Audiovisual Representation of Prosody in Expressive Speech Communication

Björn Granström and David House

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

Abstract

Prosody in a single speaking style – often read speech – has been studied extensively in acoustic speech. During the past few years we have expanded our interest in two directions: 1.) Prosody in expressive speech communication and 2.) Prosody as an audiovisual expression. Understanding the interactions between visual expressions (primarily in the face) and the acoustics of the corresponding speech presents a substantial challenge. Some of the visual articulation is for obvious reasons tightly connected to the acoustics (e.g. lip and jaw movements), but there are other articulatory movements that do not show up on the outside of the face. Furthermore, many facial gestures used for communicative purposes do not affect the acoustics directly, but might nevertheless be connected on a higher communicative level in which the timing of the gestures could play an important role. In this presentation we will give some examples of recent work, primarily at KTH, addressing these questions. We will report on methods for the acquisition and modeling of visual and acoustic data, and some evaluation experiments in which audiovisual prosody is tested. The context of much of our work in this area is to create an animated talking agent capable of displaying realistic communicative behavior and suitable for use in conversational spoken language systems, e.g. a virtual language teacher.

1. Introduction

In our interaction with others, we make use of all of our sensory modalities as we communicate and exchange information. Our senses are exceptionally well-adapted for interaction, and our neurophysiology enables us to effortlessly integrate information from different modalities fusing data to optimally meet the current communication needs. A full account of the speech communication process must therefore include multiple modalities. While the auditory modality often provides the phonetic information necessary to convey a linguistic message, the visual modality can qualify the auditory information providing segmental cues for place of articulation, prosodic information concerning prominence and phrasing and extralinguistic information such as signals for turn-taking, emotions and attitudes. Although these observations are not novel, prosody research has traditionally concentrated on describing the acoustics of prominence and phrasing in restricted speaking styles. While different speaking styles including expressive and emotional speech have received more attention during recent years, we still lack basic knowledge concerning how auditory and visual signals interact to signal communicative functions in expressive speech. One reason for this is the primary status of auditory speech. Another reason is the relatively more complicated analysis and synthesis of visual speech. Most of the work that has been done in multimodal speech perception has concentrated on segmental cues in the visual modality.

The visible articulatory movements are mainly those of the lips, jaw and tongue. However, these are not the only visual information carriers in the face during speech. Much information related to e. g. phrasing, stress, intonation and emotion are expressed by for example nodding of the head, raising and shaping of the eyebrows, eye movements and blinks. These kinds of facial actions should also be taken into account in a visual speech synthesis system, not only because they may transmit important non-verbal information, but also because they make the face look alive. These movements are more difficult to model in a general way than the articulatory movements, since they are optional and highly dependent on the speaker's personality, mood, purpose of the utterance, etc.

As we attempt to take advantage of the effective communication potential of human conversation in spoken dialogue systems, we see an increasing need to embody the conversational partner using audiovisual verbal and non-verbal communication implying the use and integration of both audio and visual modalities [1]. Effective interaction in dialogue systems involves both the presentation of information and the flow of interactive dialogue. A talking animated agent can provide the user with an interactive partner whose goal is to take the role of the human agent. An effective agent is one who is capable of supplying the user with relevant information, can fluently answer questions concerning complex information and can ultimately assist the user in a decision making process through the interactive flow of conversation. The use of the talking head also aims at increasing effectiveness by building on the user's social/communicative skills to improve the flow of the dialogue. Visual cues to feedback, turntaking and signalling the system's internal state (the thinking metaphor) are key aspects of effective interaction.

In this paper we will report on methods for the acquisition and modeling of visual and acoustic data, and review several evaluation experiments in which various aspects of audiovisual prosody have been tested. We will also briefly report on some experimental applications in which audiovisual prosody can be a key element in improving the performance of a talking face in a dialogue system context.

2. Data acquisition

For the analysis of acoustic prosodic measurements there exists well established (semi) automatic techniques operating on the audio signal. Analysis of video signals poses a much more complicated problem. To automatically extract important facial movements we have employed a motion capture procedure.

We wanted to be able to obtain both articulatory data as well as other facial movements at the same time, and it was crucial that the accuracy in the measurements was good enough for resynthesis of an animated head. Optical motion tracking systems are gaining popularity for being able to handle the tracking automatically and for having good accuracy as well as good temporal resolution. The Qualisys system that we use has an accuracy better than 1 mm with a temporal resolution of 60 Hz. The data acquisition and processing is very similar to earlier facial measurements carried out at CTT by i.e [2]. The recording set-up can be seen in Fig. 1.

Figure 1: *Data collection setup with video and IR-cameras, microphone and a screen for prompts.*

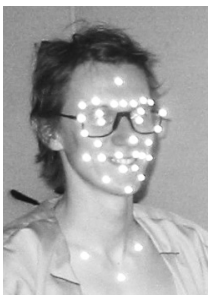

Figure 2: *Test subject with the IR-reflecting markers glued to the face.*

The subject could either pronounce sentences presented on the screen outside the window or be engaged in a (structured) dialogue with another person as shown in the figure. In the present set-up, the second person can not be recorded with the Qualisys system but is only video recorded. Audio data was recorded on DAT-tape and visual data was recorded using video and the optical motion tracking system. A synchronisation signal produced by the Qualisys system was recorded on the video and on one channel of the DAT-tape enabling audio and visual data to be matched. By attaching infrared reflecting markers to the subject's face (see Fig. 2), the system is able to register the 3D coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms. We used 30 markers to register lip movements as well as other facial movements such as eyebrows, cheek, chin and eyelids. Additionally we placed three markers on the chest to register head movements with respect to the torso. A pair of spectacles with four markers attached was used as a reference to be able to factor out head and body movements when looking at the facial movements specifically.

3. Emotions, articulation and speech segments

Most systems for visual face synthesis are modelled on non-expressive speech, i.e. the material is read with a neutral voice and facial expression. However, expressiveness might affect articulation and how we produce speech a great deal, and an articulatory parameter might behave differently under the influence of different emotions. For example Fonâgy [3] showed how intraoral articulation, e.g. tongue movement, was affected by the expression of emotions. Better knowledge about this behaviour will help us adjust the articulatory rules controlling the articulation of an animated talking head.

There have been attempts to take into consideration how articulation may change depending on speaker or style and make use of that knowledge in audiovisual synthesis. Pelachaud et al. [4] proposed a method where they could define various speaker characteristics such as speech-rate and timing issues. They also described how this model could generate emotions and expressions in the face, but the articulation was not directly affected by these rules.

In this section we will illustrate how articulation is affected in expressive speech. This work has been carried out within the EU PF-Star project which aims at establishing future activities in the field of multi-sensorial and multi-lingual communication. This will be achieved by providing technological baselines, comparative evaluations, and assessments of prospects of core technologies, which future research and development efforts can build on. In the first phase of the project, the work mainly consisted of database collection using the methods described in the preceding section. Here we will only present some striking observations from the first recordings [5]. Part of the database consisted of semantically neutral utterances, e.g. numbers all pronounced in several different expressive states. In figure 3 an example of the analysis is presented. The mean position of the left mouth corner measured in the middle of all the vowels in the material is displayed as a cross, the size of one standard deviation.

Obviously the expressive state, in some instances, has a stronger influence on the articulation than do the different vowels. It is also interesting to note that the neutral pronunciation displays a pattern different from all the (acted) expressive speech versions, with very little variation between vowels and a presumably small mouth opening. In this study we did not look into the dynamic influence on the segmental articulation in the expressive speech. How much could be described by relatively stable settings and what is best described by expressive gestures is the topic of some of our current research.

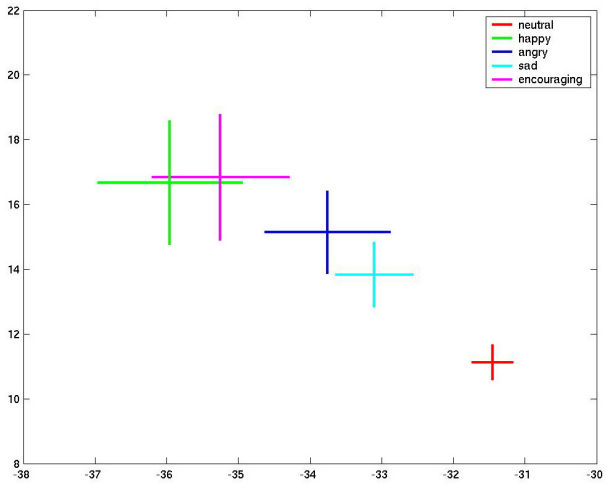

Figure 3: *Horizontal and vertical displacements of the left mouth corner for different acted expressive states. The crosses refer to the expressions: glad, encouraging angry, sad, neutral from top left to bottom right (2mm between scale markers)*

4. Implementation

4.1. Generic coding

Animated synthetic talking faces and characters have been developed using a number of different techniques and for a variety of purposes during the past two decades. Our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework [6]. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account [7]. We employ a generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation [8]. Based on concepts first introduced by Parke [9], we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. One critical difference from Parke's system, however, is that we have de-coupled the model definitions from the animation engine, thereby greatly increasing flexibility. The different models shown in figure 4 are all dynamically controlled by the same control code. Modules for connecting our model to general coding schemes included in the MPEG-4 standard have been developed.

Figure 4: *Some different versions of the KTH talking head*

4.2. Expressive wrinkles

In a recent study we wanted to include the possibility of adding real time dynamic wrinkles to the computer-generated face. The purpose is to extend an existing virtual face to make it more realistic and expressive. Wrinkles modelled as a fine grained mesh of polygons would today be totally prohibitive, from a computational point of view, so we implemented a different, but flexible solution [10]. The algorithm uses a technique called bump mapping to visualize the wrinkles. The implemented prototype shows that real time dynamic wrinkles can be implemented using existing hardware. The bump mapping algorithm has a few limitations, especially regarding the lighting model, but these limitations are hardly noticed in this fairly simple application, see figure 5, where only the wrinkles connected to raised eyebrows for e.g. surprise are modelled.

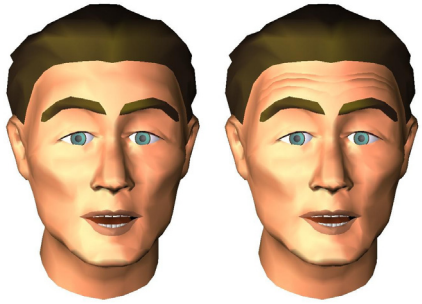

Figure 5: *The face model with and without wrinkles*

5. Visual cues for prominence

5.1. Eyebrow movements

The interaction between acoustic intonational gestures (F0) and eyebrow movements has been studied in production in e.g. [11]. A preliminary hypothesis is that a direct coupling is very unnatural, but that acoustic prominence signaling by raising F0 and visual signaling by eyebrow movement may co-occur.

In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish [12], a test sentence was created using our audio-visual text-to-speech synthesis in which the acoustic cues and lower face visual cues were the same for all stimuli. Articulatory movements were created by using the text-to-speech rule system. The upper face cues were eyebrow movement where the eyebrows were raised on successive words in the sentence. The movements were created by hand editing the eyebrow parameter. The degree of eyebrow raising was chosen to create a subtle movement that was distinctive although not too obvious. The total duration of movement was 500 ms and comprised a 100 ms dynamic raising part, a 200 ms static raised portion and a 200 ms dynamic lowering part. In the stimuli, the acoustic signal was always the same, and the sentence was synthesized as one phrase. Six versions were included in the experiment: one with no eyebrow movement and five where eyebrow raising was placed on one of the five content words in the test sentence. The words with concomitant eyebrow movement were generally perceived as more prominent than words without the movement. This tendency was even greater for a subgroup of non-native (L2) listeners. The mean increase in prominence response following an eyebrow movement was 24 percent for the Swedish native (L1) listeners and 39 percent for the L2 group. One example result is shown in figure 6.

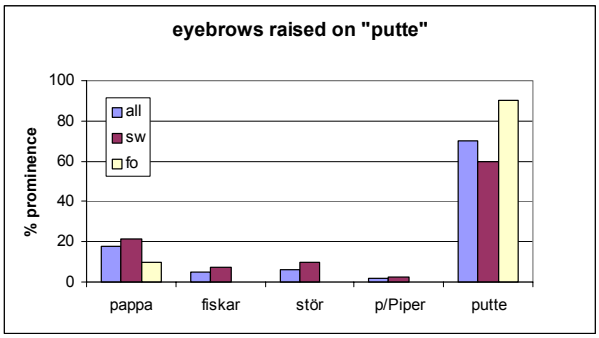

Figure 6. *Prominence responses in percent for each content word for the acoustically neutral reading of the stimulus sentence, "När pappa fiskar stör p/Piper Putte," with eyebrow movement on "Putte". Subjects are grouped as all, Swedish (sw) and foreign (fo).*

5.2. Head nods, eyebrow movement and timing

In another study [13] both eyebrow and head movements were tested as potential cues to prominence. The goal of the study was two-fold. First of all we wanted to see if head movement (nodding) is a more powerful cue to prominence than is eyebrow movement by virtue of a larger movement. Secondly, we wanted to test the perceptual sensitivity to the timing of both eyebrow and head movement in relationship to the syllable.

As in the previous experiment, our rule-based audiovisual synthesizer was used for stimuli preparation. The test sentence used to create the stimuli for the experiment was the same as that used in an earlier perception experiment designed to test acoustic cues only [14]. The sentence, *Jag vill bara flyga om vädret är perfekt* (I only want to fly if the weather is perfect) was synthesized with focal accent rises on both *flyga* (fly) (Accent 2) and *vädret* (weather) (Accent 1). The F0 rise excursions corresponded to the stimulus in the earlier experiment which elicited nearly equal responses for *flyga* and *vädret* in terms of the most prominent word in the sentence. The voice used was the Infovox 330 Ingmar MBROLA voice.

Eyebrow and head movements were then created by hand editing the respective parameters. The eyebrows were raised to create a subtle movement that was distinctive although not too obvious. In quantitative terms the movement comprised 4% of the total possible movement. The head movement was a slight vertical lowering comprising 3% of the total possible vertical head rotation. Statically, the displacement is difficult to perceive, while dynamically, the movement is quite distinct. The total duration of both eyebrow and head movement was 300 ms and comprised a 100 ms dynamic onset, a 100 ms static portion and a 100 ms dynamic offset.

Two sets of stimuli were created: set one in which both eyebrow and head movement occurred simultaneously and set two in which the movements were separated and potentially conflicting with each other. In set one, six stimuli were created by synchronizing the movement in stimulus 1 with the stressed vowel of *flyga*. This movement was successively shifted in intervals of 100 ms towards *vädret* resulting in the movement in stimulus 6 being synchronized with the stressed vowel of *vädret*. In set two, stimuli 1-3 were created by fixing the head movement to synchronize with the stressed vowel of *vädret* and successively shifting the eyebrow movements from the stressed vowel of *flyga* towards *vädret* in steps of 100 ms. Stimuli 4-6 were created by fixing the eyebrow movement to *vädret* and shifting the head movement from *flyga* towards *vädret*. The acoustic signal and articulatory movements were the same for all stimuli. A schematic illustration of the stimuli is presented in Figure 7.

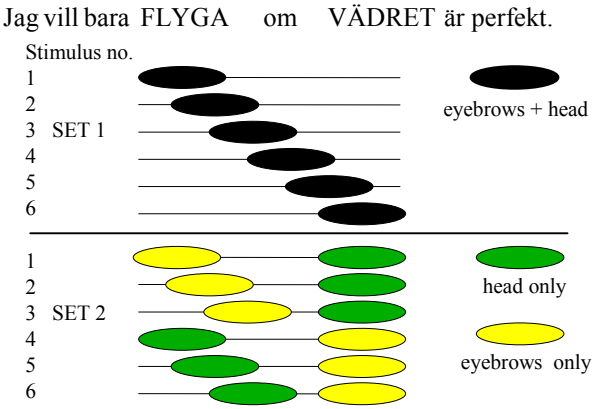

Figure 7. *Schematic illustration of face gesture timing*

The results from stimulus set 1 where eyebrow and head movements occurred simultaneously clearly reflect the timing aspect of these stimuli as can be seen in Figure 8 where percent votes for *vädret* increase successively as movement is shifted in time from *flyga* to *vädret*.

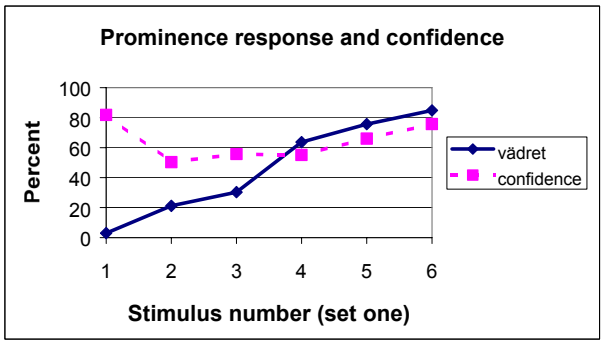

Figure 8. *Results for stimulus set one showing prominence response for* vädret *and confidence in percent.*

It is clear from the results that combined head and eyebrow movements of the scope used in the experiment are powerful cues to prominence when synchronized with the stressed vowel of the potentially prominent word and when no conflicting acoustic cue is present. Sensitivity to the timing of these movements seems to be on the order of 100 ms. However, there is a tendency for integration of the movements to the nearest potentially prominent word, thus accounting for the jump in prominence response between stimulus 3 and 4 in set 1. This integration is consistent with the results of similar experiments using visual and auditory segmental cues [15].

As could be expected, the results from set 2, where eyebrow and head movement were in conflict, showed more stimulus ambiguity. Head movement, however, demonstrated a slight advantage in signalling prominence. This advantage can perhaps be explained by the fact that the movement of the head may be visually more salient than the relatively subtle eyebrow movement. The advantage might even be increased if the head is observed from a greater distance. In an informal demonstration, where subjects were 2 to 5 meters from the screen using the same head size as in the current experiment, head-movement advantage was quite pronounced.

A number of questions remain to be answered, as a perception experiment of this type is necessarily restricted in scope. Amplitude of movement was not addressed in this investigation. If, for example, eyebrow movement were exaggerated, would this counterbalance the greater power of head movement? A perhaps even more crucial question is the interaction between the acoustic and visual cues. There was a slight bias for *flyga* to be perceived as more prominent (one subject even chose *flyga* in 11 of the 12 stimuli), and indeed the F0 excursion was greater for *flyga* than for *vädret*, even though this was ambiguous in the previous experiment. In practical terms of multimodal synthesis, however, it will probably be sufficient to combine cues, even though it would be helpful to have some form of quantified weighting factor for the different acoustic and visual cues.

Duration of the eyebrow and head movements is another consideration which was not tested here. It seems plausible that similar onset and offset durations (100 ms) combined with substantially longer static displacements would serve as conversational signals rather than as cues to prominence. In this way, non-synchronous eyebrow and head movements can be combined to signal both prominence and e.g. feedback giving or seeking. Some of the subjects also commented that the face seemed to convey a certain degree of irony in some of the stimuli in set 2, most likely in those stimuli with non-synchronous eyebrow movement. Experimentation with potential cues for feedback seeking was pursued in the study reported on in the next section.

6. Visual cues for feedback

The use of a believable talking head can trigger the user's social skills such as using greetings, addressing the agent by name, and generally socially chatting with the agent. This was clearly shown by the results of the public use of the August system during a period of six months [16]. These promising results have led to more specific studies on visual cues for feedback [17] in which smile, for example, was found to be the strongest cue for affirmative feedback. Work on turntaking regulation, feedback seeking and giving and the signalling of the system's internal state will enable us to improve the gesture library available for the animated talking head and continue to improve the effectiveness of multimodal dialogue systems.

One of the central claims in many theories of conversation is that dialogue partners seek and provide evidence about the success of their interaction [18][19][20]. That is, partners tend to follow a proof procedure to check whether their utterances were understood correctly or not and constantly exchange specific forms of feedback that can be affirmative ('go on') or negative ('do not go on') Previous research has brought to light that conversation partners can monitor the dialogue this way on the basis of at least two kinds of features not encoded in the lexico-syntactic structure of a sentence: namely, prosodic and visual features. First, utterances that function as negative signals appear to differ prosodically from affirmative ones in that they are produced with more 'marked' settings (e.g. higher, louder, slower) [21][22]. Second, other studies reveal that, in face-to-face interactions, people signal by means of facial expressions and specific body gestures whether or not an utterance was correctly understood [23].

Given that current spoken dialogue systems are prone to error, mainly because of problems in the automatic speech recognition (ASR) engine of these systems, a sophisticated use of feedback cues from the system to the user is potentially very helpful to improve human-machine interactions as well [24]. There are currently a number of advanced multimodal user interfaces using talking heads that can generate audiovisual speech along with different facial expressions. However, while such interfaces can be accurately modified in terms of a number of prosodic and visual parameters, there are as yet no formal models that make explicit how exactly these need to be manipulated to synthesize convincing affirmative and negative cues.

One interesting question, for instance, is what the strength relation is between the potential acoustic and visual cues. The goal of our study [17] was to gain more insight into the relative importance of specific prosodic and visual parameters for giving feedback on the success of the interaction. In this study, use is made of a talking head whose prosodic and visual features are orthogonally varied in order to create stimuli that are presented to subjects who have to respond to these stimuli and judge them as affirmative or negative backchanneling signals.

The stimuli consisted of an exchange between a human, who was intended to represent a client, and the face, representing a travel agent. An observer of these stimuli could only hear the client's voice, but could both see and hear the face. The human utterance was a natural speech recording and was exactly the same in all exchanges, whereas the speech and the facial expressions of the travel agent were synthetic and variable. The fragment that was manipulated, always consisted of the following two utterances:

Human: "Jag vill åka från Stockholm till Linköping."
("I want to go from Stockholm to Linköping.")
Head: "Linköping."

The stimuli were created by orthogonally varying 6 parameters, shown in table 1, using two possible settings for each parameter: one which was hypothesised to lead to affirmative feedback responses, and one which was hypothesised to lead to negative responses.

Table 1. *Settings of the different parameters, hypthesised to support affirmative or negative feedback*

	Affirmative setting	Negative setting
Smile	Head smiles	Neutral expression
Head move.	Head nods	Head leans back
Eyebrows	Eyebrows rise	Eyebrows frown
Eye closure	Eyes narrow slightly	Eyes open widely
F0 contour	Declarative	Interrogative
Delay	Immediate reply	Delayed reply

The parameter settings were largely created by intuition and observing human productions. However, the affirmative and negative F0 contours were based on two natural utterances In figure 9 an example of the all-negativ and all-affirmative face can be seen.

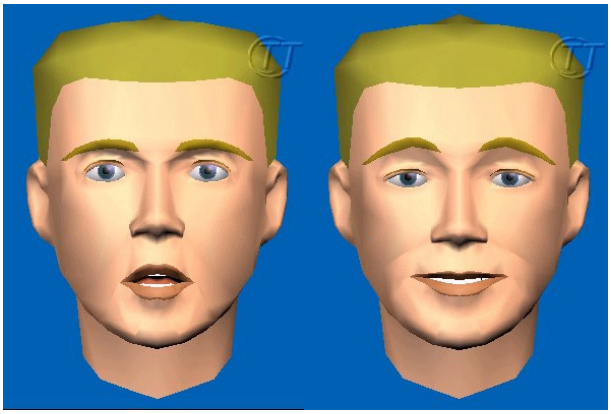

Figure 9. *The all-negative and all-affirmative faces sampled in the end of the first syllable of "Linköping"*

The actual testing was done via a group experiment using a projected image on a large screen. The task was to respond to this dialogue exchange in terms of whether the head signals that he understands and accepts the human utterance, or rather signals that the head is uncertain about the human utterance. In addition, the subjects were required to express on a 5-point scale how confident they were about their response. A detailed description of the experiment and the analysis can be found in [17]. Here, we would only like to highlight the strength of the different acoustic and visual cues. In figure 10 the mean difference in response value (the response weighted by the subjects' confidence ratings) is presented for negative and affirmative settings of the different parameters.

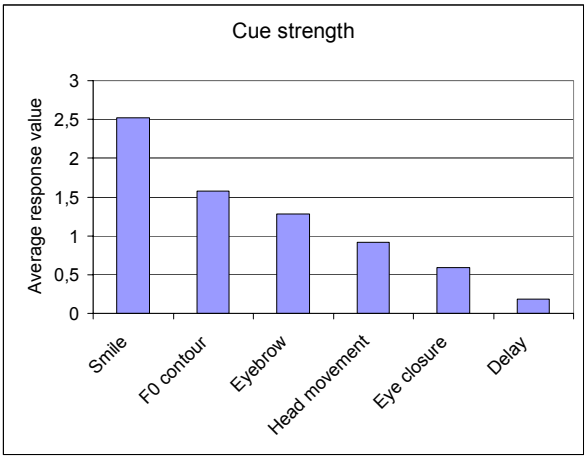

Figure 10. *The mean response value difference for stimuli with the indicated cues set to their hypothesized affirmative and negative value.*

The effects of Eye_closure and Delay are not significant, but the trends observed in the means are clearly in the expected direction. There appears to be a strength order with Smile being the most important factor, followed by F0_contour, Eyebrow, Head_movement, Eye_closure and Delay.

This study clearly shows that subjects are sensitive to both acoustic and visual parameters when they have to judge utterances as affirmative or negative feedback signals. One obvious next step is to test whether the fluency of human-machine interactions is helped by the inclusion of such feedback cues in the dialogue management component of a system.

7. Visual cues for questions

In distinguishing questions from statements, prosody has a well established role, especially in cases such as echo questions where there is no syntactic cue to the interrogative mode. Almost without exception this has been shown only for the auditory modality. Inspired by the results of the positive and negative feedback experiment presented in the previous section, an experiment was carried out to test if similar visual cues could influence the perception of question and statement intonation in Swedish [25]. Hypothesized cues for interrogative mode were in analogy to the negative feedback cues and consisted of a slow up-down head nod and eyebrow lowering. The hypothesized cues for the declarative mode were in analogy to the positive feedback cues and consisted of a smile, a short up-down head nod and eye narrowing (see figure 9). The declarative head nod was of the same type as was used in the prominence experiments reported in section 5 above. 12 different intonation contours were used in the stimuli ranging from a low final falling contour (clearly declarative) to a high final rise (clearly interrogative).

The influence of the visual cues on the auditory cues was marginal. 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 [26]. 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. This weak visual effect remained despite attempts to enhance the visual cues and make the auditory information more ambiguous.

The dominance of the auditory cues in the context of these question/statement experiments may indicate that question intonation may be less variable than visual cues for questions, or we simply may not yet know enough about the combination of visual cues and their timing in signalling question mode to successfully override the auditory cues. Moveover, a final high rising intonation is generally a very robust cue to question intonation, especially in the context of perception experiments with binary response alternatives.

8. Experimental applications

8.1. Multimodal dialogue systems

One example of using the talking head in an experimental dialogue system is the AdApt project. The practical goal of the project is to build a system in which a user can collaborate with an animated agent to solve complicated tasks [27][28]. We have chosen a domain in which multimodal interaction is highly useful, and which is known to engage a wide variety of people in our surroundings, namely, finding available apartments in Stockholm. In the AdApt project, the agent has been given the role of asking questions and providing guidance by retrieving detailed authentic information about apartments. The user interface can be seen in figure 11.

Because of the conversational nature of the AdApt domain, the demand is great for appropriate interactive signals for encouragement, affirmation, confirmation and turntaking.

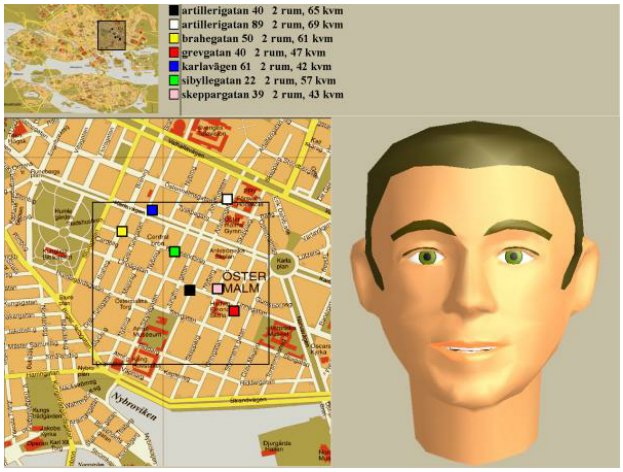

Figure 11. *The agent Urban in the AdApt domain.*

8.2. Communication aids

The speech intelligibility of talking animated agents, as the ones described above, has been tested within KTH Teleface project [29] and the EU project, Synface [30] . The projects focus on the use of multi-modal speech technology for hearing-impaired persons. The projects evaluated the increased intelligibility hearing-impaired persons experience from an auditory signal when it is complemented by a synthesised face. A demonstrator of a system for telephony with a synthetic face that articulates in synchrony with a natural voice is currently being implemented (see figure 12). While the emphasis in this project has been on intelligibility, the visual signalling of such conversational features as turntaking is used to smooth the communicative flow, and help minimize the problem of the delay that the reconstruction of the face image necessarily implies.

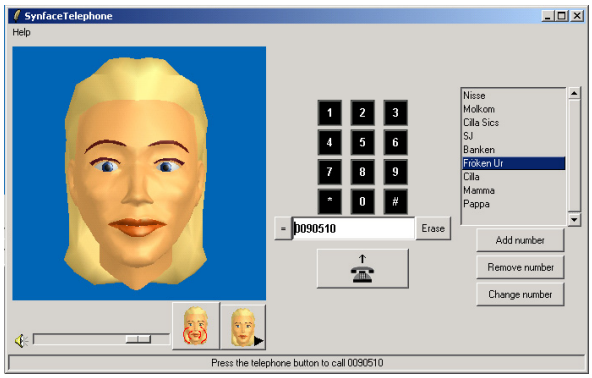

Figure 12. *Telephone interface for SYNFACE.*

8.3. Language tutor

The effectiveness of language teaching is often contingent upon the ability of the teacher to create and maintain the interest and enthusiasm of the student. The success of second language learning is also dependent on the student having ample opportunity to work on oral proficiency training with a tutor. The implementation of animated agents as tutors in a multimodal spoken dialogue system for language training holds much promise towards fulfilling these goals [31]. Different agents can be given different personalities with various attitudes and roles, which should increase the interest of the students. Many students may also be less bashful about interacting with an agent who corrects their pronunciation errors than they would be making the same errors and interacting with a human teacher. Instructions to improve pronunciation often require reference to phonetics and articulation in such a way that is intuitively easy for the student to understand. Pronunciation training in the context of a dialogue also automatically includes training of individual phonemes, sentence prosody and communication skills. In this context, visual prosody can function as an important aid to intonation training, as well as facilitating the flow of the training dialogue.

9. Acknowledgements

Much of the work presented in this overview has been done by other members of the CTT multimodal communication group including, Jonas Beskow, Loredana Cerrato, Olov Engwall, Mikael Nordenberg, Magnus Nordstrand and Gunilla Svanfeldt, which is gratefully acknowledged. The work has been supported by the EU/IST- projects SYNFACE and PF-Star, and CTT, the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA, KTH and participating Swedish companies and organizations.

10. References

[1] Massaro, D. W., 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle*. Cambridge, MA: The MIT Press.

[2] Beskow, J.; Engwall, O.; Granström, B., 2003. Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. of ICPhS 2003*. Barcelona, Spain.

[3] Fonâgy, I., 1976. La mimique buccale. *Phonetica* 33, 31–44.
[4] Pelachaud, C.; Badler, N.; Steedman, M., 1996. Generating Facial Expressions for Speech. *Cognitive Science* 20, 1–46.
[5] Nordstrand, M.; Svanfeldt, G.; Granström, B.; House, D., 2003. Measurements of Articulatory Variation and Communicative Signals in Expressive Speech. *Proc of AVSP'03,* 233-238.
[6] Carlson, R.; Granström, B., 1997. Speech Synthesis. In *The Handbook of Phonetic Sciences*, W Hardcastle and J Laver (eds.). Oxford: Blackwell Publishers Ltd,. 768-788.
[7] Beskow, J., 2003. *Talking heads – models and applications for multimodal speech synthesis*. PhD thesis, TMH/KTH.
[8] Beskow, J., 1997. Animation of talking agents. In *Proceedings of ESCA workshop on audio-visual speech processing*, Rhodes: Greece, 149-152.
[9] Parke, F.I., 1982. Parameterized models for facial animation. *IEEE Computer Graphics*. 2(9), 61-68.
[10] Nordenberg, M., 2003. *Modelling and rendering dynamic wrinkles in a virtual face.*TMH/KTH MSc thesis. (available at http://www.speech.kth.se/qpsr/masterproj/)
[11] Cavé, C.; Guaïtella, I.; Bertrand, R.; Santi, S.; Harlay, F.; Espesser, R., 1996. About the relationship between eyebrow movements and F0 variations. In *Proceedings ICSLP 96,* H.T. Bunnell, and W. Idsardi (eds.). Philadelphia, PA, USA. 2175-2178.
[12] Granström, B.; House, D.; Lundeberg, M., 1999. Prosodic Cues in Multimodal Speech Perception. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99*). San Francisco, 655-658.
[13] House, D.; Beskow, J.; Granström, B., 2001. Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proc of Eurospeech 2001*, 387-390.
[14] House, D., 2001. Focal accent in Swedish: Perception of rise properties for accent 1. In *Nordic Prosody 8,* W. van Dommelen, and Fretheim, T. (eds.). Frankfurt: Peter Lang. 127-135.
[15] Massaro, D. W., Cohen, M. M. and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am. 100*. 1777-1786.
[16] Bell, L.; Gustafson, J., 1999. Interaction with an animated agent in a spoken dialogue system, *Proc of Eurospeech '99*. Budapest, 1143-1146.
[17] Granström, B.; House, D.; Swerts, M., 2002. Multimodal feedback cues in human-machine interactions. In *Proc of the Speech Prosody 2002 Conference,* B. Bel & I. Marlien (eds.). Aix-en-Provence: Laboratoire Parole et Langage, 347-350.
[18] Clark, H.H.; Schaeffer E.F., 1989. Contributing to discourse. *Cognitive Science 13*, 259-294.
[19] Traum, D.R., 1994. *A computational theory of grounding in natural language conversation*. PhD thesis, Rochester.
[20] Brennan, S.E., 1990. *Seeking and providing evidence for mutual understanding.* Unpublished doctoral dissertation, Stanford University, Stanford, CA.
[21] Shimojima, A.; Katagiri, Y.; Koiso, H.; Swerts, M., 2002. Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. *Speech Communication 36* (1-2), 113-132.
[22] Krahmer, E.; Swerts, M.; Theune M.; Weegels M., 2002. The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication 36* (1-2), 133-145.
[23] Gill, S.P.; Kawamori, M.; Katagiri, Y.; Shimojima. A., 1999. Pragmatics of Body Moves. *The Proceedings of 3rd International Cognitive Technology Conference,* 345-358.
[24] Hirschberg, J.; Litman D.; Swerts, M., 2001. Identifying user corrections automatically in spoken dialogue systems. *Proc. NAACL 2001*. Pittsburg.
[25] House, D., 2002. Intonational and visual cues in the perception of interrogative mode in Swedish, In *Proceedings of ICSLP 2002*. Denver: Colorado, 1957-1960.
[26] Srinivasan, R.J.; Massaro, D.W., 2003. Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech* 46(1), 1-22.
[27] Edlund, J.; Beskow, J.; Nordstrand, M., 2002. GESOM - A Model for Describing and generating Multi-modal Output. *Proc of ISCA Workshop Multi-Modal Dialogue in Mobile Environments.*
[28] Gustafson, J., 2002. *Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction.* Doctoral Thesis. Department of Speech, Music and Hearing, KTH, Stockholm
[29] Agelfors, E.; Beskow, J.; Dahlquist, M.; Granström. B.; Lundeberg, M.; Salvi, G.; Spens, K-E.; Öhman, T., 1999. Synthetic visual speech driven from auditory speech. *Proc of AVSP 99*, 123-127
[30] Siciliano, C.; Williams, G.: Beskow, J.; Faulkner, A., 2003. Evaluation of a Multilingual Synthetic Talking Face. as a Communication Aid for the Hearing Impaired. *Proceedings of 15th International Congress of Phonetic Sciences*.
[31] Beskow, J.; Granström, B.; House, D.; Lundeberg, M., 2000. Experiments with verbal and visual conversational signals for an automatic language tutor. *Proc of InSTIL 2000*