INSIDE OUT – ACOUSTIC AND VISUAL ASPECTS OF VERBAL AND NON-VERBAL COMMUNICATION

Björn Granström and David House

Centre for Speech Technology, KTH, Stockholm, Sweden
{bjorn, davidh}@speech.kth.se

ABSTRACT

In face-to-face communication both visual and auditory information play an obvious and significant role. In this presentation we will discuss work done, primarily at KTH, that aims at analyzing and modelling verbal and non-verbal communication from a multi-modal perspective. In our studies, it appears that both segmental and prosodic phenomena are strongly affected by the communicative context of speech interaction. One platform for modelling audiovisual speech communication is the ECA, embodied conversational agent. We will describe how ECAs have been used in our research, including examples of applications and a series of experiments for studying multimodal aspects of speech communication.

Keywords: audiovisual speech, non-verbal communication, visual prosody, expressive speech

1. INTRODUCTION

In face-to-face communication both visual and auditory information play an obvious and significant role. Traditionally in phonetic research the auditory effects of speech production have been the primary object of study. However, when it comes to the non-verbal aspects of speech communication the primary nature of acoustics is not as evident, and understanding the interactions between visual expressions, dialogue functions and the acoustics of the corresponding speech presents a substantial challenge. Some of the visual articulation is for obvious reasons closely related to the speech acoustics (e.g. movements of the lips and jaw), but there is other articulatory movement affecting speech acoustics that is not visible on the outside of the face. On the other hand, 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. The context of much of our research regarding these questions is to be able to create an animated talking agent capable of displaying realistic communicative behavior and suitable for use in conversational spoken language systems.

Useful applications of talking heads include aids for the hearing impaired, educational software, audiovisual human perception experiments [18], entertainment, and high quality audiovisual text-to-speech synthesis for applications such as news reading. The use of the talking head aims at increasing effectiveness by building on the user's social skills to improve the flow of the dialogue [7]. Visual cues to feedback, turntaking and signalling the system's internal state are key aspects of effective interaction.

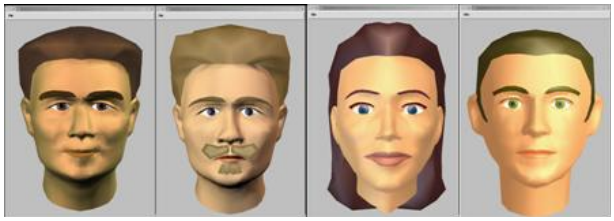
The focus of this paper is to present an overview of some of the research involved in the development of audiovisual synthesis to improve the talking head. Some examples of results and applications involving the analysis and modelling of acoustic and visual aspects of verbal and non-verbal communication are presented.

2. KTH MULTIMODAL SPEECH SYNTHESIS

The talking head developed at KTH is based on text-to-speech synthesis. Audio speech synthesis is generated from a text representation in synchrony with visual articulator movements of the lips, tongue and jaw. Linguistic information in the text is used to generate visual cues for relevant prosodic categories such as prominence, phrasing and emphasis. These cues generally take the form of eyebrow and head movements which we have termed "visual prosody" [14]. These types of visual cues with the addition of e.g. a smiling or frowning face are also used as conversational gestures to signal such things as positive or negative feedback, turntaking regulation, and the system's internal state. In addition, the head can visually signal attitudes and emotions.

Historically, our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework [8]. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account [2]. We employ a generalized parameterisation technique to adapt a static 3D wireframe of a face for visual speech animation. Based on concepts first introduced by Parke [21], we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. The parameters are designed to allow for intuitive interactive or rule-based control. The face model surfaces can be made (semi) transparent to display the internal parts of the model, including the tongue, palate, jaw and teeth [10]. The internal parts are based on articulatory measurements using MRI (Magnetic Resonance Imaging), EMA (ElectroMagnetic Articulography) and EPG (ElectroPalatoGraphy), in order to assure that the model's movements are realistic. This is of importance for language learning situations, where the transparency of the skin may be used to explain non-visible articulations [19] [11]. Several face models have been developed for different applications, some of them can be seen in Figure 1. All can be parametrically controlled by the same articulation rules.

Figure 1: Some different versions of the KTH talking head.



For stimuli preparation and explorative investigations, we have developed a control interface that allows fine-grained control over the trajectories for acoustic as well as visual parameters.

The acoustic synthesis can be exchanged for a natural utterance and synchronised to the face synthesis on a segment-by-segment basis by running the face synthesis with phoneme durations from the natural utterance. The combination of natural and synthetic speech is useful for different experiments on multimodal integration and has also been used in the SynFace project [5], where an audio controlled synthetic face is used as lip reading support for hard of hearing persons. In language learning applications this feature could be used to add to the naturalness of the tutor's voice in cases when the acoustic synthesis is judged to be inappropriate [11]. This feature is also used in a turntaking evaluation paradigm described briefly in section 6.3 [9].

3. DATA COLLECTION

More recently, we have worked on data-driven visual synthesis using an MPEG-4 compatible talking head. A data-driven approach enables us to capture the interaction between facial expression and articulation. This is especially important when trying to synthesize emotional expressions (c.f. [20]).

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.

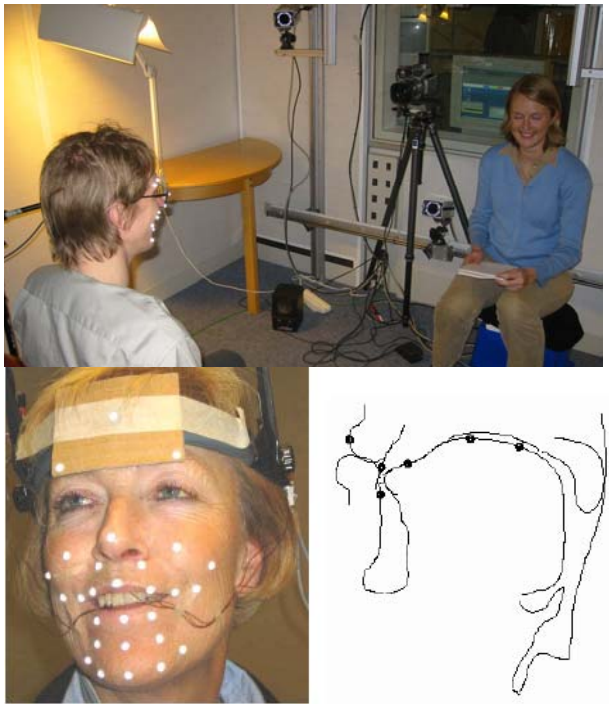
We have used an opto-electronic motion tracking system – Qualysis MacReflex – to collect multimodal corpora of expressive speech. The Qualisys system allows capturing the dynamics of emotional facial expressions by registering the 3D coordinates of a number of infrared (IR) reflective markers, with sub-millimetre accuracy, at a rate of 60 frames/second.

In a typical session, a male native Swedish amateur actor was instructed to produce 75 short sentences with the six emotions: happiness, sadness, surprise, disgust, fear and anger, plus neutral, yielding 7 x 75 recorded utterances. A total of 29 IR-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers (on the ears and on the forehead). The marker setup largely corresponds to MPEG-4 feature point (FP) configuration.

The recorded motion data was converted into an MPEG-4 facial animation parameter (FAP) representation and used as training data for the synthesis algorithms. For details, please see [6].

In addition to the above mentioned study, several other motion capture corpora have been recorded and utilised for the studies described in this paper. Two examples of the recording set-up can be seen in Fig. 2.

Figure 2: Data collection setup for a conversational recording with video and IR-cameras, microphone (top) and test subject with the IR-reflecting markers glued to the face and EMA coils for recording tongue articulation (bottom).



The database combining motion capture from IR markers and EMA [3] has been used to relate the inside (tongue and jaw) movements to movements of the face. Promising correlations [12] were found between the two sets of data. These methods are now being used in the EU project ASPI which aims at finding useful solutions to the theoretically unsolvable acoustic-to-articulation inversion problem by using e.g. visual information. If successful such methods can be used to unobtrusively estimate student articulation in e.g. articulation training [11].

The databases we have thus far collected have enabled us to analyze for example articulatory variation in expressive speech [17], visual manifestation of focal accent/prominence [20], and conversational cues in dialogue situations, in addition to providing us with data with which to develop data-driven visual synthesis. Examples of the new head displaying different emotions taken from the database are shown in Fig. 3.

Figure 3: Visual stimuli generated by data-driven synthesis from the happy database (left) and the angry database (right) using the MPEG-4 compatible talking head.

4. 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 [13] 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.

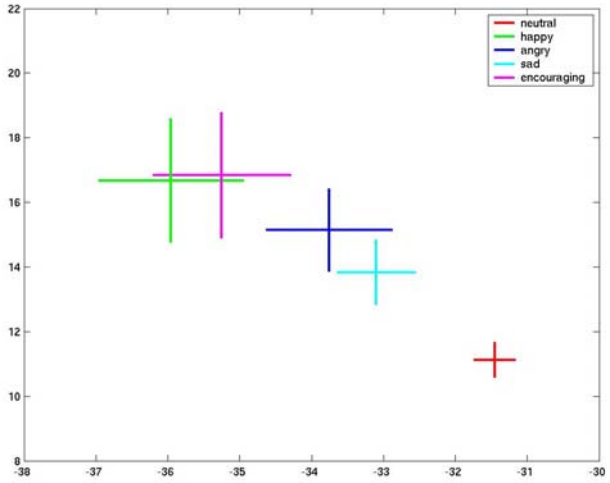
There have been attempts to take into consideration how articulation may change depending on speaker or speaking style and make use of that knowledge in audiovisual synthesis. Pelachaud et al. [22] proposed a method where they could define various speaker characteristics such as speechrate and timing 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 was carried out within the EU PF-Star project which aimed at establishing future activities in the field of multi-sensorial and multi-lingual communication. 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 [20]. Part of the database consisted of semantically neutral utterances, e.g. numbers all pronounced in several different expressive states. In Fig. 4 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 4: The left mouth corner for different acted expressive states. The crosses refer to the expressions: happy, encouraging, angry, sad and neutral from top left to bottom right (2mm between scale markers).



5. ANALYZING AND MODELLING VISUAL CUES FOR PROMINENCE

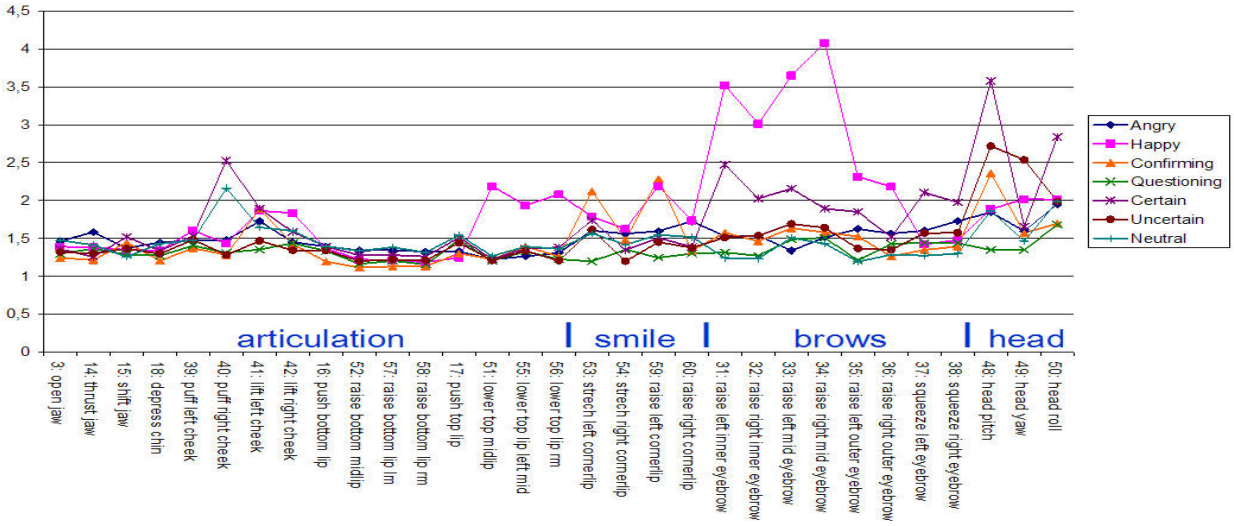
In this section we present measurement results obtained from a speech corpus in which focal accent was systematically varied in a variety of expressive modes.

The speech material used for the study consisted of 39 short, content neutral sentences such as “Båten seglade förbi” (The boat sailed by) and “Grannen knackade på dörren” (The neighbor knocked on the door), all with three content words each of which received focal accent in turn. The current study can be found in more detail in [4].

For recording the database, a total of 29 IR reflective markers were attached to the speaker’s face, of which 4 markers were used as reference markers (on the ears and on the forehead). In the present study, we chose to base our quantitative analysis of facial movement on the MPEG-4 Facial Animation Parameter (FAP) representation, because it is a compact and standardised scheme for describing movements of the human face and head. Specifically, we chose a subset of 31 FAPs out of the 68 FAPs defined in the MPEG-4 standard, including only the ones that we were able to calculate directly from our measured point data (discarding e.g. parameters for inner lip contour, tongue, ears and eyes).

We wanted to obtain a measure of how (in what FAPs) focus was realised by the recorded speaker for the different expressive modes. In an attempt to quantify this, we introduce the Focal Motion Quotient, FMQ, defined as the standard deviation of a FAP parameter taken over a word in focal position, divided by the average standard deviation of the same FAP in the same word in non-focal position. As a first step in the analysis the FMQs for all the 31 measured FAPs were averaged across the 39 sentences. These data are displayed in Fig. 5 for the analyzed expressive modes, i.e. Angry, Happy, Confirming, Questioning, Certain, Uncertain and Neutral. As can be seen, the FMQ mean is always above one, irrespective of which expressive mode or facial movement (FAP) that is studied. This means that a shift from a non-focal to a focal pronunciation on the average results in greater dynamics in all facial movements for all expressive modes. It should be noted that these are results from only one speaker and averages across the whole database. It is however conceivable that facial movements will at least reinforce the perception of focal accent. The mean FMQ taken over all expressive modes is 1.6. The expressive mode yielding the largest mean FMQ is happy (1.9) followed by confirming (1.7), while questioning has the lowest mean FMQ value of 1.3. If we look at the individual parameters and the different expressive modes, some FMQs are significantly greater, especially for the Happy expression, up to 4 for parameter 34 “raise right mid eyebrow”. While much more detailed data on facial movement patterns is available in the database, we wanted to show the strong effects of focal accent on basically all facial movement patterns. Moreover, the results suggest that the

Figure 5: The focal motion quotient, FMQ, averaged across all sentences, for all measured MPEG-4 FAPs for several expressive modes (see text for definitions and details).



specific gestures used for realization of focal accent are related to the intended expressive mode.

6. PERCEPTION EXPERIMENTS

In this section we present results from three different evaluation paradigms. The first is a conventional perception test with parametrically manipulated visual stimuli. The second is mini dialogues where the subject is an "eavesdropper" and in the final example the subject participates in mediated conversation.

6.1. Visual cues for prominence

In order to study the impact of visual cues on prominence we have carried out a series of experiments. The first concerns effects of eyebrow movements [14]. In an earlier study concerned with prominence and phrasing, using acoustic speech only, ambiguous sentences were used. In the present experiment we used one of these sentences:

1. När pappa fiskar stör, piper Putte.
When dad is fishing sturgeon, Putte is whimpering.
2. När pappa fiskar, stör Piper Putte.
When dad is fishing, Piper disturbs Putte.

Hence, "stör" could be interpreted as either a noun (1) (sturgeon) or a verb (2) (disturbs); "piper" (1) is a verb (whimpering), while "Piper" (2) is a name. In the stimuli, the acoustic signal is always the same, and synthesized as one phrase, i.e. with no phrasing prosody disambiguating the sentences. Six different versions were included in the experiment: one with no eyebrow movement and five where eyebrow rise was placed on one of the five content words in the test sentence.

The subjects were instructed to listen as well as to look at the face. Two seconds before each sentence an audio beep was played to give subjects time to look up and focus on the face. No mention was made of eyebrows. The subjects were asked to circle the word that they perceived as most stressed/most prominent in the sentence.

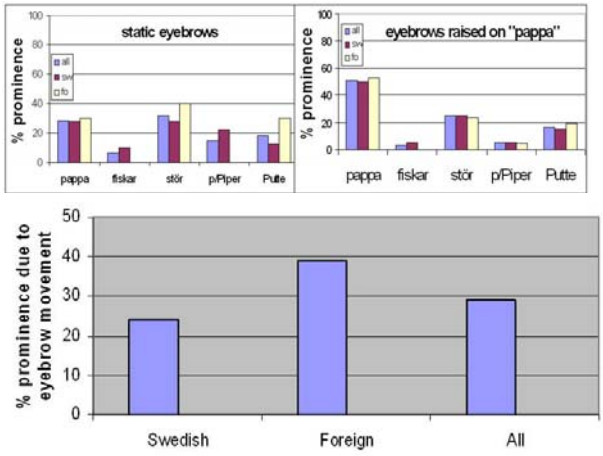
The results are shown in Fig. 6. Fig. 6 (top, left) refers to judgements when there is no eyebrow movement at all. Obviously the distribution of judgements varies with both subject group and word in the sentence and may well be related to prominence expectations.

Fig. 6 (top, right) displays the distribution when an eyebrow movement is occurring during the first word "pappa". A clear shift of responses to this word is evident. In Fig. 6 (bottom) the mean increase due to eyebrow movement across all words can be seen. It is interesting to note that the Swedish-speaking subjects with a different mother tongue showed a stronger dependence on the visual cues. It can be speculated that this group relies more on visual aspects when they communicate in Swedish due to less familiarity with Swedish prosody.

In another study [17] both eyebrow and head movements were tested as potential cues to prominence. Results indicated that combined head and eyebrow movements are effective cues to prominence when synchronized with the stressed

vowel of the potentially prominent word and when no conflicting acoustic cue is present.

Figure 6: Prominence responses in percent for each word. Subjects are grouped as all, Swedish (sw) and foreign (fo) (top). Mean prominence increase due to eyebrow movement (bottom).



6.2. The eavesdropper – visual cues for feedback

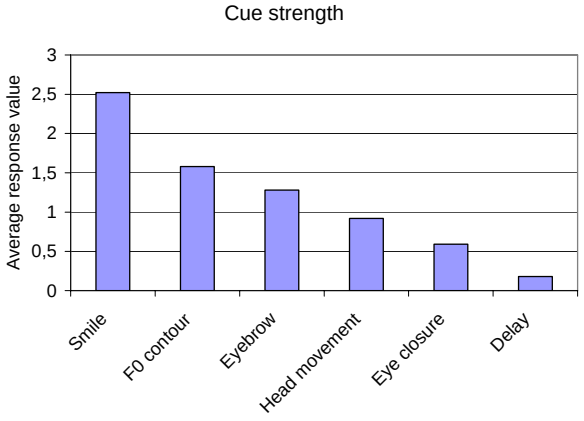
The use of a believable talking head can trigger the user's social skills such as using greetings, addressing the agent by name, and generally socially chatting with the agent. This was demonstrated by the results of the public use of the KTH August system [1][16]. These results led to more specific studies on visual cues for feedback such as [15]. The stimuli used in [15] 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, the eavesdropper, could only hear the client's voice, but could both see and hear the agent. 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:

> Client: Jag vill åka från Stockholm till Linköping.
> *(I want to go from Stockholm to Linköping.)*
> Agent: Linköping.

The stimuli were created by orthogonally varying six visual and acoustic parameters, 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. The task was to respond to this dialogue exchange in terms of whether the agent signals that he understands and accepts the human utterance, or rather signals that he is uncertain about the human utterance. A detailed description of the experiment and the analysis can be found in [15]. Here, we would only like to highlight the strength of the different acoustic and visual cues. In Fig. 7 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. 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.

Figure 7: The mean response value difference for stimuli with the indicated cues set to their affirmative and negative value.

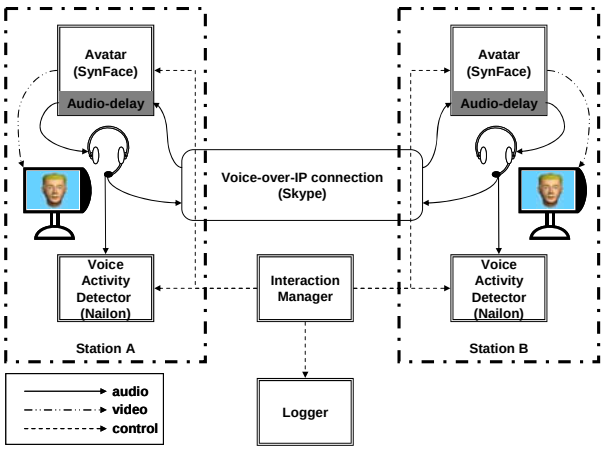


6.3. The participant – manipulating cues for feedback

Work presented in [9] describes a fully automated experimental framework for testing whether the behaviour of interlocutors can be pushed in a given direction by manipulating parts of the conversation in real time as the interlocutors participate in a computer mediated human-human dialogue. In this particular experiment, it was shown that it is possible to unobtrusively influence the turntaking behaviour of interlocutors by smoothly injecting

turntaking gestures in the conversation flow. The gestures were governed by voice activity detection and a simple set of interaction rules. In other words, the experiment showed that the facial gestures of an animated talking head could be used to make a person take the turn more or less often.

Figure 8: The experiment system.



The participants are placed in separate rooms, Fig. 8, and each participant is equipped with a headset connected to a Voice-over-IP call. On both sides, the call is enhanced with SynFace [5], a lip synchronised animated talking head representing each participant. As both talking heads represent real persons (the participants), we refer to them as *avatars* in the following.

In [9], a minimum of interaction control gestures were used:

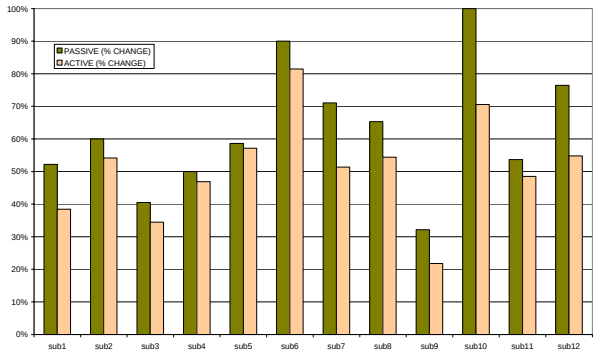
- a turntaking/keeping gesture, where the avatar makes a slight turn of the head to the side in combination with shifting the gaze away a little.
- a turn yielding/listening gesture, where the avatar looks straight forward, at the subject, with slightly raised eyebrows.
- a feedback/agreement gesture, consisting of a small nod. In the experiment described here, this gesture is not used alone, but it is added at the end of the listening gesture to add to its responsiveness.

The transition events registered in the interaction model control the gestures of the avatar. Three sets of mappings between gestures and events were used, one of which left the face without any gestures for all transitions (NEUTRAL), resulting in a relatively immobile avatar. The other two sets were designed to nudge the recipient towards a more pushy turntaking behaviour (ACTIVE) or a more meek turntaking behaviour (PASSIVE), respectively. The ACTIVE and PASSIVE gesture sets were used in pairs, so that whenever one user was confronted with the ACTIVE gestures, the other one had the PASSIVE set.

For every 10 SPEECH/SILENCE transitions, the gesture sets were shifted cycling through speaker A–speaker B mappings of ACTIVE-PASSIVE, NEUTRAL-NEUTRAL, PASSIVE-ACTIVE, and again NEUTRAL-NEUTRAL. SPEECH/SILENCE transition was chosen as a measure for governing the shifting of the gesture sets because it gives some variation in terms of time as well as number of utterances – a timer or exact utterance counter was avoided in order to make it less obvious for the participants that the gesture behaviour was varied systematically.

In Fig. 9 results from one experiment can be seen. The turntaking behaviour is obviously very different for different subjects, but it is also clear that the avatar setting provoking or suppressing turntaking is efficient. The percentage of all contributions followed by CHANGE OF TURN, denoting that the speaker lost the turn after pausing, is larger under the PASSIVE condition than under the ACTIVE condition for each participant without exception. The difference is significant at the 0.01 level with a paired two-sample t-test for means: df=11, t=4.66, P<0.01 (two-tailed).

Figure 9: Percentage of contributions followed by CHANGE OF TURN per user and condition.



We believe that this paradigm will be useful in testing different kinds of interactive behaviour in a realistic setting. While the experiment described concerns an avatar mediated human-human communication setting, we hold that the results are equally useful for human-computer interaction scenarios, in the design of human-like spoken dialogue systems.

7. FUTURE CHALLENGES

In this paper, we have presented an overview of some of the recent work in audiovisual synthesis, at KTH, regarding data collection methods, modelling and evaluation experiments, and implementation in animated talking agents for dialogue systems. From this point of departure, we can see that many challenges remain before we will be able to create a believable, animated talking agent based on knowledge concerning how auditory and visual signals interact in verbal and non-verbal communication. In terms of modelling and evaluation, there is a great need to explore in more detail the coherence between audio and visual prosodic expressions especially regarding different functional dimensions. Modelling of a greater number of parameters is also essential, such as head movement in more dimensions, eye movement and gaze, and other body movements such as hand and arm gestures. To model and evaluate how these parameters combine in different ways to convey individual personality traits while at the same time signalling basic prosodic and dialogue functions is a great challenge.

8. ACKNOWLEDGEMENTS

The work at KTH reported here was carried out by a large number of researchers at the Centre for Speech Technology which is gratefully acknowledged. The work has also been supported by the EU/IST projects SYNFACE, PF-Star, CHIL, MonAMI, MUSCLE and HaH.

9. REFERENCES

[1] Bell, L., Gustafson, J. 1999. Interacting with an animated agent: an analysis of a Swedish database of spontaneous computer directed speech. *Proc. of Eurospeech '99* Budapest, 1143-1146.

[2] Beskow, J. 1995. Rule-based Visual Speech Synthesis. *Proc. of Eurospeech '95* Madrid, 299-302.

[3] Beskow, J., Engwall, O., Granström, B. 2003. Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. 15th ICPhS* Barcelona, 431-434.

[4] Beskow, J., Granström, B., House, D. 2006. Visual correlates to prominence in several expressive modes. *Proc. of Interspeech 2006* Pittsburg, 1272-1275.

[5] Beskow, J., Karlsson, I., Kewley, J., Salvi, G. 2004. SYNFACE - A talking head telephone for the hearing-impaired. In: Miesenberger, K., Klaus, J., Zagler, W., Burger, D. (eds), *Computers Helping People with Special Needs*. Berlin: Springer-Verlag, 1178-1186

[6] Beskow, J., Nordenberg, M. 2005. Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head. *Proc. of Interspeech 2005* Lisbon, 793-796.

[7] Bickmore, T., Cassell, J. 2005. Social Dialogue with Embodied Conversational Agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. O. (eds), *Advances in Natural Multimodal Dialogue Systems.* Dordrecht, The Netherlands: Springer 23-54.

[8] Carlson, R., Granström, B. 1997. Speech Synthesis. In: Hardcastle, W., Laver, J. (eds), *The Handbook of Phonetic Sciences.* Oxford: Blackwell, 768-788.

[9] Edlund, J., Beskow, J. 2007. Pushy versus meek – using avatars to influence turn-taking behaviour. *Proc. of Interspeech 2007* Antwerp.

[10] Engwall, O. 2003. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication* 41(2-3), 303-329.

[11] Engwall, O., Bälter, O., Öster, A-M., Kjellström, H. 2006. Designing the user interface of the computer-based speech training system ARTUR based on early user tests. *Journal of Behavioural and Information Technology* 25(4), 353-365.

[12] Engwall, O., Beskow, J. 2003. Resynthesis of 3D tongue movements from facial data. *Proc. of Eurospeech 2003* Geneva, 2261-2264.

[13] Fonâgy, I. 1976. La mimique buccale. *Phonetica* 33, 31–44.

[14] Granström, B., House, D., Beskow, J., Lundeberg, M. 2001. Verbal and visual prosody in multimodal speech perception. *Proc. Nordic Prosody 2000*, Trondheim. Frankfurt am Main: Peter Lang, 77-88.

[15] Granström, B., House, D., Swerts, M.G. 2002. Multimodal feedback cues in human-machine interactions. *Proc. of Speech Prosody 2002* Aix-en-Provence, 347-350.

[16] House, D. 2005. Phrase-final rises as a prosodic feature in wh-questions in Swedish human–machine dialogue. *Speech Communication* 46, 268-283.

[17] House, D., Beskow, J., Granström, B. 2001. Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proc. of Eurospeech 2001* Aalborg, 387-390.

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

[19] Massaro, D.W., Light, J. 2003. Read My Tongue Movements: Bimodal Learning to Perceive and Produce Non-Native Speech /r/ and /l/. *Proc. of Eurospeech 2003* Geneva, 2249-2252.

[20] Nordstrand, M., Svanfeldt, G., Granström, B., House, D. 2004. Measurements of articulatory variation in expressive speech for a set of Swedish vowels. *Speech Communication* 44, 187-196.

[21] Parke, F.I. 1982. Parameterized models for facial animation. *IEEE Computer Graphics* 2(9), 61-68.

[22] Pelachaud, C., Badler, N., Steedman, M. 1996. Generating Facial Expressions for Speech. *Cognitive Science* 20, 1–46.

[23] Sjölander, K., Beskow, J. 2000. WaveSurfer - an Open Source Speech Tool. *Proc of ICSLP 2000* Beijing, Vol. 4, 464-467.