Integrating Audio and Visual Cues for Speaker Friendliness in Multimodal Speech Synthesis

David House
Department of Speech, Music and Hearing
School of Computer Science and Communication, KTH, Stockholm, Sweden
davidh@speech.kth.se

Abstract
This paper investigates interactions between audio and visual cues to friendliness in questions in two perception experiments. In the first experiment, manually edited parametric audio-visual synthesis was used to create the stimuli. Results were consistent with earlier findings in that a late, high final focal accent peak was perceived as friendlier than an earlier, lower focal accent peak. Friendliness was also effectively signaled by visual facial parameters such as a smile, head nod and eyebrow raising synchronized with the final accent. Consistent additive effects were found between the audio and visual cues for the subjects as a group and individually showing that subjects integrate the two modalities. The second experiment used data-driven visual synthesis where the database was recorded by an actor instructed to portray anger and happiness. Friendliness was correlated to the happy database, but the effect was not as strong as for the parametric synthesis.

Index Terms: audio-visual speech perception, multimodal integration, human-machine interaction, audio-visual speech synthesis

1. Introduction
Asking questions to obtain information or to establish social interaction is one of the most fundamental functions of human dialogues. In human-machine spoken language interaction, asking questions is obviously important from both a user and a system perspective. For the user, asking for information is a natural and easy way of accessing data, while the ability of the system to handle questions is fundamental if the user is to be given the initiative. For the system, posing appropriate questions is necessary for guiding the user, for maintaining the dialogue and for error handling. However, question management is crucial not only for accessing information, but also for creating a smooth flow of the dialogue between the user and the system. In this regard, not only are information-related questions important, but also the management of social-related questions. There is now considerable evidence which suggests that we relate socially to computers and especially embodied computer agents in much the same way as we relate to other humans (e.g. [1][2][3]).

The variability of intonation in asking questions is also a topic of much interest. 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 [4]. In very general terms, the most commonly described tonal characteristics for questions are high final pitch and overall higher pitch [5]. 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 [6]. In Swedish, an optional final rise can be of importance for signaling interrogative mode [7].

The prosodic features of the ends of phrases can also reveal characteristics about 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. [8][9][10]).

In a recent study of phrase-final features in a set of 200 wh-questions extracted from a large corpus of computer-directed spontaneous speech in Swedish, it was found that final rises occurred in 22 percent of the utterances [11]. Moreover, there was an indication that the questions ending in a final rise were more oriented to signaling a social interest while those with a final low were more oriented to a request for specific information. These results are consistent with a study on German spontaneous speech in which Kohler [12] proposes that “rising pitch expresses friendliness, interest and openness towards the addressee, while falling pitch focuses on routine, lack of interest and categoricalness” (p. 207).

In an effort to explore the effects of audio-visual cues on the perception of question intonation in Swedish an experiment was carried out to test if visual cues such as a smile, vertical nods, eye narrowing and eyebrow lowering could influence the perception of question and statement intonation in Swedish [7]. Results showed only a marginal influence of the visual cues. While the hypothesized cues for declarative mode (smile, short head nod and eye narrowing) reinforced declarative intonation, the hypothesized cues for interrogative mode (slow head nod and eyebrow lowering) led to more ambiguity in the responses. Similar results were obtained for English by Srinivasan and Massaro [13]. Although they were able to demonstrate that the visual cues of eyebrow raising and head tilting synthesized based on a natural model reliably conveyed question intonation, their experiments showed a weak visual effect relative to a strong auditory effect of intonation.

In view of these results, it seemed worthwhile to explore the possibility of visual cues influencing the type or category of the question such as an information question or a social question, rather than using visual cues to change declarative intonation to an interrogative percept or vice-versa. Specifically, this paper builds on the results of [11] and investigates the effect of adding visual cues to the auditory stimuli which were judged as most friendly and least friendly in the previous study. The visual cues were manifested by a talking head using two types of audio-visual synthesis, namely parametric and data-driven visual synthesis.
2. Experiment 1 (parametric synthesis)

2.1. Method

In the previous study [11], F0 peak location and F0 peak height were manipulated in the question “Vad heter du?” (What is your name?), which was also a common question posed by users when interacting with the animated agent in the dialogue system. In these experiments, a late, high peak was generally perceived as expressing more friendliness than an earlier, lower peak. Two representative stimuli from the earlier study were selected for the present study. The stimulus illustrated in Figure 1 was judged as least friendly while the one illustrated in Figure 2 was judged as most friendly.

In order to test the influence of visual cues on the perception of friendliness, two configurations combining different facial gestures were synthesized using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool. The two configurations were designed to reinforce the two audio examples described above, the low peak in the early focal position (information-oriented) and the high peak in the late focal position (friendly/social-oriented). For the information-oriented configuration, an early nod and a lowering gesture of the eyebrows were synchronized with the early focal accent (F0 peak) on the second syllable. For the friendly/social-oriented configuration, a late nod and a raising gesture of the eyebrows were synchronized with the late focal accent (F0 peak) on the final syllable. In addition, a smile was added throughout the utterance with increasing amplitude at the end of the utterance after the nod. Samples of the configurations are shown in Figure 3.

The two audio configurations were combined with the two video configurations making four stimuli. The stimuli were then converted to video files. A perception test was carried out in which the files were played using Windows Media Player and projected onto a screen in a classroom.

In order to test the influence of visual cues on the perception of friendliness, two configurations combining different facial gestures were synthesized using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool. The two configurations were designed to reinforce the two audio examples described above, the low peak in the early focal position (information-oriented) and the high peak in the late focal position (friendly/social-oriented). For the information-oriented configuration, an early nod and a lowering gesture of the eyebrows were synchronized with the early focal accent (F0 peak) on the second syllable. For the friendly/social-oriented configuration, a late nod and a raising gesture of the eyebrows were synchronized with the late focal accent (F0 peak) on the final syllable. In addition, a smile was added throughout the utterance with increasing amplitude at the end of the utterance after the nod. Samples of the configurations are shown in Figure 3.

The two audio configurations were combined with the two video configurations making four stimuli. The stimuli were then converted to video files. A perception test was carried out in which the files were played using Windows Media Player and projected onto a screen in a classroom.

The audio was played through high-quality loudspeakers. 27 native Swedish subjects were presented with three tokens of each of the four stimuli in random order. The subjects were asked to rate each stimulus on an unnumbered four-point scale where the endpoints were “friendly” and “less friendly.” Each stimulus was played twice in succession.

2.2. Results

The results showing mean scores for the subjects as a group are presented in Figure 4. It is clear that the consistent stimuli, where both audio and visual cues are intended to convey the same attitude, are perceived as conveying a low vs. a high degree of friendliness. For the inconsistent stimuli, the visual cues outweigh the auditory cues (i.e. the friendly face combined with the early, low peak received a higher rating of
friendliness than did the info-face combined with the late, high peak).

While these results tell us how the subjects performed as a group, they do not reveal if the individual subjects actually integrated the cues or if some of the subjects were more sensitive to visual cues and some were more sensitive to auditory cues. In order to examine each individual subject’s responses, the individual responses were converted to rank order. In the event of a tie, the stimuli were each given the same score halfway between the two rankings. For 18 of the 27 subjects the stimuli were ranked in the following order from least friendly to most friendly: early low peak and info-face, late high peak and info-face, early low peak and friendly face, and late high peak and friendly face. This ranking comprised the median ranking for the group as shown in Figure 5.

Two subjects showed no differences in ranking between different auditory stimuli for the same face configurations, and three subjects showed no difference in ranking between the different auditory stimuli for the info-face, but did rank the auditory stimuli differently for the friendly face. These results show evidence that most of the subjects integrated the auditory and visual modalities in perceiving friendliness. There were, however, a few subjects who showed more sensitivity to visual cues than to auditory cues.

3. Experiment 2 (data-driven synthesis)

3.1. Method

A different way of obtaining visual stimuli is by using data-driven visual synthesis. Facial movement data was collected by recording the positions of infrared markers on the face of an actor who was instructed to produce short sentences with different emotions [14]. The 3D coordinates for each marker were registered and this information was then used to drive a talking head based on the MPEG 4 facial animation standard [15]. Using the databases of different emotions results in talking head animations which differ in articulation and visual expression. For the current experiment, databases of angry, happy and neutral emotions were used to synthesize the same utterance as in the previous experiment, “Vad heter du?” (What is your name?). Samples of the three versions of the visual stimuli are presented in Figure 6. As in the previous experiment, the three versions of the visual synthesis were combined with two audio configurations: low, early pitch peak and high, late pitch peak resulting in six stimuli. There were some differences in the audio stimuli between experiment 1 and 2 in that for experiment 2 the stimuli were created using intonation rules resulting in a deaccentuation of “heter” in the temporal domain in the non-focal version. The intonation contours were otherwise very similar for the two experiments. A perception test using these six stimuli was carried out in the same way and on the same occasion as the previous experiment using the same 27 subjects.

3.2. Results

The results showing mean scores for the subjects as a group are presented in Figure 7. It is quite clear that the face synthesized from the angry database elicited the lowest friendliness score. However, there is still evidence of interaction from the audio, as the angry face with the late, high peak received a higher friendliness score than did the angry face with the early, low peak. The faces from the other databases (happy and neutral) elicited more friendliness responses, but neither combination of face and audio received as high a friendliness score as did the optimal stimulus from the parametric synthesis experiment.

As in experiment 1, the individual responses were converted to rank order. A greater variation in the order was apparent as shown in Figure 8. However, the happy face combined with the late high peak received the highest median rank of all the stimuli.
5. Acknowledgements

This research was carried out with partial support from the MUSCLE Network of Excellence (NoE) funded by the European Commission under FP6 (Contract 507752). Special thanks to Jonas Beskow, Mikael Nordenberg and Magnus Nordstrand for creating the data-driven visual synthesis.

6. References