

# INTONATIONAL AND VISUAL CUES IN THE PERCEPTION OF INTERROGATIVE MODE IN SWEDISH

*David House*

Centre for Speech Technology  
Department of Speech, Music and Hearing  
KTH, Stockholm, Sweden  
davidh@speech.kth.se

## ABSTRACT

This paper presents results from two perception experiments designed to investigate intonational cues and visual facial cues to interrogative mode in Swedish. Results from the intonation test indicate that both a widened F0 range on a final focal accent and time alignment properties of the F0 rise and peak make important contributions to the interrogative percept. Results from the audiovisual test showed that vertical head nodding and smiling tended to reinforce declarative intonation while interrogative intonation was not strengthened by hypothesized interrogative visual cues consisting of eyebrow movement and slow vertical head movement. The interaction between audio and visual cues for accentuation and interrogative mode is discussed and some implications of adding the visual modality to the traditional definition of question intonation are explored.

## 1. INTRODUCTION

The signaling of interrogative mode in speech is a topic which has long attracted interest from intonation researchers. The description of question intonation in languages has not, however, been simple and is far from uncontroversial. Different languages and different types of questions produce different kinds of question intonation. The most commonly described characteristic for questions is high final pitch and overall higher pitch [1]. In some languages, however, e.g. Neapolitan Italian [2], the time alignment of a final accent has been shown to play a decisive role in the perception of interrogative mode.

In Swedish, question intonation has been primarily described as marked by a raised topline and a widened F0 range on the focal accent [3]. An optional terminal rise has been described, 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 [4].

The role of visual facial cues in signaling the interrogative mode is an area which has not received as much attention. There has been, however, considerable research carried out on the timing and synchronization of articulator movements in audiovisual speech processing, e.g. [5], and on describing spoken and gestural conversational signals in human to human interactions [6]. There have also been exploratory investigations on visual cues for prominence and feedback signaling [7][8][9]. Work aimed at investigating the

coordination of audio and visual interrogative signals in speech perception and the implementation of this knowledge in audiovisual synthesis is not as well represented.

The purpose of this study is to investigate both intonational and visual cues to interrogative mode in Swedish. In terms of intonational cues, the object of the study is to test if the time alignment of a final focal accent (i.e. a late peak) is sufficient to signal interrogative mode in Swedish when it is not expressed by syntactic or lexical means and to test the perceptual interaction of temporal alignment and pitch range. In terms of visual cues, the objective is to test a simple set of visual cues hypothesized to signal interrogative mode and to investigate how these visual cues interact with the intonational cues.

## 2. INTONATIONAL CUES

In the first experiment, only the intonational cues were manipulated. The visual component of the stimuli was held constant and contained no head, eye or eyebrow movement.

### 2.1. Stimuli

The test sentence, *Hon vill bara flyga*, “She only wants to fly,” 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 [10]. The speech generation also included an animated talking head with parametrically generated articulatory gestures [11]. The test sentence was synthesized with the final syllable bearing a focal accent peak. Final vowel duration was set to 150ms. Two sets of six pitch-manipulated stimuli were created by systematically shifting the focal accent peak through the vowel in steps of 25ms. In one set of stimuli the accent peaks were set at 130Hz consistent with the F0 range of the entire sentence. In the second set of stimuli, the accent peaks were set at 160Hz comprising a widened F0 range on the focal accent.

The manipulated portions of the stimuli are presented schematically in Figure 1. The stimuli numbers 1-6 correspond to the timing location of the peaks in both the low-pitch and high-pitch set.

### 2.2. Subjects and procedure

A total of 11 subjects participated in the first experiment. Most of the subjects were recruited from among students and staff at KTH. No one reported any hearing loss or visual impairment

and all were native speakers of Swedish with the central Swedish (Stockholm) dialect predominating.

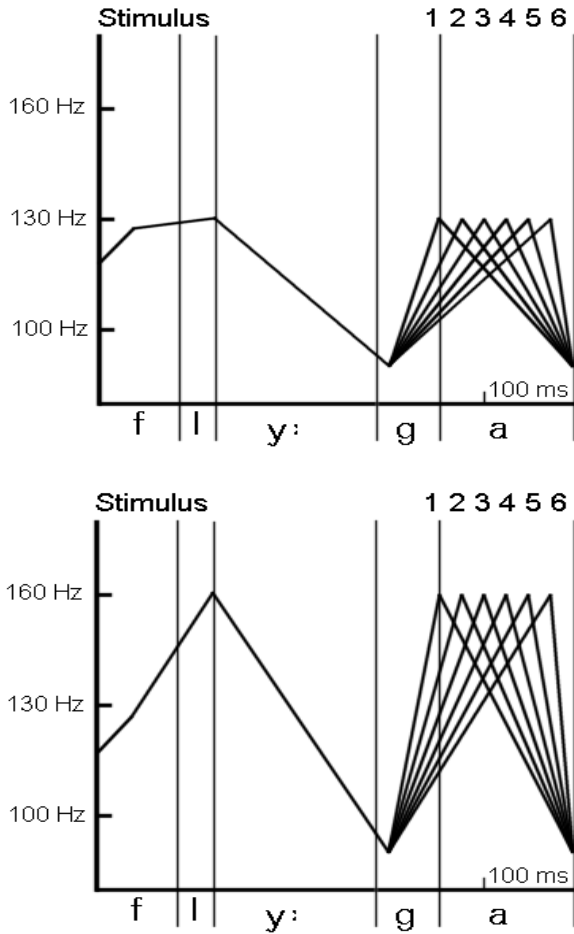


Figure 1: Schematic stimuli used in the perception test. The stimuli are numbered 1-6 in both the low-pitch set (upper panel) and the high-pitch set (lower panel).

The stimuli were presented to each subject individually using a computer interface especially designed for the experiment. The audio was presented through headphones and the face was displayed in a frame measuring 13 x 19 cm with the face itself measuring 12 x 18 cm. A 3D graphic accelerator was installed in each test computer to insure sufficient temporal resolution for audio-visual synchronization (average frame rate was at least 80 frames per second).

Subjects were asked to listen to each stimulus while looking at the face and given the task of deciding if the speaker intended to make a statement or ask a question. The subjects were also requested to indicate on a scale between 1 to 5 how confident they were of their choice where 5 was certain and 1 was guessing. The subjects were allowed to listen and look at each stimulus as many times as they wished before making their choice and proceeding to the next stimulus. The twelve stimuli were presented in two groups with the low-pitch stimuli in one group and the high-pitch stimuli in the other group. The

presentation order of the stimuli was randomized within each group. Each group was presented twice with the group presentation order reversed for the second presentation.

### 2.3. Results

Results of the first experiment indicate that both a widened F0 range on the focal accent and time alignment properties of the rise and peak make important contributions to the interrogative percept. The results are displayed in Figure 2 showing percent question response and the confidence response score in percent for each stimulus. The confidence response score is given a positive sign when a question response was chosen and a negative sign for a statement response.

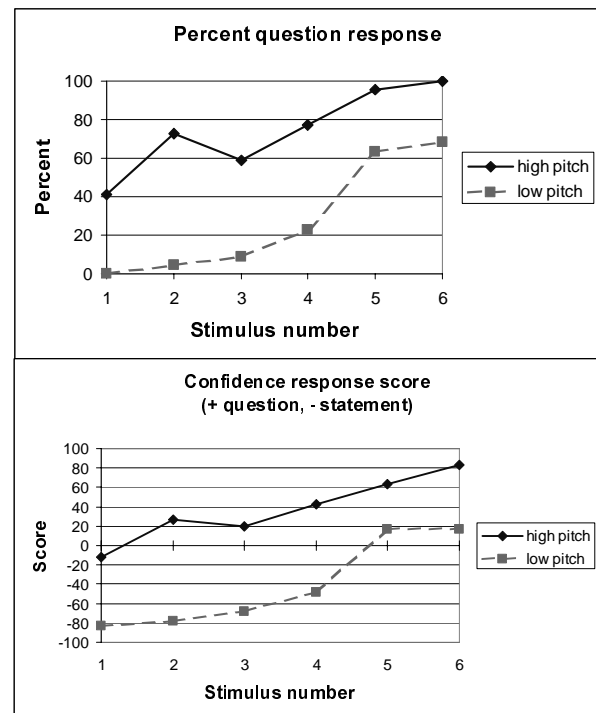


Figure 2: Results of the intonation test showing percent question responses (upper panel) and confidence response scores in percent (lower panel).

For the low-pitch stimuli, the first four peak positions signaled declarative mode while the final two signaled interrogative mode. These results were consistent between subjects: single factor ANOVA  $F(10,55)=1.09$ ,  $p=0.386$ , and showed a significant difference between stimuli:  $F(5,60)=24.12$ ,  $p<0.001$ . For the high-pitch stimuli, all peak positions except the first signaled interrogative mode. These results showed a significant between subject difference:  $F(10,55)=3.30$ ,  $p=0.002$  and a significant difference between stimuli:  $F(5,60)=6.30$ ,  $p<0.001$ . 100% question and statement response for the respective categories was obtained for the two endpoint stimuli, i.e. a low-pitch early peak gives statement responses while a high-pitch late peak gives question responses. The response curve for the low-pitch stimuli set is sharper at the category boundary than

for the high-pitch set. This difference is also reflected in the ANOVA statistics.

These results demonstrate sensitivity to time alignment on the order of 25ms, and a perceptual interaction between pitch range and temporal alignment. In terms of a trading relationship the effect of a widened F0 range is equivalent to a 75ms peak delay.

### 3. VISUAL CUES

In the second experiment the audio stimuli from the first experiment were presented with two different visual cue movement configurations involving the following facial gestures: smile, vertical head nod, eye narrowing and eyebrow lowering. The two configurations were hypothesized to convey interrogative or declarative mode.

#### 3.1. Stimuli

The twelve audio stimuli were identical to those synthesized in the first experiment. Two configurations combining different facial gestures were synthesized using the parametric manipulation tool described above. The parameter settings were inspired by those used in an earlier experiment on positive and negative feedback cues [9] with a subset of the cues being used in the current experiment.

The parameters for the hypothesized interrogative mode consisted of a slow up-down head nod and eyebrow lowering. The head nod started halfway into the sentence reaching its peak at the onset of the [y] vowel in *flyga* and ended at the end of the sentence. The eyebrows were lowered in a lowering gesture with a duration of 100 ms that began at the onset of the [y] vowel in *flyga*. The eyebrows remained lowered for the rest of the sentence.

The parameters for the hypothesized declarative mode consisted of a smile throughout the whole utterance, a short up-down head nod and eye narrowing. The smile was a gesture throughout the whole utterance, largely encompassing a widening of the mouth and a slight upwards movement of the mouth corners. The head movement was a nod which began at the onset of the [y] vowel in *flyga*, had a 100 ms onset phase, a 100 ms stationary phase, and a 100 ms offset phase. This type of nod has also been shown to be effective for conveying prominence [8]. The eye narrowing was introduced as a setting throughout the utterance and comprised slightly narrowed eyes compared to the interrogative settings. The configurations are shown in Figure 3.

#### 3.2. Subjects and procedure

A total of 27 new subjects participated in the second experiment. The subjects were students at KTH and participated in the experiment as part of a course requirement. No one reported any hearing loss or visual impairment and all were native speakers of Swedish with the central Swedish (Stockholm) dialect predominating.

The test procedure was identical to that for the first experiment using the same computer interface. Each of the twelve audio stimuli was paired with both versions of the visual configurations making a total of 24 different stimuli. The presentation order of the 12 different audio configurations was the same as in experiment 1 with the visual configurations

presented in random order within each group of six stimuli. Thus each stimulus in experiment 2 was only presented once.



Figure 3: The hypothesized interrogative configuration (left) and declarative configuration (right) sampled at the onset of the final [a] vowel in *flyga*.

#### 3.3. Results

The results of the second experiment are presented in Figure 4. They are generally quite similar to the results of the first experiment and thus do not show a great influence of the visual cues.

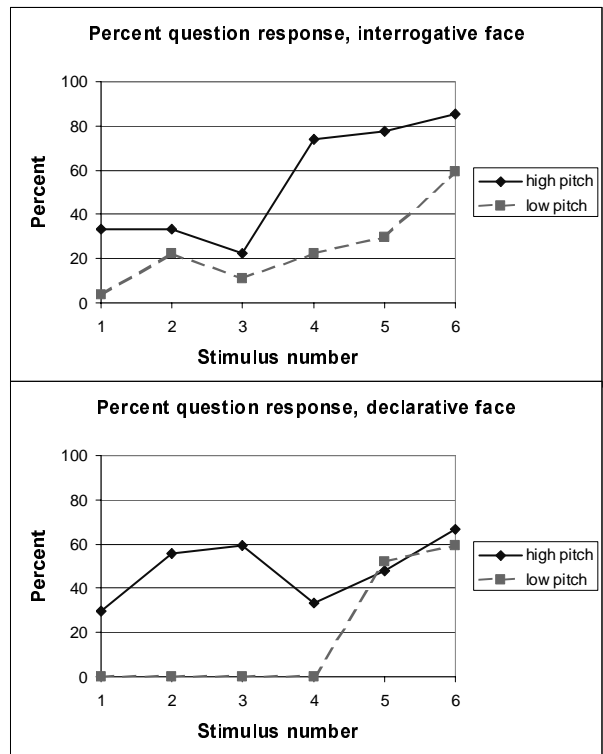


Figure 4: Results of the audiovisual test showing percent question responses for the interrogative face (upper panel) and for the declarative face (lower panel).

The influence of the visual cues is seen mainly in the results from the hypothesized declarative face. The addition of the facial cues reinforced declarative intonation especially near the perceptual boundary of the low-pitch stimuli, i.e. stimulus number 4 (low-pitch, declarative face, Figure 4). These same cues inhibited interrogative intonation in stimuli 4-6 (high-pitch, declarative face, Figure 4) and produced very low confidence scores in all the high-pitch stimuli pointing to the ambiguity of these stimuli.

The hypothesized interrogative face seemed to largely introduce more confusion to the perception of the stimuli. In none of the high-pitch stimuli do the interrogative cues enhance the interrogative percept; rather they tend to inhibit the percept. The confidence scores for these stimuli were in general much lower than those for the stimuli in the first experiment.

#### 4. DISCUSSION

The results of the intonation experiment clearly indicate that both a widened F0 range on the focal accent and time alignment properties of the rise and peak are important for the interrogative percept. While a widened F0 range was essential for obtaining 100% interrogative responses, a finding consistent with the Swedish question intonation description in [3], an early F0 peak was equally important for obtaining 100% declarative responses. This suggests a trading relationship between F0 range and peak timing in the declarative-interrogative continuum. This trading relationship is on the order of 50 to 75 ms, such that a widened F0 range has the same perceptual effect as a peak delay of 50 to 75 ms (e.g. stimulus 3 in the high F0 range receives the same response as stimulus 5 in the low F0 range as shown in Figure 2).

The peak delay necessary for signaling question intonation results in a rise in the initial part of the vowel. The rise itself is therefore a likely candidate for a perceptual correlate to interrogative intonation. The fact that the timing of the final focal accent in Swedish can be used to signal question intonation adds an extra dimension to the function of the focal accent. Since this experiment was restricted to testing the focal accent in final position it would also be interesting to test these timing cues in non-final position.

The results from the visual facial cue experiment are not as conclusive. While the hypothesized declarative visual cues reinforced declarative intonation, the hypothesized interrogative cues had little effect. As an evaluative experiment, the slow head movement and the eyebrow lowering which comprised the interrogative face did not evoke interrogative responses. These parameters were based on cues shown to signal negative feedback in an earlier experiment [9]. It is possible that negative feedback should not be equated with interrogative mode, but rather that they function as two separate dimensions in dialogue. It is quite possible to have negative feedback in declarative mode and positive feedback in interrogative mode.

On the other hand, the hypothesized declarative facial cues involving the smile and the head nod synchronized with the focal accent were evaluated as declarative. In the earlier feedback experiment [9] the smile was the strongest cue to positive feedback and it is clearly a salient, positive visual cue. The head nod has also been shown to be a strong, independent signal for prominence [8]. It could be that both the quick accent-synchronous nod and the slower nod enhanced the

percept of accentual prominence in this experiment and was interpreted as more affirmative than interrogative.

Finally, the results raise issues concerning the definition of question intonation in the context of audiovisual speech processing. Traditionally, question intonation has been seen as one category of the question-statement pair. The addition of the visual modality can introduce extra dimensions of meaning which add to the complexity of the categories of question and statement. As we gain more experience with the evaluation of audiovisual synthesis in dialogue systems, we are certain to increase our knowledge of the ways in which questions can be convincingly signaled.

#### 5. ACKNOWLEDGEMENTS

Special thanks to Jonas Beskow and Björn Granström for discussions on visual parameters. 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.

#### 6. REFERENCES

- [1] Hirst, D. and Di Cristo, A. "A survey of intonation systems," In D. Hirst and A. Di Cristo (eds.) *Intonation Systems*, 1-45. Cambridge University Press, Cambridge, 1998.
- [2] D'Imperio, M. and House, D. "Perception of questions and statements in Neapolitan Italian," In *Proceedings of Eurospeech 97*, 251-254, Rhodes, Greece, 1997.
- [3] Gårding, E. "Sentence Intonation in Swedish," *Phonetica* 36, 207-215, 1979.
- [4] Horne, M., Hansson, P., Bruce, G., Frid, J. and Jönsson, A. "Accentuation of domain-related information in Swedish dialogues," *Proceedings of ESCA International Workshop on Dialogue and Prosody*, 71-76. Veldhoven, The Netherlands, 1999.
- [5] Massaro, D. W. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, MA., 1998.
- [6] McNeill, D. *Hand and mind: What gestures reveal about thought*, University of Chicago Press, Chicago, 1992.
- [7] Granström, B., House, D. and Lundeberg, M. "Prosodic Cues in Multimodal Speech Perception," In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)*, 655-658, San Francisco, 1999.
- [8] House, D., Beskow, J. and Granström, B. "Timing and interaction of visual cues for prominence in audiovisual speech perception," In *Proceedings of Eurospeech 2001*, 387-390. Aalborg, Denmark, 2001.
- [9] Granström, B., House, D. and Swerts, M.G. "Multimodal feedback cues in human-machine interactions." In B. Bel and I. Marlien (eds.), *Proceedings of the Speech Prosody 2002 Conference*, 347-350, Aix-en-Provence, 2002.
- [10] Beskow, J. and Sjölander, K. "WaveSurfer - a public domain speech tool," In *Proceedings of ICSLP 2000*, vol. 4, 464-467, Beijing, China, 2000.
- [11] Beskow, J. Rule-based Visual Speech Synthesis. In *Proceedings of Eurospeech '95*, 299-302. Madrid, 1995.