PAPER 2.

Beskow, J. (1997). Animation of Talking Agents. In Benoît, C. and Campbell, R. (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'97)*, Rhodos, Greece, pp. 149-152.

ANIMATION OF TALKING AGENTS

Jonas Beskow

Centre for Speech Technology
KTH, Stockholm, Sweden

ABSTRACT

It is envisioned that autonomous software agents that can communicate using speech and gesture will soon be on everybody's computer screen. This paper describes an architecture that can be used to design and animate characters capable of lip-synchronised synthetic speech as well as body gestures, for use in for example spoken dialogue systems. A general scheme for computationally efficient parametric deformation of facial surfaces is presented, as well as techniques for generation of bimodal speech, facial expressions and body gestures in a spoken dialogue system. Results indicating that an animated cartoon-like character can be a significant contribution to speech intelligibility, are also reported.

INTRODUCTION

Human communication is inherently multimodal. The information conveyed through body language, facial expression, gaze, intonation, speaking-style etc. are all important components of everyday communication. If a machine can gain access to such communication channels and be taught to generate as well as interpret social nonverbal signals in an efficient and meaningful way, a completely new model of interacting with computers would be possible. Many people predict that the next big step in user interface technology will be a shift towards multimodal agent-based systems, where the user interacts with the system through several channels, including spoken natural language. Such interfaces will possess social capabilities as well as a high degree of autonomy and will be able to carry out complex requests on their own. The interaction between the user and the system will primarily be concerned with *delegation* rather than *direct manipulation* as is the case in most interfaces today.

VISIBLE AGENTS

Several aspects of the technology required for creation of multimodal agent interfaces are research areas in their own right. For example, spoken dialogue systems research has lately been a very active area, and to date, there are also several examples of animated characters in spoken dialogue interfaces.

Faces in dialogue

Katashi & Akikazu (1994) employed animated facial expressions as a backchanneling mechanism as well as an indicator of the internal state (listening, thinking etc.) in a speech-based dialogue system. Thórisson (1996) used a 2D animated character together with input from several sources, including speech and gaze, to simulate many social aspects of multimodal dialogue interaction such as turn-taking and backchanneling. Cassel et al (1994) modelled the interaction between intonation and gesture in dialogue context using two animated agents interacting with each other. In the Waxholm project at KTH (Bertenstam et.al. 1995), gaze and head movements of an animated talking face was used to visually refer to on-screen graphics, such as maps and timetables in a graphical display.

Besides the benefits of added output modalities for nonverbal communication, there are other advantages of using an animated character in a speech-based system. A visually well-articulated interface agent, using bimodal speech synthesis, will make the system more intelligible, compared to an interface with auditory-only speech synthesis as output. The addition of visual speech facilitates usage in noisy environments or by hearing-impaired people. Furthermore, a visual agent gives the system a more apparent personality and makes it anthropomorphic, which might contribute to making the dialogue experience more like human face-to-face communication.

Animation issues

Animation of human, or humanoid, faces is a key technology in this context. There exist several approaches to facial animation, ranging from display of pre-stored images to actual physical simulation. Platt & Badler (1981), Waters (1987) and Terzopoulos and Waters (1990) have all developed models that simulate the muscles and tensions in the facial tissue. Such models can provide realistic results, but determining proper muscle activation levels required to create a specific facial expression can be a difficult matter, since real muscles don't lend themselves to easy measurement. Furthermore, muscle-based models tend to be computationally expensive, which is a serious drawback if the intended application is an interactive interface agent. More suitable in this respect is the direct parameterisation approach taken by Parke(1982). Here, attention is turned solely to the outside surface of the face, and instead of muscles, a set of observation-based parameters are used to deform the surface to create facial expressions. Such surface deformations can often be of low computational complexity. Since the parameters are not required to model any real anatomical features, they can be tailored to mimic particular actions, such as speech movements (Beskow, 1995). Furthermore, parameters can relatively easily be measured from real faces, either manually or using image processing techniques.

A common drawback of direct parameterisation methods as well as some muscle-based schemes, however, is that activaion of several parameters simultaneously can yield unpredictable results. Consider a scheme with independent control parameters for upper and lower lip, as well as one parameter for jaw opening. In order to close the lips tightly together, we need to specify an exact combination of all three parameters. It's possible to solve this kind of problem using collision detection and iterative parameter adjustment, but this dramatically increases the computational demand.

A TOOLKIT FOR ANIMATION OF INTERACTIVE CHARATERS

To be able to experiment with different animated characters for dialogue systems and other purposes in a flexible way, we have developed a toolkit for character animation and visual speech synthesis.

The toolkit is useful for rendering and animation of arbitrary three-dimensional polygon objects. The main purpose, however, is modelling, animation and real-time control of parametrically controlled humanoid characters.

A general surface deformation scheme

We have developed a parameterisation scheme for nonrigid deformation of polygon surfaces, suitable for real-time animation of a 3D-character. This scheme borrows features from Parkes approach, but tries to overcome some if it's shortcomings. Particularly three main considerations guided the design:

- Predictable results of parameter combination
- Rapid parameterisation of new models
- Computational efficiency

The first requirement is handled by defining all motions in terms of *target points* and *prototype points*. For example, rather than defining tongue tip elevation using the angle of rotation, which would make the final position of the tongue tip depend on jaw opening as well, the motion is normalised with respect to a *target point* on the hard palate, that the tongue tip (the *prototype point*) is required to reach when the parameter is fully activated. This makes the final position of the articulator predictable, and independent of other parameters such as jaw rotation.

Rapid creation of new parameter sets is facilitated by a graphical *deformation editor*, where properties of each deformation parameter can be tailored interactively, and the effects on the model can be viewed instantaneously. This is possible since parameterisation data is separated from application code, so that changes in parameterisation can be done without the need for re-compilation. In fact, all data describing a character's geometry as well as deformation parameterisation can be stored together in one file.

Computational efficency is achieved by only letting each deformation parameter operate on a subset of the vertices, as opposed to schemes where all parameters modify all vertices in the model.

The deformator

The central entity in the deformation scheme is called the *deformator*. A deformator is an object that transforms a set of vertices according to some rule. Several deformators may be controlled by one parameter. The exact effect of a *deformator* is specified by a set of properties:

- *Transformation type* which is either *rotation*, *non-uniform scaling*, *translation* or *pull*
- *Influence definition*: a list of vertex-weight pairs that defines which polygon vertices should be affected by the transformation and to what extent
- A *pivot point* that defines the point of no movement for rotation and scaling operations
- A *prototype point* which typically is a central vertex in the influence area of the deformation, that acts as a "role-model" for all other vertices
- A *target point* which typically is a vertex outside of the influence area thowards which the *prototype* will move
- The *activation factor*: a control parameter with a value between zero and one, that determines the current application of the deformator.

When a deformator is applied with an activation factor A, the first step is to calculate amount of transformation, X_p, required to transform the *prototype* point to such an extent that it comes as close to the target as possible. Depending on the selected *transformation type*, X_p is either a scalar, e.g. angle of rotation, or a vector, such as a translation vector. Next, each vertex i in the influence definition is transformed by X_i, which is given by

$$X_i = X_p \cdot A \cdot w_i$$

where A is the current activation factor and w_i is the weight of vertex i.

Control points and influence definitions of a deformator can be specified interactively in a graphical editor.

Implementation aspects

The core of the system is written in C++ and it uses standard 3D-graphics libraries to perform the rendering. All control and interaction is done through a scripting language front end based on Tcl/Tk (Ousterhout, 1994) that acts both as a command line interpreter and as a general extension language. The extension language is useful for high level control of

the character's speech, facial expressions and gestures. Furthermore it facilitates customisation of the environment as well as easy integration with other components and applications without the need for re-compilation of the core system. The Tcl/Tk front-end also makes it easy to create specialised graphical interfaces for specific tasks.

THE BIRTH OF AN AGENT

This scheme makes it possible to start with any static 3D-geometry, such as a polygonal face model created in an external 3D-modelling program, import it into the toolkit, and interactively define a set of articulatory parameters, such as jaw opening and lip rounding as well as eyelid- and eyebrow parameters. To date, this process has only been completed for one character, *Olga*, but future additions to our agent inventory is planned.

A cartoon-like robot lady

The Olga project is a research project aiming to develop a multimodal spoken dialogue interface for information access (Beskow, McGlashan & Elenius, 1997). The domain deals with consumer information about microwave ovens, and the system combines spoken language input and output with character animation, 2D-graphics and direct manipulation. Within this project, the character called *Olga* was created (see Fig. 1b). Olga is an embodied cartoon-like female agent, capable of not only bimodal speech synthesis and facial expressions, but also manual gestures.

Rather than trying to create a realistic replication of a human being, Olga was intentionally modelled as a cartoon, with exaggerated proportions as well as some extravagant features such as antennas. The main reason for this has to do with user expectations; if the agent very realistically replicates a human, the user might get too high expectations of the system's social, linguistic and intellectual skills. A cartoon on the other hand, does not induce any such expectations, since the only experience most people have from cartoons comes from watching them, not *interacting* with them.

The parameter set

What control parameters need to be defined for a new character? For the purposes of agent animation, parameters can be divided in two distinct categories: those controlling the visual articulation required for speech synthesis, and those needed for non-verbal communication. The first category requires a parameter set that allows for generation of convincing mouth shapes for speech. For the Olga character, we have chosen to conform to a set of articulation parameters that was developed for a visual speech synthesiser based on an extended version of Parke's model (Beskow, 1995 and Fig. 1a). This set consists of seven parameters: *jaw rotation*, *mouth width*, *lip rounding*, *lip protrusion*, *labiodental occlusion*, *bilabial occlusion* and *tongue tip elevation*. All parameters are given in percent, relative to a well-defined target position for the articulator, except *jaw rotation* which is given in degrees.

Non-verbal controls can further be sub-divided into parameters for generation of non-speech facial expresions and parameters for body gesture control. In *Olga*, the facial parameters include *eyelid closure*, *smiling* and *eyebrow position* and *-shape*. These are all defined using the general deformation scheme described above. Body control parameters are implemented by introducing rotational joints at the neck, shoulders, elbows, wrists and fingers.

Rule-driven articulation

Temporal trajectories for the articulatory parameters are generated by a rule system for text-to-speech synthesis. Rules for synthesis of the visual modality have been developed for the extended Parke-model (Beskow, 1995). However, any model that adhere to the parameter set described above, can be controlled using this rule system, a fact that saved us a lot of

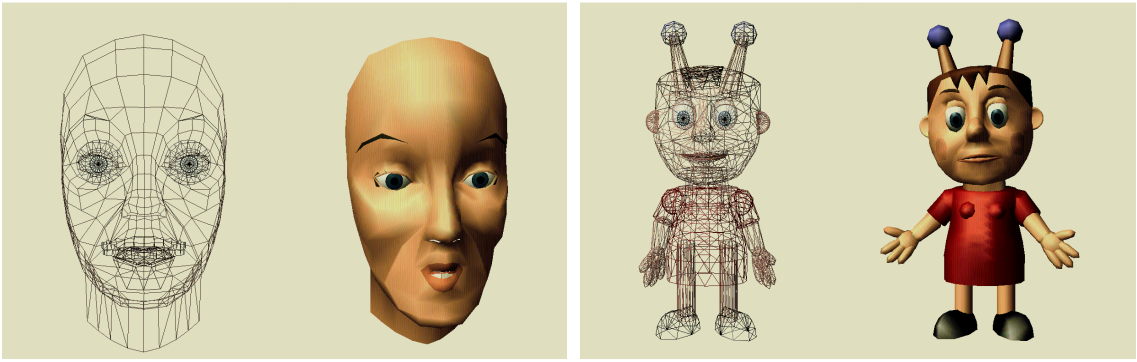

Fig. 1 (a) The extended Parke face (b) The Olga-character

work in implementing visual speech synthesis for the the Olga character. In general, sticking to a well-defined parameter set, allows for contol-startegies (such as the rule-system mentioned above) and face models to be developed independently of each other.
The rules implement a visual co-articulation strategy that is based on under-specification of target values. Typically, parameters that are essential for perceived visual *vowel* quality, such as *lip rounding* and *mouth width*, are left unspecified for most *consonants*, and will be determined by the vowel context. Calculation of a parameter trajectory is performed in three steps. First, each segment is either assigned a target value based on it's phoneme classification or is left undefined. Next, undefined target values are determined by linear interpolation between the nearest defined segments. Finally, the resultant trajectory is computed by smoothing the step function defined by the target values of all segments.

Once the parameter trajectories are calculated, the animation is carried out in synchrony with play-back of the speech waveform, which in turn is generated by a formant synthesizer controlled from the same rule-synthesis framework.

Template-based expressions and gestures

While speech synthesis is generated on an utterance-by-utterance basis, dynamic generation of body movements and non-speech facial expressions, place other requirements on the animation system, since the time frame for planning of such actions often is shorter. Say for example that we want the agent to dynamically change it's expression during a user utterance to reflect the progress of the speech understanding progress. To allow for this kind of control, the character animation toolkit makes it possible to specify a parameter trajectory as a list of time-value pairs, to be evaluated immediately. Such trajectory commands can be issued at any time from the Tcl/Tk level. Arbitrarily complex gestures can then be defined by grouping several commands together in procedures. Since a general scripting language is used, gesture templates can also be parameterised by supplying arguments to the procedure. For example, a pointing gesture might take optional arguments defining direction of pointing, duration of the movement, degree of effort etc. The scripting approach makes it easy to experiment with new gestures and control schemes. In the Olga system, template based handling of facial expressions and gestures has proven to be a simple, yet quite powerful way of managing non-speech movements.

EVALUATION

Experience from cartoons in animated films, tells us that it's clear that we can map our perception of non-verbal cues, such as emotions, to cartoon like characters. But what about

visual speech, does our perception of this modality transfer equally well? Typically, mouth movements of cartoon characters in movies are not too convincing. Is this a problem inherent to cartoons, or is it just a case of bad modelling of speech movements (and of course dubbing in many cases)?

Intelligibility test

In an ongoing study (Beskow, Dahlquist, Granström, Lundeberg, Spens & Öhman, 1997) we have been measuring audio-visual intelligibility of synthetic and natural speech. We have been using two synthetic faces - the extended Parke-model and the Olga character - as well as video recordings of a natural male face. The visual channels have been cross-combined and synchronised with two different synthetic voices (male and female), as well as one natural male voice. In the first test series we have tested 18 normal hearing subjects on a set of VCV-utterances consisting of 17 different consonants in symmetric context with the vowels /a/ and /ʊ/. The test was performed for eight different audio-visual combinations. To avoid ceiling-effects, the audio was degraded by white noise to a signal-to-noise ratio of 3 dB.

Results

With the *synthetic male voice*, intelligibility score went from *30*% in the audio only case, to *45*% in the audio-visual case when the *Parke-model* was used. The corresponding audio-visual score for *Olga* was *47*%.

With the *natural voice*, intelligibility went from *62*% (audio only) to *70*% for the *Parke-model*. Corresponding audio-visual score for the *natural face* was 76%. Olga was not tested with the natural voice.

CONCLUSIONS

The technique for surface deformation parametrisation presented in this paper provides a reasonably simple and efficient approach to facial animation that has proven to work well for the Olga character. The low computational complexity makes it useful for interactive characters in speech-based agent interfaces.

The intelligibility study shows that both synthetic faces give a substantial contribution to the intelligibility, compared with the audio-only case. The intelligibility gain given by the Parke-face is slightly more than half of the gain given by the natural face.

Furthermore, it can be noticed that the exaggerated and cartoon-like features of the Olga character does not seem to degrade it's speechreadability in comparison with the more human-like Parke face. In fact, several subjects reported that they subjectively found Olga the easier of the two to speechread.

ACKNOWLEDGEMENTS

This work has carried out at the Centre for Speech Technology with funding from NUTEK and KFB.

REFERENCES

Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. (1995), "The Waxholm system - a progress report" In *Proceedings of Spoken Dialogue Systems*, Vigsø, Denmark.

Beskow, J. (1995) "Rule-based Visual Speech Synthesis" In *Proceedings of Eurospeech '95*, Madrid, Spain.

Beskow, J. Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E. & Öhman, T. (1997), "Multimodal speech communication for the hearing impaired", In *Proceedings of Eurospeech '97*, Rhodes, Greece.

Beskow, J., Elenius, K. and McGlashan, S. (1997), "Olga - A Dialogue System with an Animated Talking Agent", In *Proceedings of Eurospeech '97*, Rhodes, Greece.

Cassel, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S. and Achorn B. (1994), "Modeling the Interaction between Speech and Gesture", In *Proceedings of 16th Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology, Atlanta, USA.

Katashi, N. and Akikazu, T (1994) "Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation", *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 102-109.

Ousterhout, J. K. (1994), *'Tcl and the Tk toolkit.'* Addison Wesley.

Platt, S. M. and Badler, N. I. (1981), "Animating Facial Expressions" *Computer Graphics*, Vol. 15, No. 3, pp. 245-252.

Terzopoulos, D., Waters, K. (1990) "Physically based facial modelling, analysis and animation" *Visualisation and Computer Animation*, 1:73-80.

Thórisson, K. R. (1997), "Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People" In *Proceedings of First ACM International Conference on Autonomous Agents*, Mariott Hotel, Marina del Rey, California, USA. pp 536-537.

Waters, K. (1987), "A muscle model for animating three-dimensional facial expressions", *Computer Graphics*, 21:17-24.