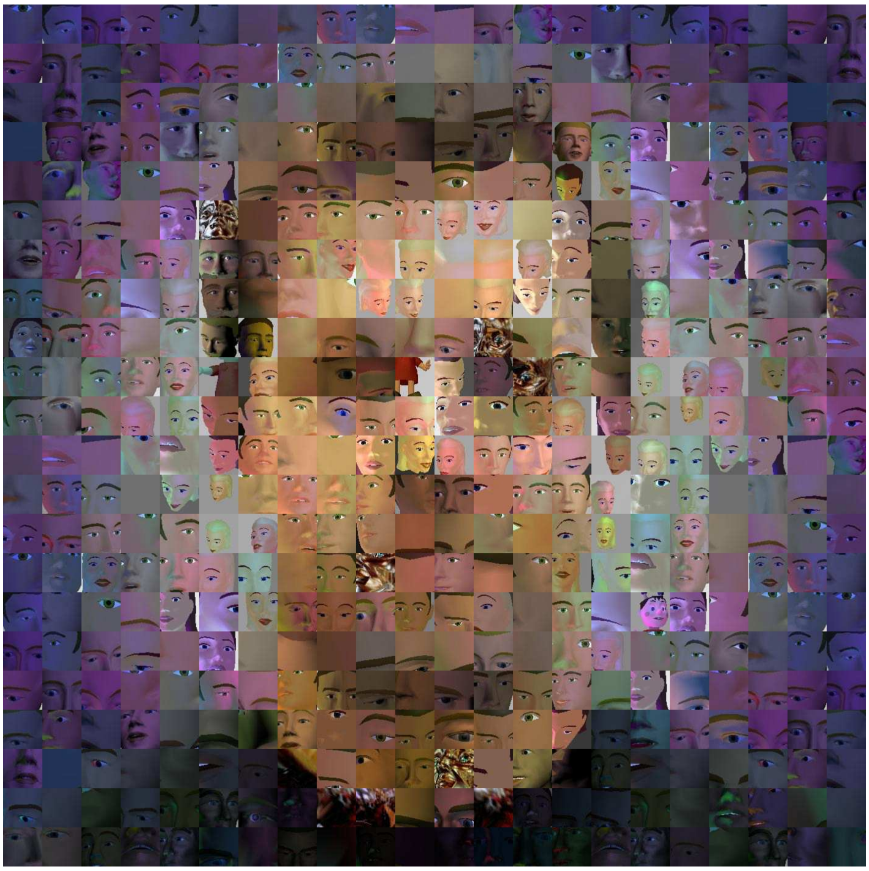
KTH
VETENSKAP
OCH KONST

Talking Heads

Models and Applications for Multimodal Speech Synthesis



Jonas Beskow

Doctoral Dissertation
Stockholm 2003

Talking Heads

Models and Applications for Multimodal Speech Synthesis

Jonas Beskow

Doctoral Dissertation
Department of Speech, Music and Hearing, KTH
Stockholm 2003

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknisk doktorsexamen onsdagen den 11 juni 2003 kl. 10.00 i sal E1, Lindstedtsvägen 3, KTH, Stockholm.

ISBN 91-7283-536-2
ISSN 1104-5787
TRITA-TMH 2003:7

©Jonas Beskow, 2003

Cover photomosaic created using Metapixel from a library of 10000 randomly generated talking head snapshots.

Universitetsservice US AB, Stockholm 2003.

ABSTRACT

This thesis presents work in the area of computer-animated talking heads. A system for multimodal speech synthesis has been developed, capable of generating audiovisual speech animations from arbitrary text, using parametrically controlled 3D models of the face and head. A speech-specific direct parameterisation of the movement of the visible articulators (lips, tongue and jaw) is suggested, along with a flexible scheme for parameterising facial surface deformations based on well-defined articulatory targets.

To improve the realism and validity of facial and intra-oral speech movements, measurements from real speakers have been incorporated from several types of static and dynamic data sources. These include ultrasound measurements of tongue surface shape, dynamic optical motion tracking of face points in 3D, as well as electromagnetic articulography (EMA) providing dynamic tongue movement data in 2D. Ultrasound data are used to estimate target configurations for a complex tongue model for a number of sustained articulations. Simultaneous optical and electromagnetic measurements are performed and the data are used to resynthesise facial and intra-oral articulation in the model. A robust resynthesis procedure, capable of animating facial geometries that differ in shape from the measured subject, is described.

To drive articulation from symbolic (phonetic) input, for example in the context of a text-to-speech system, both rule-based and data-driven articulatory control models have been developed. The rule-based model effectively handles forward and backward coarticulation by target under-specification, while the data-driven model uses ANNs to estimate articulatory parameter trajectories, trained on trajectories resynthesised from optical measurements. The articulatory control models are evaluated and compared against other data-driven models trained on the same data. Experiments with ANNs for driving the articulation of a talking head directly from acoustic speech input are also reported.

A flexible strategy for generation of non-verbal facial gestures is presented. It is based on a gesture library organised by communicative function, where each function has multiple alternative realisations. The gestures can be used to signal e.g. turn-taking, back-channelling and prominence when the talking head is employed as output channel in a spoken dialogue system. A device independent XML-based formalism for non-verbal and verbal output in multimodal dialogue systems is proposed, and it is described how the output specification is interpreted in the context of a talking head and converted into facial animation using the gesture library.

Through a series of audiovisual perceptual experiments with noise-degraded audio, it is demonstrated that the animated talking head provides significantly increased intelligibility over the audio-only case, in some cases not significantly below that provided by a natural face.

Finally, several projects and applications are presented, where the described talking head technology has been successfully employed. Four different multimodal spoken dialogue systems are outlined, and the role of the talking heads in each of the systems is discussed. A telecommunication application where the talking head functions as an aid for hearing-impaired users is also described, as well as a speech training application where talking heads and language technology are used with the purpose of improving speech production in profoundly deaf children.

ACKNOWLEDGEMENTS

I am thankful to a number of people without whom this endeavour would have been much harder, taken even longer time, been less fun, or even been downright impossible. First of all I would like to thank my supervisor Björn Granström for introducing me to, and allowing me to work in, this dynamic and fascinating field, and always following my work with great interest. Rolf Carlson also deserves credit for always being supportive and available for down-to-earth discussion.

I am much grateful to my past and present roommates at TMH: Magnus Lundeberg and Tobias Öhman, for good collaboration and great fun. Special credit goes to Magnus for continuous additions to the talking head gallery. Thanks to Olov Engwall, co-author on paper 7, for providing an excellent tongue model and a lot of know-how on articulatory modelling and measurements. Thanks also to Magnus Nordstrand, co-author on paper 5, for filling up the gesture library with communicative gestures and for quickly turning into an expert on the Qualisys system, relieving me of a lot of work during the hectic past six months. I am also grateful to Loredana Cerrato, for constructive criticism on the thesis draft, paper 8, as well as on my lunch habits, to Beata Megyesi for help and practical advice during thesis wrap-up stage, and to Gunilla Svanfeldt for tough floorball fights.

A lot of people have provided tools, inspiration and generally good ideas over the years. Kåre Sjölander deserves credit for providing rock-solid, feature-packed and constantly improving speech processing components, and sharing the vision of the ultimate speech tool infrastructure… no doubt WaveSurfer *will* conquer the world, but when will people start clicking the *PayPal*®? Thanks to Joakim Gustafson for great visions, great systems and mind-boggling code… Thanks to Jens Edlund, co-author on paper 5, for good ideas, XML-expertise and for quickly realising the virtues of the omnipotent Tcl-language. Thanks also to Mikael Nordenberg for fruitful discussions and insightful comments, and for snapping the target picture for the cover.

I would like to express particularly warm thanks to Dominic Massaro and Michael Cohen of the Perceptual Science Lab at UCSC, co-authors on papers 3 and 4, for giving me the opportunity to spend 18 months in one of the pioneering talking head labs of the world, and for making my time there so rewarding, professionally as well as personally. Not a winter goes by without me missing Santa Cruz…

A few external colleagues with whom I have enjoyed fruitful collaborations also deserve credit. In the Synface project, thanks especially to Cathrine Siciliano and Geoff Williams, co-authors of paper 6, for swiftly porting the visual speech synthesis rules to English and Dutch. Thanks also to my former colleagues in the Olga project, especially Olle Sundblad, Eva-Marie Wadman (creator of the static Olga model) and Scott McGlashan.

Furthermore, I am indebted to Anne-Marie Öster for a heroic contribution as the subject in the study of paper 7.

I am grateful to Mattias Heldner for helping me with statistical analyses and sharing my passion for old synthesisers.

David House deserves appreciation for proofreading this thesis in a great rush. Credits also to Magnus Lundeberg (again) for reading and commenting on the thesis draft, and to Rebecca Hincks and Marion Lindsey for proofreading paper 8.

The floorball players of the department are hereby acknowledged for providing the weekly fitness-keeping opportunity. In fact, everybody at the centre for speech technology and the department for speech, music and hearing deserve credit for collectively creating such a friendly and stimulating workplace atmosphere.

On a more personal note, I would like to use this space to thank the Swedish HotSynth Quintet (`shotq.se`) for providing one of my favourite extracurricular activities as well as

one of the more exotic applications for various speech technology tools. The record is long overdue…

Finally thanks to my family. A lot of love and gratitude goes to my parents, Göran and Kaarina. You always stimulated my engineering curiosity and ensured that there were plenty of electrical circuitry, lego and a C64 around to play with.

Most of all thanks to Cecilia, without your loving support and non-sentimental view of doctoral education, who knows when I would have wrapped this all up… och till Elliot, för att du är en sån gullknopp… and to the pending fourth family member for giving me a nonnegotiable deadline for this project.

Part of the presented research activities have been carried at out the Centre for Speech Technology (CTT), supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The research activities have also been funded by the Swedish National Board for Technical Development (NUTEK), Swedish Transport and Communications Research Board (KFB), the Swedish Knowledge Foundation (KK-stiftelsen), the Fulbright Commission, John och Karin Engbloms stipendiefond, University of California, National Science Foundation (NSF) and EU's Information Society Technology programme (IST).

Jonas Beskow, May 2003

INCLUDED PAPERS

Paper 1. Beskow, J. (1995). Rule-based Visual Speech Synthesis. *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95).* Madrid, Spain, pp. 299-302.

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.

.

Paper 3. Cohen, M. M., Beskow, J. Massaro, D. W. (1998). Recent developments in facial animation: an inside view. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp 201-206.

Paper 4. Massaro, D.W., Beskow, J., Cohen, M.M., Fry C.L., Rodriquez, T. (1999). Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. In D. W. Massaro (Ed.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'99)*, Santa Cruz, CA. pp. 133-138.

Paper 5. Beskow, J., Edlund, J., and Nordstrand, M. (submitted1). A Model for Generalised Multi-Modal Dialogue System Output Applied to an Animated Talking Head. To appear in *Minker, W., Bühler, D. and Dybkjær, L. (Eds) Spoken Multimodal Human-Computer Dialogue in Mobile Envirnonments.* Dordrech, The Netherlands: Kluwer Academic Publishers.

Paper 6. Siciliano, C., Williams, G., Beskow, J. and Faulkner, A. (submitted2). Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired, to appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

Paper 7. Beskow, J., Engwall, O. and Granström, B. (submitted3). Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. To appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

Paper 8. Beskow, J. (submitted4). Trainable Articulatory Control Models for Visual Speech Synthesis, submitted to *International Journal of Speech Technology.*

GLOSSARY OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AA	Animated Agent
ANN	Artificial Neural Network
ANOVA	ANalysis Of Variance
AU	Action Unit
CTT	Centre for Speech Technology at KTH *(Centrum för Talteknologi)*
CVC	Consonant-Vowel-Consonant
DM	Dialogue Manager
DMI	Direct Manipulation Interface
EMA	ElectroMagnetic Articulography
EMG	Electromyography
EPG	Electropalatography
FACS	Facial Action Coding System
FAP	Facial Animation Parameter
FAPU	Facial Animation Parameter Unit
FAT	Facial Animation Table
FDP	Facial Definition Point
F0	Fundamental frequency
GESOM	GEneric System Output Markup
GSDS	Generalised Surface Deformation Scheme
HMM	Hidden Markov Model
IR	Infrared
LED	Light Emitting Diode
LPC	Linear Predictive Coding
MPEG	Moving Picture Experts Group
MPEG-4 FA	MPEG-4 Facial Animation Standard
MRI	Magnetic Resonance Imaging
VCV	Vowel-Consonant-Vowel
VCVCV	Vowel-Consonant-Vowel-Consonant-Vowel
VHML	Virtual Human Markup Language
PCA	Principal Component Analysis
QS	Qualisys
RBF	Radial Basis Function
RMS	Root Mean Square
RMSE	Root Mean Squared Error
SNR	Signal-to-Noise Ratio
SSML	Speech Synthesis Markup Language
TDNN	Time-Delayed Neural Network
TTS	Text-to-Speech
TTAVS	Text-to-Audio-Visual-Speech
UCSC	University of California at Santa Cruz
XML	Extensible Markup Language

CONTENTS

1 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 first moments of our lives, we are all 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 have different roles in the communication process. Firstly, facial expressions are perhaps the most important way of signalling emotion. We can often tell if a person is happy, sad, scared, angry etc. by looking at his/her face. Secondly, in verbal communication situations, the face expresses information related to discourse, phrasing, emphasis and dialogue turn-taking. In this sense facial expressions are intimately related, and often complementary, to intonational features of the voice. Thirdly, the face reveals visible aspects of speech production and thus carries 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.

Technological and scientific advances have made it possible to faithfully synthesise and animate human faces and heads by computer. Combined with advances in the area of spoken dialogue systems for human-machine interaction, this opens up the possibility of leveraging the benefits of face-to-face communication in several computer-based or computer-assisted scenarios such as computer games, e-learning applications, virtual worlds as well as human computer interaction in general. Consequently, a lot of research has been conducted in the area of talking head synthesis during the past three decades. This thesis represents a part of this global effort ultimately striving to allow systems to utilise the inherent communicative power of the human face in meaningful ways, leading to systems that are friendlier, more intuitive and easier to use, as well as opening up new possibilities for communication between people.

More specifically, the thesis presents techniques for modelling, animating and evaluating synthetic talking heads, in the context of several applied scenarios in the areas of spoken dialogue systems, communication aids and language learning. The focus throughout the presented research is on communicative function rather than on video realism. In fact, *cosmetic realism* and *communicative realism* are regarded here as two more or less unrelated variables; while the former is mainly related to the static appearance of a face, the latter is essentially defined by the dynamics of facial gestures. While cosmetically realistic talking heads represent a worthy research goal in their own right, it is evident that cosmetic realism is by no means a requirement for communication - we only have to consider animated cartoons to realise this fact; while the characters are typically unrealistic, there is no doubt that their faces and bodies *communicate*. As is shown in this thesis, this situation holds for computer generated talking heads as well.

1.1 Thesis overview

This thesis consists of two parts. The first part is a summary, not only describing the essential results and methods of the thesis, but also reviewing the current state of the art in the field of talking heads, and placing the results in relation to this framework. The second part is a collection of eight internationally published scientific papers.

The summary is organised according to a number of distinct but related research areas. A number of the included papers span several of these areas, and they are referenced in all relevant places throughout the summary.

- Chapter 2 deals with facial modelling, outlining available approaches for facial image synthesis and parameterisation, and focusing on model-based techniques. Furthermore, a general approach to facial deformation is described and a speech-oriented parameterisation strategy is proposed. *Related papers: 1, 2 and 3.*
- Chapter 3 focuses on data acquisition, and reviews methodologies for static as well as dynamic, internal as well as external data gathering for talking head modelling. It also describes how different types of data sources can be integrated and applied in talking head modelling using general error minimisation techniques. *Related papers: 3, 7 and 8.*
- In chapter 4, the topic is how to drive the articulation of talking heads from phonetic input, for example in order to synchronise it with a text-to-speech system. It is described how coarticulatory effects are accounted for in different articulatory control models, and an evaluation of a number of data-driven articulatory control models is presented. *Related papers: 1 and 8.*
- Chapter 5 deals with driving the articulatory motion of a talking head directly from the acoustic speech waveform. After a review of existing methodologies, an experiment is described based on Artificial Neural Networks (ANNs). *Related paper: 4.*
- Chapter 6 discusses non-verbal aspects of facial motion: what their function is and how they can be modelled in a talking head. A formalism for describing and generating such motion for a talking head in the context of spoken dialogue systems is presented. *Related paper: 5.*
- Evaluation of talking heads is the topic of chapter 7, the focus being on evaluation of intelligibility, i.e. to what degree a talking head improves speech perception in acoustically challenging conditions, including a summary of published talking-head intelligibility evaluation studies and a description of evaluations of the talking heads developed as part of this thesis. *Related papers: 2, 4, 6, 8.*
- Chapter 8 describes the context in which essential parts of this work have been carried out, namely several applied projects in the area of spoken multimodal dialogue systems, communication aids for the hearing impaired and speech training systems. *Related papers: all.*
- Finally, chapter 9 briefly summarises the individual papers, explicitly stating the original scientific contribution of each paper, and in the case of multiple authors, also stating the role of the thesis author.

The main part of the work represented by this thesis has been carried out at the Department of Speech, Music and Hearing and the Centre for Speech Technology (CTT), at KTH in Stockholm. Papers 3 and 4 describe work that was done when the author was visiting the Perceptual Science Lab (PSL) at University of California, Santa Cruz (UCSC) on a Fulbright scholarship during 1998 and part of 1999.

He would see faces in movies, on TV, in magazines, and in books.
He thought that some of these faces might be right for him,
and through the years, by keeping an ideal facial structure fixed in his mind,
or somewhere in the back of his mind, that he might, by force of will,
cause his face to approach those of his ideal.
The change would be very subtle. It might take ten years or so...
Gradually his face would change its' shape.
A more hooked nose... wider, thinner lips... beady eyes... a larger forehead.

David Byrne (Talking Heads), Seen and not seen, 1980.

2 FACIAL MODELLING

Synthesising the human face by computer has been a topic of research for over three decades now. Existing techniques can be broadly categorised into video-based and model-based, where the former family of techniques essentially use 2D video data as raw material for producing new images, and the latter use some kind of deformable model, often in 3D, as the basis for image generation. However, the boundaries are blurring, since video-based systems are starting to incorporate elements of 3D techniques and model-based approaches make increasing use of textures and other pixel-based data sources.

It is tempting to make a comparison between the development in face synthesis and auditory speech synthesis. In the auditory synthesis domain, at least for commercial applications, model-based approaches (formant synthesis) have been almost entirely abandoned in favour of sample-based methods, where units of pre-recorded speech are drawn from huge databases and strung together into new utterances with minimal amounts of signal processing.

In the visual domain, however, there is no obvious dominating technology at this point, rather, there seems to exist a healthy balance between different approaches to facial image generation. While video-based techniques are gaining popularity, model-based facial animation seems to be more active than ever before, no doubt partly due to the MPEG-4 facial animation standard (Pandzic and Forchheimer, 2002), for the first time providing a "lingua franca" of facial modelling. Speculating further, there seems to be a potential aesthetic appeal to model-based visual synthesis that is not present in its acoustic counterpart; while there are now several full-length feature films following in the trail of Pixar/Disney's seminal *Toy Story* of 1995, where the entire cast of characters is computer animated from 3D models (figure 1), we have yet to hear a movie soundtrack composed of formant-synthesised speech.

2.1 Video-based synthesis

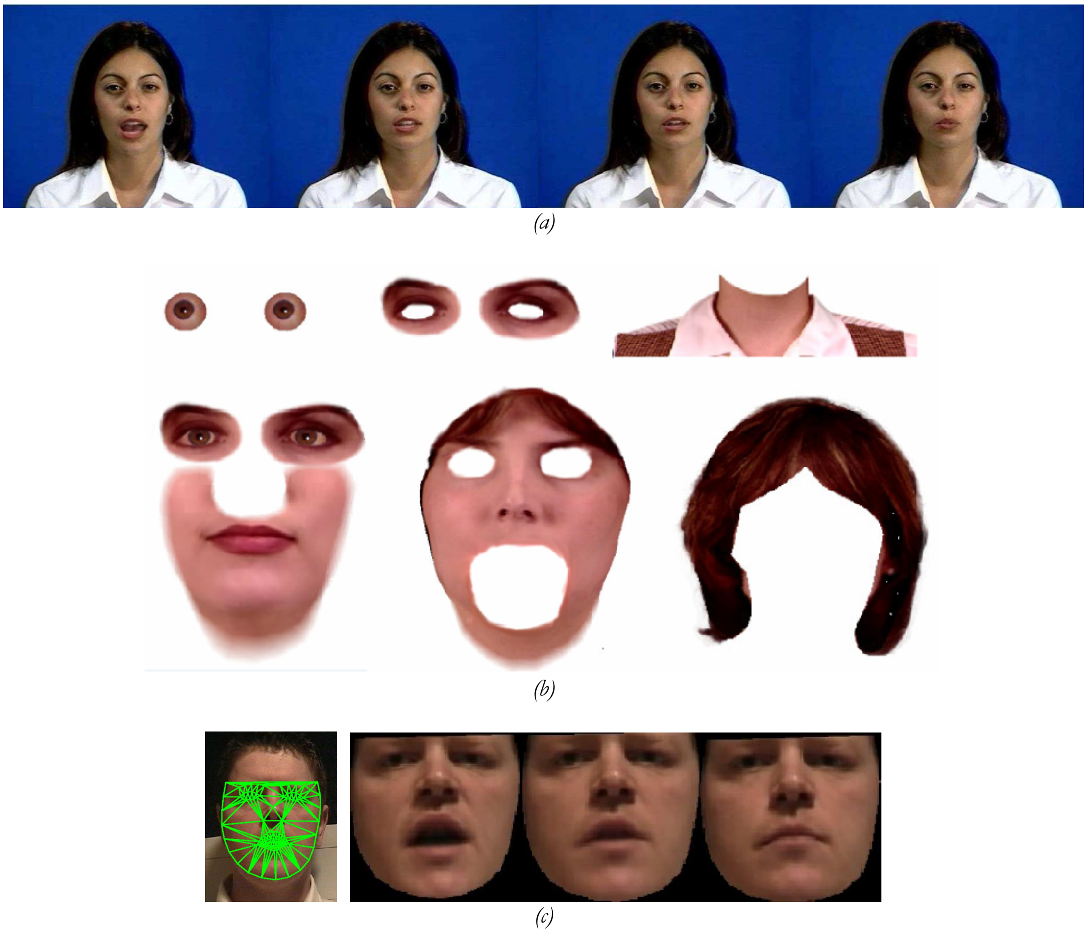
The fundamental problem to be overcome by most systems for video-based synthesis of talking heads is how to seamlessly concatenate video sequences. Even the slightest change in head posture or change in facial expression would be very noticeable if occurring from one frame to the next. Thus, pre-processing steps performed by most systems include normalisation of face position and orientation, as well as segmentation of the face into regions that can be treated separately. For example, in the Video-Rewrite system by Bregler *et al.* (1997), only the mouth-area is processed, and later re-imposed (with new articulation) into the original video sequence.

Figure 1. A model based talking head: Woody from Pixar/Disney's Toy Story. Woody's articulation and facial expression was controlled by 100 parameters, but he spoke with a natural voice, provided by the actor Tom Hanks. ©*Pixar/Disney 1995.*

Cosatto and Graf (2000) divided the face into several regions, as shown in figure 2 b, that are stored separately and can be combined with a base head to assemble new animations. Image sequences for the mouth area are generated using unit-selection techniques known from acoustic synthesis (Hunt and Black, 1996) in order to achieve photorealistic visual speech synthesis. The resulting animations are in general of very high quality, but the approach suffers from the same drawbacks as acoustic unit selection systems, i.e. it can produce undesirable results when there are no appropriate units in the database. In addition, when the system is combined with an acoustic unit selection synthesizer, the system becomes sensitive to mismatches between acoustic and visual units. The investigators report accidental reproduction of the so-called McGurk-effect[1]; synthesis of "mine" was perceived as "nine" when the selected visual unit lacked proper bilabial closure for the [m] (Cosatto, 2002).

An alternative to raw concatenation of image sequences can be to build statistical models of the image bitmaps, allowing image generation to be controlled by a compact set of parameters. Such a model was proposed by Brooke (1996). It is based on Principal Component Analysis (PCA), a technique for reducing the dimensionality of a data set by finding a small set of linear combinations of the original dimensions that optimally describes the variance of the original data. When PCA is applied to images, the grayscale level of each pixel represents one dimension, thus an *M*-by-*N* grayscale image can be represented as a vector in an *MN*-dimensional space. For colour images, each pixel represents three dimensions; red, green and blue. Brooke and Scott (1998) applied a two-level PCA by dividing the mouth region into 16 sub-blocks of 16x12 pixels, each of which was subject to PCA. Then a second, global PCA was performed on the first 30-50 principal components

[1] Conflicting auditory and visual stimuli often lead to a perceptual fusion effect, known as the McGurk effect; for example a visual [ga] paired with an auditory [ba] is often perceived as [da] (McGurk and MacDonald, 1976).



(a)

(b)

(c)

Figure 2. Three video-based systems for visual speech generation: (a) Morphing based system *Mary 101* from MIT, (b) Concatenation-based system from AT&T where the face is divided into separate parts for flexible re-assembly (Cosatto, 2002) and (c) statistically based system from Univ. of East Anglia employing separate models for shape and appearance (Theobald et al., 2002).

of the sub-blocks. Finally the first 40 global principal components were chosen as control parameters for the model.

Theobald *et al.* (2002) describe a related approach using separate models for shape and appearance. The shape model is a wireframe mesh connecting hand-labelled landmarks in the images. Using the shape model, each image is warped to a mean shape before the PCA is applied on the pixel values (the appearance model), see figure 2 c.

Another technique is to use morphing (Beier and Neely, 1992) to simultaneously warp and blend between pre-defined key images. Morphing requires a feature correspondence between the images to be established, which is usually accomplished by manual labelling. Ezzat and Poggio (1999) use optical flow to automatically derive correspondences at pixel level and then morph between a set of key viseme[2] images to form new animations. The

[2] The term viseme is used to represent a group of phonemes whose visible articulatory realisation is similar, such as [p, b, m].

technique is taken a step further in Ezzat *et al.* (2002) where the set of key images is determined automatically from data, along with a data-driven trajectory synthesis technique for calculating morph-parameters from phonemes (see also paragraph 4.2.3) to generate video-realistic animations (figure 2 a).

Several of the above systems (Cosatto, 2002; Scott and Brooke, 1998; Theobald *et al.*, 2002) also incorporate a 3D mesh, onto which the synthesised facial images are projected. This further increases the flexibility of the models, since it allows for independent control of head movements and rotations.

2.2 Model-based synthesis

In model based synthesis, the facial surface is typically described as a polygonal mesh, usually in 3D. During animation, the surface is deformed by moving the vertices of the mesh, keeping the topology of the network constant. The movement of the vertices is governed by a set of control parameters. The mapping from control parameter values to vertex displacements can be based on a number of techniques, including interpolation, direct parameterisation, pseudo-muscular deformation, physiological simulation or data-driven techniques.

2.2.1 Interpolation

Interpolation is the most straightforward and probably the most common method of animating 3D face models, since it is supported by all major commercial 3D-modelling and animation packages. The basic principle is that a number of key-shapes are defined, often corresponding to visemes or prototypical facial expressions. For each of the key-shapes, the vertex positions are stored, but the topology remains constant. Once the key-shapes are defined, they can be used as key-frames in an animation, and in-between frames can be determined by interpolation. Alternatively, if one key-shape is defined as neutral, *N* key-shapes can be considered to define an (*N*-1)-dimensional parameter space, where each parameter controls the amount of a certain key-shape to blend with the neutral shape. Thus the control parameter space represents all possible linear combinations of the key-shapes. The main attractions of the interpolation method lie in its simplicity, and the high degree of support in commercial modelling- and animation packages.

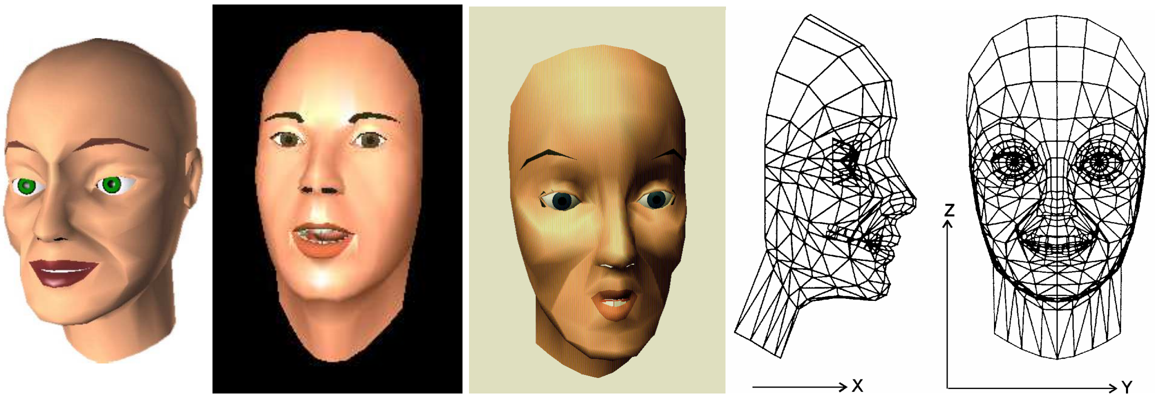


Figure 3. Parke's model and some of its descendants used for visual speech synthesis. From the left: Finnish talking head from HUT (Olivès et al., 1999), *Baldi* from UCSC (Cohen and Massaro, 1993), *Holger* from KTH (paper 1) and the original wireframe topology in side and frontal view (Parke, 1982).

One disadvantage of interpolation is that the space of expressions that can be generated is limited by the defined key-shapes. Consequently, a large number of key shapes are needed to provide enough flexibility, which makes it a rather labour-intensive method. Furthermore, facial motion is typically non-linear, hence the vertex trajectories obtained from linear interpolation between extreme shapes can sometimes yield unsatisfactory results. For example, the movement of chin points during jaw opening is better described by a curved path than a straight line. This can be remedied by introducing in-between shapes (e.g. half-open jaw), as proposed in the MPEG-4 Facial Animation Table (FAT) specification (Ostermann, 2002), at the price of increased key-shape count.

2.2.2 Direct parameterisation

Parke (1982) proposed the direct-parameterisation method to overcome the limitations of the interpolation method. In Parke's model, rather than specifying target shapes, the vertex displacements are described explicitly as basic geometric transformations, such as translations, rotations and scalings, as well as interpolation, applied to local regions of the face. In the direct parameterisation technique, no attempt is made at modelling the structures and mechanisms that lead to surface displacements, instead the observed displacements are modelled directly. To simulate elasticity of the skin, the effect of the transforms tapers at the boundaries of the regions. Parke's original parameter set was divided into *expression* and *conformation* parameters. The former category controls the facial expression and included parameters such as *eyelid opening, jaw rotation, mouth width, upper-lip position, mouth corner position, eye gaze* etc. The latter category of parameters controls the overall shape of the face and represents intra-person variations with parameters such as *jaw width*, *forehead shape, nose length* etc.

Since direct parameterisation is a concise, encapsulated and computationally efficient way of describing movements of the face, it has been a popular choice for visual speech synthesis research, and several systems based on direct descendants of Parke's model have been described, of which the most well-known is perhaps *Baldi* from UCSC (Cohen and Massaro, 1993), which is the model used in papers 3 and 4 in this thesis. LeGoff (1997) and Olivès *et al.* (1999) are other examples (figure 3).

Parke's model was also chosen as the basis for the visual speech synthesiser described in Paper 1. In this system, some additional control parameters were introduced to facilitate better modelling of articulatory movements. These are discussed in section 2.4. Furthermore a simple tongue model was added, see section 2.5.

2.2.3 Muscular and physiological models

In the direct parameterisation approach, there are no restrictions on the type of deformation that the parameters may exert on the model, since they are defined by arbitrary geometric transforms. While this is a powerful approach, it must be used with care since it easily can yield physiologically impossible results. One way of narrowing the space of allowable configurations is to study the natural anatomical limitations of the human face. More precisely, by using the facial muscle structure as a guide, it is hoped that a sensible control scheme can be obtained. One category of models sometimes referred to as pseudomuscular are essentially direct parameterised models that use the human facial muscle structure as a basis for modelling the deformations. Pseudomuscular models have been described by e.g. Thalmann and Kalra (1992), where free form deformations based on radial basis functions are used to simulate muscle actions, and Pelachaud (2002) who describes an MPEG-4 compatible pseudomuscular model.

In the simplest case, a muscle can be approximated as a linear contractor with one end affixed to the bone structure of the skull and one end attached to the skin. For such a muscle, vertex displacements resulting from muscle contraction could be modelled as translations in the direction of the muscle. In the GSDS described above, the target point would represent the bone end of the muscle and the prototype point would represent the skin end. In pseudomuscle models, the weight for each vertex is typically given as a function of the distance from the muscle attachment point.

In Waters' (1987) muscle model, a more detailed simulation of vertex movement due to muscle contraction was used; for each vertex, not only the degree of movement (weight) but also the direction of movement is determined as a function of vertex position relative to the muscle attachment area. Several types of muscles were used in Waters' model: linear muscles that contract lengthwise and attach to a single point, sheet-muscles that attach to multiple points on a line, as well as elliptical sphincter muscles that contract around an imaginary centre point. The latter type of muscle can be used in modelling for example the orbicularis oris[3] muscle that surrounds the orifice of the mouth.

Elasticity of the skin in the above models was modelled implicitly by gradual tapering of the effect of a muscular displacement as a function of distance from the muscle attachment point. An alternate approach was taken by Platt and Badler (cited by Parke and Waters, 1996), who modelled the skin surface as a tension net - a network of point masses interconnected by springs. When muscle forces are applied to one node of the net, displacements are automatically propagated to the rest of the network.

While the tension net provides an elegant model of skin elasticity, it is still an oversimplification in that it assumes the skin to be an infinitesimally thin surface with only tension forces. Lee *et al.* (1995) developed a more detailed physiological model for simulating the properties of facial tissue. Their skin model was composed of three spring-mass layers: the epidermal surface, the fascia surface and the skull, with dermal-fatty springs connecting the epidermal and fascia nodes and muscle springs acting between the skull and the fascia nodes (see figure 4). In addition to spring forces between nodes, the model included skull penetration constraint forces, ensuring that the tissue would slide over the skull rather than penetrating it, and volume preservation forces, striving to keep the volume of each tissue element constant.

One advantage of the model of Lee *et al.* is that it is capable of naturally predicting wrinkles and bulges due to skin compression, potentially resulting in more natural looking facial expressions. The main disadvantage is its computational complexity, making it unsuitable for real-time applications. Furthermore, the parameters determining the physical properties of the tissue, such as thickness of the layers and spring constants, are assumed

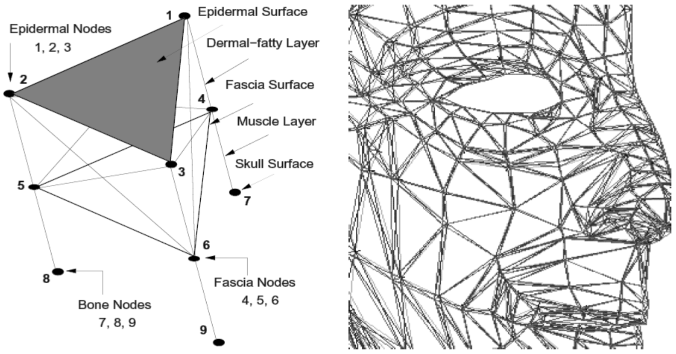


Figure 4. Physiologically based model, where the tissue is modelled as a three-layer network of springs and point masses. (Lee *et al.*, 1995).

[3] Strictly anatomically, according to Gray (1977), the Orbicluaris oris is not a sphincter muscle, but rather a collection of muscular fibres with different directions surrounding the mouth.

constant over the entire facial surface, which is an obvious simplification. Thus, a crucial problem that remains to be solved for this type of model is how to get detailed data to estimate the large number of parameters defining local properties of the tissue. The authors propose using MRI to estimate tissue thickness as a future extension, but no results of such efforts have been reported. Lucero and Munhall (1999) estimated parameters such as layer thickness, node mass (~skin density), spring compression and damping coefficients and muscle forces based on published anatomical data. Furthermore, they used EMG measurements to control muscle excitation for animation.

2.2.4 Data-driven approaches

In the data-driven (or *performance-driven*) approaches of modelling of facial surface deformation, less attention is paid to the underlying physiology of the face. Rather, just as with the direct parameterisation method, the goal is to model surface deformation directly. But instead of relying on manual observations, physical measurements form the basis of the parameterisation that is derived using statistical methods, usually some variant of Principal Component Analysis (PCA). Fundamental to these methods are the data capture techniques discussed in chapter 3. Most importantly, data must have high spatial resolution in order to provide reliable estimates for the displacements of all vertices in the target 3D-mesh model; at least several hundred vertices are used in most models, often several thousands.

Kuratate *et al.* (1998) obtained a set of eight high-resolution 3D head scans with a Cyberware scanner (see section 3.1), representing different static speech and non-speech orofacial configurations. A generic 3D facial mesh was fit to the high-resolution data of each scan. Then a PCA was applied to the vertices of the facial mesh, resulting in five linear components accounting for 99% of the variance between the eight key-configurations, that can be used to deform the mesh during animation.

Elisei *et al.* (2001) used manual 3D photogrammetry on video frames (see section 3.2) to track the position of 168 coloured beads glued to a speaker's face. The coordinates of the flesh points were subject to a modified PCA, resulting in six linear orofacial control parameters describing 97% of the variance in 40 recorded key-postures.

A similar technique, based on MRI scans, was used by Engwall (2002) in the development of the tongue model that is used in paper 7 and further described in section 2.5.

2.3 Parameterisation

One important issue that has to be addressed when developing a talking head is the choice of parameterisation. Parke (Parke and Waters, 1996) lists a number of factors to consider in choosing or developing a facial control parameterisation. These include *control range* (i.e. are all expressions possible?), *complexity*, *number of parameters* and *intuitiveness*. Pandzic and Forchheimer (2002b) add a few items to this list: *portability* (i.e. are the effect of the parameters well defined for different face models?), *ability to measure parameters* and *efficiency* (for coding, compression and low-bandwidth transmission).

There is no ideal parameterisation satisfying all of these conditions, but it is also important to note that for specific applications, not all demands are relevant. For example, if animation is going to be controlled by data-driven approaches, measurability of the parameters is important, but intuitiveness is less of an issue.

Traditionally, facial models have used a variety of different parameterisation schemes, often intimately associated with the particular surface deformation technique used. Parke (1982) developed a set of ad hoc parameters based on observation, as discussed in section 2.2.2.

Several of the muscle-oriented models adopted the Facial Action Coding System (FACS) developed by Ekman and Friesen (1978). FACS was created for psychologists as a descriptive system for coding of facial expressions, and is based around 46 action units (AUs), each of which represents an observable isolated facial muscle action. As such, it represents a natural parameterisation for muscle-based models. Since FACS was primarily developed for coding of facial expressions as opposed to articulatory movements, it provides highly detailed control over the upper parts of the face, but it has been argued that it may not provide enough detail in the oral region for proper speech articulation modelling (Pelachaud *et al.*, 1994).

2.3.1 The MPEG-4 Facial Animation Standard

In an effort to standardise face model parameterisation, originally for the purposes of efficient model-based coding of moving images, the MPEG consortium developed the MPEG-4 Facial Animation (FA) standard (Pandzic and Forchheimer, 2002a). This standard, now rapidly gaining popularity not only in the video compression field, but also in facial animation in general, defines 68 facial animation parameters (FAPs) and 84 facial feature points (FPs). The FPs are well defined landmark points on the human face, such as *bottom of the chin* and *middle point of inner upper lip contour*. The FAP set consists of two high-level FAPs and 66 low-level FAPs. The high-level FAPs, *viseme* and *expression* are is intended for simple speech animation and display of prototypical emotions (see section 6.1.1), and while these can be convenient shortcuts for simple applications, it is the low-level FAPs that make the standard truly useful. The low-level FAPs typically describe the motion of individual FPs, such as *open jaw* and *stretch left inner lip corner* and are as such comparable to AUs or Parke's ad hoc parameters. The most important difference between FAPs and these parameterisations is however the inherent portability criterion that is an essential part of the standard: the same FAP values must produce a similar result with different MPEG-4 compliant face models (otherwise the standard would be useless). This is achieved by expressing the FAPs in terms of *Facial Animation Parameter Units* (FAPUs) that are defined by distances between certain FPs of the model being animated. There are six FAPUs defined in the specification: *iris diameter*, *eye separation*, *eye-nose separation*, *mouth-nose separation*, *mouth width* and *angle unit*.

As was concluded previously, no parameterisation is ideal for all conceivable tasks. The MPEG-4 FA standard is no exception to this rule (indeed, some of its shortcomings are discussed in the following section) but the fact that there now exists a *standardised* model-independent parameterisation for facial animation probably outweighs the minor shortcomings.

2.4 A speech-oriented approach to direct parameterisation

As was mentioned, Parke's model was adopted for the visual speech synthesis system described in paper 1. One problem with Parke's implementation, was that the geometric transforms associated with the control parameters, as well as the topology of the wireframe structure, were intimately tied into the source code of the animation engine. Thus, the system was only capable of animating one particular model, and any modifications to the parameter set required re-compilation of the animation engine source code. A remedy for this inflexibility is introduced in paper 2, where a generalised surface deformation scheme is introduced.

2.4.1 Generalised surface deformation scheme

The main goal of the generalised surface deformation scheme (GSDS) described in paper 2 was to de-couple deformation parameterisation from the animation engine, thereby greatly increasing flexibility, allowing models with different topologies to be parameterised and animated, and their parameterisation data to be stored together with the model geometry. An additional goal was to facilitate efficient descriptions of articulatory targets to allow for intuitive articulatory parameterisations to be developed, as discussed in section 2.4.2. Furthermore, the computational efficiency of the Parke approach should be retained.

The fundamental entity in the GSDS is the *deformator*. A facial geometry is parameterised by defining a number of deformators, each of which applies some specific transform to a number of vertices, controlled by a scalar control parameter. The action of the deformator is determined by the following properties:

- *Activation factor* – a dimensionless control parameter value governing the degree of deformation.
- *Transform type* – one of translation, rotation, scaling or pull.
- *Influence definition* – a list of vertex indices with associated weights that determines which vertices should be affected by the transform and to what extent.
- *Target point* – index of a vertex that represents the spatial target of the deformation.
- *Prototype point* – typically a central vertex in the influence definition area. The prototype point is drawn towards the target.
- *Pivot point* – scaling and rotation transforms are performed around this point.

The activation factor describes the displacement of the prototype point towards the target. A value of zero represents no displacement, and a value of one implies that the prototype moves as close to the target as the specified transform allows. Thus all activation factors are inherently dimensionless. Once the transform that should be applied to the prototype in order to satisfy the activation factor has been determined, the same transform is applied to all other vertices included in the influence definition, but scaled with their associated weight values.

To date, a number of talking heads have been developed at CTT based on the GSDS framework. Samples from this talking head gallery can be seen in figure 5.

2.4.2 *Speech-oriented parameterisation*

One important consideration in modelling speech articulation is how to describe articulatory targets. Consonantal articulations are often based on well-defined target positions that a given articulator has to reach. In a model for visual speech synthesis, failure to attain articulatory targets, even by a few millimetres, can be highly visible and radically affect perception. If a parameterisation is to be portable across different face models, as is the goal of the MPEG-4 FA specification, there has to be a way of unambiguously expressing articulatory targets, independent of model geometry.

A very general solution to the problem is the paradigm used by Bailly (1998) for controlling an articulatory vocal tract model for speech synthesis, which is to use two separate parameter spaces; one prototypical space where articulatory targets and constraints can be easily and unambiguously defined, and one space for the actual model control parameters that govern the articulators movements, corresponding to e.g. MPEG-4 FAPs. This representation requires a mapping between the two spaces to be developed, capable of dealing with incomplete target specifications, which results in a very complex model.

A more straightforward approach is taken in paper 1 and further developed in the GSDS in paper 2; namely to declare dedicated parameters for primitive articulatory gestures. Paper 1 introduces such parameters in the Parke model, most notably *bilabial occlusion*, *labiodental occlusion* and *lip rounding*.

Bilabial occlusion requires accurate coordination of jaw, upper lip and lower lip, making it cumbersome to control in a precise way in traditional parameterisations (such as Parke's parameterisation or FACS). In the MPEG-4 FA standard this problem is solved by defining jaw opening in such a way that it (somewhat counterintuitive) does *not* affect lip opening, and postulating that lip closure occurs when the upper and lower vertical lip displacement parameters sum to zero. By introducing a single dedicated parameter for bilabial closure, as

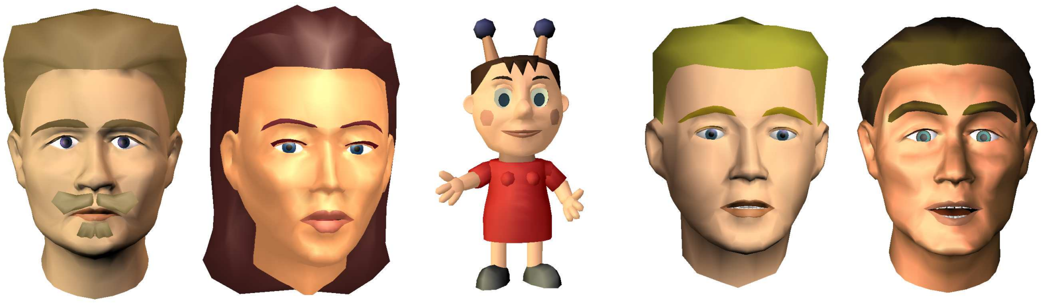


Figure 5. A selection of the talking characters developed at CTT using the generalised surface deformation scheme (GSDS). From the left: *August*, *Kattis*, *Olga*, *Sven* and *Vikthor*.

is done in paper 1, the lower lip can follow the jaw opening motion (as one would expect), and occlusion is always guaranteed when the *bilabial occlusion* parameter is set to one, regardless of the state of the jaw.

For labiodental fricatives, the lower lip is pressed against the upper front teeth to form a constriction. This is problematic in the MPEG-4 FA standard, as there is no FAPU that specifies *teeth-to-lip* distance. Thus the displacement required for the lower lip to meet the upper front teeth will vary from model to model, so the parameterisation is non-portable in this respect. Again, the definition of a dedicated parameter for *labiodental occlusion* is a reliable way of ensuring target attainment for this articulation.

Lip rounding is another complex articulatory gesture that is difficult to attain in many parameterisation schemes. In the MPEG-4 FA specification, it can involve as much as 20 of the 66 MPEG-4 low-level FAPs[4]. Introducing a single, normalised control parameter for lip rounding, that pulls the lip-vertices towards an imaginary centre point in the mouth orifice, greatly simplifies articulatory target specification.

The definition of this type of target oriented, normalised control parameter is facilitated by the generalised surface deformation scheme introduced in paper 2 and summarised in section 2.4.1. In addition to facilitating modelling of articulatory movements, it has the advantage of being portable across different models (just as the MPEG-4 FA standard), as the parameters are normalised and dimensionless. This fact has been exploited throughout this thesis work, as a wide range of face models have shared the same parameterisation and been driven from the same articulatory control model.

It should be stated that the benefit of controlling basic articulatory gestures with dedicated control parameters is most evident when working with rule-based articulatory control models or hand-edited animations. As soon as data from a real speaker are used as a basis for driving the articulation, the interpretation of individual parameters becomes less of an issue, as long as the parameter set as such allows for the full articulatory range to be expressed.

2.5 Models of the tongue

Any model aimed at realistic speech animation requires a tongue model. The Swedish apically articulated consonants [t, ʈ, d, ɖ, n, ɳ, r, s, ʂ, l, ɭ] are often associated with clearly visible tongue motion that is an important cue in speechreading. Additionally, if the model is to be used for multimodal speech training, a realistic tongue model capable of the full

[4] There are 20 lip-related FAPs affecting different segments of the outer and inner lip contours, all of which would be displaced during extreme lip rounding.

articulatory register is likely to be of great pedagogical value if tongue movements can be made visible by rendering the skin semi-transparent or by removing parts of the face.

In the domain of articulatory synthesis, tongue models have been developed in 2D as well as in 3D; the reader is referred to Engwall (2002) for an excellent survey of the field. However, the requirements of a tongue model for auditory articulatory synthesis are quite different from those placed on a model for visual synthesis. While the former must provide a good approximation of vocal tract geometry, to facilitate calculation of the acoustic properties, it does not have to provide a visually interpretable display. In so-called tube-like models, for example, only the surface delimiting the air channel in the vocal tract is modelled, which is adequate for sound generation, but less well suited for visual presentation.

Tongue models for visual synthesis, on the other hand, have typically been less anatomically elaborate, since it is often sufficient to provide a view from the outside, through the mouth opening, where mainly tongue tip motion is visible. Cohen and Massaro (1993) modelled the tongue as a rigid object that could be rotated, translated and scaled to simulate apical articulation. Pelachaud et al. (1994) describes an articulated tongue model based on implicit surface modelling around a jointed skeleton.

In the studies included in this thesis, three different tongue models have been put to use: one simple and two more complex. Paper 1 describes the creation of a tongue model for apical articulation. The tongue was created as a complement to the Parke model, and was reused in subsequent models based on the GSDS (paper 2). Based around a deformable mesh comprised of a mere 64 polygons, the design was kept simple to allow for real-time animation on 1995 year's hardware. The primary control parameters of the model are *apex elevation* and *tongue length*. The former controls the vertical position of the tongue tip relative to a target corresponding to the alveolar ridge, to simulate dental articulations. The latter parameter is a scale factor to attain for example retroflexed targets.

In paper 3, a more elaborate tongue model is employed, with the goal being realistic modelling of the full tongue articulation (as opposed to only apical articulation) for visible speech synthesis. The model consists of a polygon surface defined by four b-spline curves[5]; one controlling the sagittal contour and three controlling the coronal cross-sections at front, middle and back. The shape of the curves is determined by 30 parameters (9 for the sagittal curve and 7 for each of the coronal curves), yielding a very flexible model. Proper parameter values for each phoneme were estimated by comparing the surface shape of the model to three-dimensional ultrasound measurements and employing an automatic minimisation procedure, further described in section 3.4.

In paper 7, the tongue model developed by Engwall (2002) is employed. This model also aims at providing realistic articulatory modelling of the whole tongue, but rather than being defined by generic curves, it consists of a polygon mesh re-constructed from MRI measurements of 43 static articulations. Using linear component analysis, six control parameters were defined that account for 78% of the variation in the measured three-dimensional tongue shapes. These parameters can then be used to drive the articulation for running speech, as is described in section 6.3.

2.6 Physiological constraints

To maintain the realism of a three-dimensional talking head model during animation it is essential to consider physical constraints imposed by the impenetrable components of the oral cavity. For certain parameter value combinations, any model can be made to produce physiologically unrealistic shapes. Typical examples of such model misbehaviour are the

[5] B-splines are smooth interpolation curves commonly used in computer graphics modelling of natural objects.

teeth and gums showing through the lips when they retract, lips that intersect each other and the tongue penetrating through closed teeth and lips. Even the physiologically based models discussed in section 2.2.3 suffer from this problem, since they typically only model the forces between tissue elements that are directly connected to each other. Lee *et al.* (1995) simulate forces to prevent tissue from penetrating the skull, but not from penetrating other tissue. There are essentially two ways of dealing with this problem. The first one is straightforward: avoid all parameter combinations resulting in un-physiological configurations. In speech production, this is often a natural outcome of attaining articulatory targets, which commonly represent extreme points in the articulatory space, hence going beyond the targets often results in disallowed configurations. Thus the goals of plausible physiology and proper articulation are interrelated. To achieve these goals, it helps to adopt a parameterisation where individual articulatory targets have dedicated parameters, such as the parameters for bilabial and labiodental occlusion that are discussed in section 2.4.2. Furthermore, when using data-driven control models to drive the articulation (see section 4.2), proper articulatory targets and avoidance of un-physiological configurations can be learnt from data, making the choice of parameterisation less critical.

2.6.1 Modelling tongue-palate contact

In some cases, however, there might still be a need for explicit detection and correction of physiological violations. One such case is in simulating the contact between tongue and palate, as is done paper 3. In this case, there is significant interaction between the two structures, i.e. the tongue is deformed as it is pressed against the palate. Unconstrained collision detection for arbitrary polygon-based objects is computationally very expensive, and once a collision has been detected a strategy to correct the colliding structures is needed, which in the general case requires an iterative solution of constraints in a physiologically based model.

In paper 3, a more specialised and significantly faster approximate method is proposed to deal with this problem. The basic idea is to perform the collision detection and correction against a simplified regular structure rather than the actual polygons making up the visual representation of the palate. Pelachaud (1994) used a semi-sphere to approximate the palate. In paper 3, a regular grid is constructed during a separate pre-processing stage. This grid represents the bounding surface of the upper oral cavity, with the grid points located at regular intervals in a spherical coordinate system centred in the oral cavity. Thus by transforming tongue vertices to this coordinate system, an instantaneous comparison between grid and vertex can be made to determine whether the vertex should be corrected. Correction is a simple matter of adjusting the radial component of the transformed vertex in such a way that it ends up on the grid surface. While this obviously is an approximation of the actual deformations that take place as the tongue is pressed against the palate, it provides visually satisfying results and is sufficiently fast to provide real-time display on standard PCs.

2.6.2 Volume preservation

Another physical constraint that is addressed in paper 3 as well as in paper 7 is that of constant tongue volume. It is however difficult to enforce this constraint in real-time during animation, since it requires an iterative solution; when one parameter causes a change in tongue volume, then the other parameters need to be iteratively adjusted to compensate for this change. During model fitting on the other hand, when tongue parameters are automatically adjusted by a minimisation algorithm to find the optimal fit with a given target, then the tongue volume criterion can stabilise the fitting process to avoid ending up with impossible tongue configurations. Including the volume constraint in the minimisation is a simple matter of augmenting the error function with a term representing the squared deviation between current volume and the reference volume. The volume error term is given a weight, which in both studies was determined empirically. Setting this weight too high caused the resulting tongue to appear stiff, in the extreme case never deviating from its reference position.

And all I see, is little dots…

David Byrne (Talking Heads), Drugs, 1979.

3 DATA SOURCES FOR TALKING HEADS

As we have seen in the previous chapter, a 3D talking head is based around a static geometry, typically a polygon mesh, that is deformed to produce animation. There are several ways to obtain this initial mesh. Papers 1, 3 and 4 in this thesis use the Parke model, whose geometry came from 3D photogrammetry (see below). In paper 2, the underlying geometry for the Olga character was sculpted using commercial 3D modelling tools (by graphical artist Eva-Marie Wadman). In papers 5-8 facial geometries were based on a generic head mesh from Viewpoint Datalabs[6] that has been interactively deformed and re-modelled to produce a gallery of characters, see Lundeberg and Beskow (1999) for more details on this process (see also figure 5). As was argued in section 1, static geometry is essentially related to cosmetic values, alas, while the main focus has been on the communicative function of the models, a certain degree of "artistic freedom" has been allowed in shaping the static geometries. For the communicative functions, however, ground truth data from several sources has been an important ingredient in the modelling. The tongue shape in paper 3 is modelled from static 3D ultrasound measurements, and dynamical articulatory measurements of the face and the tongue are used in resynthesis (paper 7) as well as training of models for coarticulation (paper 8; only face data in this case).

Different kinds of data are needed at different stages of the modelling process, and different techniques are called for to obtain these data; there is no single method that provides everything. Often there is a trade-off is between spatial and temporal resolution. Based on temporal resolution, methods can be divided into static and dynamic, where dynamic implies that *at least* video frame-rate can be obtained (25-30 Hz) for capturing dynamic aspects of articulation. Static data sources are primarily used in initial construction

Table 1. Overview of data sources for talking head modelling.

	Temporal resolution	**Dimension-ality**	**Internal/ external**	**Data type**	**Comments**
3D photo-grammetry	static or dynamic	3D	external	fleshpoints (+ texture)	Requires manual intervention
Cyberware	static	3D	external	shape + texture	
Ultrasound	static	3D	internal	shape	Can be used dynamically in 2D
MRI	static	3D	internal	shape/ volume	
Video tracking	dynamic	2D	external	distance measures or 2D shapes	
Optical motion tracking	dynamic	3D	external	fleshpoints	
EMA	dynamic	2D	internal	fleshpoints	

[6] `http://www.viewpoint.com`

of a model, and in some cases in development of the parameterisation. Dynamic data, on the other hand, are used in modelling speech articulation and dynamic communicative facial expressions.

Another distinction is made between external and internal methods. While talking head modelling in general is concerned with the externally visible parts of the face and head, tongue modelling requires measurements of the internal structures as well, which calls for a different palette of techniques (usually from the medical field), only partially covered in this overview. For a complete survey of intra-oral measurement techniques, the reader is referred to Engwall (2002).

Other aspects that need to be considered are whether a method provides 2D or 3D data, and whether the data source provides the entire shape (and/or texture) of objects, or only the coordinates of a number of individual fleshpoints.

An overview of static and dynamic data sources is presented in the following sections and summarised in table 1. Once data has been obtained, it needs to be put in relation to the model in order to be useful. This process is discussed in the final section of this chapter.

3.1 Static methods

3.1.1 3D photogrammetry

In his pioneering facial modelling work, Parke (1982) used a static 3D photogrammetry method, where manual measurements on photographs showing multiple simultaneous views of a human face were combined to calculate 3D coordinates of the vertices of a polygonal face mesh painted on the face of the subject. A mirror was used to make it possible to contain the multiple views in the same photograph, and ensuring perfect temporal synchrony. This method requires makeup on the face to enable identification of the points in the different views. It is also labour intensive, as identification and measurement of points in the different views is a manual process. The main advantage is that no specialised or expensive hardware is required; an ordinary camera and a mirror are sufficient. Another advantage is that it can be used as a dynamic method if video registration is made, as was done by Elisei *et al.* (2001) although the degree of required manual labour increases by several orders of magnitude in this case.

A variant of the standard photogrammetric approach, yielding photo-realistic models, is proposed by Pighin *et al.* (1998). In their work, no makeup is used on the subject, which makes it possible to use the photographs as textures on the final model. Based on manual identification of a small number of feature points or curves in each view, a generic face mesh is reshaped using scattered data interpolation and fitted with a texture derived from the original photographs.

3.1.2 Laser-based scanning

Specialised 3D range scanning hardware, such as that produced by Cyberware Inc[7] can be used to obtain highly detailed geometry and texture information of a static object, for example a human head. A laser-based scanner is moved along a circular path around the subject, producing a large cylindrical range map, where each "pixel" corresponds to the radial distance from scanning device to object surface and an associated texture map provides the colour of each pixel. The scanning process typically takes a few seconds. Although the resulting data sets are textured 3D models in their own right, they are seldom used directly for animation, since they have a regular topology that does not match the features of the face. Instead, the most common approach is to fit a generic face or head

[7] `http://www.cyberware.com`

mesh model to the range map after automatic or manual identification of a number of landmark points or curves (c.f. Cohen *et al.*, 2002; Kuratate *et al.*, 1998; Lee *et al*, 1995;).

3.1.3 Internal static methods

To gather detailed data about the shape of the internal articulators, there are several techniques that can be used. In paper 3, the tongue model shape is trained by matching it against 3D surface meshes obtained from ultrasound measurements of the tongue during sustained articulations of 18 English phonemes (Stone and Lundberg, 1996). The surface meshes are assembled from a series of 2D slices produced by a rotary ultrasound transducer attached under the chin. The data represent the upper surface of the tongue. One limitation of using ultrasound for tongue measurements is that normally the tip of the tongue is not captured. This is due to the fact that the air cavity under the tongue tip causes the ultrasound rays to be reflected before they reach the tongue.

Engwall (2002) used Magnetic Resonance Imaging (MRI) scans to construct the 3D model of the tongue used in paper 7. The MRI scans represent a series of cross-sectional slices of the vocal tract, in which the tongue outlines were manually traced and combined into a 3D mesh, see also section 2.5. MRI-data has also been proposed as the geometrical basis for muscle-based face modelling (Sams *et al.*, 2000). The acquisition time required for a complete vocal tract scan is far beyond real-time; the subject in Engwall's study had to sustain the articulations for 40 seconds.

3.2 Dynamic methods

While static geometry data is mainly important for creating or reshaping the underlying geometry model, and in some cases obtaining the parameterisation, modelling facial movements such as speech articulation, requires dynamic data.

3.2.1 Video-based methods

A number of methods have been proposed to obtain dynamic facial measurements based on ordinary video images. Benoît *et al.* (1996) and Öhman (1997) both applied blue makeup to the lips of a speaker and used chroma-key based image analysis to automatically measure a number of parameters from frontal and profile views of the lips. As discussed in section 3.1.1, the 3D photogrammetry method can also be used dynamically, but this is a tedious method unless an automatic point tracking algorithm is used. Elisei *et al.* (2001) manually identified and measured 168 coloured beads glued to the face of a speaker in two views for each frame in the recorded corpus.

Other investigators have developed methods for automatic tracking of unmarked faces, which simplifies the recording process. Such techniques can employ model-based analysis-by-synthesis, where an underlying model of the face or part of the face is fitted to each video frame (Ahlberg, 2002; Basu *et al.*, 1998; Reveret and Benoît, 1998) or bottom-up heuristically based feature tracking (Petajan and Graf, 1996). While these methods can provide useful data for talking head modelling, their primary area of application is where special face makeup is not possible or realistic, such as in systems for audiovisual speech recognition or model based video compression. In modelling dynamics of talking heads, the overshadowing goal of data collection is often to obtain as accurate data as possible, even if that implies applying make-up or other artefacts to the speakers face and articulators.

3.2.2 *Systems for optical tracking*

There exists several systems for optical motion tracking employing specialised hardware and software to track markers in 3D, for example *ELITE* (from BTS[8]) *OPTOTRAK* (from Northern Digital[9]), *MacReflex* or *ProReflex* (from Qualisys[10]).

Such systems are gaining popularity in areas related to visual speech and talking head research, since they provide a number of advantages over video-based methods: the tracking is fully automatic, accuracy is good (typically well below 1 mm) and the temporal resolution is high (60 Hz or more). The principle behind these systems is the same as for 3D photogrammetry: 3D coordinates are re-constructed from multiple views, however, to ensure stable automatic tracking, special infrared (IR) sensitive cameras are used, and the markers are either active light-emitting diodes (LEDs) or passive reflective beads, reflecting the light emitted from IR-flashes. The use of IR ensures high-contrast images, where marker positions appear as bright dots on a dark background, that can be reliably processed automatically. Marker positions are estimated with sub-pixel accuracy using a centre-of-gravity approach taking advantage of the greyscale level of the pixels.

3.2.3 *Non-optical internal dynamical methods*

Alternative techniques for measuring the dynamics of articulation include electromyography (EMG), electropalatography (EPG) and electromagnetic articulography (EMA). The main advantage of these methods is that they can provide information about movements of the hidden articulatory structures.

EMG measures muscular activity by inserting electrodes directly into the muscles. Lucero and Munhall (1999) report on using EMG to control the muscle activations in a physiologically based face model, adapted from the model of Lee *et al.* (1995).

EPG is a technique for dynamically measuring patterns of tongue-palate contact. The method produces binary maps indicating whether or not contact is made at a number of electrodes affixed to a plastic insert in the palate. EPG data were used by Engwall (2002) to fine-tune articulatory targets in a 3D tongue model.

EMA is another technique developed specifically for articulatory measurements. In the *Movetrack* system (Branderud, 1985), the 2D position of a number of small coils (1,5x4 mm) glued to the tongue can be measured. Two stationary transmitter coils mounted on a lightweight headmount worn by the subject produce a variable magnetic field, and the currents induced in the coils on the tongue can be used to infer their positions relative to the transmitter coils on the headmount. One limitation is that the system only works in two dimensions, so all coils must lie in the same plane, normally the mid-sagittal plane.

3.3 Corpora for internal and external dynamic articulatory modelling

Two papers in this study are based on dynamic data gathered using optical motion tracking and EMA. In paper 8, 30 markers glued to the face of a non-professional Swedish male speaker were tracked using a four-camera *Qualisys MacReflex* (QS) system for 287 phonetically rich Swedish sentences extracted from a large newspaper corpus to optimise triphone coverage. The database was subsequently analysed and articulatory parameters were extracted using resynthesis (see section 3.4.2) and used in the training of articulatory control models (section 4.4).

To overcome the limitations that are present when each method is used in isolation, multiple data sources can be combined. In paper 7, simultaneous recordings were carried

[8] `http://www.bts.it`

[9] `http://www.ndigital.com`

[10] `http://www.qualisys.se`

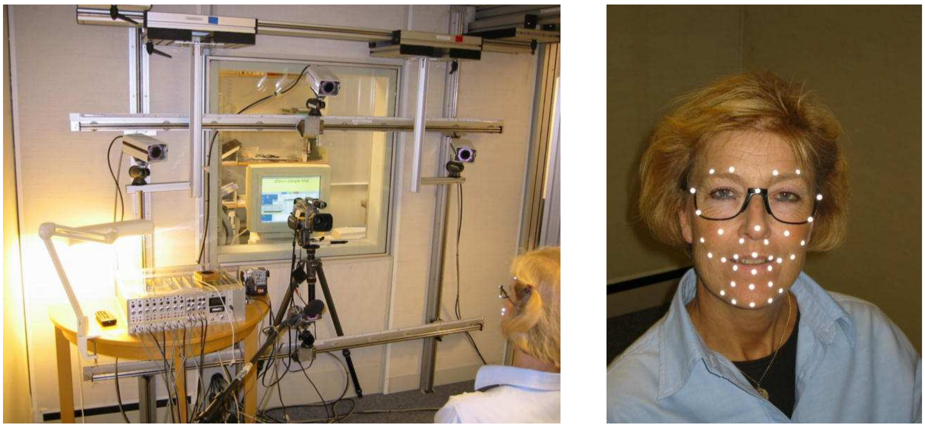


Figure 6. Four-camera *Qualisys* measurement setup (left) and subject with reflective markers glued to the face (right).

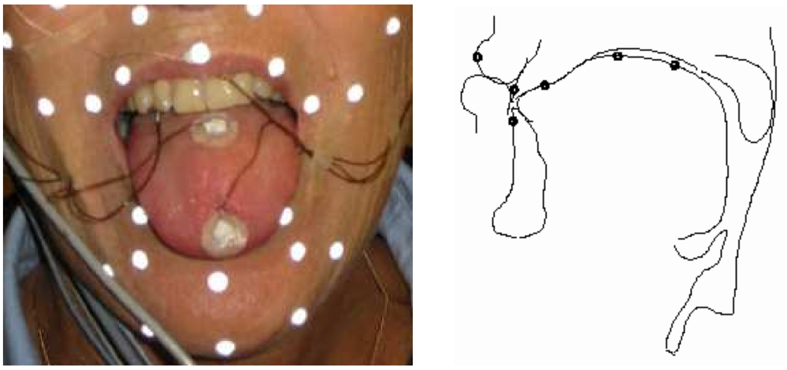


Figure 7. EMA-coils glued to the tongue of the subject, only the two frontmost coils are seen (left). Coil positioning in the mid-sagittal plane is shown schematically to the right.

out with the QS and *Movetrack* (MT) systems. The QS system tracked 28 markers in the face, depicted in figure 6, while the MT system measured the positions of six coils placed on the teeth, upper lip, and front, middle and back of the tongue, see figure 7. The recorded subject was a female native speaker of Swedish, who has received high intelligibility ratings in audio-visual tests.

The recorded material consisted of an unbalanced CVC corpus, a balanced VCV corpus and a sentence corpus. There were 41 asymmetric C$_1$VC$_2$ words, with the Swedish vowels V=[uː, oː, ɑː, iː, eː, ϵː, øː, ʊ, ɔ, a, ɪ, e, ϵ, ʏ, ø] in C$_1$=[k], C$_2$=[p] and C$_1$=[p], C$_2$=[k] context. The [r] allophones V=[æː, œː, æ, œ] were collected with C$_1$=[k] and C$_2$=[r].

The 138 VCV and VCC{C}V words consisted of the consonants [p, t, k, ʈ, b, d, g, ɖ, m, n, ɳ, ŋ, l, ɭ, f, s, ɧ, ç, j, r, v, h] and the consonant clusters [ɟk, rk, pl, bl, kl, gl, pr, br, kr, gr, kt, nt, tr, dr, st, sp, str, spr, sk, fl, fr, sl, skl, skr] in symmetric vowel context with V=[a, ɪ, ʊ].

The sentence corpus consisted of 270 Swedish everyday sentences developed specially for audio-visual speech perception tests, see section 7.2.2.

In order to be able to combine and align the two data sets, one point (mid-point on the upper lip) was co-registered with the two systems. An additional point (lower teeth) was co-registered during a special alignment recording.

Since the MT receiver coils are fixed with respect to the head, the MT data are independent of the subject's head movement. The QS data on the other hand represent absolute coordinates relative to the cameras. To be able to express facial marker positions independently of head movements, three reference markers were affixed to a piece of cardboard on the MT headmount. A similar strategy was used in paper 8 where reference markers were glued to a spectacle frame worn by the subject.

3.3.1 Data alignment and synchronisation

To combine the two data sets, the QS coordinates were first normalised with respect to global head movements by projection onto the axes of a head-local coordinate system expressed by the three reference markers on the headmount. Then they were roto-translated in such a way that the mid-sagittal plane coincided with the XY-plane.

The QS system emits a sync pulse for each frame that is captured. This sync signal was recorded together with the Movetrack data and was used during post-processing to extract one frame of Movetrack coordinates for each Qualisys frame. The Movetrack data were then translated and rotated in the XY-plane to align the co-registered points in the two data sets, eventually forming a coherent set of facial and intra-oral movement data.

3.4 Model fitting and resynthesis

Once data has been collected, it has to be put in relation to the model in order to be useful. In the case of Parke's 3D-photogrammetry, data were used to actually *define* the model; the measured coordinates were used directly as vertices of the model. In other cases, there already exists a model, and the objective is to use the data to train the model to incorporate some aspect of this data. The general solution to this type of problem is error-minimisation. If we can provide an error function, defining how well the model fits the data, then we can let a minimisation algorithm iteratively adjust the parameters of the model to obtain the best fit. This is the approach used in papers 3, 7 and 8. There are several general-purpose algorithms for minimisation of multivariate functions. Paper 3 uses STEPIT (Chandler, 1969) while Powell's direction set method (Press *et al.*, 1992) is used in papers 7 and 8.

The main problem, then, boils down to how to define the error between model and data. There are often many possible ways to do this, and sometimes several alternative error functions have to be evaluated before the most suitable one is found. The nature of the error function will be different from case to case depending on the formats of data and model. Below, two examples of error functions used in the papers in this thesis are given.

3.4.1 Static shape fitting

In paper 3, ultrasound measurements of the tongue are used to estimate tongue model parameters for a number of static articulations, with the goal of using these as target shapes in a visual speech synthesis system. The ultrasound measurements represent the upper surface of the measured tongue body as regular quadrilateral mesh. There is one such mesh for each articulation. The tongue model to be fitted (described in section 2.5) is defined by a triangle mesh and its shape is controlled by 30 parameters. The error metric must capture how closely the surface of the model matches the ultrasound mesh. For each vertex in the ultrasound mesh, a line is drawn to a common point located centrally below the tongue. The squared distances along each of these lines between the ultrasound vertex and the point where the line intersects the tongue surface are summed to produce the total error.

Finding the intersection between each line and the tongue mesh requires a search through the polygons of the tongue. Exhaustively searching the entire polygon mesh for each vertex in the ultrasound surface would however make the error calculation prohibitively slow. A neighbouring-polygon algorithm is used to speed up the search. For each triangle in the tongue mesh, indices are stored to its neighbouring triangles. Given an initial estimate of at which polygon the intersection might occur, the point of intersection in the plane of this polygon is calculated. If the intersection occurs inside the polygon, the search is terminated. Otherwise, the search continues with the neighbouring triangle in the direction of the intersection. Typically, if the initial estimate is reasonable, only a few triangles need to be searched before the intersecting triangle is located.

In addition, as discussed in section 2.6.2, a tongue volume preservation term is added to the error function. This term represents the squared difference between current tongue volume and reference (neutral) tongue volume.

3.4.2 Dynamic fleshpoint fitting

In paper 7, dynamic data from optical motion tracking and EMA are used to drive the talking head and tongue to reproduce the articulatory movements of the recorded speaker. Measurement data are organised into a sequence of frames, each containing 3D coordinates of a number of points in the face and on the tongue, sampled at a rate of 60 Hz (see section 3.3). The basic approach is to carry out the error minimisation on a frame-by-frame basis to find the vector of articulatory parameters that best matches each frame.

Matching these data against the model would be straightforward if we could assume that 1) each marker corresponded exactly to one vertex on the model, and 2) the model geometry matches the overall facial shape of the recorded subject. In that case, the error function could be calculated as the sum of the squared distances between markers and corresponding vertices on the model. Unfortunately, if there are static shape differences between model and subject (as is the case here), the minimisation algorithm will try to compensate for these using articulatory parameters, leading to inaccurate articulation. Instead, the strategy chosen is to define a set of points - *virtual markers* - on the model, each of which corresponds to one measured marker. For a reference frame, the model is manually adjusted to match the measured subject's face and tongue configuration, and the virtual markers are "locked" to the model by freezing the position of a real marker. This is done by expressing the position of each virtual marker in terms of three non-colinear model vertices (see figure *8*). Thus, when the model moves, the virtual markers move with it. The error function is given as the summed squared distances between corresponding real and virtual markers. The main advantages of this procedure are that 1) marker-placement is less critical than if the markers were matched directly against existing vertices, and 2) it is possible to drive models whose geometries differ from the recorded subjects (for example cartoon-like faces).

A similar procedure, but without tongue data, was used in paper 8 to obtain training data for the articulatory control models described in section 4.4.

3.4.3 Data extrapolation

There are certain apparent limitations to the optical motion tracking systems when applied to articulatory measurements. For example, it is not possible to place markers on the inner lip contour, since the markers would obstruct the subject's articulation, and bilabial occlusion would cause the markers to be obscured, making them difficult to track. Still, lip closure is an important property to model correctly in a talking head, since bilabial occlusion is among the most highly visible articulatory movements.

The lips are highly deformable, and our current models are not always able to attain correct lip closure based only on the outer lip contour measurements, especially in rounded and protruded contexts. Therefore, in paper 7, a pre-processing stage is performed prior to model fitting, where points on the *inner* lip contour ($\mathbf{i}_1$ and $\mathbf{i}_2$ in figure 9) are estimated from the points measured on the outer contour ($\mathbf{o}_1$ to $\mathbf{o}_8$).

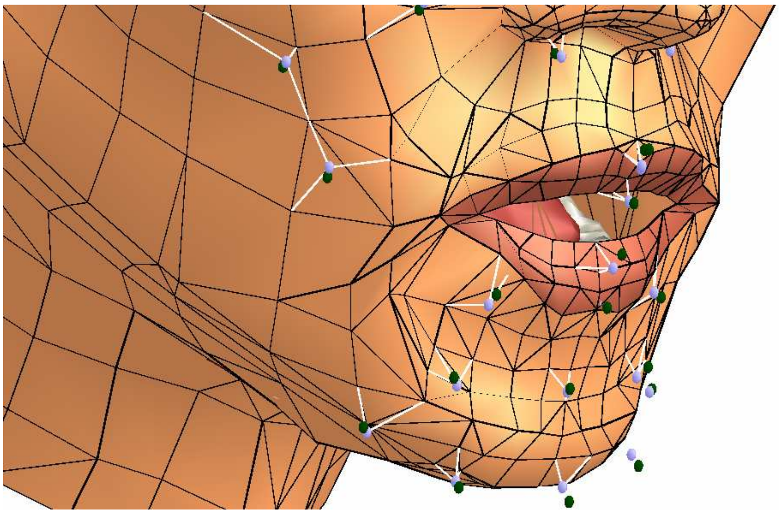


Figure 8. Each of the virtual markers (light blue spheres) is connected to three model vertices (white lines) and moves with the model as it deforms. They are compared against the position of corresponding measured markers (dark green spheres).

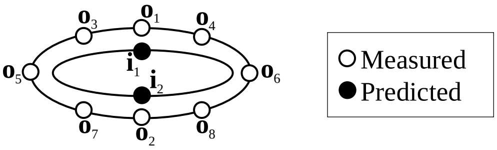


Figure 9. Two points on the inner lip contour are predicted from the measurements on the outer contour.

The estimation is based on linear regression. First a training set was assembled, by picking an equal number of frames from bilabial stops and from non-labial phonemes, for which the positions of the inner lip points were estimated using heuristics based on the phonetic labelling (see paper 7 for details). Then a linear estimator was trained on this data to predict the positions of $\mathbf{i}_1$ and $\mathbf{i}_2$ from the outer points. The reason for using a linear estimator and not using the heuristic method directly to predict the inner points is that the linear estimator preserves the dynamical properties of lip motion.

3.5 Correlating the data sources

When simultaneous measurements of tongue and face motion are made, it is of general interest to investigate the correlation between the two data sets. Furthermore, since EMA measurements are somewhat invasive and may have a slight impact on articulation, the prospect of *predicting* tongue movements from facial data seems like an interesting option. Yehia *et al.* (1998) proposed a procedure to derive an association between vocal-tract and face data for one English and one Japanese subject, for which non-simultaneous OPTOTRAK and EMA data had been recorded for a small sentence material (multiple repetitions of two English and five Japanese sentences). Jiang *et al.* (2000) reported on a similar study, where CV syllables were recorded with simultaneous QS and EMA registration for four subjects. In paper 7, the simultaneously recorded QS and EMA datasets were subject to the same type of analysis. The procedure followed in all these studies (described in more detail in paper 7) involves partitioning the material into training and test parts, and then using multiple linear regression to compute an estimator matrix, that can be used to express the vocal-tract point coordinates as a linear combination of face point coordinates, in a least-squares sense.

After predicting vocal-tract from face for the test part of the material, the correlation between predicted and measured tongue point trajectories is calculated. This correlation gives a measure of to what degree the tongue data can be recovered from face movements only. The inverse analysis, prediction of face from vocal-tract data, was also carried out in the mentioned studies, and in Yehia *et al.* and Jiang *et al.*, predictions from and to acoustic features were made as well.

The correlation measures for the vocal-tract to face and face to vocal-tract predictions in the different studies are summarised in table 2. One observation to be made about these data is that the number of coils (and their locations) have great impact on the resulting correlations. In paper 7, the estimations and correlations were calculated with 3 coils (tongue only), 4 coils (tongue + jaw) and 5 coils (tongue, jaw and upper lip), see figure 7 for an illustration of coil placement. The upper lip coil was co-registered optically. Yehia *et al.* used seven coils; four on the tongue, two on the lips and one on the jaw (lower incisors). Both lip points were registered optically as well as with EMA. Quite naturally, points that are directly co-registered show the largest predictability with correlations close to 1.0, and therefore they improve the mean correlation values significantly. The same is true for the jaw coil, which is highly correlated with optical markers on the chin. The results from Yehia et al are placed in the 5+ coil column in the table, but it should be kept in mind that they include two co-registered points, whereas paper 7 uses one, which is part of the explanation for the differences in correlation obtained in the studies. Jiang *et al.* only used three coils, all of them on the tongue, equivalent to the 3 coil case in paper 7.

When three coils are used, tongue is recovered from face better than the opposite, but when lips and jaw are included in the vocal-tract data set (5+ coils), face data is recovered from vocal-tract better than vocal-tract from face.

Another observation regards the different types of speech tokens used in the studies. In paper 7, VCV and CVC as well as sentences were used. VCVs produce much higher correlations than sentences and CVCs. One possible explanation for this could be that the VCVs are temporally dominated by three carrier vowels [a, ɪ, ʊ] for which there is likely to exist a strong association between facial and vocal-tract configuration. The same is true of the CV corpus used by Jiang *et al.* where 18 consonants were paired with 3 vowels.

Also the sizes and diversity of the corpora differ between the studies. While Yehia *et al.* used five repetitions of a small number (2-5) of sentences, and Jiang *et al.* used 69 CV-syllables, the study in paper 7 used 138 VCVs, 41 CVCs and 270 sentences. It is reasonable to assume that a more diverse corpus exhibits less predictability, entailing lower average correlations between measurements and predictions.

Table 2. Results from three studies where vocal-tract motion is predicted from facial motion and vice versa: correlation coefficients between measurement and prediction.

		Vocal-tract from face			**Face from vocal-tract**			**Comment**
	Corpus	3 coils	4 coils	5+ coils	3 coils	4 coils	5+ coils	
Paper 7	VCV	0.658	0.738	0.763	0.539	0.684	0.815	
	CVC	0.511	0.624	0.693	0.298	0.484	0.820	
	Sentences	0.525	0.636	0.690	0.357	0.581	0.724	
	All	0.520	0.624	0.670	0.385	0.595	0.737	
Yehia *et al.* (1998)	Sentences			0.81			0.91	Averaged over two subjects. 7 coils
Jiang *et al.* (2000)	CV	0.75			0.66			Averaged over four subjects.

4 PHONEME-DRIVEN ARTICULATION

Given a face model, and a way to animate the model to re-synthesise the motion of a real speaker, one important step remains, namely that of automatically generating animation from some form of input other than direct measurements. In a Text-to-Audio-Visual-Speech (TTAVS) synthesis system, there is typically a pre-processing stage common to both the acoustic and visual modalities that converts orthographic text into a phonetic representation. An articulatory control model is responsible for production of visible articulatory movements from the phonetic representation: a sequence of time-labelled phonemes. Other types of control models for generating articulatory movements directly from speech as well as generation of non-verbal gestures are discussed in the following two chapters.

For the resulting animation to appear realistic, the articulatory control model has to take a number of factors into account: critical articulatory targets have to be attained, dynamical properties and the timing of the articulator movements should be proper, and coarticulation must somehow be taken into account. Coarticulation refers to the way in which the realisation of a phonetic segment is influenced by neighbouring segments.

Two papers in this study deal with the subject of articulatory control models. Paper 1 presents a rule-based approach, based on introspection and phonetic knowledge, while paper 8 deals with data-driven articulatory control, evaluating a number of different approaches. The distinction of rule-based, or knowledge-based versus data-driven can be applied to most articulatory control models presented in the literature.

4.1 Rule-based models

Pelachaud *et al.* (1996) describe a rule-based control model where phonemes are clustered into visemes that are classified into different deformability ranks, which serve to indicate to what degree each viseme should be influenced by its context. Visemes with low deformability serve as key-shapes that influence the shape of the more deformable ones.

The model described by Cohen and Massaro (1993) is based on Löfqvist's (1990) gestural theory of speech production. Here, the trajectory of an articulator is modelled as being composed of overlapping articulatory gestures. Each segment is associated with a target value and a temporal blending function with exponential rise and fall, called the dominance function, that dictates to what extent the target influences its surroundings. At each point in time, the value of a parameter is given as the normalised weighted sum of all target values and their associated dominance levels at that instant. The height of each dominance function at the peak and the rate of rise and fall are free parameters that can be adjusted for each phoneme and articulatory control parameter.

In paper 1 of this thesis, a rule-based coarticulation model is presented. In this model, each phoneme is assigned a target vector of articulatory control parameters. To allow the targets to be influenced by coarticulation, the target vector may be under-specified, i.e. some parameter values can be left undefined. In this case, the value is inferred from context using interpolation, followed by smoothing of the resulting trajectory. As an example, consider the lip rounding parameter in a V_1CCCV_2 utterance where V_1 is unrounded and V_2 is rounded. Lip rounding would be unspecified for the consonants, leaving these targets to be determined from the vowel context by linear interpolation from the unrounded V_1, across the consonant cluster, to the rounded V_2.

An alternative to using rules or mathematical equations to combine context-independent phoneme descriptions is to consider context directly in the units used for synthesis. Breen *et al.* (1996) manually analysed a VCV video corpus to determine a set of static context-dependent visemes that were used as interpolation targets during animation. From context-dependent static targets, a natural extension is concatenation of dynamic units, sequences of video or trajectories based on measurements of a real speaker, which is the most direct way to incorporate substantial amounts of data into a model, in pursuit of more natural animations.

4.2 Data-driven models

4.2.1 *Concatenative approaches*

In analogy with the development in acoustic speech synthesis, where concatenative approaches are currently predominant, a number of researchers have proposed concatenation of dynamic context dependent units of visual speech. This technique is mostly applied in video-based systems, such as Video Rewrite by Bregler et al. (1997) where the unit of concatenation is triphones. Cosatto and Graf (2000) implemented an audio-visual variant of the unit selection strategy known from acoustic speech synthesis (Hunt and Black, 1996), where arbitrary sized units are selected from a large database of continuous speech, to minimise certain target and concatenation costs.

Concatenative approaches have also been applied to model based systems; Hällgren and Lyberg (1998) describe a system where control point trajectories, measured using an optical tracking system, are concatenated on a demi-syllable basis to drive a 3D talking head. Engwall (2002) employed unit selection on an EMA database to drive a 3D tongue model.

4.2.2 *Trainable coarticulation models*

Direct concatenation of parameter trajectories is one way to take advantage of data in articulation modelling. Trainable coarticulation models represent another. Coarticulation models with any number of free parameters can be automatically trained to reproduce patterns observed in data. The technique is to employ optimisation techniques to minimise the error between the predicted trajectory and the measurement by adjusting the parameters of the model. This approach has been taken by Le Goff (1997), who trained the dominance model of Cohen and Massaro (1993) for French, based on trajectories extracted from video recordings of a speaker uttering nonsense words. Cosi *et al.* (2002) employed a similar procedure to train a modified version of the Cohen-Massaro model on an Italian VCV corpus of optical tracking data. Cohen *et al.* (2002) also describe training their model for English based on a sentence corpus recorded with an optical tracking system.

Reveret *et al.* (2000) adopt Öhman's (1967) model of coarticulation. In this model parameter trajectories are modelled as a composition of a slowly changing vowel gesture and more rapid consonantal articulations. Consonants are described by a target value and an emergence function, dictating the temporal blending of vowel track and consonant target. The free parameters of the model were estimated from a corpus derived using video analysis of 24 VCV words.

4.2.3 *Trainable general-purpose models*

The coarticulation models described in the previous section have their origin in speech production theory and were originally conceived as tools for gaining more insight into the process of human speech production. In modelling articulatory movements for a talking head, there is however no a priori reason to be restricted to speech-specific models. If we view coarticulation as a general trajectory modelling problem, there is a large toolbox of mathematical and statistical methods that can be applied, as has been shown by several investigators.

Pelachaud (2002) models the trajectories of articulatory parameters for VCV-words as a weighted sum of radial basis functions (RBFs). The RBFs used by Pelachaud share some properties with the dominance functions of the Cohen-Massaro model, in that they are negative exponential functions, but in Pelachaud's model three RBFs per segment and parameter are used as opposed to one single dominance function. The free parameters of the RBFs were estimated from a training corpus of optically tracked VCV-sequences.

Hidden Markov Models (HMMs), a stochastic modelling technique mostly known from the speech recognition field, has been applied to the problem of trajectory modelling by Scott & Brooke (1998) and Tamura *et al.* (1998). In the latter system, each syllable is modelled as a state sequence where each state is associated with a gaussian probability density function for each parameter and its time derivative. The optimal (most probable) parameter trajectory is obtained by solving a set of linear equations.

A similar approach is taken in the pixel-based system of Ezzat *et al.* (2002), where parameter trajectory generation is posed as a mathematical regularisation problem (Girosi *et al.*, 1993). A trajectory is synthesised by finding the curve that optimises a combination of a target cost and a smoothness criterion. The target cost is based on how closely the trajectory matches a gaussian distribution of parameter values for each segment. The smoothness criterion is to minimise the 4th order derivative of the synthesised trajectory. To compensate for target undershoot caused by averaging across phones, the mean and standard deviations of the gaussian distributions were re-trained by an iterative error minimisation procedure.

Galanes *et al.* (1998) report on using regression trees, trained to predict one target vector per segment by looking at the immediate left and right contexts.

In general it can be argued that data-driven methods provide several significant advantages over the rule-based approaches. While manual rule building or parameter tweaking is a very time consuming task, data driven methods are automatic, which means that re-training a system to model a different speaker or even a new language can be accomplished with less manual labour. More importantly, data-driven methods can capture subtle details and patterns in the data that are difficult to model by rule, resulting in a (potentially) increased level of naturalness. On the other hand, a rule-based system offers a way to make incremental improvements by refining the articulation of individual segments. In addition, as we shall see in the next section, the increased naturalness of data-driven systems does not automatically lead to increased intelligibility.

4.3 An ANN-based control model with low-latency

In paper 8, an articulatory control model based on Artificial Neural Networks (ANNs) is suggested. Recurrent Time-Delayed ANNs (c.f. Ström, 1997) were trained on trajectories resynthesised from optical measurements for a corpus of 200 Swedish sentences. Input to the networks consisted of a series of feature vectors, one per time frame, where each vector contained 17 phonetic features uniquely describing the present phone. The three-layer ANNs had recursive connections in the hidden layer and 15 frames forward and 15 frames backward context provided by time delay in the input layer to model the dynamic properties of articulation. Frame rate was 60 Hz, which makes the corresponding context window +/- 250 milliseconds. Four separate ANNs were used, each predicting a group of related parameters. For example, one ANN predicted parameters related to bilabial occlusion (jaw rotation, upper lip raise and lower lip depression). One advantage of the ANN method over the other models described above, is that it operates in a time-synchronous mode, i.e. it uses a constant number of time frames look-ahead, which makes it suitable for real-time applications such as the one envisioned in the Synface project discussed in section 8.2, while other models typically require access to the full utterance before synthesis can begin.

Specifically with this type of application in mind, a second set of ANNs was trained, identical to the first except that the context window was shifted asynchronously to provide

2 frames forward and 28 frames backward context, yielding a low-latency control model with only 33 milliseconds look-ahead.

4.4 Control model comparison

Given the plethora of different models for articulatory control, a relevant question to ask is which one to prefer, faced with the task of building a system for visual speech synthesis. This is the basic question posed in paper 8, where a head-on comparison of a number of different articulatory control models is performed.

A direct comparison of the results presented in the various studies mentioned above is not possible to do for several reasons. In many of the studies, no evaluation of the results is reported. In (Cohen *et al.*, 2002) the goodness of fit of the model is reported as the average root mean squared (RMS) difference between synthesised and measured trajectories, normalised by parameter range. This measure can be used to compare the fit of different models, but a meaningful comparison cannot be made unless the models are trained on the same corpus. And while several studies report results from perceptual intelligibility evaluations (Geiger *et al.*, 2003; Le Goff, 1997; Cohen *et al*, 1996) these results are hard to compare since the experimental conditions vary greatly between the studies, as is discussed further in section 7.2.

In paper 8, four different control models were trained on an identical corpus and then evaluated objectively as well as perceptually. The four models were chosen to cover the different classes of trainable articulatory control models presented above: two speech-specific coarticulation models represented by the Cohen-Massaro and Öhman models, and two general purpose models based on ANNs, where one of the ANN models used a very short look-ahead time (33 ms) designed specifically to allow for real-time applications.

4.4.1 Training

Training data consisted of control parameter trajectories for a 10-parameter talking head model for a corpus of 200 phonetically rich sentences. The trajectories were resynthesised from optical tracking of 30 points on the face of a non-professional native Swedish speaker. Input to the control models consisted of time-labelled phone sequences with a phone set of 76 phones composed of 30 consonants, 23 unstressed vowels and 23 stressed vowels. In addition to the 200 training sentences, 89 sentences were set aside for testing.

In the Cohen-Massaro model, for each of the articulatory parameters each phoneme was modelled using five free parameters: a target value and a dominance function defined by the four parameters height, rise-time, fall-time and peak shape. Given 10 articulatory parameters and 76 phonemes, the model has a total of 3800 free parameters to be estimated. The model was trained in Matlab using the Gauss-Newton error minimisation algorithm. To improve training performance, the gradient of the error function was used. After each iteration, the error over the test set was computed. Training was terminated when this error stopped decreasing, to prevent over training and loss of generalisation.

In the Öhman model, fewer free parameters were used. Vowels were characterised only by their target value, and consonants by target value and a coarticulation factor, comparable to the dominance height in the Cohen-Massaro model, yielding a total of 1040 free parameters. Training was carried out in the same way as for the Cohen-Massaro model.

The two ANN-based models, as described in the previous section, were identical except for the input context window. The first model (ANN 1) used a symmetrical context of 15 frames forward and backward, plus current frame. The other model (ANN 2), was modified to provide lower latency time and used an asymmetrically shifted context window with 2 frames forward and 28 frames backward context, yielding the same total number of input nodes as for ANN1. Frame length was 1/60th second. The ANNs were trained using error back-propagation with the same termination criterion as for the other models.

4.4.2 *Objective evaluation*

An objective evaluation of the four control models was carried out by comparing target and predicted parameter trajectories over the 89-sentence test set. The average RMSE was calculated as percentages of the full parameter range. In addition the Pearson product-moment correlation was calculated between targets and predictions. Results are summarised in table 3. As the table indicates, the Cohen-Massaro model performs the best, having the lowest average RMSE, followed by the Öhman model and the two ANN models. Regarding correlation coefficients, the Cohen-Massaro model gives the highest correlation, followed by ANN1, Öhman and ANN2. It has been suggested (Yehia *et al.*, 1998) that correlation is a more relevant measure when comparing articulatory trajectories, since RMSE is too strongly influenced by areas of large amplitude (where larger errors are more likely to occur). Even though the Cohen-Massaro model appears to be better at predicting articulatory trajectories, the differences between the models are rather small, and it is not obvious how the objective measures relate to the quality of the resulting animations. In order to obtain a rating of this, a perceptual evaluation was carried out.

4.4.3 *Intelligibility evaluation*

A perceptual evaluation was carried out where 25 native Swedish subjects were presented with animations generated by the four data-driven control models, as well as the rule-based control model described in paper 1. In addition, there was an audio-only condition to provide a baseline for intelligibility, yielding six conditions. The material consisted of 90 noise-degraded sentences from a Swedish audio-visual testing corpus described in more detail in section 7.2. Each subject was presented with 15 sentences in each of the six conditions, the task being to repeat what they had perceived. Results were scored by counting percentage of correctly identified keywords within each condition. The results are shown in the lower part of table 3. Analysis of variance (ANOVA) and pairwise comparisons indicated that all face conditions gave significantly higher intelligibility than the audio-alone condition (p < 0,05), and that the rule-based control model provided higher intelligibility than the data-driven ones, but no significant difference could be found between the four data-driven models.

The perhaps unintuitive result that the data-driven models provide lower intelligibility than the rule-based model is likely due to the fact that while the data-driven models are trained to reproduce the speaking style of the target speaker, that in this case could be characterised as rather relaxed, the rule-based model was developed with clear articulation and high intelligibility as the primary goal. So while the rule-based method may provide less naturalistic articulation, its stylised, slightly exaggerated articulation actually leads to higher audiovisual intelligibility than its data-driven counterparts.

It is envisioned, however, that re-training the data-driven models on speech from a highly intelligible target-speaker could make the data-driven models equal to or better than the rule-based.

The fact that none of the data-driven models outperforms the others is encouraging from the point of view of real-time applications, since it indicates that the low-latency ANN model can be used without sacrificing performance.

Table 3: Control model comparison in terms of objective and perceptual measures.

	Cohen-Massaro	Öhman	ANN 1	ANN 2 (low latency)	Rule-based	Audio only
	Objective score					
RMSE(%)	9,04	9,50	9,61	9,61		
Correlation	0,6626	0,6188	0,6342	0,6065		
	Perceptual score					
Keywords correct (%)	74,8	75,3	72,1	72,8	81,1	62,7

5 AUDIO-DRIVEN ARTICULATION

In some situations, a phonetic transcript of the audio signal is not available, rather the goal is to drive the articulation of a talking head directly from an acoustic signal. One application of such a technique is discussed in section 8.2: talking heads providing automated lip-reading support for hearing-impaired telephone users. Other scenarios where audio-driven animation is desirable can be multi-user communication in virtual environments, where each user is represented by an avatar controlled from his/her own voice (Morishima, 1998) or in traditional computer animation, where lip-synchronisation is considered a tedious and time-consuming task and automated procedures can significantly ease the animator's burden.

The problem of audio-driven articulation is studied in paper 4 of this thesis. Numerous researchers have investigated the mapping of acoustic features into visual parameters for facial animation (Arslan and Talkin, 1998; Brand 1999; Hong *et al.*, 2002; Lavagetto, 1996; McAllister *et al.* 1997; Morishima, 1998; Parke and Waters, 1996; Tamura *et al.*, 1998; Yamamoto *et al.*, 1998; Yehia *et al.*, 1998; Öhman and Salvi, 1999). The approaches can broadly be divided into two categories: symbol based and regression based. In the former category, the audio signal is first translated into an intermediate discrete representation, before it is converted to visual parameters. In the latter category, into which paper 4 belongs, a direct continuous association is made between acoustic and visual features. Öhman and Salvi (1999) compared two approaches, one symbol-based (using HMMs) and one regression-based (using ANNs).

5.1 Symbol-based mapping

In 1987, Lewis and Parke (as cited in Parke and Waters, 1996), classified acoustic frames into one of 12 viseme classes (9 vowels and 3 consonants) by matching the LPC spectrum of the incoming frame against that of visemes prototypes, each of which had a facial model parameter vector associated with it.

Arslan and Talkin (1998) proposed using an audiovisual codebook approach for estimating face point trajectories from speech. Each codebook entry contains acoustic as well as visual features. Incoming acoustic feature vectors are rated in terms of their similarity to the acoustic parts of the codebook entries, and the corresponding visual parts are combined as a weighted sum, where the weight of each entry is based on the acoustic similarity, to generate visual output vectors.

5.1.1 HMM-based methods

Several investigators have proposed solutions based around an Hidden Markov Model (HMM) framework. HMMs are a natural choice for this type of problem, which can be viewed as a form of acoustic recognition, outputting visual features values rather than words. What differs between the HMM-based approaches are mainly two things: what the basic unit for recognition is (i.e what the HMM states represent), and how parameter trajectories are synthesised from the recognised state sequence.

Öhman and Salvi (1999) trained an HMM to recognise Swedish phonemes, which were used as input to a rule based articulatory control model (the model described in paper 1, outlined in section 4.1). The performance was evaluated for three different symbol sets: monophones (context-independent phonemes), visemes and triphones (context-dependent phonemes). Best performance was achieved with triphones.

Yamamoto *et al* (1998) used monophones as the basic unit for recognition. For trajectory synthesis, visual parameter vectors for each frame came from table lookup indexed by context-dependent Japanese visemes.

Tamura *et al.* (1998) modelled Japanese syllables as HMM state sequences, and used an integrated synthesis procedure where each state had an associated gaussian probability distribution for each visual parameter and its time derivative, from which the resulting parameter trajectory was obtained as the solution to a linear optimisation problem.

Brand (1999) used a similar technique for trajectory synthesis, but the basic units in Brand's study were automatically derived during training based on entropy minimisation in the HMM models.

5.2 Regression-based mapping

One disadvantage of using an intermediate discrete representation is that errors potentially can be more severe. If a velar stop is misrecognised as a bilabial, the resulting animation is likely to be very misleading. If the intermediate symbolic level can be eliminated, there is a possibility that errors could be gradual, allowing for graceful degradation.

The regression-based approaches are all based on the concept of training some form of continuous estimation algorithm to learn the complex relationship between corresponding acoustic and visual parameters in the speech signal. Yehia *et al.* (1998) demonstrated that, for a limited number of sentences from the same speaker (in English or Japanese), a linear estimator could predict the facial movements from acoustics features, yielding a correlation between prediction and measurement of around 0.7. It is not clear whether this predictive performance would be enough to drive a face model with sufficient quality for speechreading, but it is nonetheless interesting that a linear association can recover a fair amount of the information.

McAllister *et al.* 1997 employed non-linear predictor surfaces, trained on a small English vowels-only corpus, to map three acoustic features into three visual parameters for vowels.

5.2.1 ANN-based methods

Artificial Neural Networks (ANNs) provide a learning framework that is well suited for the task of acoustic to visual feature mapping: they can model non-linear associations, and through the use of hidden units they are also capable of modelling complex input-output relationships.

Lavagetto (1996) used one four-layered Time-delayed Neural Network (TDNN) to predict each one of five articulatory parameters based on acoustic feature input (12 cepstral coefficients). A TDNN considers context by not only using the current frame as input, but also a number of previous and succeeding frames. The networks were trained on 25 minutes of Italian speech. The resulting animations were evaluated through a speechreading experiment. Morishima (1998) trained an ANN to map LPC cepstrum parameters to mouth shape parameters for Japanese vowels. Different networks were trained on speech from 75 different talkers. Faced with a new talker not in the training material, the system automatically picked the network producing the best output for a small adaptation material of five vowels. In the study by Öhman and Salvi (1999), visual training data was generated for a large phonetically labelled acoustic database using a rule-based articulatory control model, after which speaker-independent recurrent ANNs were trained. Hong *et al.* (2002) used a Gaussian classifier to classify each frame into one of 44 groups based on acoustic properties, each of which was modelled by a separate TDNN.

5.3 Experiments with an ANN-based method

The study covered in paper 4 employs ANN's to map acoustic features into talking head control parameters for Baldi. The paradigm used is similar to that of Öhman and Salvi (1999) in that rather than using facial measurements, training data is generated using an articulatory control model from a TTAVS system. The advantage is mainly that existing audio-only speech corpora can be utilised, not requiring any task-specific data collection. In paper 4, TDNNs were trained on two different corpora: isolated monosyllabic words (sampled 16 kHz) and extemporaneous telephone speech, both in English. The word corpus was taken from an audiovisual speechreading corpus (Bernstein and Eberhardt, 1986). The acoustic data were phonetically labelled using Viterbi-based forced alignment, and 39 articulatory control parameters for the talking head were generated from the time-aligned phonetic transcriptions using the Cohen-Massaro articulatory control model (see section 4.1). There were 400 words for training and 68 for testing. Acoustic feature vectors consisting of 13 cepstral coefficients were generated at 50 frames per second. The TDNN used an input context window of eleven frames (five frames back + current + five frames forward), yielding 143 input nodes and 39 output nodes, one per control parameter. The network used 600 hidden units and was able to learn the mapping from acoustic to visual features with a correlation between target and prediction of 0.77 for the training set and 0.64 for the test set.

A perceptual evaluation was carried out with 17 subjects, using a paradigm similar to that of Cohen *et al.* 1996 (see section 7.2), but without the natural face condition. Silent animations were generated for 66 words not included in the training corpus by the TDNN (from acoustics) as well as by the Cohen-Massaro control model (from phonetic labels, serving as a reference condition). Average percentage of correct viseme identification from the TDNN animations was 46%, compared to 72% for the phonetically driven system, i.e. well above chance level.

The monosyllabic word corpus represents a well-structured task, and it is reasonable to assume that the uniform temporal characteristics of the words makes the TDNN perform better for this task than it would for more diverse patterns. In order to evaluate this aspect, a more challenging corpus was chosen, consisting of extemporaneous speech recorded over the telephone, where each speaker speaks about a topic of his or her choice for one minute. The data was taken from the CSLU stories corpus[11]. Speech from ten speakers was used to train ten different speaker-dependent networks. Data from each speaker was restricted to about 50 seconds, 40 of which were used for training and 10 for testing. Due to the paucity of training data, fewer hidden nodes (100) were used than in the word experiment (600). Using the restricted material, the TDNNs were able to learn to reproduce the training data (average correlation between target and prediction around 0.8), but generalisation to unseen data was poor (average correlation around 0.2) clearly indicating that more data is required for this task.

[11] `http://cslu.cse.ogi.edu/corpora/stories/`

When I have nothing to say, my lips are sealed
Say something once - why say it again?

David Byrne (Talking Heads), Psycho Killer, 1977

6 NON-VERBAL SIGNALS

While much of the discussion in this thesis so far has been geared towards visible speech and modelling the movements of the visible articulators, it is important to consider other roles of the face in communication. Signals such as eyebrow movements, eye gaze, head movements and smiling all have fundamental roles in communication, and any communicatively realistic talking-head model needs to take this into account. In this chapter, we will consider these non-verbal cues (where the term non-verbal is used to denote all movements that are not directly linked to the production of speech sounds). A general categorisation of different types of non-verbal signals is followed by a review of proposed methods and systems for generating non-verbal communicative signals in text-to-speech and dialogue systems. Paper 5 in this thesis presents a gesture-library approach to generating non-verbal gestures in the context of spoken dialogue systems.

6.1 Signal categories

6.1.1 Emotion

One important use of facial cues is to signal emotion. Ekman and Friesen (1975) described six facial expressions of emotion that have been characterised as universal in that they occur in all cultures on earth (happiness, sadness, fear, disgust, surprise and anger). While such a prototypical categorisation of static displays can be useful in studying the psychological and evolutionary aspects of human emotion, it is likely that the practical use of emotion in applications of talking heads in man-machine communication requires a more nuanced and dynamic encoding of emotional expression. While such a description is outside the scope of this thesis, there are other areas where facial expressions and gestures can play an important role in practical applications.

6.1.2 Communicative signals

According to Ekman (as cited by Pelachaud, 1991), non-emotional facial expressions can be divided into the following categories: *manipulators, emblems, illustrators, punctuators* and *regulators*. While *manipulators* are gestures motivated by direct biological needs, such as blinking to wet the eye, the other categories are communicative signals related to conversation and face-to-face interaction.

Emblems are visual replacements for, or augmentations to, spoken words, such as nodding for agreement. The role of *illustrators* is to augment the speech signal, for example by some movement on emphasised words, while *punctuators* signal pauses and phrase boundaries in the speech. These signals are typically instantiated by eyebrow or head motion, but it has also been shown that eye-blinks often tend to be synchronised with events in the speech signal.

Regulators are signals used in dialogue situations to direct the flow and turn taking in the conversation. Signals for taking and giving up the turn in a dialogue, as well as acknowledgements and back-channelling while the other party is speaking are examples of regulators.

6.1.3 *Visual prosody*

Prosody refers to suprasegmental variation in the speech signal correlated with stress, prominence and phrasing. Typical acoustic correlates of prosody are F0 contour, loudness contour and segmental durations. There is a clear analogy between acoustic prosody and the visual gestures correlated with speech (mainly illustrators and punctuators). Visual prosody has been studied to a much lesser extent than acoustic prosody, but it is reasonable to assume that prosody in the two modalities manifest the same underlying code, and as such are highly correlated. Cavé *et al.* (1996) studied the occurrence of eyebrow movements in relation to F0 variation for ten speakers recorded with an optical tracking system. They found that F0-rises were accompanied by rapid eyebrow movements in 71% of the cases. Graf *et al.* (2002) automatically extracted head movement parameters from several hours of video recorded speech, and studied their correlation with events output by a prosodic prediction tool from a text-to-speech system. They found several patterns of head movements that frequently co-occurred with prosodic events.

Perception of visual prosodic cues has been studied using the visual speech synthesis tools developed as part of this thesis work. In Granström *et al.* (2001) and House *et al.* (2001) it was shown that timing of eyebrow movements affected the perception of prominence in a sentence.

6.2 Gesture generation

Given the presumed link between acoustic and visual prosody, it seems natural to treat generation of prosodic patterns in the two modalities in a similar way. This is the approach taken by Pelachaud *et al.* (1996). A complete rule-system is devised for automatic generation of voice intonation, articulation and facial expression from an underlying representation of text that has been bracketed into elements and with accents marked. The system predicts occurrences of illustrators, punctuators, manipulators and regulators.

Cassel *et al.* (2000) describe BEAT - a toolkit for automatic generation of facial and bodily gestures from text. To obtain a basis for gesture generation, it includes a language-tagging module that identifies e.g. old and new information (theme and rheme) that can be used to produce contrastive emphasis.

6.2.1 *Gestures in conversational systems*

Incorporating talking heads as animated agents in conversational dialogue systems opens up new possibilities in terms of the modelling of non-verbal gestures, when compared to the plain text-to-speech case. Firstly, a richer input description than plain text can be used, since the system generating the output has knowledge about the underlying semantics of the utterance as well as the dialogue history, which can be taken advantage of in generating appropriate gestures. Secondly, proper use of regulator gestures for turn-taking and back-channelling has the potential of making the system appear more responsive and making the dialogue run smoother.

Several dialogue systems incorporating facial and other gestures at various levels have been described. Nagao and Takeuchi (1994) used static facial displays to convey conversational signals in a dialogue system. Cassell *et al.* (1994) modelled a virtual conversation between two virtual characters, where each of the characters' speech, gestures and facial expressions were automatically generated using rules similar to those of Pelachaud *et al.* (1996). Poggi and Pelachaud (2000) model the intention and communicative goal of a virtual agent and use that information to generate appropriate facial displays. Thórisson (1999) and Cassell *et al.* (2000) both describe complete frameworks for conversational dialogue systems that incorporate animated agents capable of generating deictic gestures, turn-taking signals, and emblematic gestures, relying on input from several sources.

6.2.2 *Output markup*

To pass information from the response generating module of a dialogue system (typically the dialogue manager) to the animated agent, it is common to use a markup language to annotate the text to be spoken with other information that can be relevant to the agent. The XML-based Speech Synthesis Markup Language (SSML) (Burnett *et al.*, 2002) or the Virtual Human Markup Language (VHML) (Gustavsson *et al.*, 2001), including tagsets for speech as well as for facial and body animation, can be used to this effect. With these languages, it is possible to specify realisation of facial expressions, gestures and intonation at a rather detailed level. One problem with using such a detailed markup as output from the dialogue manager is that it can lead to loss of generality and flexibility in the dialogue system, since it requires the dialogue manager to know about the features of the output device: How should emphasis be realised in this device? Does it have eyebrows? Etc.

6.3 Generic System Output Markup

To allow for a high-level abstract encoding of non-verbal output from the dialogue system, the Generic System Output Markup (GESOM), is proposed in paper 5, along with a method of converting such markup into gestures for a talking head. The intent of GESOM is to specify communicative functions, but not the exact realisation of these functions. Instead, realisation is left to the output device, for example an animated talking agent. Indeed, one of the goals is to provide an output specification that is valid for multiple types of output devices e.g. web-browsers, cell phones as well as animated talking characters.

It needs to be stressed that the scope of GESOM is limited to providing a high-level abstraction of dialogue system output; not an alternative to SSML or VHML. Indeed, it is fully possible to use these languages to realise GESOM output for animated talking characters.

6.3.1 *States, events and background attributes*

At different times during a dialogue, each party can be said to be in different conversational states: waiting for speech, listening, thinking, speaking. Such states are the foundation for GESOM encoding of output. At each point in time, the output device is in a specific state, which can be changed through the **`state`** markup. One state lasts until another state begins. Thus states can be used to encode output of which the duration is not known beforehand, such as listening to a user (it is not possible to know beforehand when to stop listening) or searching for information over a network connection. At other times, the duration of the output is finite and predictable, such as an emphasis signal associated with a particular word or phrase. This type of output is marked as an **`event`**.

In addition to the general states and events, GESOM provides a way to supply additional information that may optionally be used during realization of a particular function. This information is encoded using the **`background`** attribute that can accompany any tag. An example of usage is to choose a proper realization of emphasis in a talking head based on whether the current statement is affirmative (in which case a head nod might be appropriate) or negative (in which case an eyebrow lowering or a head shake works better). Background attributes could also be used to encode emotional information, so that gesture or voice realizations could be affected appropriately.

In order to maximize flexibility, the GESOM specification does not define a strict set of states and events that should be used in a given application. However, to illustrate the intended use, the set of states, events and background attributes used to control the agent in the AdApt system (Gustafson *et al.*, 2000) are listed in table 4. As can be seen, the state set represents a rather straightforward encoding of the different stages that occur during a dialogue turn, and it is likely that this set can be used unmodified in other applications as

Table 4: States, events and background attributes in the AdApt system GESOM implementation.

	States
Name	Description
idle	System is inactive
attentive	System is ready for input
cont_attentive	System has received input that is not sufficient to prepare a response
busy	System is busy preparing a response
text_presentation	System is presenting a response
sleep	System is off-line
	Events
Name	Description
break	System should pause the presentation at this event
emphasis	Marked text should be emphasised
	Background attributes
Name	Description
attitude:negative **attitude:positive** **attitude:neutral** **attitude:question**	Relates to type of response

well, but there might be cases where additions or alterations are necessary. It can be noted that only two events are defined: **emphasis** and **break**. These events have the same meaning as their SSML counterparts, and can thus be easily mapped into SSML tags if the output device employs an SSML-compatible TTS system.

6.4 The gesture library

It is the task of the animated agent module to convert the states, events and background information present in the GESOM-encoded message into facial gestures to accompany the speech. This process is based around a gesture library. The gesture library contains entries for each state and event, and holds groups of semantically equivalent or similar gestures. The gestures for the parameterised talking head are encoded as parameter trajectories and include the time offset to the stroke of the gesture, i.e. the part that carries the meaning of the gesture, which needs to be synchronised to the timing of other events such as stressed syllables.

For each event, the library holds a list of candidate gestures that can be used to realise this event. States cannot be realised with a single gesture, since they are of arbitrary length. Therefore, states are divided into three phases: entry, sustain and exit. For each of the phases, the gesture library holds a list of candidate gestures. Enter and exit gestures are executed once, while the sustain gestures are executed repeatedly on random intervals based on an intensity level that can be set on a per state basis. The use of multiple candidate realisations makes it possible to avoid the repetitiveness that may occur if a particular function (e.g. emphasis) is realised using the same gesture every time.

6.4.1 Choosing between gesture candidates

Gesture selection is done in a weighted-random fashion. Each candidate gesture has an associated weight that indicates how likely it is to be chosen. These weights may be static or may be dynamically updated to reflect changes in background attributes or as a function of time. Letting the background attributes affect the weights of gestures can be a way of modelling attitudes or emotions; the gesture repertoire is altered or completely replaced to reflect the attitude.

Time-varying weights allow for the possibility of modelling a gradual change in the behaviour of the agent during a state. This is useful in the idle state, where long idle periods can cause the agent to exhibit signs of boredom such as yawning or even falling asleep.

6.5 Representation of communicative signals in GESOM

We started this chapter by looking at Ekman's categorisation of non-verbal communicative gestures. We will now consider how these categories map into the proposed framework of the gesture library and the Generic System Output Markup with the tag set shown in table 4.

The first category, *manipulators*, are compactly represented by sustain-gestures in the gesture library; they are executed repeatedly with randomly varying time intervals, and are not synchronised with any other events.

Emblems are probably best represented by their own event types in GESOM, although they have not been implemented in the present system.

Illustrators essentially correspond to the **`emphasis`** event in GESOM, while the *punctuator* gestures are represented by the **`break`** event.

In GESOM, dialogue flow information is encoded in terms of states, and consequently, the *regulator* gestures occur when the agent enters different conversational states: a take-turn-signal occurs when the **`text_presentation`** state is entered, while entering the **`attentive`** state results in a give-turn-signal. Back-channelling signals can occur when the **`cont_attentive`** state is entered, indicating that the system has received input that is incomplete or otherwise insufficient to prepare a response.

7 EVALUATION

As we have seen, computer-based talking heads can be generated using a variety of techniques, and with different goals and motives. However, there is one thing uniting all talking heads: they are all in the end intended for human consumption. Thus, an essential part of talking head development is perceptual evaluation with human subjects, to measure the degree to which the talking head is successful in its pursuit. Depending on the application, it may be relevant to measure the degree of visual realism, or it may be more interesting to study how well it communicates the intended message, be it spoken or non-verbal.

7.1 Visual realism versus communicative function

There are many aspects of talking heads that can be subject to evaluation. One aspect concerns the visual realism of the talking head, i.e. to what degree it is mistakable for a real human character, often referred to as photo-realism. For obvious reasons, video-based systems commonly obtain a much higher degree of photo-realism than 3D-model based systems, since they are indeed based on photos. It is often a problem, however, that the illusion given by the still face can break down quickly as the face starts to speak, so any evaluation of photo-realism has to be based on moving sequences, preferably speech. Geiger *et al.* (2003) performed such an evaluation with the video-based system *Mary 101* (Ezzat *et al.*, 2002) (see figure 2a) . In their visual "Turing test", subjects were presented with synthetically generated sequences and words as well as their natural counterparts, and were asked to judge after each presentation whether it was real or synthetic. Average identification for the 22 subjects did not differ significantly from 50% (chance) which means subjects were not able to tell the synthetic face from the real. While this is an impressive result, it is interesting to note that in a second experiment, when faced with a speechreading task, the same subjects performed much worse with the synthetic face than with the natural one (see table 5). This tells us something important: We may be seduced by video-realistic animations, but when it comes to speech perception we are much more discriminating. This result is in line with one of the underlying assumptions in this thesis: that the visual realism and communicative function of a talking head are not necessarily correlated.

In the applications primarily targeted in this thesis (and further discussed in chapter 8), communicative function has been considered to be of primary importance, while visual realism of the talking heads have been a subordinate goal. Indeed, in some cases deliberately unrealistic cartoon-like faces have been employed. The main communicative function that has been in focus throughout this thesis is the audiovisual intelligibility.

7.2 Intelligibility studies

It is obvious that the intelligibility of a talking head is of crucial importance in applications such as that envisioned in the *Teleface* and *Synface* projects discussed in section 8.2, but it is also relevant in many other situations such as dialogue systems in public environments where the audio signal may be degraded by noise.

A number of investigators have reported results from intelligibility experiments with talking heads, summarised in table 5 (Agelfors *et al.*, 1998; Beskow *et al.*, 1997; Cohen *et al.*, 1996; Geiger *et al.*, 2003; Le Goff *et al.*, 1994; Olivès *et al.*, 1999; Pandzic *et al.*, 1999). Typically, the basic structure of these experiments is borrowed from audiovisual perception studies (c.f. Erber, 1969; Sumby and Pollack, 1954) and aims to measure either phoneme identification (representing bottom-up processing) or speechreading performance (top-down processing).

Intelligibility evaluation has been an integral part of the talking head development that this thesis represents. Consequently, four of the eight papers (papers 2, 4, 6 and 8) report results from perceptual intelligibility evaluations, and in paper 6, evaluation comprises the main theme of the study. A series of such intelligibility studies has been performed: phoneme identification (Beskow *et al.* 1997, Beskow *et al.*, 2002), speechreading and phoneme identification with hearing impaired subjects (Agelfors *et al.*, 1998), as well as a multilingual speechreading study (paper 6). Thus paper 6 represents the latest result in this series of assessments.

In paper 2, a comparative study between different face models is reported; the modified Parke-model of paper 1 is compared to the more cartoon-like Olga-model (finding the two models equally intelligible, see table 5). This study was piggybacked onto the evaluation reported in (Beskow *et al.* 1997), but that paper never reported the Olga scores.

In papers 4 and 8, intelligibility studies are included as part of the evaluation scheme for new control models. In paper 4, the ANN-based acoustic-to-visual-speech mapping is evaluated and compared to the text-to-visual-speech case. Here, a pure speechreading test was done, i.e. no audio signal was available.

In paper 8, a speechreading comparison of four data-driven models and one rule-based articulatory control model is carried out, with an auditory alone condition providing the baseline level.

What all evaluations of this type have in common is that subjects are presented with animations/video sequences of speech, and are asked to report what they perceive. Their responses are scored according to some pre-defined scheme. Apart from that, the details of the studies can differ in many ways.

7.2.1 *Purpose*

As illustrated by the papers in this thesis, the purpose of conducting an intelligibility study varies across studies. Often, the intent is a general assessment of the quality of a talking head without a particular application in mind. In some cases, an additional goal is a diagnostic evaluation, in order to pinpoint problem areas to guide further development efforts (Beskow *et al.*, 1997; Cohen *et al.*, 1996). This is often a fruitful strategy when developing rule-based systems. Sometimes the evaluation is done with a well-defined application in mind, in which case the study can be designed to simulate the given application scenario, as was done in the *Teleface*-related study of Agelfors *et al.* (1998) where hearing impaired subjects were used and speech was filtered to telephone bandwidth. In other cases, the goal of an evaluation is to compare the performance of a number of candidate technologies for a certain component, for example different face models (paper 2) or different control models (papers 4 and 8). Pandzic *et al.* (1999) compared a 3D-based talking head to a sample-based one (Cosatto *et al.*, 1998). The study by Beskow *et al.* (2002) had the specific goal of finding out whether a simple linear amplification of the articulatory parameters in the rule-based synthesis could lead to improved intelligibility.

7.2.2 *Items and scoring*

The type of items used in the tests varies between nonsense words (VCV, VCVCV), words and sentences. Consonant identification tasks with nonsense VCV words (Agelfors *et al.*, 1998; Beskow *et al.*, 1997), provide a sharp diagnostic tool to pinpoint problematic segment realisations but represent a rather unnatural condition. Sentence speechreading tasks (Agelfors *et al.*, 1998; Paper 6) are better suited for measuring overall intelligibility, but make it difficult to study individual phoneme confusions, since lexical and semantic constraints affect the subjects' responses. A compromise can be to use isolated monosyllabic words as was done by Cohen *et al.* (1996) and in paper 4, or digit sequences (Pandzic *et al.*, 1999).

Scoring of the results obviously differs according to the type of items used. With VCV words, the common way of scoring is counting percentage of correctly identified consonants. With the monosyllabic word corpus, scoring is typically done on initial consonant, final consonant or vowel. Cohen *et al.* (1996) also cluster the responses according to visemes prior to scoring, which leads to higher scores.

Scoring of sentence-based materials can be done on phoneme, syllable, word or keyword level. McLeod and Summerfield (1990) proposed a set of sentences and a keyword-based scoring procedure for audio-visual speech intelligibility testing in English. The sentences were short everyday sentences, 5 to 9 words in length, with three words marked as keywords. Scoring is based on number of correctly identified keywords, ignoring errors of morphology. A corresponding sentence material has been developed for Swedish by G. Öhngren (unpublished, the sentences are listed in Öhman, 1998). This material was used for the sentence intelligibility tests in Agelfors *et al.* (1998) and in papers 6 and 8.

7.2.3 Methods of audio manipulation

Another variable between studies is whether or not audio is presented along with the visual speech. "Pure" visual speechreading tasks (Cohen *et al.*, 1996; Geiger *et al.* 2003, paper 4) are normally very difficult, and the scores at word level are typically rather low, but nevertheless informative.

More common, however, are speech-in-noise experiments, following the paradigm of audiovisual intelligibility experiments for natural speech (Sumby and Pollack, 1954; Erber, 1969; Benoît *et al.*, 1994), where the acoustic speech signal is degraded by various levels of background noise, in order to prevent ceiling effects, since clear audio would normally yield intelligibility scores close to 100%. An exception is when hearing-impaired subjects are used (Agelfors *et al.* 1998), where clean audio can be used. Le Goff *et al.* (1994) used additive white noise at six Signal-to-Noise Ratio (SNR) levels ranging from -18 to 6 dB (only the -18 and 0 dB cases are shown in table 5). Olivès *et al.* (1999) used pink noise at four levels from -18 to 0 dB SNR. Other studies have used noise at a single SNR level (Beskow *et al.*, 1997; Pandzic *et al.*, 1999). A potential problem with static additive noise is that local energy perturbations in the signal can cause the effective SNR to change; portions with low amplitude can easily "drown" in noise. A different type of audio degradation is the noise-excited vocoder proposed by Shannon *et al.* (1995) that is used in papers 6 and 8 in this thesis. Here, the signal is reconstructed as a sum of bandpass-filtered channels of white noise, distributed across 100-5000 Hz, where the amplitude of each channel is modulated by the energy in the corresponding frequency band of the original signal. The ranges of the frequency bands are selected to cover equal parts of the basilar membrane. The advantage of this method over static additive noise is that it is robust to local energy perturbations in the signal. The degree of difficulty can be controlled by varying the number of channels used in the reconstruction; in paper 6, two and three channels were used. In paper 8, three channels were used.

7.2.4 Subjects

A third factor that differs between studies is the subjects used. Most studies use normal-hearing subjects, as this has some obvious methodological advantages: normal-hearing subjects are easy to find, and there is one less variable to control for: the degree of hearing impairment. When hearing-impaired subjects are used, as in Agelfors *et al.* (1998), the level of performance varies significantly, because each subject's hearing impairment is unique, and there is a larger risk of experiencing floor and ceiling effects for some subjects. Still, when evaluating technology with the goal of being used by hearing-impaired users, there is no real substitute for using hearing-impaired subjects.

Table 5: Overview of intelligibility studies with synthetic talking heads (ordered by publication date)

	Audio	**Items**	**Scoring**	**Lang-uage**	**Main results (% correct)**			**Comment**
					No face	Synt. face	Nat. face	
Le Goff *et al.* (1994)	-18 dB SNR	VCVCV	Cons.	French	0	42	62	Also: -12, -6 & +6 dB SNR
	0 dB SNR	VCVCV	Cons.	French	64	84	84	
Cohen *et al.* (1996)	No audio	Monosyl. Words	Intl.cons viseme	English	-	55	73	
Beskow *et al.* (1997) + *Paper 2*	3 dB SNR	VCV	Cons.	Swedish	63	70	76	Parke model
	Synt. speech, 3 dB SNR	VCV	Cons.	Swedish	31	45	-	
		VCV	Cons.	Swedish	31	47	-	Olga model
Agelfors *et al.* (1998)	Clean, telephone filtered	VCV	Cons.	Swedish	30	55	58	Hearing impaired subjects
		Sentences	Keywrds	Swedish	57	66	83	
Pandzic *et al.* (1999)	-2 dB SNR	Digits sequences	Digits	English	83	90	-	3D face
					83	91	-	sample based face
Olivès *et al.* (1999)	-18 dB SNR	VCV	Cons	Finnish	5	19	39	Also tested: -12 & -6 dB SNR
	0 dB SNR	VCV	Cons	Finnish	64	68	77	
Paper 4	No audio	Monosyl. words	Intl.cons viseme	English	-	42	-	Audio-driven
					-	76	-	Text-driven
Paper 6	2 ch. vocoder	Sentences	Keywrds	Swedish	6	24	28	
				English	14	37	68	
				Dutch	2	15	32	
	3 ch. vocoder	Sentences	Keywrds	Swedish	32	61	66	
				English	37	58	83	
				Dutch	19	40	62	
	No audio	VCV	Cons	English	-	14	23	
Geiger *et al.* (2003)	No audio	words & sentences	Words	English	-	7	15	
Paper 8	3 ch. vocoder	Sentences	Keywrds	Swedish	63	75	-	Data-driven
					63	81	-	Rule-based

7.2.5 Results

Table 5 summarises the main variables and results from the intelligibility evaluations reported on in this chapter. The first conclusion that can be drawn from the table is that synthetic faces improve intelligibility over audio-alone conditions in all studies. The second is that they typically fall short of natural faces; no study has reported synthetic faces achieving higher intelligibility than a natural face, although in paper 6, there was no significant difference found between the synthetic and natural face condition for Swedish, and LeGoff *et al* (1994) report equal scores for synthetic and natural face at 0 dB SNR. A natural face is however no absolute calibration level for intelligibility; there is large intra-talker variability. It is likely that many of the synthetic faces could prove more intelligible than *some* natural talkers whose articulatory movements are restricted, or obscured, for example by facial hair.

This intra-talker variability is one reason that the intelligibility scores of different studies are difficult to compare. In the cross-lingual study of paper 6, there were significant intelligibility differences between natural and synthetic faces for English and Dutch, but not for Swedish, as noted above. Even though the experimental conditions were identical except for the natural speaker used (and language, obviously) it is not possible to tell whether this fact could be explained by a more refined articulatory control model for Swedish than for English and Dutch or by lower audiovisual intelligibility of the Swedish talker.

8 APPLICATIONS

To conclude the thesis, this chapter will review how talking heads have been and are being applied in other research projects and systems developed at CTT. In fact, much of the development represented in this thesis was directly driven by immediate needs in these systems. There are three main areas of applications that have been targeted during the course of development: spoken man-machine dialogue systems, communication aids for the hard of hearing and speech training for hearing-impaired children.

8.1 Embodied agents in spoken dialogue systems

Spoken dialogue systems for human-machine communication have been an increasingly popular area for research over the past decade. Of particular relevance here is the branch of such research focusing on multimodal systems incorporating animated talking agents capable of speech and non-verbal behaviour. There are several compelling reasons to include animated agents (in the form of talking heads or full-bodied characters) into spoken dialogue systems: the non-verbal signals reviewed in chapter 6 can be of significant value in a dialogue system if used properly. Furthermore, the agent provides a sense of presence to the system, potentially making the dialogue situation more comfortable, and the lip movements of the agent boosts the intelligibility of the speech signal, which can be an important side effect in kiosk-type systems in noisy environments. At KTH (CTT), there is a long tradition of research in the area; the first spoken dialogue system research project, Waxholm, was initiated in 1993. The reader is referred to Gustafson (2002) for a thorough review of the state of the art in multimodal dialogue systems in general, and the systems developed at KTH in particular. Below is a brief summary, focusing on the role of the agent in the various systems.

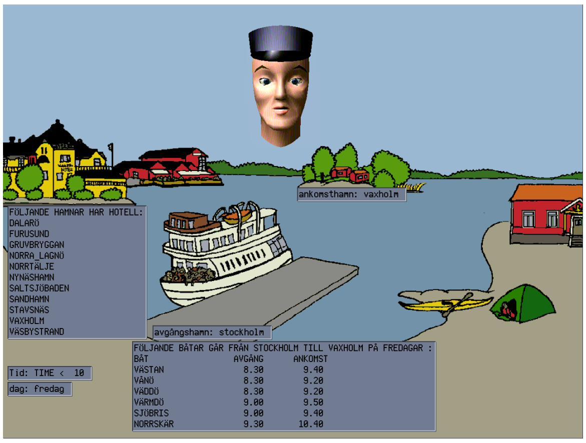
8.1.1 Waxholm

The goal of the Waxholm project (Bertenstam *et al.*, 1995) was to build a mixed-initiative, speaker-independent dialogue system for tourist information in the Stockholm archipelago. Users were able to ask questions about timetables for boats, information about restaurants and accommodation and location of islands and harbours. In addition to speech, the system featured graphical output in the form of tables, charts and a map. The graphical interface is shown in figure 10.

The first generation of the visual speech synthesis system (paper 1) was integrated into this system. Other than providing visible lip movements to accompany the synthesised voice output, the head was capable of deictic movements: when information (e.g. a timetable) was presented somewhere in the graphical interface, the face would look and turn towards that location on the screen, thereby guiding the user's attention.

8.1.2 Olga

The Olga project was initiated in 1995 as a collaboration between several Stockholm based research partners: KTH CTT, KTH CID (Centre for user oriented IT design), the Linguistics department at Stockholm University and Swedish Institute for Computer Science (SICS). The initial goal of the project was ambitious: a multimodal spoken dialogue system, featuring a direct manipulation interface and a full-bodied agent capable of natural verbal as well as non-verbal communication, acting in a consumer information domain to provide guidance regarding microwave ovens. The system was based on a modular design in which the basic modules IM (input manager), DM (dialogue manager,) DMI (direct manipulation interface) and AA (animated agent) communicated over sockets via a hub



```
ankomsthamn: vaxholm

FÖLJANDE HAMNAR HAR HOTELL:
DALARÖ
FURUSUND
GRUVBRYGGAN
NORRA_LAGNÖ
NORRTÄLJE
NYNÄSHAMN
SALTSJÖBADEN
SANDHAMN
STAVSNÄS
VAXHOLM
VÄSBYSTRAND

avgångshamn: stockholm

FÖLJANDE BÅTAR GÅR FRÅN STOCKHOLM TILL VAXHOLM PÅ FREDAGAR :
BÅT                 AVGÅNG     ANKOMST
VÄSTAN                8.30        9.40
VÅNÖ                  8.30        9.20
VÄDDÖ                 8.30        9.20
VÄRMDÖ                9.00        9.50
SJÖBRIS               9.00        9.40
NORRSKÄR              9.30       10.40

Tid: TIME <  10
dag: fredag
```

Figure 10. The graphical interface of the Waxholm system.

through a simple ASCII-based protocol. This strict modularity was a key feature since the DM, DMI and AA were developed at different sites, in different programming languages and ran under different operating systems.

Due to lack of funding, the system only made it to the initial prototype stage, which implied a very limited IM only capable of dealing with a handful of phrases. The DM and AA modules were however fully functional (Beskow and McGlashan, 1997) albeit not thoroughly evaluated. The AA module featured audiovisual speech synthesis (with a female version of the KTH formant synthesis) and the embodied Olga agent (see figure 11). The agent had a small repertoire of facial expressions and gestures (shrugs, nods, deictic gestures, idle gestures etc.) that were triggered from the DM level through the **`state`** directive. For example, during presentation of data in the DMI, the message **`state(attention_display)`** would cause the agent to turn towards the DMI (on a different screen) and make a gesture with the arm. If the DM failed to satisfy a user request, the message **`state(regret)`** resulted in the agent performing a regretful shrug (hands turned up, eyebrows raised, mouth-corners turned down and head slowly shaking sideways).

Despite the fact that Olga never was turned into a fully functional dialogue system, the

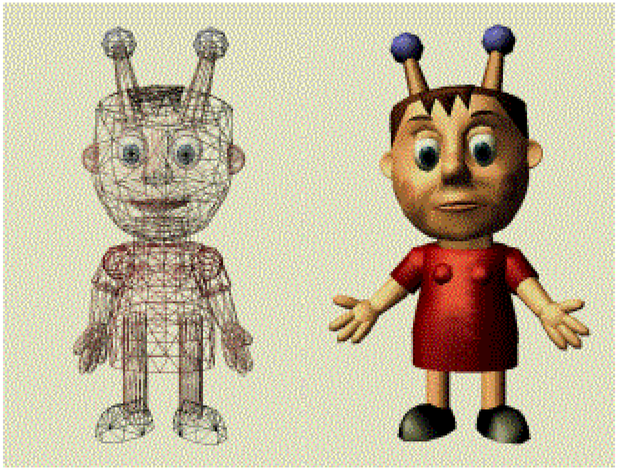


Figure 11. Olga - a cartoon-like full-bodied animated agent performing a regretful gesture (right).

project was still very influential on the talking head technology development represented by this thesis. The project demands of a full-bodied agent with a cartoon-like appearance and articulated body required a new architecture for talking head modelling, where arbitrary polygon geometries could be parameterised and articulated, forcing abandonment of the hard-wired Parke-approach previously used (where the face model definition is intertwined into the animation engine code). This fuelled the development of the GSDS described in paper 2.

8.1.3 August

The August system (Gustafson *et al.*, 1999) developed in 1998, represented new challenges in terms of system robustness: whereas previous systems were developed and tested in well-controlled lab settings, August was a publicly available, unsupervised system where users could freely walk up to the system (located in the centre of culture in downtown Stockholm) and engage in conversation without prior instruction or guidance regarding the systems capabilities apart from those provided by the system itself. The system featured a talking head modelled after the famous Swedish 19th century author August Strindberg (see figure 12). The system had knowledge in multiple shallow domains, each handled by its own dialogue manager: restaurants, KTH, the life of August Strindberg and general social conversation (greetings etc.). In order to be engaging and entertaining to the general public, August was given a rich non-verbal repertoire, including some supernatural behaviours such as rotating the head 360 degrees or twisting the moustache. For more information on the creation of the August model and the associated gestures the reader is referred to Lundeberg and Beskow (1999).

8.1.4 AdApt

The most recent in the line of dialogue systems incorporating animated agents at CTT is *AdApt* (Gustafson *et al.*, 2000). The system is a real-estate browsing system for downtown Stockholm, featuring multimodal input (speech and an interactive map) and output (audiovisual speech, maps and tables). A screenshot from the AdApt system, showing the virtual real-estate broker *Urban* can be seen in figure 13.

In AdApt, a unified and structured way of specifying and generating spoken as well as non-verbal output for the animated agent was needed. This requirement led to the development of the XML-based GESOM formalism and the gesture library approach described in paper 5 (and summarised in chapter 6).

8.2 Synthetic faces as aids in communication

The second category of applications that has been pivotal in driving this research is communication aids for hearing-impaired people. Two projects have framed this research over the years: the national *Teleface* project, lasting from 1995 to 2001, and the EU/IST project *Synface* scheduled from 2001 to 2004.

The fundamental idea behind both projects is the ambitious (but realistic) goal of a communication device that, in a speaker independent fashion, translates telephone-quality speech signals into visible articulatory movements in a synthetic talking head with sufficient accuracy to provide significant speechreading support to the hearing-impaired user, improving his/her ability to communicate over the telephone.

Synface can essentially be viewed as a continuation of the efforts initiated in the Teleface project, expanding the scope into multiple languages. The Synface prototypes will be developed for Swedish, English and Dutch.

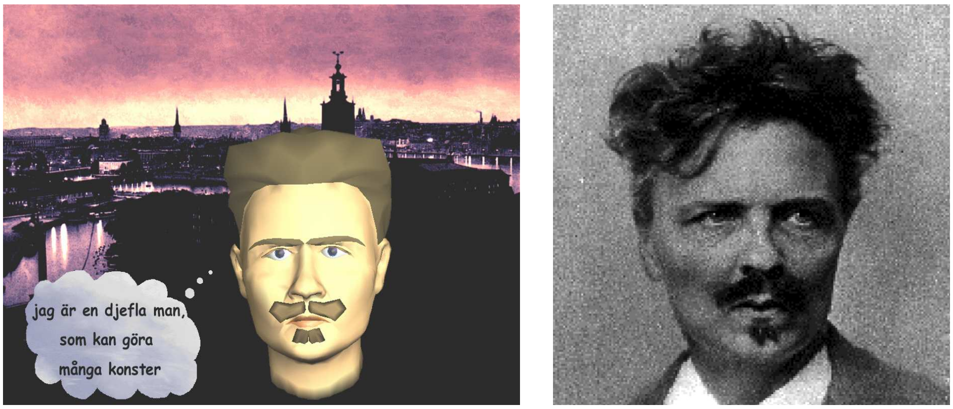


Figure 12. August the talking head and the 19th century Swedish author August Strindberg.

In the Teleface project, it was realised that before attempting actual construction of the device, the effectiveness of the talking head for the task had to be assessed separately. Thus a series of intelligibility evaluation studies were conducted, where the talking head was controlled by ideal conditions, i.e. manually corrected phonetic alignments. The same type of experiments were also run in the Synface project in a uniform way for all three languages (paper 6, see section 7.2 for further details).

The actual construction of the communication device represents a significant technical challenge. In chapter 5 the state of the art in the field was reviewed, and paper 4 describes one possible approach. However, not all approaches are suited for the demands of the Synface application. In particular, there are two factors making the Synface scenario difficult: 1) the device has to be speaker independent, and 2) it has to work in (close to) real time, with no more than about 100 milliseconds delay.

In fact, with the exception of Öhman and Salvi (1999) and Morishima (1998), all of the systems reviewed in chapter 5 are speaker dependent, and in the case of Morishimas system, new users had to undergo an audiovisual enrolment procedure, which would be difficult to implement in the proposed device. It seems that symbol-based AV-mappings have an advantage over the regression based, in that the acoustics-to-symbols stage (e.g. phoneme

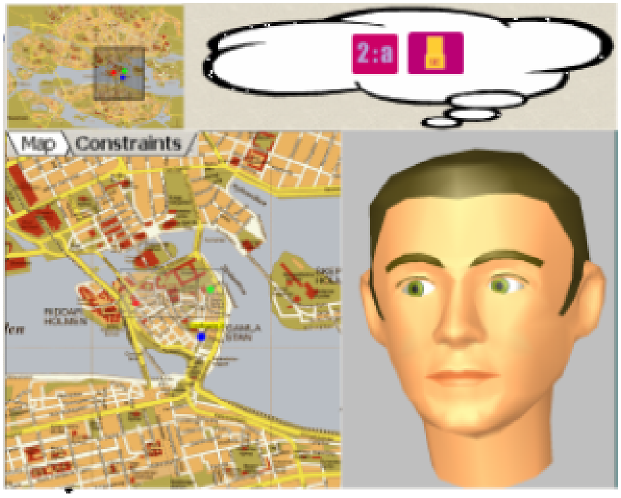


Figure 13. The graphical interface of the AdApt system includes an interactive scrollable map, icons for visualisation of system beliefs, and the animated agent Urban.

recognition) can be trained on very large quantities of audio-only material, which is required if the system is to generalise well across talkers.

The real-time criterion is usually not considered in the symbol-based approaches; typical HMM-based recognition systems are unable to output phonemes incrementally, since the standard Viterbi algorithm uses a backtracking stage over the whole utterance before any output is produced. Furthermore, in the symbol-based methods, an additional delay is introduced in mapping the recognised symbols into visual parameter trajectories. In the general case, many articulatory control models require information about the full utterance before any output can be produced. To deal with this problem, paper 8 presents a low-latency ANN-based articulatory control model that only introduces 33 milliseconds delay.

The regression-based approaches have an advantage with respect to the real-time criterion. Since they use a small, fixed number of frames look-ahead, the introduced delay will be small and predictable.

In the Synface project, a first working prototype is currently under development. Experiments have been performed with speaker independent phoneme recognisers based on HMM and hybrid ANN/HMM (Salvi, 2003; Seward, 2003). To allow for real-time performance, a new low-latency recognition engine has been developed especially for this purpose.

8.3 Language training with talking heads

The work described in paper 3 was carried out as part of a project aiming at utilising language and speech technology components including talking heads in speech training for profoundly deaf children (Cole *et al.*, 1998). The project was a collaboration between several US research institutes: University of California Santa Cruz (UCSC), Oregon Graduate

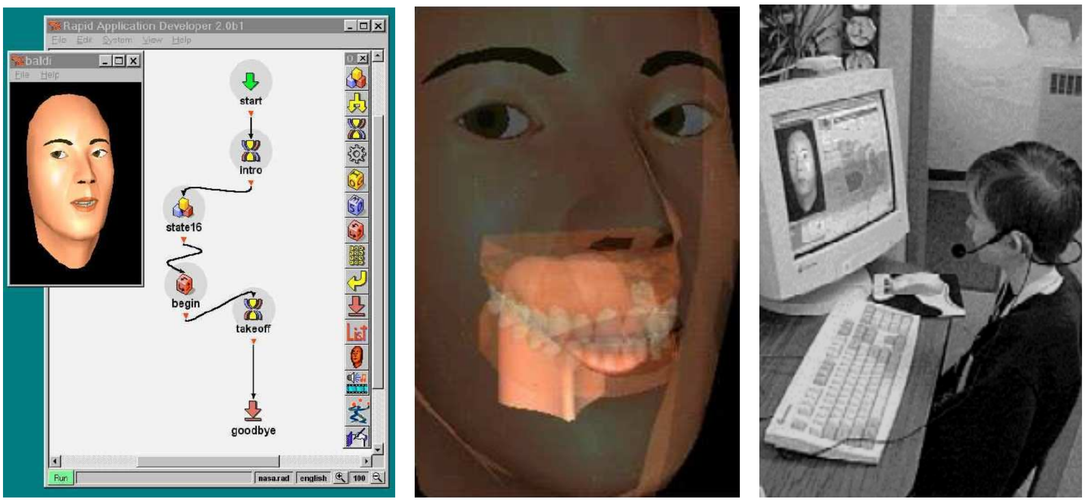


Figure 14. Speech therapy using talking heads. CSLU Rapid Application Development tool (left) that includes the animated talking head Baldi with a tongue model trained on ultrasound data (middle). To the right a student at the Tucker-Maxon Oral School interacting with a speech training application.

Institute (OGI) as well as Carnegie Mellon University (CMU). Associated with the project was also the Tucker-Maxon Oral School (TMOS) in Portland, Oregon. The students of TMOS are profoundly deaf, many use cochlear implants to enhance their hearing. Rather than aiming at building a fixed set of speech training applications, the project focused on integrating a number of relevant technologies into an interactive, easy to use environment, making it possible for teachers, parents and other interested parties to construct applications involving multimodal speech technology. This environment was based around the CSLU Speech Toolkit[12], a rapid-prototyping tool for construction of simple spoken dialogue systems. The improved tongue model and tongue-palate contact detection developed in paper 3 was integrated into the toolkit, and could easily be integrated into applications developed by the teachers at TMOS. Using simple GUI controls, users could select different ways of viewing the face and tongue (different angles, semi-transparent skin, half-face cut-away etc.) as well as highlighting of tongue-palate contact.

[12] **http://cslu.cse.ogi.edu/toolkit**

9 SUMMARY OF PAPERS

Below, each of the included papers is briefly summarised. The original scientific contribution is explicitly stated for each paper (original contribution), and when there is more than one author, the contribution of the author of this thesis is described (author contribution).

9.1 Paper 1

Rule-based Visual Speech Synthesis
Jonas Beskow
Eurospeech'95

The paper covers the initial implementation of a complete rule-based system for text-to-audiovisual-speech (TTAVS) synthesis. The components of the system have since then been continuously refined. The rule-based articulatory control model was employed in many of the studies reported on in this thesis.

An existing three-dimensional parameterised face model (Parke, 1982) was extended with a tongue as well as a novel set of articulatorily motivated control parameters, including dedicated parameters for lip rounding, bilabial occlusion and labiodental occlusion. The face model was integrated into the KTH TTS system. A rule-based articulatory control model was developed in order to drive the articulation of the model from phonetic information in the TTS system. 45 Swedish phonemes were clustered into 21 visemes, for which proper parameter settings were estimated based on introspection and phonetic knowledge. The control model accounts for forward- and backward coarticulation in a simple yet effective way through under-specification of the viseme targets, where unspecified targets are determined using linear interpolation between nearest neighbouring segments where the target is specified. Each resulting articulatory parameter trajectory is independently filtered to obtain suitable dynamical properties. The paper also describes the incorporation of the audiovisual text-to-speech system into a spoken human-machine dialogue system developed at the department.

Original contribution

The introduction of articulatory motivated control parameters and the rule-based articulatory control model where coarticulation is modelled by target under-specification. In addition the paper describes the first audiovisual speech synthesis system for Swedish.

9.2 Paper 2

Animation of Talking Agents
Jonas Beskow
AVSP'97

This paper describes an approach to facial animation based on parametric deformation of polygon meshes. The proposed deformation scheme (GSDS) is a generalisation of Parke's (1982) approach, over which it can be considered to have two main advantages: 1) parameterisation data is separated from the animation engine code, which greatly increases flexibility and allows for rapid parameterisation of new geometries, and 2) it offers a way to define the spatial target of a deformation, to ensure that e.g. an articulator reaches a certain position, regardless of other deformations affecting the same region.

A technique based on parameterised templates for controlling the facial expressions and body gestures to supplement the speech output of an animated agent in a spoken dialogue system is also outlined. Finally, a cartoon-like character implemented using the presented deformation scheme was evaluated in a visual speech intelligibility test, where it was shown to provide a substantial increase in speech intelligibility when compared to an audio-alone condition.

Original contribution

The generalised surface deformation scheme (GSDS), as well as the evaluation results for the cartoon-like Olga character.

9.3 Paper 3

Recent Developments in Facial Animation: an Inside View
Michael M. Cohen, Jonas Beskow and Dominic W. Massaro
AVSP'98

This paper reports on developments within the *Baldi* facial animation system, developed at UCSC. The face is augmented with a highly realistic palate, teeth, and an improved tongue model. The aim of these additions is twofold: to improve realism and accuracy of visible speech production, and to provide an anatomically valid and pedagogically useful display that can be used in speech training of children with hearing loss. The general approach taken is to use both static and dynamic observations of natural speech to guide the facial modelling. High-resolution models of palate and teeth were reduced to a relatively small number of polygons for real-time animation. For the tongue model, 3D ultrasound data was used in combination with error minimisation algorithms to train the parametric b-spline based tongue model to simulate realistic speech. In addition, a high-speed algorithm was developed for detection and correction of collisions, to prevent the tongue from protruding through the palate and teeth, and to enable the real-time display of synthetic electropalatography (EPG) patterns.

Author contribution

The author of this thesis collaborated with MMC on the development of the algorithms used in collision detection/correction, synthetic EPG display and error metrics for three-dimensional tongue shape training. The b-spline based tongue model was developed by MMC.

Original contribution

The use of ultrasound data to tune static articulatory targets of a parameterised tongue model, as well as the algorithms for collision detection/correction and display of synthetic EPG patterns.

9.4 Paper 4

Picture my Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks
Dominic W. Massaro, Jonas Beskow, Michael M. Cohen, Christopher L. Fry and Tony Rodriguez
AVSP'99

This paper presents an initial implementation and evaluation of a system that synthesises visual speech directly from the acoustic waveform. An artificial neural network (ANN) was trained to map the cepstral coefficients of an individual's natural speech to the control parameters of an animated synthetic talking head. Training data for the ANNs in terms of visual parameter trajectories were generated from phonetically labelled audio data. ANNs were trained on two data sets; one was a set of 400 words spoken in isolation by a single speaker and the other a subset of extemporaneous speech from ten different speakers, for which ten speaker-specific ANNs were trained. The ANN trained on words was able to generalise to new speech from the same speaker. A perceptual evaluation test indicated that the produced animations provided significant visible information, but significantly below that given by a text-to-speech algorithm. The ANNs trained on extemporaneous speech were poor at generalising to new speech from the same speaker, which is probably due to lack of training data.

Author contribution

The author of this thesis was responsible for preparing test and training data and for the objective and perceptual evaluation of the results. ANN training was done collaboratively by CLF and the author of this thesis. TR assisted in perceptual testing. The manuscript was prepared by DWM in collaboration with MMC and the author of this thesis.

Original contribution

The use of ANNs trained on parameter trajectories generated from phonetic transcriptions, rather than direct visual measurements.

9.5 Paper 5

A Model for Generalised Multi-modal Dialogue System Output Applied to an Animated Talking Head
Jonas Beskow, Jens Edlund and Magnus Nordstrand
Spoken multimodal human-computer dialogue in mobile envirnonments

This paper deals with the problem of encoding non-verbal output, such as facial gestures in an animated talking head, from a multimodal dialogue system. A high-level XML-based output specification provides information about communicative functions of the output, without detailing the realisation of these functions. The aim is to let dialogue systems generate the same output for a wide variety of output devices and modalities. Examples are given from an implementation in the multimodal spoken dialogue system AdApt, and it is described how facial gestures for signalling of turn-taking, back-channelling and emphasis are implemented in the 3D-animated talking head module used within this system. A gesture library holds a set of multiple alternative realisations of each communicative function to be signalled, and a system for selecting between realisations based on background information is outlined as well as a facial gesture coarticulation scheme.

Author contribution

The author of this thesis was responsible for the initial idea, design and implementation of the XML-based output specification and the gesture library. JE improved and generalised the output specification to cater for other output devices than the animated talking head and was responsible for integration into the AdApt system. MN developed the facial gestures in the gesture library. The manuscript was prepared in collaboration by JE and the author of this thesis.

Original contribution

The XML-based output specification, and the use of a gesture library for realisation of facial gestures.

9.6 Paper 6

Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired
Catherine Siciliano, Geoff Williams, Jonas Beskow and Andrew Faulkner
ICPhS'03

This paper represents the latest results in a series of intelligibility evaluations performed within the Teleface and Synface projects, two projects aimed at developing a synthetic talking face to aid the hearing impaired in telephone conversation. This report investigates the gain in intelligibility from the synthetic talking head when controlled by phonetic annotations of acoustic speech. Audio from Swedish, English and Dutch sentences was degraded to simulate the information losses typical for severe-to-profound hearing impairment. 12 normal-hearing native speakers for each language took part. Auditory signals were presented alone, with the synthetic face, and with a video of the original talker. Purely auditory intelligibility was low. With the addition of the synthetic face, average intelligibility increased by 20%. Scores with the synthetic face were significantly lower than for the natural face for English and Dutch, but not for Swedish.

Author contribution

The author of this thesis was responsible for implementing the testing environment used in all experiments, as well as designing the Swedish sentence intelligibility experiment.

Original contribution

The multilingual intelligibility study, where identical experiments are performed for three languages.

9.7 Paper 7

Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements
Jonas Beskow , Olov Engwall and Björn Granström
ICPhS'03

Simultaneous measurements of tongue and facial motion, using a combination of electromagnetic articulography (EMA) and optical motion tracking, are analysed to improve the articulation of an animated talking head and to investigate the correlation between facial and vocal tract movement. The recorded material consists of VCV and CVC words and 270 short everyday sentences spoken by one Swedish subject. The recorded articulatory movements are re-synthesised by a parametrically controlled 3D model of the face and tongue, using a procedure involving minimisation of the error between measurement and model. In order to obtain an estimate of the inner lip contour, that is crucial to visual speech perception but difficult to measure using the employed techniques, linear estimators are used to predict the inner lip points from the outer. Linear estimators are also employed for prediction of tongue data from the face and vice versa, and the correlation between the measurement and prediction is computed.

Author contribution

The author of this thesis was responsible for data analysis and resynthesis of the talking head and tongue. Linear estimation and correlation analysis, as well as manuscript preparation was performed by OE and the author of this thesis in collaboration. All three authors participated in the data collection.

Original contribution

Combined resynthesis of tongue and facial motion in a talking head from simultaneous measurements, and a novel resynthesis procedure, where a model can be driven by measurements from a face with a different overall shape than the model. The idea of predicting the inner lip contour from the outer is also new. Furthermore, the recorded database contains a more diverse and extensive corpus than has been reported for other similar studies, and represents the first reported simultaneous optical and intra-oral (EMA) articulatory measurements for Swedish.

9.8 Paper 8

Trainable Articulatory Control Models for Visual Speech Synthesis
Jonas Beskow
Submitted to International Journal of Speech Technology

This paper deals with the problem of modelling the dynamics of articulation for a parameterised talking head based on phonetic input. Four different models are implemented and trained to reproduce the articulatory patterns of a real speaker, based on a corpus of optical measurements. Two of the models are based on coarticulation models from speech production theory and two are based on artificial neural networks, one of which is specially intended for streaming real-time applications. The different models are evaluated through comparison between predicted and measured trajectories, as well as through a perceptual intelligibility experiment. Results show that all models give significantly increased speech intelligibility over the audio-alone case, but none of the models can be said to outperform the others.

Original contribution

The use of ANNs to drive the articulation of a talking head from phonetic input, and the objective and subjective evaluation of several articulatory control models trained on a common database.

LIST OF PUBLICATIONS

The following is a list of articles that have been published or submitted as part of this thesis work, ordered according to publication date. Articles marked with ● are included in this thesis, those marked with · are not.

● Beskow, J. (submitted4). Trainable Articulatory Control Models for Visual Speech Synthesis, submitted to *International Journal of Speech Technology*.

● Beskow, J., Engwall, O. and Granström, B. (submitted3). Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. To appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

● Siciliano, C., Williams, G., Beskow, J. and Faulkner, A. (submitted2). Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired, to appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

● Beskow, J., Edlund, J., and Nordstrand, M. (submitted1). A Model for Generalised Multi-Modal Dialogue System Output Applied to an Animated Talking Head. To appear in Minker, W., Bühler, D. and Dybkjær, L. (Eds) *Spoken Multimodal Human-Computer Dialogue in Mobile Envirnonments*. Dordrech, The Netherlands, Kluwer Academic Publishers.

· Siciliano, C., Williams, G., Beskow, J., Faulkner, A. (2002). Evaluation of a synthetic talking face as a communication aid for the hearing impaired. *Speech, Hearing and Language: Work in Progress*, 14, 2002, pp. 51-61.

· Beskow, J., Edlund., J. and Nordstrand, M. (2002). Specification and Realisation of multimodal output in dialogue systems, in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, CO.

· Beskow, J., Granström, B. and Spens, K-E. (2002). Articulation strength - readability experiments with a synthetic talking face, in *Proceedings of Fonetik 2002*, Stockholm, Sweden.

· Granström, B., House, D. and Beskow, J. (2002). Speech and gestures for talking faces in conversational dialogue systems. In Granström, B., House, D., and Karlsson, I. (eds.) *Multimodality in language and speech systems*. Kluwer Academic Publishers. Dordrecht, The Netherlands. pp 209-241.

· House, D., Beskow, J. and Granström, B. (2001). Timing and Interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, Aalborg, Denmark.

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

· 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. (eds.) *Nordic Prosody VIII*, Trondheim, Norway. pp. 77- 87.

- Sjölander, K. and Beskow, J. (2000). WaveSurfer - an Open Source Speech Tool. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China.

- Gustafson, J. Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000). AdApt–a multimodal conversational dialogue system in an apartment domain. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 134-137.

- Massaro, D.W., Cohen, M. M., Beskow, J., Cole, R.A. (2000). Developing and Evaluating Conversational Agents, in Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.) *Embodied Conversational Agents.* Cambridge, MA: MIT Press.

- Lundeberg, M. and Beskow, J. (1999). Developing a 3D-agent for the August dialogue system. In D. W. Massaro (Ed.) *Proceedings of AVSP'99*, Santa Cruz, CA. pp. 151-156.

- 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. In D. W. Massaro (Ed.) *Proceedings of AVSP'99*, Santa Cruz, CA. pp. 123-127.

● Massaro, D.W., Beskow, J., Cohen, M.M., Fry C.L., Rodriquez, T. (1999). Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. In D. W. Massaro (Ed.) *Proceedings of AVSP'99*, Santa Cruz, CA. pp. 133-138.

- Sjölander, K., Beskow, J., Gustafson, J., Levin, E., Carlson, R., Granström, B. (1998). Web-based educational tools for speech technology, *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia.

- Aglefors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1998). Synthetic faces as a lipreading support, *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, pp. 3047-3050.

● Cohen, M. M., Beskow, J. Massaro, D. W. (1998). Recent developments in facial animation: an inside view. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 201-206.

- Beskow, J. (1998). A Tool for Teaching and Development of Parametric Speech Synthesis, *Proceedings of Fonetik '98*, Stockholm, Sweden.

● Beskow, J. (1997). Animation of Talking Agents, In Benoît, C. and Campbell, R. (Eds.) *Proceedings of the ESCA/ESCOP Workshop on Audio-Visual Speech Processing (AVSP'97)*, Rhodes, Greece, pp. 149-152.

- Beskow, J. and McGlashan, S. (1997). Olga - a conversational agent with gestures. In André, E. (Ed.) *Proceedings of the IJCAI'97 Workshop on Animated Interface Agents - Making Them Intelligent*, Nagoya, Japan., pp 39-44.

- Beskow, J., Elenius, K. & McGlashan, S. (1997). Olga - A dialogue system with an animated talking agent. *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Rhodos, Greece.

- Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1997). The Teleface project - Multimodal Speech Communication for the Hearing Impaired, *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Rhodos, Greece.

- Beskow, J., Elenius, K. & McGlashan, S. (1997). Olga - A dialogue system with an animated talking agent. *Proceedings of Fonetik 97*, Umeå, Sweden.
- Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E. & Öhman, T. (1997). The Teleface project - disability, feasibility and intelligibility. *Proceedings of Fonetik'97*, Umeå, Sweden.
- Beskow, J. (1996). Talking Heads - communication, articulation and animation. In *TMH-QPSR 2/1996, Proceedings of Fonetik '96,* Nässlingen, Sweden.
- ● Beskow, J. (1995). Rule-based Visual Speech Synthesis. *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH'95)*, Madrid, Spain.
- Beskow, J. (1995). *Regelstyrd Visuell Talsyntes.* Master of science thesis, Department of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden.

REFERENCES

Aglefors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1998). Synthetic faces as a lipreading support, *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, pp. 3047-3050.

Ahlberg, J. (2002). Extracting MPEG-4 FAPS from Video. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, , Chichester, England: John Wiley & Sons. pp. 103-112

Arslan, L.M. & Talkin, D. (1998) 3-D Face Point Trajectory Synthesis using an Automatically Derived Visual Phoneme Similarity Matrix. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 175-180.

Bailly, G. (1998) Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22(2-3): 251-267.

Basu, S., Oliver, N. and Pentland, A. (1998). 3D lip shapes from video: A combined physical-statistical model. *Speech Communication*, 26(1-2): 131-148.

Beier, T. and Neely, S. (1992). Feature-based image metamorphosis. *Computer Graphics*, 26(2):35-42.

Benoît, C., Guiard-Marigny, T., Le Goff, B., and Adjoudani, A. (1996). Which components of the Face Do Humans and Machines Best Speechread? In D. G. Stork and M. E. Hennecke (Eds.) *Speechreading by Humans and Machines: Models, Systems and Applications.* Berlin: Springer. pp. 315-330.

Bernstein, L.E. & Eberhardt, S.P. (1986). *Johns Hopkins lipreading corpus videodisk set.* Baltimore, MD: The Johns Hopkins University.

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. and Ström, N. (1995). The Waxholm system - a progress report, *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, Vigsø, Denmark..

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T (1997). The Teleface project: Multimodal speech communication for the hearing impaired, *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Rhodos, Greece.

Beskow, J. and McGlashan, S. (1997). Olga - a conversational agent with gestures. In André, E. (Ed.) *Proceedings of the IJCAI'97 Workshop on Animated Interface Agents - Making Them Intelligent*, Nagoya, Japan., pp 39-44.

Beskow, J., Granström, B. and Spens, K-E. (2002). Articulation strength - readability experiments with a synthetic talking face, in *Proceedings of Fonetik 2002*, Stockholm, Sweden.

Brand, M. (1999). Voice puppetry. *Proceedings of SIGGRAPH 99*, Los Angeles, CA, pp. 21-28.

Branderud, P. (1985). Movetrack - a movement tracking system, *Proceedings of the French-Swedish Symposium on Speech*, Grenoble, France, pp. 113-122.

Breen, A. P., Bowers, E., and Welsh, W. (1996). An Investigation into the Generation of Mouth Shapes for a Talking Head, *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, PA.

Bregler, C., Covell, M, Slaney, M. (1997). Video Reqrite: Driving Visual Speech with Audio. *Proceedings of SIGGRAPH'97*, Los Angeles, CA, pp. 353-360.

Brooke, M. N. (1996). Talking Heads and Speech Recognisers That Can See: The Computer Processing of Visual Speech Signals. In D. G. Stork and M. E. Hennecke (Eds.) *Speechreading by Humans and Machines: Models, Systems and Applications*. pp. 351-372. Berlin: Springer.

Brooke, N. M. and Scott, S. D. (1998). Two- and three-dimensional audio-visual speech synthesis. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 213-218.

Burnett, D. C., Walker, M. R., and Hunt, A. (2002). Speech Synthesis Markup Language Specification. W3CWorking Draft, updated 2002-12-02, accessed 2003-05-07. **http://www.w3.org/TR/2002/WD-speech-synthesis-20021202/**

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. (1994). Animated Converstaion: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. *Proceedings of SIGGRAPH 94*, Orlando, FL, pp. 413-420.

Cassell *et al.*, J. (2000). Human Conversation as a System Framework: Designing Embodied Conversational Agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, Cambridge, MA: MIT Press. pp 29–63.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, PA.

Chandler, J. P. (1969). Subroutine STEPIT - Finds local minima of a smooth function of several parameters. *Behavioral Science,* 14: 81-82.

Cohen, M. M. and Massaro, D. W. (1993). Modelling Coarticulation in Synthetic Visual Speech. Magnenat-Thalmann N., Thalmann D. (Eds), *Models and Techniques in Computer Animation*, Springer Verlag, Tokyo, pp. 139-156.

Cohen, M. M., Walker, R. L, and Massaro, D. W. (1996). Perception of Synthetic Visual Speech. In D. G. Stork and M. E. Hennecke (Eds.) *Speechreading by Humans and Machines: Models, Systems and Applications*. Berlin: Springer. pp. 153-168.

Cohen, M. M., Massaro, D. W. and Clark, R. (2002). Training a talking head. *Proceedings of 4th IEEE International Conference on Multimodal Interfaces (ICMI'02)*, Pittsburgh, PA.

Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J. de Villiers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow. J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., and Soland, C. (1998). Animated agents for interactive language training. *Speech Technology in Language Learning ESCA workshop*. Marholmen, Sweden.

Cosatto, E. and Graf, H. P. (2000). Photo-realistic talking-heads from image samples, *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 152-163.

Cosatto, E. (2002). *Sample-Based Talking-Head Synthesis*. Doctoral Dissertation. Signal Processing Lab, Swiss Federal Institute of Techology, Lausanne, Switzerland.

Cosi P., Magno Caldognetto E., Perin G. and Zmarich C. (2002) Labial Coarticulation Modeling for Realistic Facial Animation, *Proceedings of 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*, Pittsburgh, PA, pp. 505-510.

Ekman, P. and Friesen, W. (1975). *Unmasking the face: A guide to recognising emotion from facial clues*. Prentice-Hall.

Ekman, P. and Friesen, W. (1978). *Manual for the Facial Action Coding System*, Palo Alto, CA: Consulting Psychologists Press.

Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001). Creating and controlling video-realistic talking heads. In Massaro, D. W., Light, J., and Geraci, K. (Eds.) *Proceedings of the Workshop on Audiovisual Speech Processing (AVSP 2001).* Scheelsminde, Denmark, pp. 90-97.

Engwall (2002). *Tongue Talking - Studies in Intraoral Speech Synthesis.* Doctoral Dissertation, KTH, Stockholm, Sweden.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of speech stimuli. *Journal of Speech and Hearing Research*, 12:423-425.

Ezzat, T. and Poggio, T. (1999). Visual Speech Synthesis by Morphing Visemes. MIT AI Memo No 1658/CBCL Memo No 173.

Ezzat, T, Geiger, G., Poggio, T. (2002). Trainable Videorealistic Speech Animation. *Proceedings of ACM SIGGRAPH 2002*, San Antonio, TX, pp. 388-398.

Galanes, F. M., Unverferth, J., Arslan, L., and Talkin, D. (1998). Generation of lip-synched faces from phonetically clustered face movement data. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 191-194.

Geiger, G., Ezzat, T. and Poggio, T. (2003). Perceptual Evaluation of Video-.Realistic Speech. *AI Memo 2003-003*, MIT AI Lab.

Girosi, F., Jones, M. and Poggio, T. (1993). Priors, stabilizers, and basis functions: From regularization to radial, tensor, and additive splines. Tech. Rep. 1430, MIT AI Lab, June.

Graf, H.P., Cosatto E., Strom, V., Huang, F.J. (2002). Visual Prosody; Facial Movements Accompanying Speech. *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR 02),* Washington, DC.

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. (eds.) *Nordic Prosody VIII*, Trondheim, Norway. pp. 77- 87.

Gray, H. (1977). *Gray's anatomy.* New York, NY: Random House.

Gustafson, J. Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000). AdApt–a multimodal conversational dialogue system in an apartment domain. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 134-137.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems: Empirical Studies of Spoken Human-Computer Interaction.* Doctoral Dissertation, KTH, Stokholm, Sweden.

Gustavsson, C., Strindlund, L., Wiknetrz, E., Beard, S., Huynh, Q., Marriot, A., and Stallo, J. (2001). *VHML. Working Draft v0.3*, updated 2001-10-21, accessed 2003-05-07.
`http://www.vhml.org/documents/VHML/2001/WD-VHML-20011021/`

Hong, P., Wen, Z., Huang, T. S. (2002). Real-Time Speech-Driven Face Animation. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 115-124.

House, D., Beskow, J. and Granström, B. (2001). Timing and Interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, Aalborg, Denmark.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database, *Proceedings of ICASSP'96*, pp. 373-376.

Hällgren, Å. and Lyberg, B. (1998). Visual Speech Synthesis with Concatenative Speech. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 181-183.

Jiang, J., Alwan, A., Bernstein, L., Keating, P., and Auer, E., On the correlation between facial movements, tongue movements and speech acoustics, *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 42-45.

Kuratate, T., Yehia, H., Vatikiotis-Bateson, E. (1998) Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 185-190.

Lavagetto, F. and Lavegetto, P. (1996). Time Delay Neural Networks for Articulatory Estimation from Speech: Suitable Subjective Evaluation Protocols. In D. G. Stork and M. E. Hennecke (Eds.) *Speechreading by Humans and Machines: Models, Systems and Applications.* Berlin: Springer. pp. 437-444.

Lee, Y., Terzopolous, D., and Waters, K. (1995). *Proceedings of SIGGRAPH 95*, Los Angeles, CA, pp. 55-62.

Le Goff, B. (1997). Automatic Modeling of Coarticulation in Text-to-Visual Speech Synthesis. *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Rhodos, Greece, pp. 1667–1670.

Lucero, J. C., and Munhall, K. G. (1999). A model of facial biomechanics for speech production. *J. Acoust. Soc. Am.* 106: 2834–2842.

Lundeberg, M. and Beskow, J. (1999). Developing a 3D agent for the August dialogue system. In D. W. Massaro (Ed.) *Proceedings of AVSP'99*, Santa Cruz, CA. pp. 151-156.

Löfqvist, A. (1990). Speech as audible gestures. In Hardcastle, W. J. and Marchal, A. (Eds.) *Speech Production and Speech Modelling.* Dordrecht: Kluwer Academic Publishers. pp. 289-322.

McAllister, D.F., Rodman, R. D., Bitzer, D. L. and Freeman, A. S. (1997). Lip synchronisation of speech. In Benoît, C. and Campbell, R. (Eds.) *Proceedings of the ESCA/ESCOP Workshop on Audio-Visual Speech Processing (AVSP'97)*, Rhodes, Greece, pp. 133-136.

McGurk, H. & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature*, 264: 746-748.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology,* 24: 29-43.

Morishima, S. (1998). Real-time talking head driven by voice and its application to communication and entertainment. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 195-200.

Nagao, K. and Takeuchi, A. (1994). Speech dialogue with facial displays: Multimodal human computer conversation. In *Proceedings of the 32nd ACL'94*, pp. 102–109.

Olivès, J.-L., Möttönen, R., Kulju, J. and Sams, M. (1999). Audio-visual speech synthesis for Finnish. In D. W. Massaro (Ed.) *Proceedings of AVSP'99*, Santa Cruz, CA. pp. 157-162.

Ostermann, J. (2002). Face Animation in MPEG-4. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 17-56.

Pandzic, I. S., Ostermann, J., and Millen, D. (1999). User evaluation: Synthetic talking faces for interactive services. *The Visual Computer*, 15:330-340.

Pandzic, I. S. and Forchheimer, R. (2002a). *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons.

Pandzic, I. S. and Forchheimer, R. (2002b). The Origins of the MPEG-4 Facial Animation Standard. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 1-16.

Parke, F. I. (1982). Parametrized models for facial animation, *IEEE Computer Graphics*, 2(9), pp 61-68.

Parke, F. I. and Waters, K. (1996). *Computer Facial Animation*. Wellesley, Massachusetts: A K Peters.

Pelachaud, C. (1991). *Communication and Coarticulation in Facial Animation*. Doctoral Dissertation, University of Pensylvania, PA.

Pelachaud, C. (2002). Visual Text-to-Speech. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 125-140.

Pelachaud, C., van Overveld, C. W. A. M., and Seah, C. (1994). Modelling and Animating the Human Tongue during Speech Production., Proceedings of Computer Animation '94, Geneva, pp. 40-47.

Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating Facial Expressions for Speech, *Cognitive Science, 20* (1): 1-46.

Pelachaud, C., Badler, N. I., and Viaud, M.-L. (1994). Final Report to NSF of the Standards for Facial Animation Workshop. University of Pensylvania, PA.

Petajan, E. and Graf, H. P., (1996). Robust Face Feature Analysis for Automatic Speechreading and Character Animation. In D. G. Stork and M. E. Hennecke (Eds.) *Speechreading by Humans and Machines: Models, Systems and Applications*. Berlin: Springer. pp. 425-436.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin., D. H. (1998). Synthesizing realistic facial expressions from photographs. *Proceedings of SIGGRAPH 98*, Orlando, FL, pp. 75-84.

Poggi, I. and Pelachaud, C. (2000). Performative Facial Expressions in Animated Faces. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), *Embodied Conversational Agents*, MIT Press, Cambridge, MA. pp. 155–188.

Press, W. H., Teukolsky, S. A., Vetterling, W., T. and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing – 2nd ed.* Cambridge University Press.

Reveret, L. and Benoît, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 207-212.

Reveret, L., Bailly, G. and Badin, P. (2000). Mother: a New Generation of Talking Heads Providing a Flexible Articulatory Control for Video-Realistic Speech Animation. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 755-758.

Salvi, G. (2003). Truncation Error and Dynamics in Very Low Latency Phonetic Recognition, *To appear in proceedings of ISCA Workshop on Non-linear speech processing (NOLISP'03)*, Le Croisic, France.

Sams, M., Kulju, J., Möttönen, R. Jussila, V., Olivès, J-L., Zhang, Y., Kaski, K., Majaranta, P., and Rih, K-J. (2000). Towards a High-Quality and Well-Controlled Finnish Audio-Visual Speech Synthesizer. In *Proceedings of 4th World Multiconference on Systemics, Cybernetics and Informatics and 6th International Conference on Information Systems Analysis and Synthesis*, Orlando, FL.

Seward, A. (2003). Low-Latency Incremental Speech Transcription in the Synface Project. Submitted to the *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland.

Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Stone, M. and Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *J. Acoust Soc. Am.* 99, 3728-3737.

Ström, N. (1997). Phoneme Probability Estimation with Dynamic Sparsely Connected Artificial Neural Networks. *The Free Speech Journal*, vol. 1(5).

Sumby, W. H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *J. Acoust Soc. Am.*, vol. 26, pp. 212-215.

Tamura, M., Masuko, T., Kobayashi, T., and Tokuda, K. (1998). Visual speech synthesis based on parameter generation from HMM: speech-driven and text-driven approaches. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 219-224.

Magnenat-Thalmann, N. and Kalra, P. (1992). A Model for Creating and Visualizing Speech and Emotion, in Dale R. etal. (eds) *Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence,* Springer-Verlag, Heidelberg, pp. 1-12.

Theobald, B. J., Bangham, J. A., Matthews, I. A., and Cawley, G. C (2002). Towards video realistic synthetic visual speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, Orlando, FL, pp. 3892--3895.

Thòrisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures and Humanoids. *International Journal of Applied Artificial Intelligence*, 13(4-5):449–486.

Waters, K. (1987). A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH '87)*, 21(4):17-24.

Yamamoto, E., Nakamura, S. and Shikano, K. (1998). Lip movement syntheis from speech based on Hidden Markov Models. *Speech Communication*, 26(1-2): 105-116.

Yehia, H., Rubin, P. and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior, *Speech Communication*, 26(1-2): 23-43.

Öhman, T. (1998). An audio-visual speech database and automatic measurements of visual speech, *TMH-QPSR*, 1-2/1998, pp. 61-76.

Öhman, T. and Salvi, G. (1999). Using HMMs and ANNs for mapping acoustic to visual speech. *TMH-QPSR*, 1-2/1999.

Öhman, S. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41*, 310-320.