

Dept. for Speech, Music and Hearing
**Quarterly Progress and
Status Report**

**Talking heads -
communication, articulation
and animation**

Beskow, J.

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 053-056



**KTH Computer Science
and Communication**

<http://www.speech.kth.se/qpsr>

Talking heads - communication, articulation and animation

Jonas Beskow

Department of Speech, Music and Hearing, KTH

Abstract

Human speech communication relies not only on audition, but also on vision, especially during poor acoustic conditions. The face is an important carrier of both linguistic and extra-linguistic information. Using computer graphics it is possible to synthesize faces and do audio-visual text-to-speech synthesis, a technique that has a number of interesting applications for example in the area of man-machine interfaces. At KTH, a system for rule-based audio-visual text-to-speech synthesis has been developed. The system is based on the KTH text-to-speech system which has been complemented with a three-dimensional parametric model of a human face, that is animated in real time in synchrony with the auditory speech. The audio-visual text-to-speech synthesis has also been incorporated into a system for spoken man-machine dialogue.

Introduction

The human face, regarded as a medium for visual communication, is extremely expressive, and in many situations an invaluable complement to the acoustic speech signal. Since we are confronted with faces from the very first moment of our lives, we are all more or less experts in both interpreting and producing all kinds of subtle facial cues. Thus, faces play a natural and significant role in everyday communication.

Various types of facial cues are present on different levels of the communication process. Firstly, facial expressions are perhaps the most important way of signaling emotion. We can immediately tell if a person is happy, sad, scared, angry etc. by simply looking at his/her face. Secondly, in verbal communication situations, the face express information related to discourse, phrasing, emphasis and dialogue turn-taking. In this sense facial expressions are intimately related, and often complementary, to prosodic features of the voice.

Thirdly, the face reveals some visible aspects of the speech production and thus also carries much information about the phonetic content of a spoken utterance. Visual speech information can greatly increase speech intelligibility especially during acoustically degraded conditions, regardless of whether the acoustic degradation is due to external noise or to hearing impairment, as has been shown by for example Erber (1969). Benoît et al., (1994) showed an audio-visual intelligibility score of over 60% for

French nonsense words at -24 dB SNR, where the auditory-alone score was close to zero.

In the last decade, the rapid development of high performance graphics hardware has made it possible to process and synthesize faces using computers. More specifically, computer-animated talking heads, that can be used in multi-modal human-computer interfaces, applications for hearing impaired people and perceptual experiments, are starting to appear.

The following section gives a brief overview of the history and the state of the art of this relatively new field of research.

Talking faces over the world

There exist a number of approaches to computer-controlled synthesis of human facial motion. One possibility is to use pre-stored images of real faces, carefully chosen to represent all the shapes that one is interested in reproducing, for example a set of visemes (A viseme is a visually distinguishable speech shape, often defined in terms of a phoneme cluster). While this method can be quite successful for limited applications, such as creating stimuli for perceptual experiments, a system like this will never be very flexible, since there is no way to control different facial features independently of each other.

Facial modelling

One way to obtain more flexibility, is to move to tree dimensional parametric facial models. Such a model can be viewed as a geometric description of the facial surface that can be

deformed using a limited set of parameters, and rendered using standard computer graphics techniques.

This approach was first taken by Parke (1982). He constructed a facial surface using planar polygons, and assigned parameters to deform it in various ways. The parameters in Parke's model have little or no anatomical motivation, they are instead based on observation. Other models, such as the one described by Waters (1987), on the other hand, use facial muscle activation parameters as the control scheme. While muscle based methods may seem more elegant, there are significant difficulties involved in modeling all the muscles needed to simulate for example articulate lip motion. To this end, the general parameterisation technique (as used by Parke) is more feasible, since parameters can be tailored to suit the particular purpose. Guiard-Marigny et al., (1994) have designed a model of the lips with parameters optimised for speech articulation, based on analysis of a real speaker's lip movements.

Bimodal speech synthesis and animated agents

In order to synthesize visual speech, a facial model can be combined with a text-to-speech synthesis system. Audio-visual text-to-speech synthesizers, employing parametric 3D facial models, have been developed at the University of California Santa Cruz, at ICP in Grenoble, at KTH in Stockholm and at the University of Pennsylvania (Benoît et al., 1995; Pelachaud & Prevost, 1994).

There are several possible applications of an audio-visual text-to-speech system. As mentioned in the introduction, the greatest benefit from the visual modality in speech perception can be expected during acoustically degraded conditions. Thus, bimodal speech synthesis has many applications to hearing impaired people. It can for example be used as a tool for interactive and adaptive training of speechreading. Also a face with semi-transparent skin and a well modelled tongue can be used to visualize tongue positions in speech training for deaf children.

Another promising application is speech-based multimodal dialogue systems employing synthesized audio-visual speech. This technology is well suited for information systems in public and noisy areas, such as airports, train stations and shopping centres. The next section describes the integration of a talking face into a

spoken dialogue system that is being developed at KTH.

A closely related concept is that of autonomous virtual agents, serving as personal assistants, with whom we communicate using speech, gesture, facial expressions etc. It has been predicted that such agents will be the next big step in user interface technology (Negroponte, 1989).

Cassel et al. (1994) have developed rule-based models for the interaction between intonation and gesture, and implemented these rules in a conversation simulation system, with two animated talking and gesturing agents. The agents are functionally identical, but they have been given different goals, beliefs and knowledge about the world, and they need to communicate and carry out a conversation with each other in order to reach their respective goals.

Audio-visual synthesis at KTH

A system for audio-visual text-to-speech synthesis has been implemented (Beskow, 1995), based on the KTH rule-based formant synthesis (Carlson et al., 1982; 1991) and the parametric face model developed by Parke (1982).

The Parke facial model consists of a mesh of about 800 polygons, that approximate the surface of a human face including eyes, eyebrows, lips and teeth. The polygon surface can be deformed using 50 parameters, that move the vertices of the polygon network in different ways.

A number of modifications were made to the Parke model, in order to make it suitable for speech synthesis. Special rules were written for the generation of facial control parameter trajectories from phonetic text, and implemented in the speech synthesis system.

The tongue

In order to allow synthesis of visual speech without omitting any potentially important visual information carrier, a tongue model was created and added to the face. The tongue model could be kept simple, since it would only be seen through the opening of the lips. Thus, only apical tongue motion needed to be modelled. The result was a tongue shape made up by 64 polygons, controlled by two parameters: length and apical articulation. The apical articulation parameter controls the raising of the tongue tip relative to the hard palate. The parameter is normalised with respect to jaw opening and should be given a value between zero and one.



Figure 1. Activation of lip parameters. Left to right: bilabial occlusion, labiodental occlusion, rounding and neutral lips.

The lips

Synthesis of realistic lip motion required a new set of lip control parameters to be developed. The three most important parameters in this set are the parameters for lip rounding, bilabial occlusion and labiodental occlusion (Figure 1). These parameters are all normalised in such a way that the value is in the range between zero and one, where zero represents neutral (no activation), and one means full activation. In this way the parameters become independent of the actual geometry of the model, as well as the activation of other parameters.

Lip rounding is achieved by pulling the lip vertices in the direction of the centre of the mouth. Bilabial occlusion is done by pulling the lips towards each other, and labiodental occlusion involves pulling of the lower lip towards the edge of the upper front teeth.

The articulation

Ten parameters are used in the animation of speech movements. Temporal evolution of these parameters is controlled by rules, implemented in *Rulsys* (Carlson et al., 1982). *Rulsys* is a generic system for parametric synthesis by rule, that also controls the sound generation module, *GLOVE*, a formant-based speech synthesizer (Carlson et al., 1991).

The face can be viewed as an extension of the existing synthesis system into a new modality. *Rulsys* generates parameters for the two modalities according to the same principles, and the higher level stages of the rule processing, that involve conversion of the orthographic text into a phonetic representation, are shared by both modalities.

The rules used for generation of parameters for visual speech from the phonetic string, follow three steps, in order to take coarticulation effects and dynamic properties of the articulators into account.

- At each segment, each parameter is either assigned a target value or left undefined. The parameter value is determined from the viseme category that the current phoneme falls into. (45 Swedish phonemes were clus-

tered to 21 visemes.) If the parameter is left undefined, it means that the viseme is independent of that particular parameter.

- When assignment is complete, the string is searched for undefined target values. The missing target values are calculated by linear interpolation between the nearest forward and backward segment where targets are defined for that parameter. This accounts for forward and backward coarticulation.
- Finally, the parameters are filtered by the output module of *Rulsys*, using different time constants that were selected on basis of the mass of the corresponding speech organ.

The above algorithm is a simple, yet practically usable way of modeling coarticulation in visual speech. Other methods have been proposed, for example by Cohen & Massaro (1993). In their model, each segment is not only assigned a target value, but also an exponentially growing and decaying temporal dominance function, used for weighting of the target value.

Future improvements of the KTH visual synthesis should certainly include a refined coarticulation model. It must however be kept in mind that a higher degree of control, inevitably leads to more parameters, that all need to be set appropriately. Compared with the current algorithm, the dominance model described above requires four times as many parameters, whose interpretation is not always trivial.

One solution to this problem is statistic determination of parameters from real visual speech data. A corpus of such data can be gathered automatically using image processing techniques (LeGoff et al., 1994).

Recent improvements and future plans

The facial animation system has recently been completely rewritten using object-oriented programming techniques. The new system allows animation of not only the facial model, but also of for example general jointed character bodies, as needed for animation of gesturing agents. In addition, there is flexibility regarding the actual facial model used; future inclusion and creation of parametric models other than the Parke face is facilitated.

Facing the user

At KTH, a system for spoken man-machine dialogue is currently being developed (Bertenstam et al., 1995). The dialogue system, also known as *Waxholm*, deals with information

on boat traffic, restaurants and accommodation in the Stockholm archipelago. Input to the system is speech and output is in the form of audio-visual synthetic speech and graphics, such as charts, tables and maps.

In this multimodal dialogue environment, there are many ways of utilising the facial display other than for plain text-to-audio-visual-speech synthesis purposes. Since the dialogue system has a deeper knowledge about for example discourse and information structure than a plain text-to-speech system does, it provides a good base for modeling of facial actions at the prosodic level. Ultimately one would probably like to use one single generation model for intonation, facial expressions and perhaps even gesture. Work in this direction has been done by Cassel et al. (1994) and Pelachaud & Prevost (1994).

In human dialogue, all available channels are normally used in the communication, including, but not restricted to, speech, gesture and facial expressions. This should ultimately be the case also in a man-machine dialogue. In the Waxholm project, one concrete step in this direction has been taken so far: When some piece of information, such as a timetable, appears on the screen, the face will turn towards this new item, thereby directing the users attention to it.

There are several other possible ways in which the face can improve the communication. Actions such as head nods, eyebrow movements, gaze direction etc. can support turntaking and serve as backchanneling signals, indicating the status of the speech understanding process. The dialogue system provides an ideal framework for experiments with this kind of nonverbal communication.

Conclusions

The area of multimodal speech synthesis is still quite new, and a lot of research and development can be expected in the near future. As personal computers grow more powerful, it will be possible to incorporate audio-visual speech synthesis in user interfaces, alongside with automatic speech recognition.

Work on bimodal synthesis at KTH has passed it's first stage and will continue, with focus on multimodal dialogue issues and applications for the hard of hearing.

References

- Benoît C, Beskow J, Cohen M, Granström B, Le Goff B & Massaro D (1995). Text-to-Audio-Visual Speech Synthesis over the world, *Advanced topics for speech mapping*, Speech Maps workshop, Grenoble.
- Benoît C, Mohamadi T, Kandel S (1994). Effects of Phonetic Context on Audio-Visual Intelligibility of French, *Journal of Speech and Hearing Research*, 37: 1195-1203.
- Bertenstam J, Beskow J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Nord L, de Serpa-Leitao A & Ström N (1995). The Waxholm system - a progress report, *Proc ESCA Workshop on Spoken Dialogue Systems*, Vigsø, Denmark.
- Beskow J (1995). Rule-Based Speech Synthesis, *Proceedings of Eurospeech '95*, Madrid, Spain, 1: 299-302.
- Carlson R, Granström B & Hunnicutt S (1982). A multi-language text-to-speech module, *Proc ICASSP-Paris*, Paris, 3: 1604-1607.
- Carlson R, Granström B & Karlsson I (1991). Experiments with voice modeling in speech synthesis, *Speech Communication* 10: 481-489.
- Cassel J, Steedman M, Badler N, Pelachaud C, Stone M, Douville B, Prevost S & Achorn B (1994). Modeling the Interaction between Speech and Gesture, *Proceedings of 16th Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology, Atlanta, USA.
- Cohen MM & Massaro DW (1993). Modeling co-articulation in synthetic visual speech, In: Thalmann NM & Thalmann D, eds. *Models and Techniques in Computer Animation*. Tokyo: Springer-Verlag.
- Erber NP (1969). Interaction of audition and vision in the recognition of speech stimuli, *Journal of Speech and Hearing Research*, 12: 423-425.
- Guiard-Marigny T, Adjouani A & Benoît C (1994). A 3-D model of the lips for visual speech synthesis, *Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, USA, 49-52.
- Le Goff B, Guiard-Marigny T, Cohen M & Benoît C (1994). Real-Time Analysis-Synthesis and Intelligibility of Talking Faces, *Proceedings of 2nd International conference on Speech Synthesis*, Newark, NY, USA.
- Negroponte N (1989). From Bezel to Proscenium, *Proceedings of SigGraph '89*.
- Parke FI (1982). Parametrized models for facial animation, *IEEE Computer Graphics*, 2(9): 61-68.
- Pelachaud C & Prevost S (1994). Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context, *Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, USA, 216-219.
- Waters K (1987). A muscle model for animating three-dimensional facial expressions, *Computer Graphics*, 21: 17-24.