Phrase-final rises as a prosodic feature in wh-questions in Swedish human–machine dialogue

David House *

Department of Speech, Music and Hearing, Centre for Speech Technology, KTH, Lindstedtsvägen 24, 100 44 Stockholm, Sweden

Received 13 September 2004; received in revised form 1 March 2005; accepted 25 March 2005

Abstract

This paper examines the extent to which optional final rises occur in a set of 200 wh-questions extracted from a large corpus of computer-directed spontaneous speech in Swedish and discusses the function these rises may have in signalling dialogue acts and speaker attitude over and beyond an information question. Final rises occurred in 22% of the utterances, primarily in conjunction with final focal accent. Children exhibited the largest percentage of final rises (32%), with women second (27%) and men lowest (17%). The distribution of the rises in the material is examined and evidence relating to the final rise as a signal of a social interaction oriented dialogue act is gathered from the distribution. Two separate perception tests were carried out to test the hypothesis that high and late focal accent peaks in a wh-question are perceived as friendlier and more socially interested than low and early peaks. Generally, the results were consistent with these hypotheses when the late peaks were in phrase-final position. Finally, the results of this study are discussed in terms of pragmatic and attitudinal meanings and biological codes.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Intonation; Spontaneous speech; Human–machine dialogue; Question intonation; Prosodic features

1. Introduction

Asking questions to obtain information is one of the most important functions of human–machine interaction in current spoken dialogue systems. This is obvious from both a user and a

system perspective. For the user, asking for information is a natural and easy way of accessing data, and the ability of the system to handle questions is indeed fundamental if the user is to be given the initiative. For the system, generating appropriate questions is necessary for guiding the user, for maintaining the dialogue and for error handling. Question management is crucial not only for accessing information, but also for making possible the smooth flow of a spoken dialogue where

* Tel.: +46 8 790 7565; fax: +46 8 790 7854.

E-mail address: davidh@speech.kth.se

asking and answering questions can be directly related to the state of the dialogue itself.

The prosodic features of the ends of phrases can reveal characteristics of questions which are important for the dialogue flow. For example, there has been recent interest in the automatic analysis of phrase-final tones and short utterances with the objective of categorizing and extracting dialogue acts such as agreement, acknowledgement, backchannels, turntaking and speaker attitude (see e.g. Ferrer et al., 2002; Ishi et al., 2003; Caspers, 2003; Bhagat et al., 2003). Such information is also highly useful in building and improving spoken dialogue systems using animated agents and spoken human–system interaction. A greater understanding of different variations in question intonation and the meanings associated with these variations could prove valuable for recognizing user intentions and for improving expressiveness and appropriateness of system output through the speech of an animated agent (Syrdal and Jilka, 2004; Jilka et al., 2003).

The signaling of interrogative mode in speech through intonation is a topic which has long attracted interest from intonation researchers. Question intonation, however, has remained somewhat elusive in both descriptive phonetics and intonation models. Not only does question intonation vary in different languages but also different types of questions (e.g. *wh*, *yes/no* or *echo* questions) 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 (Hirst and Di 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. *Wh*-questions can, however, often be associated with a large number of various contours (Cruttenden, 1986). Bolinger (1989), for example, presents various contours and combinations of contours which he relates to different meanings in *wh*-questions. In Dutch, 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. There are also languages which

have no morphosyntactic differences between *yes/no* questions and statements and therefore make use of intonation to mark questions. For example in Neapolitan Italian (D’Imperio and House, 1997), a late time alignment of a final accent has been shown to play a decisive role in the perception of interrogative mode.

In a recent study of pitch patterns in German word order (*yes/no*) questions and question-word questions (*wh*-questions) taken from a large corpus of spontaneous speech, Kohler (2004), tested the hypothesis that final rising and final falling intonation occur in both syntactic structures. He found that both pitch patterns occur in both structures, but that word-order questions had predominantly rising patterns (with more high-rising than low-rising patterns) while question word questions had mostly falling patterns (but a substantial number of low-rising patterns). By resynthesising complementary pitch patterns in the two structures, Kohler establishes that in “both syntactic structures, rising pitch expresses friendliness, interest and openness towards the addressee, while falling pitch focuses on routine, lack of interest and categoricalness” (p. 207). He further proposes that the difference in distribution between the two structures can be explained by the different semantic and pragmatic functions of the two structures. The question-word question is generally information and fact oriented while the *yes/no* question is generally asking for a decision from the addressee and is thus more addressee oriented.

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 (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 *F0* range on the focal accent (Gårding, 1979, 1998; Bredvad-Jensen, 1984). In Gårding (1979), it is proposed that this description also applies to *yes/no* questions with inverted word order, with the difference that the topline does not

need to be raised as much as in a question with declarative word order. Most of the work cited here has been done on read or elicited speech. However, studies of spontaneous speech in Swedish have demonstrated that in general prosodic differences between spontaneous speech and read speech are not necessarily fundamental (Bruce and Touati, 1992).

An optional terminal rise has also been described for questions in Swedish (Hadding-Koch and Studdert-Kennedy, 1964), but the time alignment of the focal accent rise has not generally been associated with question intonation. Instead, a rightward shift of the focal accent peak has been associated with lending prominence to given domain-specific information in a dialogue context (Horne et al., 1999). In two recent perception studies, however, House (2002, 2003), 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. Perception results confirmed the importance of timing where an early peak followed by a final falling contour was perceived as a statement while a late peak resulting in a rise through the final syllable was perceived as a question. This applied to test material where focal accent was carried on words with accent 1 and words with accent 2. 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 50–75 ms.

It is clear that listeners perceive a final rise as an interrogative signal in Swedish even though a rise has not been described as a necessary component of interrogative intonation. These perception study results called for a study of speech production data to see to what extent final rises can be found in spontaneous speech. The availability of an annotated corpus of spontaneous computer-directed speech presented very suitable speech material for analyzing phrase-final intonation in questions directed to an animated agent. We may conjecture that the optionality of a final rise is represented in the data and that an optional rise may in fact signal more than question intonation in Swedish.

There may also be a relationship between a final rise and a final focal accent. This paper examines the extent to which final rises occur in wh-questions in a corpus of computer-directed spontaneous speech and discusses the function these rises may have in signaling dialogue acts and speaker attitude over and beyond an information question. One such social function of the phrase-final rise, that of signaling friendliness or social interest, was also tested in two separate perception experiments.

2. Production study

2.1. Material and analysis

The utterances used in this study were taken from the August database recorded in 1998 and 1999 (Gustafson et al., 1999). The August spoken-dialogue system was a multimodal system using a talking head as an animated agent. The system was a fully automatic dialogue system using modules for speech recognition and audio-visual speech synthesis with a broker architecture that enabled a distributed system. The animated agent was modeled on the Swedish author August Strindberg, and the system had several simple domains including facts about Strindberg, Stockholm, KTH and local information such as the locations of restaurants in Stockholm. The system was built into a kiosk and placed in public in central Stockholm for a period of six months. The system was equipped with a push-to-talk button to ensure that the recorded utterances were actually directed to the system. The speech recordings resulted in a database consisting of 10,058 utterances. The utterances were transcribed orthographically and labeled for speaker characteristics and utterance types by Bell and Gustafson (1999a,b) (see also Gustafson, 2002; Bell, 2003 for recent reviews of this work). The total number of speakers represented in the database was 2685 of which 50% were men, 26% women and 24% were children. Most of the speakers had dialects from central Sweden although other dialects and non-native speakers were also represented in the database. No specific dialect information was

obtained. Bell and Gustafson (1999a) further analyzed the utterances according to the presumed intentions of the speakers. Two major categories were used: socializing or information seeking. In other words, was the user mainly interested in a social interaction with the agent, or was the user genuinely trying to retrieve information?

To obtain material for the present study, utterances containing the question words *vad* “what” and *var* “where” were extracted from the corpus. A total of 334 utterances contained the word “what” and 161 contained the word “where”. For this analysis, 100 questions containing “what” and 100 questions containing “where” were randomly extracted from the subcorpus. The 200 utterances were analyzed auditively and visually using the WaveSurfer speech analysis package (Sjölander and Beskow, 2000).

Each utterance was marked for presence or absence of final rise, presence or absence of final focal accent, and whether the utterance was produced by a child, woman or man. These decisions were made on the basis of an auditory analysis and the inspection of the waveforms, spectrograms and *F0* traces of each utterance.

Waveforms, spectrograms and *F0* traces of two typical utterances are shown in Figs. 1 and 2. The utterance in Fig. 1 displays a characteristic final rise while the utterance in Fig. 2 displays a focal accent ending in a final low.

2.2. Results

2.2.1. Final rises

Twenty two percent of the questions in this material ended in a phrase-final rise. Proportionally, children produced more questions with final rises (32%), with women second (27%) and men the lowest percentage of questions with a final rise (17%). This breakdown is presented for both question words in Fig. 3.

2.2.2. Distribution by speaker type and question word

Of the 200 questions analyzed for this study, 17% were produced by children, 26% by women, and 57% by men. This compares to the distribution of the entire corpus which was 24% children, 26% women and 50% men (Bell and Gustafson, 1999a). Children asked a proportionally larger

Fig. 1. Waveform, spectrogram and *F0* contour of the utterance “Hej, jag heter Peter vad heter du?” (Hi, my name is Peter what is your name?) showing final rise.

Fig. 2. Waveform, spectrogram and F0 contour of the utterance “Hej, vad heter du?” (Hi, what is your name?) showing final low.

Fig. 3. Percentage of questions with a final rise distributed by speaker type and question word.

Fig. 4. Distribution of the total number of questions by speaker type and question word.

number of “what” questions (22) than “where” questions (12), while men asked a proportionally larger number of “where” questions (62) than “what” questions (52). This is displayed in Fig. 4.

“What” questions were generally more oriented to social interaction such as “What is your name?” “What are you doing?” and “What time is it?”. “Where” questions were of a more strictly information retrieval character pertaining to the location of streets and restaurants and public facilities.

2.2.3. Final focal accent

Seventy seven percent of the questions ending in a final rise also ended in a focal accent. Of the questions ending in a final low, 48% ended in a focal accent. The distribution for each question word is presented in Fig. 5.

2.2.4. Speaker dialect

As in the database as a whole, most of the speakers represented in the material had dialects

Fig. 5. Distribution of questions with final focus.

from central Sweden. There were a couple of speakers of dialects of West Sweden where a final rise is characteristic of focal accent, and there were also a few non-native speakers of Swedish. These speakers were not eliminated as this is representative of the Stockholm population, and moreover, there is no exact dialect information pertaining to each speaker in the database.

2.3. Discussion

2.3.1. Final rise as interrogative marker

The results of this study are consistent with the traditional description of interrogative intonation in Swedish (Gårding, 1998), where a final rise is not a necessary component for questions. However, as an optional interrogative marker, a final rise does occur in a considerable number of spontaneous wh-questions and therefore should be considered as an important component of interrogative intonation when describing and modeling Swedish intonation.

In this material, final rises occur mostly in conjunction with final focal accent. The rise can be seen as a replacement accent (i.e. an interrogative focal accent) or a deformation of the focal accent where the focal peak is delayed as in the perception experiments cited above (House, 2002, 2003). The fact that the final rise can also occur on a non-focal accent can be seen as evidence that the rise is an extra intonational factor which either delays the focal peak or surfaces as an extra rise. This

extra intonation factor may not only be an optional reinforcement for interrogative mode, it may also signal pragmatic meaning.

2.3.2. Final rise and pragmatic meaning

There has been much discussion about the role of the phrase-final tone, or boundary tone, in intonational meaning (see Ladd, 1996 for a review). A phrase-final high or rise can convey speaker attitude such as uncertainty, or be a signal of dialogue act such as feedback seeking or turngiving (e.g. Caspers, 2003; Cerrato, 2002).

Using the utterance categories proposed by Bell and Gustafson (1999a), socializing and information seeking, there is evidence that the final rise in this material is a signal of socializing while the final low is a signal of information seeking. In the complete August corpus, children produced more socializing utterances than did women and men. In this material, children produced proportionally more questions with final rises than did women and men. Furthermore, children asked proportionally more “what” questions than “where” questions with the “what” questions being more socially oriented. This interpretation is consistent with a final rise being a signal of feedback seeking rather than purely information question intonation. This interpretation also follows Kohler’s (2004) division of interrogatives into fact-oriented and addressee-oriented questions.

2.3.3. Biological codes and intonation

An explanation for the function of a final rise as a socializing marker in question intonation can be seen in the framework of biological codes for universal meanings of intonation, proposed by Gussenhoven (2002). Gussenhoven describes three codes or biological metaphors: a frequency code, an effort code and a production code. The frequency code, originally proposed by Ohala (1983, 1984), implies that a raised F_0 is a marker of submissiveness or non-assertiveness and hence question intonation. The effort code implies that articulation effort is increased to highlight important focal information producing a higher F_0 . The production code associates high pitch with phrase beginnings (new topics) and low pitch with phrase endings.

In this account of the frequency code, the final rise as a cue to submissiveness or non-assertiveness can be interpreted as an invitation to socialize and engage in conversation. A final low in a question is more a signal of assertiveness and command, thus characterizing a more information-seeking question.

In terms of the effort code, a rise is already associated with focal accent to highlight important information. Therefore, to signal the intention to socialize, the production code needs to be exploited. In the production code, high pitch is normally associated with new topics at phrase beginnings. In the case of the final rise, however, the production code is reversed with the high coming at the end of the phrase and signaling the invitation to continue the social interaction. In this interpretation, the final rise used in social oriented questions is not the same code as a focal accent rise. The final rise is part of the frequency code which interacts with the production code, while the focal rise is part of the effort code. The final rise can therefore be found on non-focal final words, or, when coinciding with a final focal accent, may result in a focal peak delay.

2.3.4. Focal structure of question utterances

The fact that final rises occur mostly in conjunction with final focus is also interesting. In this material the focal accent in final position seems to facilitate the final rise. The timing of the focal accent rise has been shown to be more variable than the timing of the word accent (Bruce, 1987). A more detailed study of the focal peak timing in the questions with non-final focal accents would be interesting from the perspective of focal accent variability. However, it is clear that the final rise in these questions contributes to the variability of the timing of the focal accent rise.

It also appears that in many of the instances in this material of final lows with non-final focal accent, final focus is avoided by an optional sentence structure containing a final non-focal word or phrase, such as *Vet du var Kulturhuset ligger?* “Do you know where the Culture Center is located?” or *Ja, men var bor du, någonstans?* “Yes, but where do you live, somewhere?” (translated more idiomatically as, “Yes, but where do you live, anyway?”). In this way, the tendency of a final rise coupled to final focus is avoided and the information-seeking intention of the speaker is

Fig. 6. Waveform, spectrogram and F_0 contour of the utterance “Var bor du?” (Where do you live?) uttered by a child showing final high.

Fig. 7. Waveform, spectrogram and F_0 contour of the utterance “Ja, men var bor du, någonstans?” (Yes, but where do you live, anyway?) uttered by an adult male showing final low.

emphasized. This is illustrated in Figs. 6 and 7 where a child asks, “Where do you live?” with a final rise (Fig. 6) and an adult male asks, “Yes, but where do you live, somewhere (anyway)?” with a final low (Fig. 7).

3. Perception study I

The distribution of the rises in the production material indicated the possible pragmatic function of the rises as signaling intended social interaction rather than purely information seeking. Evidence relating to the final rise as a signal of a social interaction oriented dialogue act is thus gathered from the production study. To test this interpretation, a perception experiment where subjects were asked to select the friendlier sounding utterance of a pair of question word questions was carried out. Variables were peak height, focal accent placement and final peak location where the hypothesis was that a higher and later peak would be perceived as friendlier. The concept of “friendliness” as synonymous with “socially interested” was tested by having an independent group of four subjects participate in the same experiment with a different set of instruc-

tions in which the subjects were asked to select the utterance which sounded more “socially interested” of the pair.

3.1. Stimuli

The test sentence, *Vad heter du?* “What’s your name?” was synthesized using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool (Sjölander and Beskow, 2000).

Two sets of four pitch-manipulated stimuli were created by temporally shifting the focal accent rise in a low-peak configuration (set one) and a high-peak configuration (set two). The first stimulus in both sets contained a focal accent rise through the stressed vowel [e:] of *heter* followed by a plateau and a falling contour into the final word *du* “you”. Stimuli 2, 3 and 4 were created with an early, mid and late focal accent rise on the final word *du* “you”. The peaks of stimuli 2–4 are time-shifted in intervals of 100 ms. In the low-peak set of stimuli the accent peak was set at 130 Hz. In the high-peak set of stimuli, the accent peak was set at 160 Hz comprising a widened F_0 range on

Fig. 8. Schematic presentations of the stimuli used in perception experiment I. The left panel represents the low pitch set and the right panel the high pitch set.

the focal accent. All versions are naturally sounding utterances comprising possible versions of focal accent timing in final position. However, a final focal accent will often result in a deaccentuation of the earlier accent also regarding duration. In these stimuli, duration was held constant while in study II (Section 4) duration was shortened in the earlier accent when the final word was focussed.

The manipulated portions of the stimuli are presented schematically in Fig. 8. The peak position numbers correspond to the timing location of the peaks in both the low-pitch and high-pitch set.

3.2. Subjects

Fourteen subjects participated in the experiment. All the subjects were students at KTH and were native speakers of Swedish with the central Swedish (Stockholm) dialect predominating. None of the subjects reported any hearing loss. The students participated in the experiment as part of a course requirement.

3.3. Test procedure and task

The experiment was conducted using a pairwise test paradigm in which subjects were presented with pairs of utterances and asked to decide which utterance of the pair was the friendlier sounding question. Within each pitch level, all stimuli were paired with each other, i.e. 1–2, 1–3, 1–4, 2–3, 2–4, and 3–4 separately for the low pitch set and for the high pitch set. Across the two pitch levels, each peak position was paired with the cor-

responding peak position of the other level, i.e. 1L–1H, 2L–2H, 3L–3H and 4L–4H. The order of each stimulus pair was also reversed making a total of 32 stimulus pairs comprising 6×2 (within pitch levels) + 4 (across levels) = 16×2 (reversed order) = 32. The stimuli were presented over loudspeakers to the subjects as a group. Each stimulus pair was presented twice in succession with a 2-s pause between each utterance and between the two successions. Subjects then had 4 s in which to mark their responses on an answer sheet before the next stimulus pair was presented. The test was divided into four groups of eight stimuli pairs with a longer pause between each group. A buffer group of two pairs was used to introduce the test.

3.4. Results

Results obtained by plotting the percentage “friendly” response for each utterance in the entire test are shown in Fig. 9. Most of the utterances obtain a score of around 50%. However, two

Fig. 9. Percent “friendly” responses for each stimulus in all pair combinations in perception experiment I.

utterances are noteworthy. These are utterance 2 in the low-pitch configuration, which received only 33% “friendly” responses, and utterance 3 in the high-pitch configuration, which received 61% “friendly” responses.

An examination of the results of each pair (both presentation orders combined) showed a significant difference in judgments for three pair combinations within pitch levels and for three pair combinations across pitch levels (binomial distribution, $p < 0.05$). The three within-pitch-level pairs were pairs 2L–3L, 2L–4L, and 2H–3H, where in all cases stimulus 2 was regarded as the less friendly (see also Fig. 9). The three across-pitch-level pairs were 1L–1H, 2L–2H and 4L–4H where the high pitch member of the pair was judged as friendlier. Focal accent placement, i.e. the differences between stimuli 1 and the others did not have a significant effect. There was a slight, but non-significant difference in the results between stimuli 1 and stimuli 3 (for both pitch levels). These results are illustrated in Fig. 10 where the upper panel shows the percent “friendly” re-

sponses for pairs differing in peak position (combined for both pitch levels), and the lower panel shows percent “friendly” responses for all pairs across pitch levels. When combining both pitch levels in peak position pairs, the only pair showing significant differences is pair 2–3.

3.5. The concept of friendliness as expressing social interest

The concept of “friendliness” as synonymous with “socially interested” was tested by having an independent group of four subjects participate in the same experiment with a different set of instructions in which the subjects were asked to select the utterance which sounded more “socially interested” of the pair. The same general tendencies were seen in the results of this test as can be seen in Figs. 11 and 12. Some differences between the two test can be noted. The final peak position (stimulus 4) elicited more “socially interested” responses than the other stimuli except for stimulus 1 with high pitch which was also perceived as “socially interested”. Stimulus 1 with low pitch did not, however, receive high scores for “socially interested”. Significant differences between pairs were found between pairs 1–2, 1–4 and 2–4.

3.6. Discussion

The results of perception study I provide an indication that both peak position and peak height can contribute to the impression of friendliness in a question utterance. Generally, a mid or late focal

Fig. 10. Results from test pairs in perception experiment I. The upper panel shows pairs differing in peak position (combined for both pitch levels) and the lower panel shows pairs differing in peak height. Asterisks indicate significant differences (binomial distribution, $p < 0.05$).

Fig. 11. Percent “socially interested” responses for each stimulus in all pair combinations in the “socially interested” task experiment.

Fig. 12. Results from test pairs in the “socially interested” task experiment. The upper panel shows pairs differing in peak position (combined for both pitch levels) and the lower panel shows pairs differing in peak height. Asterisks indicate significant differences (binomial distribution, $p < 0.05$).

peak in the final syllable was associated with friendliness while an early final focal peak was perceived as less friendly. In all peak positions except position 3, the higher peak was perceived as friendlier when paired with the lower peak in the same position. The fact that the higher and lower peak in position 3 were not judged as differing in friendliness is noteworthy. This would indicate that position 3, a focal rise in the middle of the vowel followed by a final fall, is in this context the strongest peak position for signalling friendliness, since in this position, the high peak does not increase the percept of friendliness. The fact that position 3 is the strongest position rather than position 4, which is only a final rise, is also interesting. It may be that the rise in position 4 is too late to be prominent and is therefore not as salient as the rise in position 3. However, the results of the additional experiment showed a preference for position 4 with the final rise as being the most “socially interested”. Although the number of listeners is not sufficient to draw conclusions about differences between the interpretation of

“friendly” and “socially interested”, continued exploration of shades of meaning related to fine differences in final timing could be worthwhile. Generally, friendliness or social interest are two closely related attitudes that seem to be marked by a phrase-final high or rise. These results are generally consistent with results for friendliness obtained for English and Dutch by Chen et al. (2001) where higher $F0$ range and a final rising contour elicited more responses for friendliness.

4. Perception study II

As a follow-up to the perception study presented above, a second study was designed to test more explicitly if there is a difference in perceived friendliness between a late focal peak in non-final position and a late focal peak in phrase-final position.

4.1. Stimuli

The same test sentence as in study I, *Vad heter du?* “What’s your name?” was synthesized using an experimental version of the Infovox 330 di-phone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool (Sjölander and Beskow, 2000).

As in study I, two sets of four pitch-manipulated stimuli were created by temporally shifting the focal accent rise and peak in a low-peak configuration (set one) and a high-peak configuration (set two). The first stimulus in both sets contained a focal accent rise through the stressed vowel [e:] of *heter* followed by a plateau and a falling contour into the final word *du* “you”. The focal accent high on *heter* was 134 Hz for the low-peak configuration and 161 Hz for the high-peak configuration. The waveform, spectrogram and $F0$ contour of stimulus 1 in the low peak configuration is shown in Fig. 13, upper panel. This rise and peak (in both high and low pitch configurations) was time shifted to the right by 100 ms (delayed) to create stimulus 2.

Stimulus 4 in this study corresponds to stimulus 3 of study I. The focal accent high on *du* “you” was 125 Hz for the low-peak configuration and

Fig. 13. Waveform, spectrogram and F_0 contour of the synthesized utterance “Vad heter du?” (What is your name?) used in perception experiment II. The upper panel represents stimulus 1L with a low, early peak. The lower panel represents stimulus 4H with a late high peak.

152 Hz for the high-peak configuration. It was created with a late peak in the final word *du*, but not as late as stimulus 4 of study I. This was based on the fact that stimulus 3 received the highest “friendliness” rating in study I. Stimulus 3 in this

study was created by shifting the rise and peak of stimulus 4 to the left (earlier in time) by 100 ms. Thus stimulus 3 in this study corresponds to stimulus 2 in study I and was expected to result in the least number of “friendly” responses.

A difference between the stimuli with final focus in this study and those with final focus in study I is that in this study, *heter* was deaccentuated in the temporal domain by decreasing duration. This was done to strengthen the final focal accent on *du*. Another result of deaccentuation is that the final [r] in *heter* is not realized, rather the initial consonant of *du* becomes a retroflex. The waveform, spectrogram and *F0* contour of stimulus 4 with a high peak is shown in Fig. 13, lower panel. The final word *du* was not shortened when carrying the non-focal accent, as this would have restricted the normal durational characteristics of final lengthening.

4.2. Subjects

A new group of 25 subjects participated in the experiment. All the subjects were students at KTH and were native speakers of Swedish with the central Swedish (Stockholm) dialect predominating. None of the subjects reported any hearing loss. The students participated in the experiment as part of a course requirement.

4.3. Test procedure and task

The test procedure and task were identical to that of perception study I. See Section 3.3 above.

4.4. Results

Results obtained by plotting the percentage “friendly” response for each utterance in the entire test are shown in Fig. 14. It is clear that the final

Fig. 14. Percent “friendly” responses for each stimulus in all pair combinations in perception experiment II.

peak position for both high and low pitch configuration elicited the most “friendly” responses. The early-low peak in the final word (stimulus 3L) obtained the least number of “friendly” responses. The delayed peak in the non-final focal accent (stimulus 2) did not receive high scores for friendliness in either high-pitch or low-pitch configuration.

An examination of the results of each pair (presentation orders combined) showed a significant difference in judgments for all pair combinations within pitch levels and across pitch levels except for pair 2L–3L (binomial distribution, $p < 0.05$). Combined results are shown in Fig. 15 where the upper panel shows the percent “friendly” responses for pairs differing in peak position (combined for both pitch levels), and the lower panel shows percent “friendly” responses for all pairs across pitch levels. In these combined results, there are significant differences for all pair combinations differing in peak position. Again it is clear that the

Fig. 15. Results from test pairs in perception experiment II. The upper panel shows pairs differing in peak position (combined for both pitch levels) and the lower panel shows pairs differing in peak height. Asterisks indicate significant differences (binomial distribution, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

final peak position is a strong cue for “friendliness”. It is also noteworthy that the test pairs 1–4 showed a highly significant result in favor of peak position 4 for “friendliness” indicating a preference for final focus, late peak as a signal of friendliness.

4.5. Discussion

The results of perception study II are much more conclusive than the results from study I but are still consistent with the results of study I in that the late peak position and raised pitch contribute to the impression of friendliness in a question utterance. The results of study II also demonstrate a clear preference for final focus compared to early focus for the impression of friendliness, with the rise and peak delay in early focus (stimulus 2) actually decreasing the impression of friendliness. It may be that this peak delay approaches the early peak on the final word and is therefore perceived as less friendly for that reason.

The major difference between the stimuli in study I and study II lies in the deaccentuation in the temporal domain of the earlier accent when the final accent is focal in the stimuli of study II. This temporal deaccentuation may have strengthened the perceived late focal accent thereby rendering it even more friendly sounding than in study I. The interplay between intonation and duration may also have contributed to a more natural sounding final focal accent.

Although the results from study II are more conclusive, they are based on only one test sentence. The sentence is semantically suitable in that it can be easily interpreted as information seeking (focus on “name”) or social interest (focus on “you”). This kind of interaction between phonetic and semantic focus and phrase-final position clearly needs to be investigated further.

5. General discussion and conclusions

The results of the production study show final rises to occur in about 20% of wh-questions extracted from a corpus of spontaneous questions directed to an animated agent in a spoken dialogue

system. The results confirm the optionality of a final rise in Swedish interrogative intonation. The distribution of the rises in the material also gave an indication of the pragmatic function of the rises as signalling intended social interaction rather than being purely information seeking. Final rises can also serve as a signal of interest on the part of the speaker for the addressee as proposed by Kohler (2004) in his study of spontaneous German question intonation. The results of the perception tests are consistent with this interpretation where friendliness and social interest is interpreted as interest or concern for the addressee.

Although the questions in the production study were directed to a machine and not a human, it is interesting that pragmatic functions of human-to-human communication seem to be carried over to the human-machine interaction, in this case possibly contingent upon the interface being an animated agent in the form of a personified talking head which elicits utterances with a social function (cf. Bell and Gustafson, 1999a). Spoken communication does, however, have a biological basis, and it seems to be quite probable that the biological codes discussed above apply not only to speech between humans, but also to expressive speech with machines when the interface resembles a human, especially regarding codes which have become conventionalised as part of an intonational grammar as discussed by Gussenhoven (2002). The age and gender differences found in the production study may also be a reflection of this fact. Children can be much more oriented to the social aspects of communication. It would also be interesting to study a similar database with a female or child talking head agent to see if gender differences in user question intonation would be affected by the gender of the agent.

Finally the results of this study indicate the importance of further research on the prosodic aspects of spontaneous speech in human-machine interaction. A greater understanding of prosodic features and variability in interaction will help us better model both speech input and output aspects of spoken dialogue systems. The further study of phrase-final characteristics and the meanings they convey can have considerable importance for our understanding of human-computer interaction

and for improving the responsiveness of animated agents.

Acknowledgments

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. This work has also been supported in part by a grant from the Swedish Research Council (VR) to the project: “Boundaries and groupings—the structuring of speech in different communicative situations” (Carlson et al., 2002). Special thanks to Linda Bell and Joakim Gustafson for providing the transcribed August database, to Rolf Carlson for extracting the question utterances, and to two anonymous reviewers for helpful comments.

References

- Bell, L., 2003. Linguistic adaptations in spoken human–computer dialogues; empirical studies of user behavior, Doctoral dissertation, Department of Speech, Music and Hearing, KTH, Stockholm.
- Bell, L., Gustafson, J., 1999a. Utterance types in the August system. In: Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, pp. 81–84.
- Bell, L., Gustafson, J., 1999b. Interaction with an animated agent in a spoken dialogue system. In: Proc. Eurospeech '99, Budapest, pp. 1143–1146.
- Bhagat, S., Carvey, H., Shriberg, E., 2003. Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings. In: Proc. 15th ICPhS, Barcelona, pp. 2961–2964.
- Bolinger, D., 1989. *Intonation and Its Uses: Melody in Grammar and Discourse*. Edward Arnold, London.
- Bredvad-Jensen, A.C., 1984. Tonal geography. Geographical variation in declarative and interrogative intonation along the west coast of Sweden. In: Elert, C.C., Johansson, I., Stangert, E. (Eds.), *Nordic Prosody III*. University of Umeå, pp. 31–41.
- Bruce, G., 1987. How floating is focal accent? In: Gregersen, K., Basbøll, H. (Eds.), *Nordic Prosody IV*. Odense University Press, Odense, pp. 41–49.
- Bruce, G., Touati, P., 1992. On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Comm.* 11, 453–458.
- Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E., Swerts, M., 2002. Boundaries and groupings—the structuring of speech in different communicative situations: A description of the GROG project. In: Proc. Fonetik 2002, TMH-QPSR 44, Vol. 1, pp. 65–68.
- Caspers, J., 2003. On the function of low and high boundary tones in Dutch dialogue. In: Proc. 15th ICPhS, Barcelona, pp. 1771–1774.
- Cerrato, L., 2002. Some characteristics of feedback expressions in Swedish. In: Proc. Fonetik 2002, TMH-QPSR 44, Vol. 1, pp. 101–104.
- Chen, A., Rietveld, T., Gussenhoven, C., 2001. Language-specific effects of pitch range on the perception of universal intonational meaning. In: Proc. Eurospeech 2001, Aalborg, Denmark, pp. 1403–1406.
- Cruttenden, A., 1986. *Intonation*. Cambridge University Press, Cambridge.
- D’Imperio, M., House, D., 1997. Perception of questions and statements in Neapolitan Italian. In: Proc. Eurospeech 97, Rhodes, Greece, pp. 251–254.
- Ferrer, L., Shriberg, E., Stolcke, A., 2002. Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody. In: Proc. ICSLP 2002, Denver Colorado, pp. 2061–2064.
- Gårding, E., 1979. Sentence intonation in Swedish. *Phonetica* 36, 207–215.
- Gårding, E., 1998. Intonation in Swedish. In: Hirst, D., Di Cristo, A. (Eds.), *Intonation Systems*. Cambridge University Press, Cambridge, pp. 112–130.
- Gussenhoven, C., 2002. Intonation and interpretation: Phonetics and phonology. In: Bel, B., Marlien, I. (Eds.), *Proc. Speech Prosody 2002 Conf.*, Aix-en-Provence, pp. 47–57.
- Gustafson, J., 2002. Developing multimodal spoken dialogue systems; empirical studies of spoken human–computer interaction. Doctoral dissertation, Department of Speech, Music and Hearing, KTH, Stockholm.
- Gustafson, J., Lindberg, N., Lundeberg, M., 1999. The August spoken dialogue system. In: Proc. Eurospeech 99, Budapest, pp. 1151–1154.
- Hadding-Koch, K., Studdert-Kennedy, M., 1964. An experimental study of some intonation contours. *Phonetica* 11, 175–185.
- Hirst, D., Di Cristo, A., 1998. A survey of intonation systems. In: Hirst, D., Di Cristo, A. (Eds.), *Intonation Systems*. Cambridge University Press, Cambridge, pp. 1–45.
- Horne, M., Hansson, P., Bruce, G., Frid, J., Jönsson, A., 1999. Accentuation of domain-related information in Swedish dialogues. In: Proc. ESCA International Workshop on Dialogue and Prosody, Veldhoven, The Netherlands, pp. 71–76.
- House, D., 2002. Intonational and visual cues in the perception of interrogative mode in Swedish. In: Proc. ICSLP 2002, Denver, Colorado, pp. 1957–1960.
- House, D., 2003. Perceiving question intonation: The role of pre-focal pause and delayed focal peak. In: Proc. 15th ICPhS, Barcelona, pp. 755–758.
- Ishi, C.T., Mikhtari, P., Campbell, N., 2003. Perceptually-related acoustic-prosodic features of phrase finals in spon-

- taneous speech. In: Proc. Eurospeech 2003, Geneva, pp. 405–408.
- Jilka, M., Syrdal, A.K., Conkie, A., Kapiłow, D.A., 2003. Effects on TTS quality of methods of realizing natural prosodic variations. In: Proc. 15th ICPhS, Barcelona, pp. 2549–2552.
- Kohler, K.J., 2004. Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In: Fant, G., Fujisaki, H., Cao, J., Xu, Y. (Eds.), *From Traditional Phonology to Modern Speech Processing*. Foreign Language Teaching and Research Press, Beijing, pp. 205–214.
- Ladd, D.R., 1996. *Intonation Phonology*. Cambridge University Press, Cambridge.
- Ohala, J.J., 1983. Cross-language use of pitch: An ethological view. *Phonetica* 40, 1–18.
- Ohala, J.J., 1984. An ethological perspective on common cross-language utilization of *F0* in voice. *Phonetica* 41, 1–16.
- Sjölander, K., Beskow, J., 2000. WaveSurfer—a public domain speech tool. In: Proc. ICSLP 2000, Vol. 4, Beijing, China, pp. 464–467.
- Syrdal, A.K., Jilka, M., 2004. Acceptability of variations in question intonation in natural and synthesized American English. *J. Acoust. Soc. Amer.* 115 (5, Pt. 2), 2543(A).
- van Heuven, V.J., Hann, J., Kirsner, R.S., 1999. Phonetic correlates of sentence type in Dutch: Statement, question and command. In: Proc. ESCA Internat. Workshop on Dialogue and Prosody, Veldhoven, The Netherlands, pp. 35–40.