Chapter 3

MODELLING AND EVALUATING VERBAL AND NON-VERBAL COMMUNICATION IN TALKING ANIMATED INTERFACE AGENTS

Björn Granström and David House
KTH (Royal Institute of Technology)
Stockholm, Sweden
bjorn@speech.kth.se, davidh@speech.kth.se

Abstract The use of animated talking agents is a novel feature of many multimodal experimental spoken dialogue systems. The addition and integration of a virtual talking head has direct implications for the way in which users approach and interact with such systems. Established techniques for evaluating the quality, efficiency, and other impacts of this technology have not yet appeared in standard textbooks. The focus of this chapter is to look into the communicative function of the agent, both the capability to increase intelligibility of the spoken interaction and the possibility to make the flow of the dialogue smoother, through different kinds of communicative gestures such as gestures for emphatic stress, emotions, turntaking, and negative or positive system feedback. The chapter reviews state-of-the-art animated agent technologies and their applications primarily in dialogue systems. The chapter also includes examples of methods of evaluating communicative gestures in different contexts.

Keywords Audio-visual speech synthesis; Talking heads; Animated agents; Spoken dialogue systems; Visual prosody.

1 Introduction

In our interaction with others, we easily and naturally use all of our sensory modalities as we communicate and exchange information. Our senses are exceptionally well adapted for these tasks, and our brain enables us to effortlessly integrate information from different modalities fusing data to optimally meet the current communication needs. As we attempt to take advantage of

L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 65–98.

the effective communication potential of human conversation, we see an increasing need to embody the conversational partner using audio-visual verbal and non-verbal communication in the form of animated talking agents. The use of animated talking agents is currently a novel feature of many multimodal experimental spoken dialogue systems. The addition and integration of a virtual talking head has direct implications for the way in which users approach and interact with such systems (Cassell et al., 2000). However, established techniques for evaluating the quality, efficiency, and other impacts of this technology have not yet appeared in standard textbooks.

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 user requirements, and can ultimately assist the user in a decision-making process through the interactive flow of conversation. One way to achieve believability is through the use of a talking head where information is transformed through text into speech, articulator movements, speech-related gestures, and conversational gestures. Useful applications of talking heads include aids for the hearing impaired, educational software, audio-visual human perception experiments (Massaro, 1998), entertainment, and high-quality audio-visual 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 (Bickmore and Cassell, 2005). Visual cues to feedback, turntaking, and signalling the system's internal state are key aspects of effective interaction. There is also currently much interest in the use of visual cues as a means to ensure that participants in a conversation share an understanding of what has been said, i.e. a common ground (Nakano et al., 2003).

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" (Granström et al., 2001). These types of visual cues with the addition of, for example, 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. Recently, we have been exploring data-driven methods to model articulation and facial parameters of major importance for conveying social signals and emotion.

The focus of this chapter is to look into the communicative function of the agent, both the capability to increase intelligibility of the spoken interaction and the possibility to make the flow of the dialogue smoother, through different kinds of communicative gestures such as gestures for emphatic stress, emotions, turntaking, and negative or positive system feedback. The chapter reviews state-of-the-art animated agent technologies and their applications primarily in dialogue systems. The chapter also includes some examples of methods of evaluating communicative gestures in different contexts.

2 KTH Parametric Multimodal Speech Synthesis

Animated synthetic talking faces and characters have been developed using a number of different techniques and for a variety of purposes for more than two decades. Historically, our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework (Carlson and Granström, 1997). The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account (Beskow, 1995). We employ a generalised parameterisation technique to adapt a static 3D wireframe of a face for visual speech animation (Beskow, 1997). Based on concepts first introduced by Parke (1982), 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 decoupled the model definitions from the animation engine. The animation engine uses different definition files that are created for each face. This greatly increases flexibility, allowing models with different topologies to be animated from the same control parameter program, such as a text-to-speech system.

The models are made up of polygon surfaces that are rendered in 3D using standard computer graphics techniques. The surfaces can be articulated and deformed under the control of a number of parameters. The parameters are designed to allow for intuitive interactive or rule-based control. For the purposes of animation, parameters can be roughly divided into two (overlapping) categories: those controlling speech articulation and those used for non-articulatory cues and emotions. The articulatory parameters include jaw rotation, lip rounding, bilabial occlusion, labiodental occlusion, and tongue tip elevation. The non-articulatory category includes eyebrow raising, eyebrow shape, smile, gaze direction, and head orientation. Furthermore, some of the articulatory parameters such as jaw rotation can be useful in signalling non-verbal elements such as certain emotions. The display can be chosen to show only the surfaces or the polygons for the different components of the face. The surfaces can be made (semi-)transparent to display the internal parts of the model,

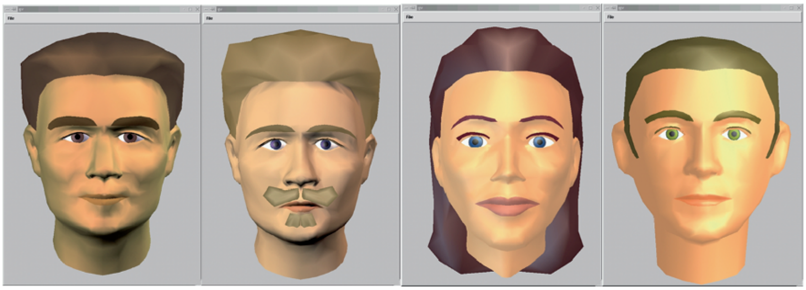
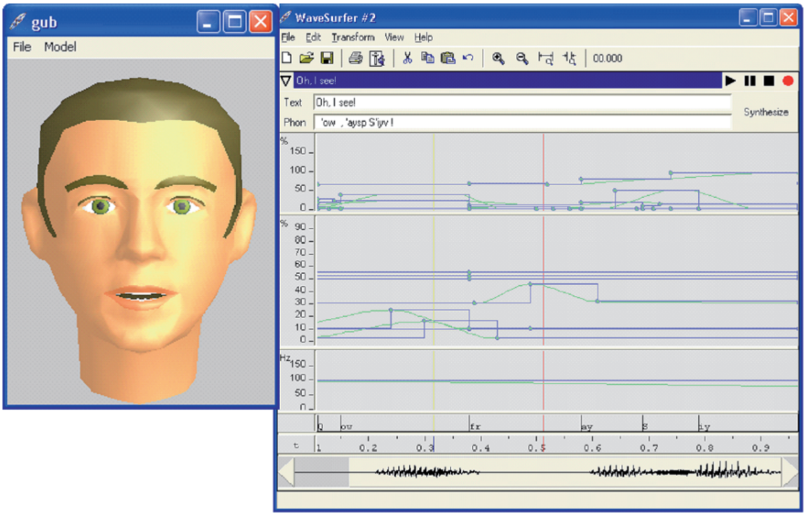


Figure 1. Some different versions of the KTH talking head.

including the tongue, palate, jaw, and teeth (Engwall, 2003). The internal parts are based on articulatory measurements using magnetic resonance imaging, electromagnetic articulography, and electropalatography, in order to ensure that the model's physiology and 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 (Cole et al., 1999; Massaro et al., 2003; Massaro and Light, 2003; Bälter et al., 2005). Several face models have been developed for different applications, and some of them can be seen in Figure 1. All can be parametrically controlled by the same articulation rules.

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 interface is implemented as an extension to the WaveSurfer application (http://www.speech.kth.se/wavesurfer) (Sjölander and Beskow, 2000), which is a freeware tool for recording, playing, editing, viewing, printing, and labelling audio data.

The interface makes it possible to start with an utterance synthesised from text, with the articulatory parameters generated by rule, and then interactively edit the parameter tracks for F0, visual (non-articulatory) parameters as well as the durations of individual segments in the utterance to produce specific cues. An example of the user interface is shown in Figure 2. In the top box a text can be entered in Swedish or English. The selection of language triggers separate text-to-speech systems with different phoneme definitions and rules, built in the Rulsys notation (Carlson and Granström, 1997). One example of language-dependent rules are the rules for visual realisation of interdentals in English which do not apply to Swedish. The generated phonetic transcription can be edited. On pushing "Synthesize", rule-generated parameters will be created and displayed in different panes below. The selection of parameters is user-controlled. The lower section contains segmentation and the acoustic waveform. A talking face is displayed in a separate window. The acoustic synthesis can be exchanged for a natural utterance and synchronised to the face



Text: Oh, I see!

Phon: 'ow , 'aysp S'iyv !

Figure 2. The WaveSurfer user interface for parametric manipulation of the multimodal synthesis.

synthesis on a segment-by-segment basis by running the face synthesis with phoneme durations from the natural utterance. This requires a segmentation of the natural utterance which can be done (semi-)automatically in, for example, WaveSurfer. The combination of natural and synthetic speech is useful for different experiments on multimodal integration and has been used in the Synface/Teleface project (see below). 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. The parametric manipulation tool is used to experiment with, and define, gestures. Using this tool we have constructed a library of gestures that can be invoked via XML markup in the output text.

3 Data Collection and Data-Driven Visual Synthesis

More recently, we have begun experimenting with data-driven visual synthesis using a newly developed MPEG-4 compatible talking head (Beskow and Nordenberg, 2005). 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 (cf. Nordstrand et al., 2004).

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 of 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 by Beskow et al. (2003). The recording set-up can be seen in Figure 3.

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 cannot be recorded with the Qualisys system but is only video recorded. By attaching infrared reflecting markers to the subject's face, see Figure 3, the system is able to register the 3D coordinates for each marker at a frame rate of 60 Hz, i.e. every 17 ms. 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.

The databases we have thus far collected have enabled us to analyse speech movements such as articulatory variation in expressive speech (Nordstrand et al., 2004) in addition to providing us with data with which to develop data-driven visual synthesis. The data has also been used to directly drive synthetic 3D face models which adhere to the MPEG-4 Facial Animation (FA) standard (Pandzic and Forchheimer, 2002) enabling us to perform comparative

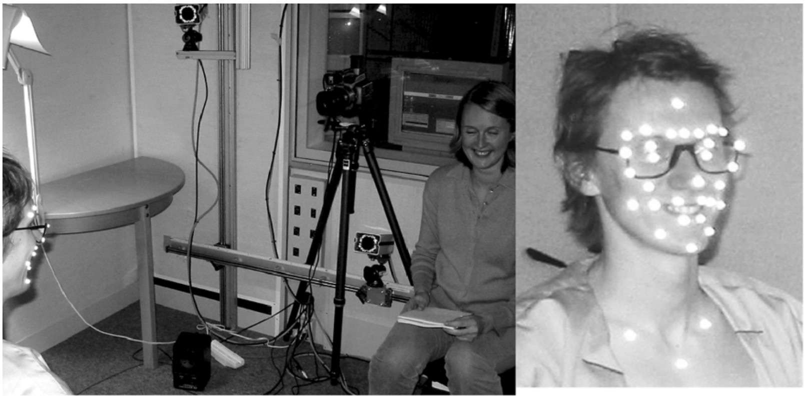


Figure 3. Data collection set-up with video and IR-cameras, microphone and a screen for prompts (left), and test subject with the IR-reflecting markers glued to the face (right).

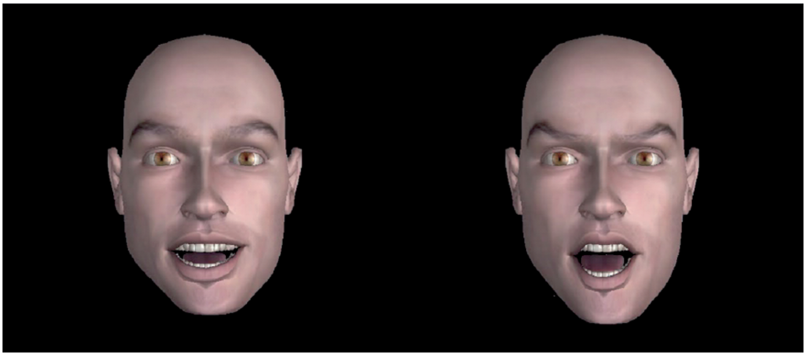


Figure 4. Visual stimuli generated by data-driven synthesis from the happy database (left) and the angry database (right) using the MPEG-4 compatible talking head.

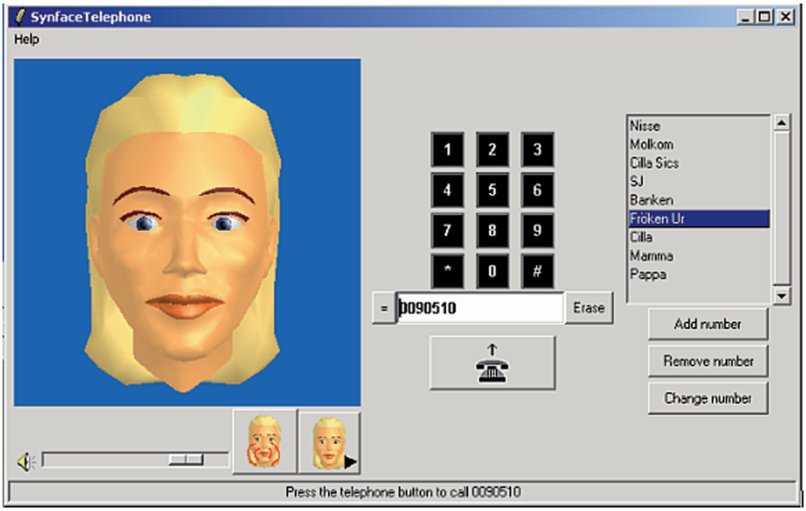
evaluation studies of different animated faces within the EU-funded PF-Star project (Beskow et al., 2004a, b).

The new talking head is based on the MPEG-4 FA standard and is a textured 3D model of a male face comprising around 15,000 polygons. Current work on data-driven visual synthesis is aimed at synthesising visual speech articulation for different emotions (Beskow and Nordenberg, 2005). The database consists of recordings of a male native Swedish amateur actor who was instructed to produce 75 short sentences with the six emotions happiness, sadness, surprise, disgust, fear, and anger plus neutral (Beskow et al., 2004c). Using the databases of different emotions results in talking head animations which differ in articulation and visual expression. The audio synthesis used at present is the same as that for the parametric synthesis. Examples of the new head displaying two different emotions taken from the database are shown in Figure 4.

4 Evaluating Intelligibility and Information Presentation

One of the more striking examples of improvement and effectiveness in speech intelligibility is taken from the Synface project which aims at improving telephone communication for the hearing impaired (Agelfors et al., 1998). A demonstrator of the system for telephony with a synthetic face that articulates in synchrony with a natural voice has now been implemented as a result of the project. The telephone interface used in the demonstrator is shown in Figure 5.

Evaluation studies within this project were mainly oriented towards investigating differences in intelligibility between speech alone and speech with the addition of a talking head. These evaluation studies were performed offline: e.g. the speech material was manually labelled so that the visible speech



0090510

Press the telephone button to call 0090510

Figure 5. Telephone interface for Synface.

synthesis always generated the correct phonemes rather than being generated from the Synface recogniser, which can introduce recognition errors. The results of a series of tests using vowel-consonant-vowel (VCV) words and hearing-impaired subjects showed a significant gain in intelligibility when the talking head was added to a natural voice. With the synthetic face, consonant identification improved from 29% to 54% correct responses. This compares to the 57% correct response result obtained by using the natural face. In certain cases, notably the consonants consisting of lip movement (i.e., the bilabial and labiodental consonants), the response results were in fact better for the synthetic face than for the natural face. This points to the possibility of using overarticulation strategies for the talking face in these kinds of applications. Recent results indicate that a certain degree of overarticulation can be advantageous in improving intelligibility (Beskow et al., 2002b).

Similar intelligibility tests have been run using normal hearing subjects where the audio signal was degraded by adding white noise (Agelfors et al., 1998). Similar results were obtained. For example, for a synthetic male voice, consonant identification improved from 31% without the face to 45% with the face.

Hearing-impaired persons often subjectively report that some speakers are much easier to speech-read than others. It is reasonable to hypothesise that this variation depends on a large number of factors, such as rate of speaking, amplitude and dynamics of the articulatory movements, orofacial anatomy of the speaker, presence of facial hair, and so on. Using traditional techniques, however, it is difficult to isolate these factors to get a quantitative measure of their relative contribution to readability. In an attempt to address this issue, we employ a synthetic talking head that allows us to generate stimuli where each variable can be studied in isolation. In this section we focus on a factor that we will refer to as articulation strength, which is implemented as a global scaling of the amplitude of the articulatory movements.

In one experiment the articulation strength has been adjusted by applying a global scaling factor to the parameters marked with an x in Table 1. They can all be varied between 25% and 200% of normal. Normal is defined as the default articulation produced by the rules, which are hand-tuned to match a target person's articulation.

The default parameter settings are chosen to optimise perceived similarity between a target speaker and the synthetic faces. However, it is difficult to know whether these settings are optimal in a lip-reading situation for hearing-impaired persons. An informal experiment was pursued to find out the preferred articulation strength and its variance. Twenty-four subjects all closely connected to the field of aural rehabilitation either professionally or as hearing impaired were asked to choose the most intelligible face out of eight recordings of the Swedish sentence "De skrattade mycket högt" (They laughed very loudly). The subjects viewed the eight versions in eight separate windows on a computer screen and were allowed to watch and compare the versions as many times as they wished by clicking on each respective window to activate the recordings. The recordings had 25%, 50%, 75%, 100%, 112%, 125%, 150% and 175% of the default strength of articulation. The default strength of articulation is based on the phoneme parameter settings for visual speech synthesis as developed by Beskow (1997). The different articulation strengths were implemented as a global scaling of the amplitude of the articulatory movements. The amount of co-articulation was not altered.

The average preferred hyperarticulation was found to be 24%, given the task to optimise the subjective ability to lip-read. The highest and lowest preferred

Table 1. Parameters used for articulatory control of the face. The second column indicates which ones are adjusted in the experiments described here.

Parameter	Adjusted in experiment
Jaw rotation	x
Labiodental occlusion	
Bilabial occlusion	
Lip rounding	
Lip protrusion	x
Mouth spread	x
Tongue tip elevation	x

values were 150% and 90% respectively with a standard deviation of 16%. The option of setting the articulation strength to the user's subjective preference could be included in the Synface application. The question of whether or not the preferred setting genuinely optimises intelligibility and naturalness was also studied as is described below.

Experiment 1: Audio-visual consonant identification. To test the possible quantitative impact of articulation strength, as defined in the previous section, we performed a VCV test. Three different articulation strengths were used: 75%, 100%, and 125% of the default articulation strength for the visual speech synthesis. Stimuli consisted of nonsense words in the form of VCV combinations. Seventeen consonants were used: /p, b, m, f, v, t, d, n, s, l, r, k, g, ng, sj, tj, j/ in two symmetric vowel contexts /a, U/ yielding a total of 34 different VCV words. The task was to identify the consonant. (The consonants are given in Swedish orthography – the non-obvious IPA correspondences are: ng=/ŋ/, sj=/ɧ/, tj= /ç/.) Each word was presented with each of the three levels of articulation strength. The list was randomised. To avoid starting and ending effects, five extra stimuli were inserted at the beginning and two at the end.

Stimuli were presented audio-visually by the synthetic talking head. The audio was taken from the test material from the Teleface project recordings of natural speech from a male speaker (Agelfors et al., 1998). The audio had previously been segmented and labelled, allowing us to generate control parameter tracks for facial animation using the visual synthesis rules.

The nonsense words were presented in white masking noise at a signal-to-noise ratio of 3 dB.

Twenty-four subjects participated in the experiment. They were all undergraduate students at KTH. The experiments were run in plenary by presenting the animations on a large screen using an overhead projector. The subjects responded on pre-printed answer sheets.

The mean results for the different conditions can be seen in Table 2. For the /a/ context there are only minor differences in the identification rate, while the

Table 2. Percent correct consonant identification in the VCV test with respect to place of articulation, presented according to vowel context and articulation strength.

Articulation strength (%)	/aCa/	/UCU/
75	78.7	50.5
100	75.2	62.2
125	80.9	58.1

results for the /U/ context are generally worse, especially for the 75% articulation rate condition. A plausible reason for this difference lies in the better visibility of tongue articulations in the more open /a/ vowel context than in the context of the rounded /U/ vowel. It can also be speculated that movements observed on the outside of the face can add to the superior readability of consonants in the /a/ context. However, we could not find evidence for this in a study based on simultaneous recordings of face and tongue movements (Beskow et al., 2003; Engwall and Beskow, 2003). In general, the contribution of articulation strength to intelligibility might be different with other speech material such as connected sentences.

Experiment 2: Rating of naturalness. Eighteen sentences from the Teleface project (Agelfors et al., 1998) were used for a small preference test. Each sentence was played twice: once with standard articulation (100%) and once with smaller (75%) or greater (125%) articulation strength. The set of subjects, presentation method, and noise masking of the audio was the same as in experiment 1 (the VCV test). The subjects were asked to report which of the two variants seemed more natural or if they were judged to be of equal quality. The test consisted of 15 stimuli pairs, presented to 24 subjects. To avoid starting and ending effects, two extra pairs were inserted at the beginning and one at the end. The results can be seen in Table 3. The only significant preference was for the 75% version contrary to the initial informal experiment. However, the criterion in the initial test was readability rather than naturalness.

The multimodal synthesis software together with a control interface based on the WaveSurfer platform (Sjölander and Beskow, 2000) allows for the easy production of material addressing the articulation strength issue. There is a possible conflict in producing the most natural and the most easily lip-read face. However, under some conditions it might be favourable to trade off some naturalness for better readability. For example, the dental viseme cluster [r, n, t, d, s, and l] could possibly gain discriminability in connection with closed vowels if tongue movements could be to some extent hyperarticulated and well rendered. Of course the closed vowels should be as open as possible without jeopardising the overall vowel discriminability.

The optimum trade-off between readability and naturalness is certainly also a personal characteristic. It seems likely that hearing-impaired people would

Table 3. Judged naturalness compared to the default (100%) articulation strength.

Articulation strength(%)	Less natural	Equal	More natural
75	31.67	23.33	45.00
125	41.67	19.17	39.17

emphasise readability before naturalness. Therefore it could be considered that in certain applications like in the Synface software, users could be given the option of setting the articulation strength themselves.

Experiment 3: Targeted audio. A different type of application which can potentially benefit from a talking head in terms of improved intelligibility is targeted audio. To transmit highly directional sound, targeted audio makes use of a technique known as "parametric array" (Westervelt, 1963). This type of highly directed sound can be used to communicate a voice message to a single person within a group of people (e.g., in a meeting situation or at a museum exhibit) without disturbing the other people. Within the framework of the EU project CHIL, experiments have been run to evaluate intelligibility of such targeted audio combined with a talking head (Svanfeldt and Olszewski, 2005).

Using an intelligibility test similar to the ones described above, listeners were asked to identify the consonant in a series of VCV words. The seven consonants to be identified were [f, s, m, n, k, p, t] uttered in an [aCa] frame using both audio and audio-visual speech synthesis. Four conditions were tested. Two audio conditions with and without the talking head were tested, one which targeted the audio directly towards the listener (the 0° condition) and one which targeted the audio 45° away from the listener (the 45° condition). The subjects were seated in front of a computer screen with the target audio device next to the screen. See Figure 6.

The addition of the talking head increased listener recognition accuracy from 77% to 93% in the 0° condition and even more dramatically from 58% to 88% in the 45° condition. Thus the talking head can serve to increase intelligibility and even help to compensate for situations in which the listener may be moving or not located optimally in the audio beam.

A more detailed analysis of the data revealed that consonant confusions in the audio-only condition tended to occur between [p] and [f] and between [m] and [n]. These confusions were largely resolved by the addition of the talking

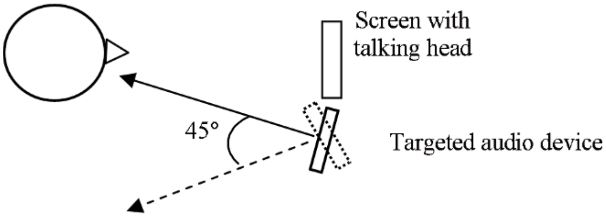


Figure 6. Schematic view of the experimental set-up (Svanfeldt and Olszewski, 2005).

head. This is as could be expected since the addition of the visual modality provides place of articulation information for the labial articulation.

5 Evaluating Visual Cues for Prominence

Another quite different example of the contribution of the talking head to information presentation is taken from the results of perception studies in which the percept of emphasis and syllable prominence is enhanced by eyebrow and head movements. In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish (Granström et al., 1999), 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% for the Swedish native (L1) listeners and 39% for the L2 group. One example result is shown in Figure 7. Similar results have also been obtained for Dutch by Krahmer et al. (2002a).

In another study (House et al., 2001) both eyebrow and head movements were tested as potential cues to prominence. The goal of the study was twofold. 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 audio-visual synthesiser 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 (House, 2001). 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

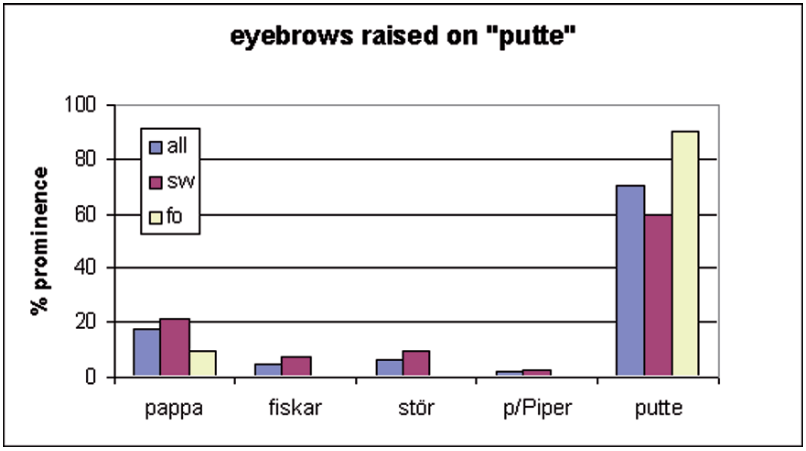


Figure 7. Prominence responses in percent for each content word for the acoustically neutral reading of the stimulus sentence “När pappa fiskar stör p/Piper Putte”, with eyebrow movement on “Putte”. Subjects are grouped as all, Swedish (sw), and foreign (fo).

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 1 in which both eyebrow and head movement occurred simultaneously, and set 2 in which the movements were separated and potentially conflicting with each other. In set 1, 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 2, 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 8.

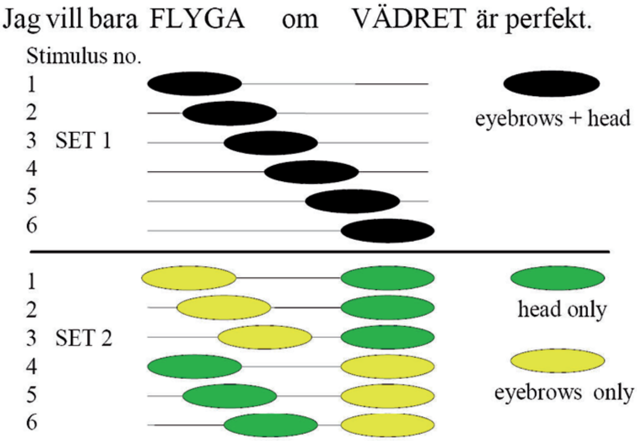


Figure 8. Schematic illustration of face gesture timing.

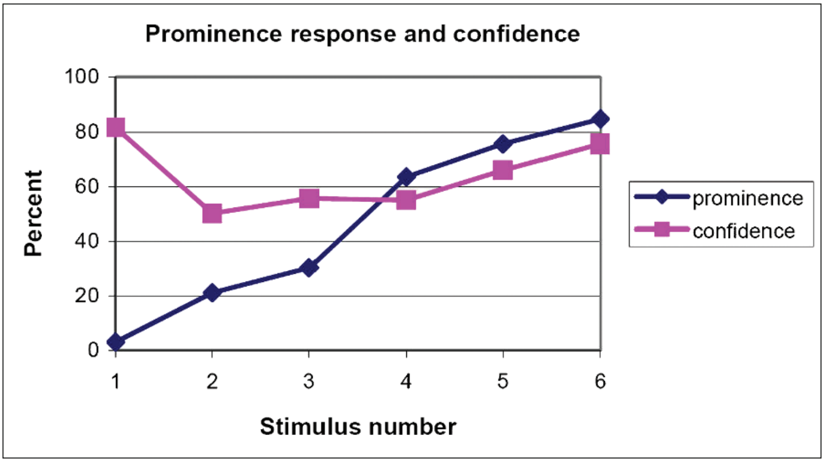


Figure 9. Results for stimulus set 1 showing prominence response for *vädret* and confidence in percent.

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 9 where percent votes for *vädret* increase successively as movement is shifted in time from *flyga* to *vädret*.

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. The results demonstrate a general sensitivity to the timing of these movements at least on the order of 100 ms as the prominence response moves successively from the word *flyga* to the word *vädret*. 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 (Massaro et al., 1996).

As could be expected, the results from set 2, where eyebrow and head movements 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 observers were further away from the computer screen than the subjects in the 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, for example 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, and evaluation of, potential cues for feedback seeking was pursued in the study reported on in Section 6.

6 Evaluating Prosody and Interaction

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 (Bell and Gustafson, 1999a) during a period of 6 months (see Section 9). These promising results have led to more specific studies on visual cues for feedback (e.g., Granström et al., 2002), in which smile, for example, was found to be the strongest cue for affirmative feedback. Further detailed work on turntaking regulation, feedback seeking and giving, and 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 (Clark and Schaeffer, 1989; Traum, 1994; Brennan, 1990). 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) (Shimojima et al., 2002; Krahmer et al., 2002b). 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 (Gill et al., 1999).

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 (e.g., Hirschberg et al., 2001). There are currently a number of advanced multimodal user interfaces in the form of talking heads that can generate audiovisual speech along with different facial expressions (Beskow, 1995, 1997; Beskow et al., 2001; Granström et al., 2001; Massaro, 2002; Pelachaud, 2002; Tisato et al., 2005). 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 synthesise convincing affirmative and negative cues.

One interesting question, for instance, is what the strength relation is between the potential prosodic and visual cues. The goal of one study (Granström et al., 2002) 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 backchannelling 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 six parameters, shown in Table 4, 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 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 10 an example of the all-negative and all-affirmative face can be seen.

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 Granstroöm et al. (2002). Here, we would only like to highlight the strength of the different acoustic and visual cues. In Figure 11

Table 4. Different parameters and parameter settings used to create different stimuli.

	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 wide
F0 contour	Declarative	Interrogative
Delay	Immediate reply	Delayed reply

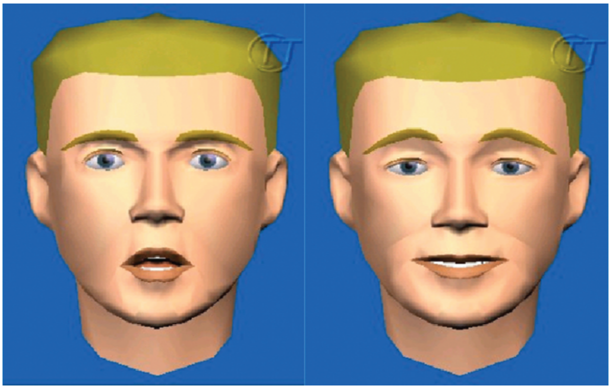


Figure 10. The all-negative and all-affirmative faces sampled in the end of the first syllable of Linköping.

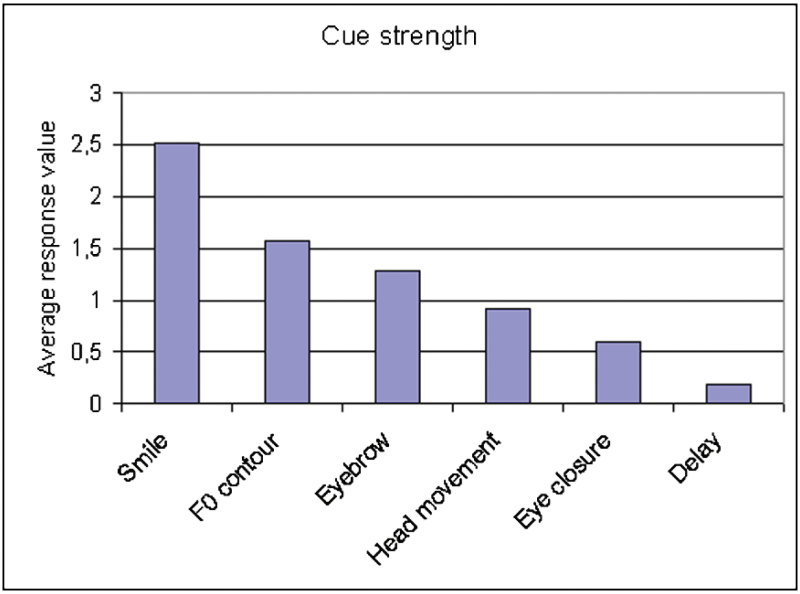


Figure 11. The mean response value difference for stimuli with the indicated cues set to their affirmative and negative value.

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. 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 Evaluating Visual Cues to Sentence Mode

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 Section 6, an experiment was carried out to test if similar visual cues could influence the perception of question and statement intonation in Swedish (House, 2002). Parameters were hand manipulated to create two configurations: one expected to elicit more interrogative responses and the other expected to elicit more declarative responses. These configurations were similar, although not identical, to the positive and negative configurations used in the feedback experiment. Hypothesised cues for the interrogative mode consisted of a slow up–down head nod and eyebrow lowering. Hypothesised cues for the declarative mode consisted of a smile, a short up–down head nod, and eye narrowing. The declarative head nod was of the same type as was used in the prominence experiments reported in Section 5. 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). A separate perception test using these audio-only stimuli resulted in 100% declarative responses for the low falling contour and 100% interrogative responses for the high final rise with a continuum of uncertainty in between.

The influence of the visual cues on the audio cues was only marginal. While the hypothesised cues for the declarative mode (smile, short head nod, and eye narrowing) elicited somewhat more declarative responses for the ambiguous and interrogative intonation contours, the hypothesized cues for the interrogative mode (slow head nod and eyebrow lowering) led to more uncertainty in the responses for both the declarative intonation contours and the interrogative intonation contours (i.e., responses for the declarative contours were only slightly more interrogative than in the audio-only condition while responses for the interrogative contours were actually more declarative). Similar results were obtained for English by Srinivasan and Massaro (2003). Although they were able to demonstrate that the visual cues of eyebrow raising and head tilting synthesised based on a natural model reliably conveyed question intonation, their experiments showed a weak visual effect relative to a strong audio effect of intonation. This weak visual effect remained despite attempts to enhance the visual cues and make the audio information more ambiguous.

The dominance of the audio 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 audio cues. Moreover, 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 Evaluation of Agent Expressiveness and Attitude

In conjunction with the development of data-driven visual synthesis as reported in Section 3, two different evaluation studies have been carried out. One was designed to evaluate expressive visual speech synthesis in the framework of a virtual language tutor (cf. Granström, 2004). The experiment, reported in detail (Beskow and Cerrato, 2006), used a method similar to the one reported on in Section 6. The talking head had the role of a language tutor who was engaged in helping a student of Swedish improve her pronunciation. Each interaction consisted of the student's pronunciation of a sentence including a mispronounced word. The virtual tutor responds by correcting the mispronunciation after which the student makes a new attempt in one of three ways: with the correct pronunciation, with the same mistake, or with a new mistake. The test subjects hear the student's pronunciation and both see and hear the tutor. The task was to judge which emotion the talking head expressed in its final turn of the interaction.

Visual synthesis derived from a happy, angry, sad, and neutral database was used to drive the new MPEG-4 compatible talking head as described in Section 3. For the audio part of the stimuli, a pre-recorded human voice was used to portray the three emotions since we have not yet developed suitable audio data-driven synthesis with different emotions. All possible combinations of audio and visual stimuli were tested. The results indicated that for stimuli where the audio and visual emotion matched, listener recognition of each emotion was quite good: 87% for neutral and happy, 70% for sad, and 93% for angry. For the mismatched stimuli, the visual elements seemed to have a stronger influence than the audio elements. These results point to the importance of matching audio and visual emotional content and show that subjects attend to the visual element to a large degree when judging agent expressiveness and attitude.

In another experiment reported on in House (2006), the new talking head was evaluated in terms of degrees of friendliness. Databases of angry, happy, and neutral emotions were used to synthesise the utterance *Vad heter du?* (What is your name?). Samples of the three versions of the visual stimuli are presented in Figure 12. The three versions of the visual synthesis were combined with two audio configurations: low, early pitch peak; and high, late pitch

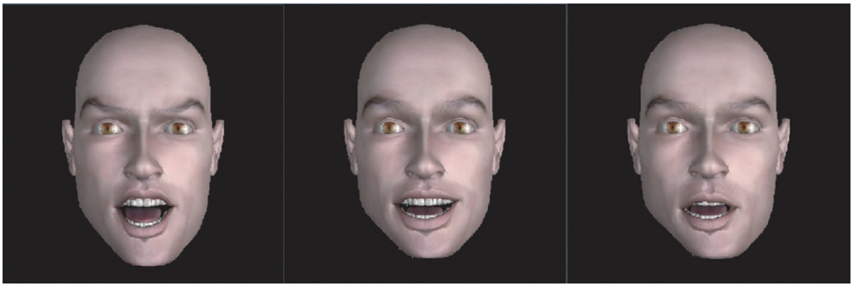


Figure 12. Visual stimuli generated by data-driven synthesis from the angry database (left), the happy database (middle), and the neutral database (right). All samples are taken from the middle of the second vowel of the utterance *Vad heter du?* (What is your name?).

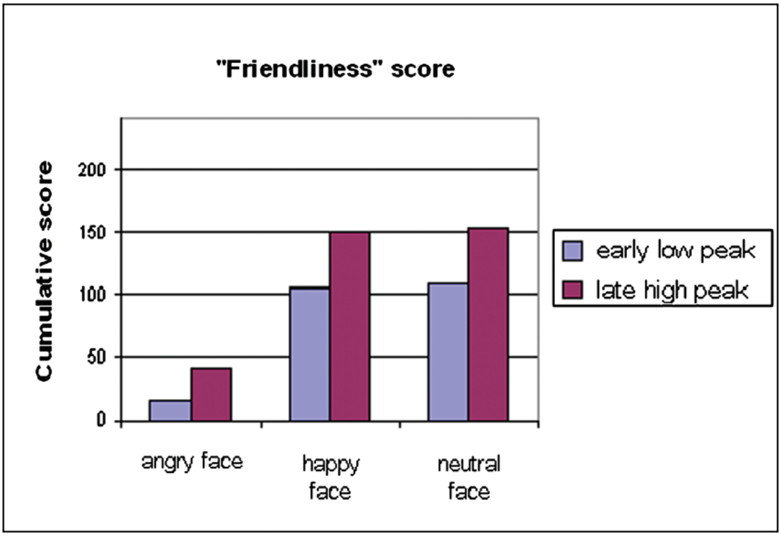


Figure 13. Results from the data-driven synthesis test showing the cumulative "friendliness score" for each stimulus.

peak, resulting in six stimuli. Previous experiments showed that the high, late pitch peak elicited more friendly responses (House, 2005). A perception test using these six stimuli was carried out by asking 27 subjects to indicate on a 4-point scale how friendly they felt the agent was.

The results are presented in Figure 13. It is quite clear that the face synthesised 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 a particularly high friendliness score. The happy face did not elicit more friendliness responses than did the neutral face, but the influence of the audio stimuli remained consistent for all the visual stimuli. Nevertheless, the results show that the visual modality can be a powerful signal of attitude. Moreover, the effects of the audio cues for friendliness indicate that subjects make use of both modalities in judging speaker attitude. These results stress the need to carefully consider both the visual and audio aspects of expressive synthesis.

9 Agent and System Evaluation Studies

The main emphasis of the evaluation studies reported on in this chapter has been the evaluation of the intelligibility and the dialogue functions of the talking head agent as presented to subjects in experimental test situations. During the last decade, however, a number of experimental applications using the talking head have been developed at KTH (see Gustafson, 2002 for a review). Two examples that will be mentioned here are the August project, which was a dialogue system in public use, and the Adapt multimodal real-estate agent. Finally, we will also report on some studies aimed at evaluating user satisfaction in general during exposure to the August and the Adapt dialogue systems.

9.1 The August System

The Swedish author, August Strindberg, provided inspiration to create the animated talking agent used in a dialogue system that was on display during 1998 as part of the activities celebrating Stockholm as the Cultural Capital of Europe (Gustafson et al., 1999). The system was a fully automatic dialogue system using modules for speech recognition and audio-visual speech synthesis. This dialogue system made it possible to combine several domains, thanks to the modular functionality of the architecture. Each domain had its own dialogue manager, and an example-based topic spotter was used to relay the user utterances to the appropriate dialogue manager. In this system, the animated agent “August” presents different tasks such as taking the visitors on a trip through the Department of Speech, Music, and Hearing, giving street directions, and also reciting short excerpts from the works of August Strindberg, when waiting for someone to talk to. The system was built into a kiosk and placed in public in central Stockholm for a period of 6 months. One of the main challenges of this arrangement was the open situation with no explicit instructions being given to the visitor. A simple visual “visitor detector” made August start talking about one of his knowledge domains.

To ensure that the recorded user utterances were actually directed to the system, a push-to-talk button was used to initiate the recordings. The speech recordings resulted in a database consisting of 10,058 utterances from 2,685 speakers. The utterances were transcribed orthographically and labelled for speaker characteristics and utterance types by Bell and Gustafson (1999a,b; see also Gustafson, 2002 and Bell, 2003 for recent reviews of this work). The resulting transcribed and labelled database has subsequently been used as the basis for a number of studies evaluating user behaviour when interacting with the animated agent in this open environment.

Gustafson and Bell (2000) present a study showing that about half the utterances in the database can be classified as socially oriented while the other half is information-seeking. Children used a greater proportion of socialising utterances than did adults. The large proportion of socialising utterances is explained by the presence of an animated agent, and by the fact that the system was designed to handle and respond to social utterances such as greetings and queries concerning some basic facts about the life of Strindberg. Furthermore, it was found that users who received an accurate response to a socialising utterance continued to use the system for a greater number of turns than did those users who were searching for information or those who did not receive an appropriate response to a socially oriented utterance.

In another study concerning phrase-final prosodic characteristics of user utterances comprising wh-questions, House (2005) found that final rises were present in over 20% of the questions. Final rises can indicate a more friendly type of question attitude and were often present in social-oriented questions. However, rises were also found in information-oriented questions. This could indicate that the intention to continue a social type of contact with the agent may not be restricted to questions that are semantically categorized as social questions. The social intention can also be present in information-oriented questions. Finally children's wh-question utterances as a group contained the greatest proportion of final rises followed by women's utterances, with men's utterances containing the lowest proportion of final rises. This could also reflect trends in social intent. These results can be compared to findings by Oviatt and Adams (2000) where children interacting with animated undersea animals in a computer application used personal pronouns with about one-third of the exchanges comprising social questions about the animal's name, birthday, friends, family, etc.

9.2 The Adapt Multimodal Real-Estate Agent

The practical goal of the Adapt project was to build a system in which a user could collaborate with an animated agent to solve complicated tasks (Gustafson et al., 2000). We chose 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 was given the role of asking questions and providing

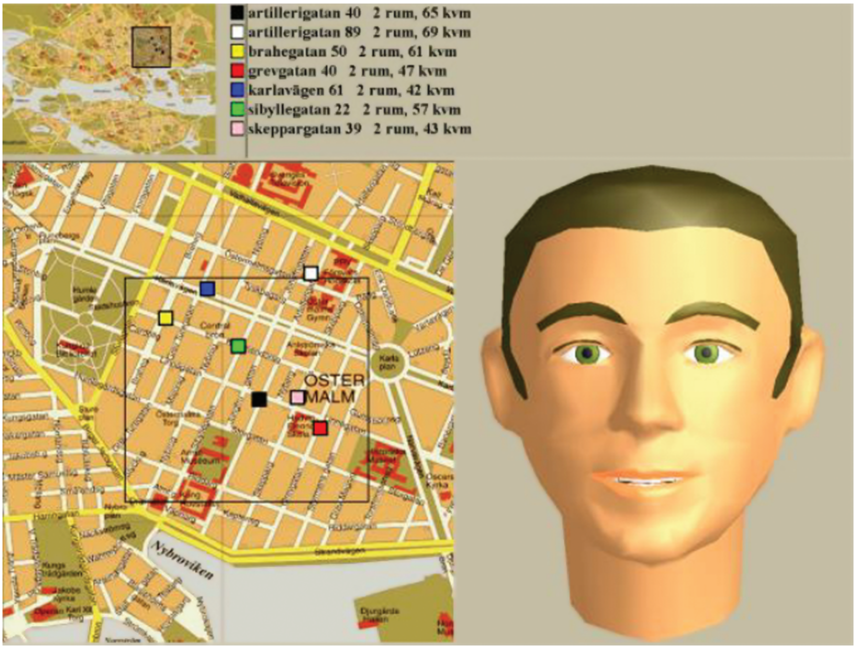

■ artillerigatan 40 2 rum, 65 kvm
□ artillerigatan 89 2 rum, 69 kvm
🟨 brahegatan 50 2 rum, 61 kvm
🟥 grevgatan 40 2 rum, 47 kvm
🟦 karlavägen 61 2 rum, 42 kvm
🟩 sibyllegatan 22 2 rum, 57 kvm
□ skeppargatan 39 2 rum, 43 kvm
OSTER MALM
Nybroviken

Figure 14. The agent Urban in the Adapt apartment domain.

guidance by retrieving detailed authentic information about apartments. The user interface can be seen in Figure 14.

Because of the conversational nature of the Adapt domain, the demand was great for appropriate interactive signals (both verbal and visual) for encouragement, affirmation, confirmation, and turntaking (Cassell et al., 2000; Pelachaud et al., 1996). As generation of prosodically grammatical utterances (e.g., correct focus assignment with regard to the information structure and dialogue state) was also one of the goals of the system, it was important to maintain modality consistency by simultaneous use of both visual and verbal prosodic and conversational cues (Nass and Gong, 1999). In particular, facial gestures for turntaking were implemented in which the agent indicated such states as attention, end-of-speech detection, continued attention, and preparing an answer (Beskow et al., 2002a; Gustafson, 2002).

Two different sets of data were collected from the Adapt system. The first collection was carried out by means of a Wizard-of-Oz simulation to obtain data for an evaluation of the prototype system under development. This first database represents 32 users and contains 1,845 utterances. The second database was collected in a study where 26 users interacted with a fully automated Adapt system. The second database comprises 3,939 utterances (Gustafson, 2002).

The study used to generate the second database was carried out in order to evaluate user reactions to the use of the agent's facial gestures for feedback (Edlund and Nordstrand, 2002). The users were split up into three groups and exposed to three different system configurations. One group was presented with a system which implemented facial gestures for turntaking in the animated agent, the second group saw an hourglass symbol to indicate that the system was busy but were provided with no facial gestures, and the third group had no turntaking feedback at all. The results showed that the feedback gestures did not produce an increase in efficiency of the system as measured by turntaking errors where the subjects started to speak during the time in which the system was preparing a response. However, users were generally more satisfied with the system configuration having the facial feedback gestures as reflected by responses in a user satisfaction form based on the method described in PARADISE (Walker et al., 2000).

In another evaluation of the Adapt corpus, Hjalmarsson (2005) examined the relationship between subjective user satisfaction and changes in a set of evaluation metrics over the approximately 30-minute time span of each user interaction. She found that users with high subjective satisfaction ratings tended to improve markedly during the course of the interaction as measured by task success. Users with low subjective satisfaction ratings showed a smaller initial improvement, which was followed by a deterioration in task success.

9.3 A Comparative Evaluation of the Two Systems

In a study designed to test new metrics for the evaluation of multimodal dialogue systems using animated agents, Cerrato and Ekeklint (2004) compared a subset of the August corpus with the Adapt Wizard-of-Oz simulation. Their hypothesis was that the way in which users ended their dialogues (both semantically and prosodically) would reveal important aspects of user satisfaction and dialogue success. The general characteristics of the dialogues differed substantially between the systems. The August dialogues were characterised by a small number of turns and frequent dialogue errors, while the dialogues in the Adapt simulations were much longer and relatively unproblematic. The final user utterances from both systems were analysed and classified as social closures or non-social closures. The social closures were then grouped into subcategories such as farewell, thanks, other courtesy expressions such as "nice talking to you", and non-courtesy expressions such as insults.

The comparison presented in Figure 15 shows a much greater percent of thanking expressions in the Adapt interactions than in those from the August corpus. While there were no insults ending Adapt interactions, insults and non-courtesy expressions comprised a fair proportion of the final utterances in the August interactions.

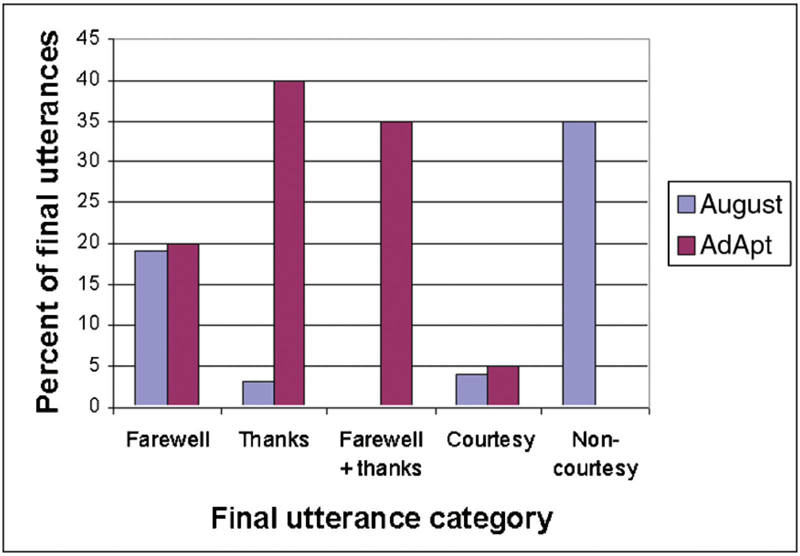


Figure 15. Distribution of social subcategories of the final utterances in the August and Adapt corpus. (Adapted from Cerrato and Ekeklint, 2004).

In addition to the category of final utterance, Cerrato and Ekeklint (2004) also analysed prosodic characteristics of the final utterances from the farewell and thanks category. They found a tendency for a farewell or thanks to have a rising intonation contour following a successful interaction with the system. They also found a tendency for users to end with a falling intonation, higher intensity, or greater duration in those cases where there had not been a successful interaction with the system.

10 Future Challenges in Modelling and Evaluation

In this chapter, we have presented an overview of some of the recent work in audio-visual synthesis, primarily 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 audio 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. As we demonstrated in the section on prominence above, head nods which strengthen the percept of prominence tend to be integrated with the nearest candidate syllable resulting in audio-visual coherence. However, head nods which indicate dialogue functions such as feedback or turntaking may not be integrated with the audio in the same way. Visual gestures can even be used to contradict or qualify the verbal message, which is often the case in ironic expressions. On the other hand, there are other powerful visual communicative cues such as the smile which clearly affect the resulting audio (through articulation) and must by definition be integrated with the speech signal. 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.

Acknowledgements The work reported here was carried out by a large number of researchers at the Centre for Speech Technology 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 organisations. Marc Swerts collaborated on the feedback study while he was a guest at CTT.

References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., and Öhman, T. (1998). Synthetic Faces as a Lipreading Support. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 3047–3050, Sydney, Australia.

Bälter, O., Engwall, O., Öster, A.-M., and Kjellström, H. (2005). Wizard-of-Oz Test of ARTUR – a Computer-Based Speech Training System with Articulation Correction. In *Proceedings of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility*, pages 36–43, Baltimore, Maryland, USA.

Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues; Empirical Studies of User Behavior*. Doctoral dissertation, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Bell, L. and Gustafson, J. (1999a). Interacting with an Animated Agent: An Analysis of a Swedish Database of Spontaneous Computer Directed Speech. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1143–1146, Budapest, Hungary.

Bell, L. and Gustafson, J. (1999b). Utterance Types in the August System. In *Proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems (IDS)*, pages 81–84, Kloster Irsee, Germany.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 299–302, Madrid, Spain.

Beskow, J. (1997). Animation of Talking Agents. In *Proceedings of ESCA Workshop on Audio-Visual Speech Processing (AVSP)*, pages 149–152, Rhodes, Greece.

Beskow, J. and Cerrato, L. (2006). Evaluation of the Expressivity of a Swedish Talking Head in the Context of Human-Machine Interaction. In *Proceedings of Gruppo di Studio della Comunicazione Parlata (GSCP)*, Padova, Italy.

Beskow, J., Cerrato, L., Cosi, P., Costantini, E., Nordstrand, M., Pianesi, F., Prete, M., and Svanfeldt, G. (2004a). Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces. In André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors, *Affective Dialogue Systems. Proceedings of the Irsee Tutorial and Research Workshop on Affective Dialogue Systems*, volume 3068 of *LNAI*, pages 301–304, Springer.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., and Svanfeldt, G. (2004b). Expressive Animated Agents for Affective Dialogue Systems. In André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors, *Affective Dialogue Systems. Proceedings of the Irsee Tutorial and Research Workshop on Affective Dialogue Systems*, volume 3068 of *LNAI*, pages 240–243, Springer.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordstrand, M., and Svanfeldt, G. (2004c). The Swedish PF-Star Multimodal Corpora. In *Proceedings of the LREC Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, pages 34–37, Lisbon, Portugal.

Beskow, J., Edlund, J., and Nordstrand, M. (2002a). Specification and Realisation of Multimodal Output in Dialogue Systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 181–184, Denver, Colorado, USA.

Beskow, J., Engwall, O., and Granström, B. (2003). Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. In *Proceedings of the International Congresses of Phonetic Sciences (ICPhS)*, pages 431–434, Barcelona, Spain.

Beskow, J., Granström, B., and House, D. (2001). A Multimodal Speech Synthesis Tool Applied to Audio-Visual Prosody. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M., editors, *Improvements in Speech Synthesis*, pages 372–382, John Wiley, New York, USA.

Beskow, J., Granström, B., and Spens, K.-E. (2002b). Articulation Strength - Readability Experiments with a Synthetic Talking Face. In *The Quarterly Progress and Status Report of the Department of Speech, Music and Hearing (TMH-QPSR)*, volume 44, pages 97–100, KTH, Stockholm, Sweden.

Beskow, J. and Nordenberg, M. (2005). Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head. In *Proceedings of the European Conference on Speech Communication and Technology (Interspeech)*, pages 793–796, Lisbon, Portugal.

Bickmore, T. and Cassell, J. (2005). Social Dialogue with Embodied Conversational Agents. In van Kuppevelt, J., Dybkjær, L., and Bernsen, N. O., editors, *Advances in Natural Multimodal Dialogue Systems*, pages 23–54, Springer, Dordrecht, The Netherlands.

Brennan, S. E. (1990). *Seeking and Providing Evidence for Mutual Understanding*. Unpublished doctoral dissertation, Stanford University, Stanford, California, USA.

Carlson, R. and Granström, B. (1997). Speech synthesis. In Hardcastle, W. and Laver, J., editors, *The Handbook of Phonetic Sciences*, pages 768–788, Blackwell Publishers, Oxford, UK.

Cassell, J., Bickmore, T., Campbell, L., Hannes, V., and Yan, H. (2000). Conversation as a System Framework: Designing Embodied Conversational Agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 29–63, MIT Press, Cambridge, Massachusetts, USA.

Cerrato, L. and Ekeklint, S. (2004). Evaluating Users' Reactions to Human-like Interfaces: Prosodic and Paralinguistic Features as Measures of User Satisfaction. In Ruttkay, Z. and Pelachaud, C., editors, *From Brows to Trust: Evaluating Embodied Conversational Agents*, pages 101–124, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Clark, H. H. and Schaeffer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, 13:259–294.

Cole, R., Massaro, D. W., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., and Solcher, D. (1999). New Tools for Interactive Speech and Language Training: Using Animated Conversational Agents in the Classrooms of Profoundly Deaf Children. In *Proceedings of the ESCA/Socrates Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)*, pages 45–52, University College London, London, UK.

Edlund, J. and Nordstrand, M. (2002). Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. In *Proceedings of the ISCA Workshop on Multi-Modal Dialogue in Mobile Environments*, pages 181–184, Kloster Irsee, Germany.

Engwall, O. (2003). Combining MRI, EMA and EPG Measurements in a Three-Dimensional Tongue Model. *Speech Communication*, 41(2–3): 303–329.

Engwall, O. and Beskow, J. (2003). Resynthesis of 3D Tongue Movements from Facial Data. In *Proceedings of the European Conference on Speech*

Communication and Technology (Eurospeech), pages 2261–2264, Geneva, Switzerland.

Gill, S. P., Kawamori, M., Katagiri, Y., and Shimojima, A. (1999). Pragmatics of Body Moves. In *Proceedings of the Third International Cognitive Technology Conference*, pages 345–358, San Francisco, USA.

Granström, B. (2004). Towards a Virtual Language Tutor. In *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, pages 1–8, Venice, Italy.

Granström, B., House, D., Beskow, J., and Lundeberg, M. (2001). Verbal and Visual Prosody in Multimodal Speech Perception. In von Dommelen, W. and Fretheim, T., editors, *Nordic Prosody: Proceedings of the Eighth Conference, Trondheim 2000*, pages 77–88, Peter Lang, Frankfurt am Main, Germany.

Granström, B., House, D., and Lundeberg, M. (1999). Prosodic Cues in Multimodal Speech Perception. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pages 655–658, San Francisco, USA.

Granström, B., House, D., and Swerts, M. G. (2002). Multimodal Feedback Cues in Human-Machine Interactions. In Bel, B. and Marlien, I., editors, *Proceedings of the Speech Prosody Conference*, pages 347–350, Laboratoire Parole et Langage, Aix-en-Provence, France.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems; Empirical Studies of Spoken Human-Computer Interaction*. Doctoral dissertation, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Gustafson, J. and Bell, L. (2000). Speech Technology on Trial: Experiences from the August System. *Natural Language Engineering*, 6(3–4):273–296.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., and Wirén, M. (2000). Adapt – a Multimodal Conversational Dialogue System in an Apartment Domain. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 134–137, Beijing, China.

Gustafson, J., Lindberg, N., and Lundeberg, M. (1999). The August Spoken Dialogue System. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1151–1154, Budapest, Hungary.

Hirschberg, J., Litman, D., and Swerts, M. (2001). Identifying User Corrections Automatically in Spoken Dialogue Systems. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 208–215, Pittsburg, USA.

Hjalmarsson, A. (2005). Towards User Modelling in Conversational Dialogue Systems: A Qualitative Study of the Dynamics of Dialogue Parameters.

In *Proceedings of the European Conference on Speech Communication and Technology (Interspeech)*, pages 869–872, Lisbon, Portugal.

House, D. (2001). Focal Accent in Swedish: Perception of Rise Properties for Accent 1. In van Dommelen, W. and Fretheim, T., editors, *Nordic Prosody 2000: Proceedings of the Eighth Conference*, pages 127–136, Trondheim, Norway.

House, D. (2002). Intonational and Visual Cues in the Perception of Interrogative Mode in Swedish. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1957–1960, Denver, Colorado, USA.

House, D. (2005). Phrase-final Rises as a Prosodic Feature in Wh-Questions in Swedish Human-Machine Dialogue. *Speech Communication*, 46:268–283.

House, D. (2006). On the Interaction of Audio and Visual Cues to Friendliness in Interrogative Prosody. In *Proceedings of the Second Nordic Conference on Multimodal Communication*, pages 201–213, Gothenburg University, Sweden.

House, D., Beskow, J., and Granström, B. (2001). Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 387–390, Aalborg, Denmark.

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002a). Perceptual Evaluation of Audiovisual Cues for Prominence. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1933–1936, Denver, Colorado, USA.

Krahmer, E., Swerts, M., Theune, M., and Weegels, M. (2002b). The Dual of Denial: Two Uses of Disconfirmations in Dialogue and their Prosodic Correlates. *Speech Communication*, 36(1–2):133–145.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle*. MIT Press, Cambridge, Massachusetts, USA.

Massaro, D. W. (2002). Multimodal Speech Perception: A Paradigm for Speech Science. In Granström, B., House, D., and Karlsson, I., editors, *Multimodality in Language and Speech Systems*, pages 45–71. Kluwer Academic Publishers, The Netherlands.

Massaro, D. W., Bosseler, A., and Light, J. (2003). Development and Evaluation of a Computer-Animated Tutor for Language and Vocabulary Learning. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pages 143–146, Barcelona, Spain.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1996). Perception of Asynchronous and Conflicting Visual and Auditory Speech. *Journal of the Acoustical Society of America*, 100(3):1777–1786.

Massaro, D. W. and Light, J. (2003). Read My Tongue Movements: Bimodal Learning to Perceive and Produce Non-Native Speech /r/ and /l/.

In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2249–2252, Geneva, Switzerland.

Nakano, Y., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a Model of Face-to-Face Grounding. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 553–561, Sapporo, Japan.

Nass, C. and Gong, L. (1999). Maximized Modality or Constrained Consistency? In *Proceedings of Auditory-Visual Speech Processing (AVSP)*, pages 1–5, Santa Cruz, USA.

Nordstrand, M., Svanfeldt, G., Granström, and House, D. (2004). Measurements of Articulatory Variation in Expressive Speech for a Set of Swedish Vowels. In *Speech Communication*, volume 44, pages 187–196.

Oviatt, S. L. and Adams, B. (2000). Designing and Evaluating Conversational Interfaces with Animated Characters. In *Embodied Conversational Agents*, pages 319–343, MIT Press, Cambridge, Massachusetts, USA.

Pandzic, I. S. and Forchheimer, R., editors (2002). *MPEG Facial Animation – The Standard, Implementation and Applications*. John Wiley Chichester, UK.

Parke, F. I. (1982). Parameterized Models for Facial Animation. *IEEE Computer Graphics*, 2(9):61–68.

Pelachaud, C. (2002). Visual Text-to-Speech. In Pandzic, I. S. and Forchheimer, R., editors, *MPEG-4 Facial Animation – The Standard, Implementation and Applications*. John Wiley, Chichester, UK.

Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating Facial Expressions for Speech. *Cognitive Science*, 28:1–46.

Shimojima, A., Katagiri, Y., Koiso, H., and Swerts, M. (2002). Informational and Dialogue-Coordinating Functions of Prosodic Features of Japanese Echoic Responses. *Speech Communication*, 36(1–2):113–132.

Sjölander, K. and Beskow, J. (2000). WaveSurfer – an Open Source Speech Tool. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 464–467, Beijing, China.

Srinivasan, R. J. and 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.

Svanfeldt, G. and Olszewski, D. (2005). Perception Experiment Combining a Parametric Loudspeaker and a Synthetic Talking Head. In *Proceedings of the European Conference on Speech Communication and Technology (Interspeech)*, pages 1721–1724, Lisbon, Portugal.

Tisato, G., Cosi, P., Drioli, C., and Tesser, F. (2005). INTERFACE: A New Tool for Building Emotive/Expressive Talking Heads. In *Proceedings of the European Conference on Speech Communication and Technology (Interspeech)*, pages 781–784, Lisbon, Portugal.

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Rochester, USA.

Walker, M. A., Kamm, C. A., and Litman, D. J. (2000). Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*, 6(3–4):363–377.

Westervelt, P. J. (1963). Parametric Acoustic Array. *Journal of the Acoustical Society of America*, 35:535–537.