

Prosodic Features in the Perception of Clarification Ellipses

Jens Edlund, David House, and Gabriel Skantze
Centre for Speech Technology, KTH, Sweden

Abstract

We present an experiment where subjects were asked to listen to Swedish human-computer dialogue fragments where a synthetic voice makes an elliptical clarification after a user turn. The prosodic features of the synthetic voice were systematically varied, and subjects were asked to judge the computer's actual intention. 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.

Introduction

Detection of and recovery from errors is important for spoken dialogue systems. To this effect, system hypotheses are often verified explicitly or implicitly: the system makes a clarification request or repeats what it has heard. These error handling techniques are often perceived as tedious, and one of the reasons for this is that they are often constructed as full propositions, verifying the complete user utterance. In contrast, humans often use short elliptical constructions for clarification – Purver et al. (2001) show that 45% of the clarification requests in British National Corpus (BNC) are elliptical. A dialogue system using word level confidence scores could use elliptical clarifications to focus on problematic fragments, making the dialogue more efficient (Gabsdil, 2003). However, the interpretation of ellipses is often dependent both on context and on prosody, and the prosody of clarification requests has not been greatly studied.

We present an experiment in which subjects were asked to listen to Swedish dialogue fragments where the computer makes elliptical clarifications after user turns, and to judge what was actually intended by the computer. The study is connected to the HIGGINS spoken dialogue system (Edlund et al., 2004). The primary

domain of HIGGINS is pedestrian navigation, as seen in Table 1.

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.

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	[...] on the right I see a red building.
System	Red (?)

Clarification

Clarification is part of a process called grounding (Clark, 1996) or interactive communication management (Allwood et al., 1992). 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. (2001) 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 (1996) and Allwood et al. (1992) 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 below 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 Schlangen (2004) a more finegrained analysis of the understanding level is given. In Ginzburg & Cooper (2001), 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.

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 (Ladd, 1996). In very general terms, the most commonly described tonal characteristic for questions is high final pitch and overall higher pitch. In many languages, yes/no questions are reported to have a final rise, while wh-questions typically are associated with a final low. 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, question intonation has been primarily described as marked by a raised top-line and a widened F0 range on the focal accent (Gårding, 1998). In recent perception studies, however, House (2003), demonstrated that a raised fundamental frequency (F0) 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 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.

Metod

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 (Filipsson & Bruce, 1997), implemented as a plug-in to the WaveSurfer speech tool (Sjölander & Beskow, 2000).

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 duration 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.

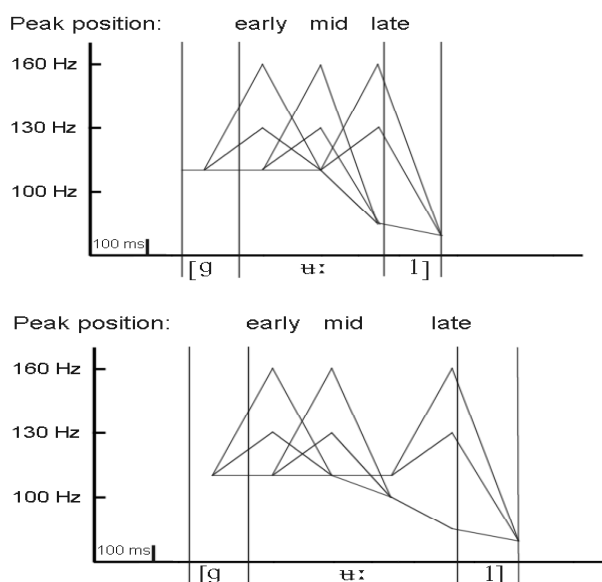


Figure 1. Stylized representations of the stimuli “gul” (“yellow”), showing the F0 peak position. The top panel shows normal duration, the bottom lengthened duration.

The subjects were 8 Swedish speakers, all of which have some experience of speech technology, although none of them are involved in this research.

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. They were informed that they could only listen to each stimulus once.

During the experiment, the subjects were played each of the 42 stimuli once, in random order. After each stimulus, they chose a paraphrase for the system utterance. The different paraphrases were (X stands for a color):

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

Results

There were no significant differences in the distribution of votes between the different colors (“red”, “blue”, and “yellow”) ($\chi^2=3.65$, $dF=4$,

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

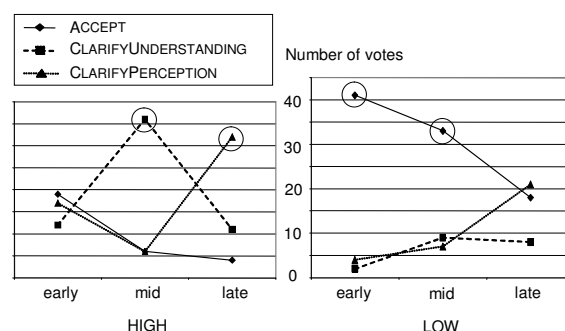


Figure 2: The distribution of votes for the three interpretations as a function of position: where HEIGHT is “high” on the left, and “low” on the right. The circles mark distributions that are significantly overrepresented.

$p>0.05$), nor were there any significant differences for any of the eight subjects ($\chi^2=19.00$, $dF=14$, $p>0.05$). Neither had the DURATION parameter any significant effect on the distribution of votes ($\chi^2=5.72$, $dF=2$, $p>0.05$). Both POSITION and HEIGHT had significant effects on the distribution of votes, which is shown in Table 3 ($\chi^2=70.22$, $dF=4$, $p<0.001$ 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 3. Figure 2 shows the distribution of votes for the three interpretations as a function of position for both high and low HEIGHT. Results from the double-peak stimuli were generally more complex and are not presented here.

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 utter-

ances 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 (House, 2003) and for German clarification requests in (Rodriguez & Schlangen, 2004). 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 2. This category can also be seen as a type of wh-question similar to the “reclamatory” question discussed in (Bolinger, 1989).

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 CLARIFY-UNDERSTANDING 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.

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).

References

- Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9, 1-26.
- Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. London: Edward Arnold.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Edlund, J., Skantze, G., & Carlson, R. (2004). Higgins – a spoken dialogue system for investigating error handling techniques. In *Proceedings of ICSLP*, 229-231.
- 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.
- Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the AAAI spring symposium on natural language generation in spoken and written dialogue*.
- Gårding, E. (1998). Intonation in Swedish, In D. Hirst and A. Di Cristo (eds.) *Intonation Systems*. Cambridge: Cambridge University Press, 112-130.
- Ginzburg, J. & Cooper, R. (2001). Resolving ellipsis in clarification. In *Proceedings of the 39th meeting of the ACL*.
- House, D. (2003). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. In *Proc 15th ICPhS*, Barcelona, 755-758
- Ladd, D. R. (1996). *Intonation phonology*. Cambridge: Cambridge University Press.
- Purver, M., Ginzburg, J., & Healey, P. (2001). On the means for clarification in dialogue. In *Proceedings of SIGDial*.
- 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.
- Schlangen, D. (2004). Causes and strategies for requesting clarification in dialogue. In *Proceedings of SIGDial*.
- Sjölander, K. & Beskow, J. (2000). WaveSurfer – a public domain speech tool, In *Proceedings of ICSLP 2000*, 4, 464-467, Beijing, China.