Multimodality and speech technology: Verbal and non-verbal communication in talking agents

Björn Granström and David House

Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Stockholm, Sweden

{bjorn|davidh}@speech.kth.se

Abstract

This paper presents methods for the acquisition and modelling of verbal and non-verbal communicative signals for the use in animated talking agents. This work diverges from the traditional focus on the acoustics of speech in speech technology and will be of importance for the realization of future multimodal interfaces, some experimental examples of which are presented at the end of the paper.

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 neurophysiology enables us to effortlessly integrate information from different modalities fusing data to optimally meet the current communication needs.

As advances are made in systems for human-machine interaction, it has become increasingly important to make use of multiple modalities. The exploitation of several modalities can increase the naturalness and ease of human-system communication. Ease of communication, however, is not the only benefit of using several modalities in language and speech systems. From a system-design point of view, different modes of presentation can complement each other and provide different types of information. They can also reinforce each other and provide more complete information for the user during a difficult task, letting the user make use of different sensory modalities. The use of multiple channels is important for being able to adapt the system to different environments, terminals and possible sensory impairments.

As we contribute to advances in spoken dialogue systems and see them being integrated into commercial products, we are witnessing a transformation of the user interface, a transition from the desktop metaphor to the person metaphor. As we attempt to take advantage of the effective communication potential of human conversation, we see an increasing need to embody the conversational partner using audiovisual verbal and non-verbal communication implying the use and integration of both audio and visual modalities [1].

Effective interaction in dialogue systems involves both the presentation of information and the flow of interactive dialogue. A talking animated agent can provide the user with an interactive partner whose goal is to take the role of the human agent. An effective agent is one who is capable of supplying the user with relevant information, can fluently answer questions concerning complex information and can ultimately assist the user in a decision making process through the interactive flow of conversation. The use of the talking head also aims at increasing effectiveness by building on the user's social skills to improve the flow of the dialogue. Visual cues to feedback, turntaking and signalling the system's internal state (the thinking metaphor) are key aspects of effective interaction.

This paper addresses three key challenges that we face as we develop and refine interactive talking agents for use in future applications, namely 1) how to obtain data, 2) how to model it and 3) how to exploit it in dialogue systems. In the context of a current EU-funded research project, PF-Star, effort is being placed on these challenges, particularly on verbal and non-verbal features of audio-visual spoken language communication. This paper presents some of the current methods used to gather data and model it and also presents some examples of experimental applications in which an expressive talking head is used.

2. Data acquisition

2.1. Internal and external articulation

In our work on three-dimensional models for articulatory and visual speech synthesis at KTH, we have exploited several kinds of data sources. For the externally visible articulators and the facial surface, we have used optical motion tracking. For modelling of the tongue and internal vocal tract, three-dimensional data from magnetic resonance imaging (MRI) and kinematic data from electropalatography (EPG) and electromagnetic articulography (EMA) have been used [2]. While each of these methods in isolation can provide useful information, none yields complete 3D data with good temporal resolution, and they hence need to be combined. We have recently performed simultaneous measurements of vocal tract and facial motion using EMA and optical motion tracking. The data is used to improve and extend the articulation of an animated talking head.

Previous related studies [3][4][5] concluded that information from the face supplies information on the articulation of the speech organs, but [3] warned that the information is insufficient to recover the lingual constriction. The most important difference is that none of these studies aims directly at applying the results to articulatory speech synthesis of the face, jaw and the entire tongue.

The EMA data is collected with the Movetrack system [6] using two transmitters on a light-weight head mount and six receiver coils (1.5x4 mm) positioned in the midsagittal plane

as depicted in figure 1: three coils on the tongue (around 8 mm, 20 mm and 52 mm from the tip of the tongue) and two coils above and below the upper and lower incisors respectively. One coil was placed on the upper lip for co-registration with the optical system.

The optical motion tracking is done using a Qualisys system (http://www.qualisys.se) with four cameras (see figure 2). The system tracked 28 small reflectors (4 mm diameter) glued to the subject's jaw, cheeks, lips, nose and eyebrows and the Movetrack headmount (to serve as reference for head movements) and calculated their 3D-coordinates at a rate of 60 frames per second. The coherent data gives us possibilities of increasing the realism of the model [7]. Potentially the articulatory 3D model could be used for direct generation of the acoustic speech, though this has not been our focus so far.
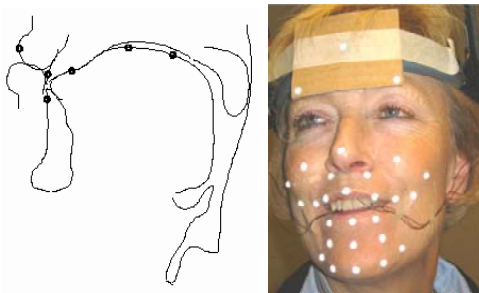

Figure 2. Data collection setup with video, micro-phone and four IR-cameras and a screen for prompts (left). Facial markers used for the recording (right).

3. Implementation

3.1. Generic coding

Animated synthetic talking faces and characters have been developed using a number of different techniques and for a variety of purposes during the past two decades. Our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework [12]. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account [13]. We employ a generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation [14]. Based on concepts first introduced by Parke [15], we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. One critical difference from Parke's system, however, is that we have de-coupled the model definitions from the animation engine, thereby greatly increasing flexibility. The different models shown in figure 3 are all dynamically controlled by the same control code. Modules for connecting our model to general coding schemes included in the MPEG-4 standard have been developed.

Figure 1. Marker placement for Movetrack (left) and Qualisys (right) measurements.

2.2. Speech in communication

Communicatively important movements of parts of the face occur both when talking and when silent. Also articulation is affected by attitudes, emotions and the communicative state. In an effort to obtain hard data on these interactions we used a recording technique similar to the one just described. As can be seen in figure 2 the placements of the reflective spots are quite different. Eyebrow, eyelid and forehead spots were added for possibly picking up non-articulatory movements. The eyeglass spots were used as a skull-based reference and the three lowest points as reference points for the chest. Thus head movements can be related to the torso [8]. The sentences to be read and acted were shown on a screen. We chose to record 15 different expressions. Together with the six universal prototypes for emotions: anger, fear, surprise, sadness, disgust and happiness [9], we also had the subject act worried, satisfied, insecure, confident, questioning, encouraging, doubtful, confirming and neutral. The sentences were kept neutral with respect to content. The material was automatically transcribed and transcriptions were then manually corrected to match what was actually read by the actor. An automatic aligner [10] was used to pair the phonetically transcribed speech with the sound signal to retrieve the time for phoneme and word boundaries. This information together with the sync-signal was then used to match 3D-data with speech for analysis. This method is also being used in conjunction with traditional video recordings in face-to-face dialogues [11].

Figure 3. Some different versions of the KTH talking head

3.2. Expressive wrinkles

In a recent study we wanted to include the possibility of adding real time dynamic wrinkles to the computer-generated face. The purpose is to extend an existing virtual face to make it more realistic and expressive. Wrinkles modelled as a fine grained mesh of polygons would today be totally prohibitive, from a computational point of view, so we implemented a different, but flexible solution [16]. The solution consists mainly of two parts. The first is to calculate the compression of the skin when the face is deformed. This is achieved through geometric calculations on the triangles the face is composed of. The compression is then converted to wrinkle intensity using a translation function. The wrinkle intensity is then used to graphically render wrinkles on the face. The

algorithm uses a technique called bump mapping to visualize the wrinkles. The implemented prototype shows that real time dynamic wrinkles can be implemented using existing hardware. The bump mapping algorithm has a few limitations, especially regarding the lighting model, but these limitations are hardly noticed in this fairly simple application, see figure 4, where only the wrinkles connected to raised eyebrows for e.g. surprise are modelled.
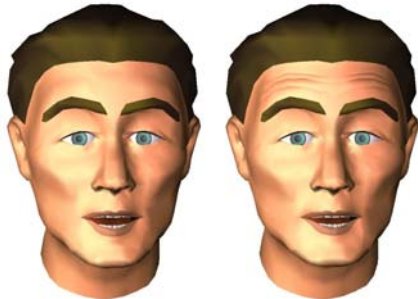

Figure 4. The face model with and without wrinkles

3.3. Gesture libraries

To be able to use our derived gestures in applications we are at present developing an XML-based representation of such cues that facilitates description of both verbal and visual cues at the level of speech generation. These cues can be of varying range covering attitudinal settings appropriate for an entire sentence or conversational turn or be of a shorter nature like a qualifying comment to something just said. Cues relating to turntaking or feedback need not be associated with speech acts but can occur during breaks in the conversation. Also in this case, it is important that there is a one-to-many relation between the symbols and the actual gesture implementation to avoid stereotypic agent behaviour. [17].

4. Experimental applications

4.1. Multimodal dialogue systems

One example of using the talking head in an experimental dialogue system is the AdApt project. The practical goal of the project is to build a system in which a user can collaborate with an animated agent to solve complicated tasks [17][18]. We have chosen a domain in which multimodal interaction is highly useful, and which is known to engage a wide variety of people in our surroundings, namely, finding available apartments in Stockholm. In the AdApt project, the agent has been given the role of asking questions and providing guidance by retrieving detailed authentic information about apartments. The user interface can be seen in figure 5.

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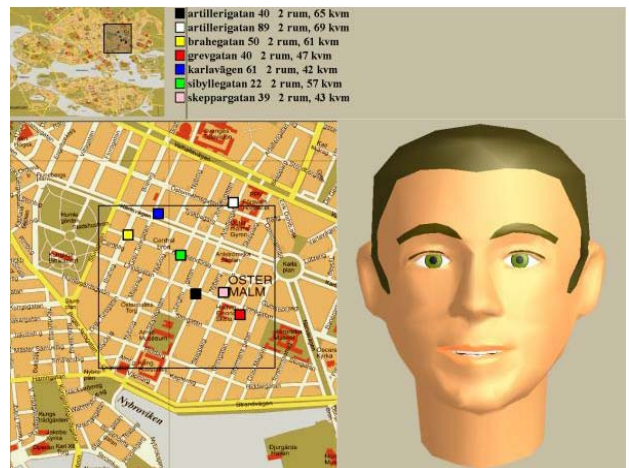

Figure 5. The agent Urban in the AdApt domain.

4.2. Communication aids

The speech intelligibility of talking animated agents, as the ones described above, has been tested within KTH Teleface project [19] and the EU project, Synface [20]. The projects focus on the use of multi-modal speech technology for hearing-impaired persons. The projects evaluated the increased intelligibility hearing-impaired persons experience from an auditory signal when it is complemented by a synthesised face. A demonstrator of a system for telephony with a synthetic face that articulates in synchrony with a natural voice is currently being implemented (see figure 6). While the emphasis in this project has been on intelligibility, non-verbal features like turn taking signalling could be used to smooth the communication, despite the delay that the reconstruction of the face image necessarily implies.
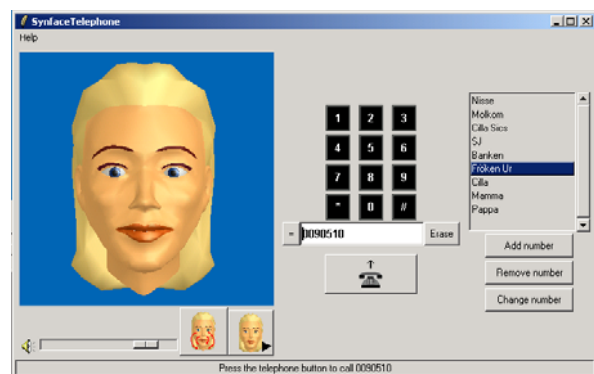

Figure 6. Telephone interface for SYNFACE.

4.3. Language tutor

The effectiveness of language teaching is often contingent upon the ability of the teacher to create and maintain the interest and enthusiasm of the student. The success of second language learning is also dependent on the student having ample opportunity to work on oral proficiency training with a tutor. The implementation of animated agents as tutors in a multimodal spoken dialogue system for language training holds much promise towards fulfilling these goals. Different agents can be given different personalities and different roles,

which should increase the interest of the students. Many students may also be less bashful about interacting with an agent who corrects their pronunciation errors than they would be making the same errors and interacting with a human teacher. Instructions to improve pronunciation often require reference to phonetics and articulation in such a way that is intuitively easy for the student to understand. An agent can demonstrate articulations by providing sagittal sections which reveal articulator movements normally hidden from the outside. This type of visual feedback is intended to both improve the learner's perception of new language sounds and to help the learner in producing the corresponding articulatory gestures by internalising the relationships between the speech sounds and the gestures. The articulator movements of such an agent can also be synchronised with natural speech at normal and slow speech rates. Furthermore, pronunciation training in the context of a dialogue automatically includes training of both individual phonemes, sentence prosody and communication skills. As can be seen in figure 7, the agent also provides different display possibilities for showing internal articulations.
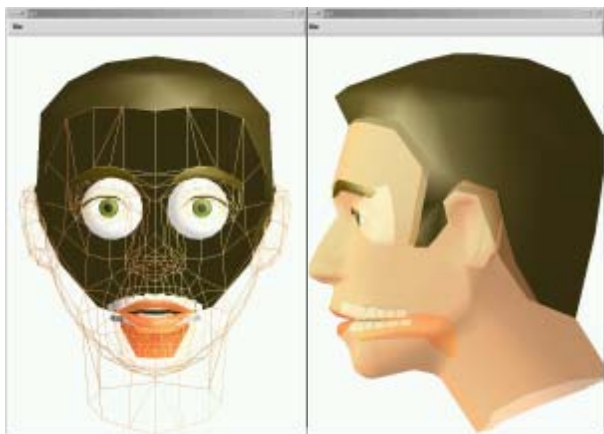

Figure 7. *Different displays of the talking head.*

5. Acknowledgements

6. References

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

[2] Engwall O. (2002) Tongue Talking – Studies in Intraoral Speech Synthesis, PhD Thesis, KTH, Sweden

[3] Bailly G. and P. Badin. "Seeing tongue movements from outside", *Proc. of ICSLP 2002,* pp 1913-1916, 2002.

[4] Jiang J., A. Alwan, L. Bernstein, P. Keating and E. Auer (2000) On the correlation between facial movements, tongue movements and speech acoustics", Proc of ICSLP2000, vol.1, pp. 42-45

[5] Yehia H., P. Rubin and E. Vatikiotis-Bateson, (1998) Quantitative association of vocal-tract and facial behaviour, Speech Communication, vol. 26, pp. 23-43

[6] Branderud P (1985) Movetrack - a movement tracking system", Proc of the French-Swedish Symposium on Speech, Grenoble, pp. 113-122

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

[8] Svanfeldt, G., Nordstrand, M., Granström, B. and House, D. "Measurements of articualtory variation in expressive speech" In Proc Fonetik 2003, Phonum 9: 53-56, Umeå, 2003

[9] Ekman, P. (1982) Emotion in the human face. New York: Cambridge University Press.

[10] Sjölander, K. (2003) .An HMM-based system for automatic segmentation and alignment of speech. In Proceedings of Fonetik 2003, Umeå, Sweden.

[11] Cerrato, L. and Skhiri, M. "Quantifying gestures in face-to-face human dialogues", to appear in Proceedings of the 1st Nordic Symposium on Multimodal Communication 25-26 September, 2003, Copenhagen.

[12] Carlson R and Granström B "Speech Synthesis", In Hardcastle W and Laver J (Eds.). The Handbook of Phonetic Sciences, 768-788, Oxford: Blackwell Publishers Ltd., 1997.

[13] Beskow, J. *Talking heads – models and applications for multimodal speech synthesis*. PhD thesis, TMH/KTH, 2003.

[14] Beskow, J. (1997) Animation of talking agents. In Proceedings of ESCA workshop on audio-visual speech processing, Rhodes, Greece, pp. 149-152.

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

[16] Nordenberg, M. "Modelling and rendering dynamic wrinkles in a virtual face" TMH/KTH MSc thesis , 2003 (available at http://www.speech.kth.se/qpsr/masterproj/)

[17] Edlund J, Beskow J and Nordstrand M (2002). GESOM - A Model for Describing and generating Multi-modal Output. Proc of ISCA Workshop Multi-Modal Dialogue in Mobile Environments.

[18] Gustafson J (2002). Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction. Doctoral Thesis. Department of Speech, Music and Hearing, KTH, Stockholm

[19] Agelfors E, Beskow J, Dahlquist M, Granström B, Lundeberg M, Salvi G, Spens K-E & Öhman T (1999). Synthetic visual speech driven from auditory speech. Proc of AVSP 99, 123-127

[20] Siciliano, C., Williams, G., Beskow, J. & Faulkner, A. (2003) "Evaluation of a Multilingual Synthetic Talking Face. as a Communication Aid for the Hearing Impaired." Proceedings of 15th International Congress of Phonetic Sciences.