clarification requests in [14]. However, the further clear division between the interrogative categories CLARIFYUNDERSTANDING and CLARIFYPERCEPTION 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 CLARIFYPERCEPTION. 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 CLARIFYUNDERSTANDING which dominates as is clearly seen in the left panel of Figure 4. This category can also been seen as a type of wh-question similar to the "reclamatory" question discussed in [11].

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 CLARIFYUNDERSTANDING 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 [13] 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.

5. Conclusions and future work

The results of this preliminary study can be seen in terms of a tentative model for the intonation of clarification ellipses in Swedish. A low-early peak would function as an ACCEPT statement, a mid-high peak as a CLARIFYUNDERSTANDING question, and a late high peak as a CLARIFYPERCEPTION question. This would hold for single-syllable accent I words. Accent II words may be more complex. We intend to test this model and extend this research in two ways. By implementing these prototypical patterns in the Higgins dialogue system, we will study responses of actual users to the different prototypes. We also plan to study these types of clarification ellipses in a database of Swedish human-human dialogue.

6. 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).

7. References

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

[2] Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the AAAI spring symposium on natural language generation in spoken and written dialogue.*

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

[6] Schlangen, D. (2004). Causes and strategies for requesting clarification in dialogue. In *Proceedings of SIGDial.*

[7] Ginzburg, J. & Cooper, R. (2001). Resolving ellipsis in clarification. In *Proceedings of the 39th meeting of the ACL.*

[8] Ladd, D. R. (1996). *Intonation phonology*. Cambridge: Cambridge University Press.

[9] 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.

[10] 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.

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

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

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

[14] 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.

[15] 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.

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