Image-based Talking Heads using Radial Basis Functions

James D. Edge and Steve Maddock
University of Sheffield
Regent Court
211 Portobello st
Sheffield S1 4DP
j.edge@dcs.shef.ac.uk

Abstract

In recent years talking heads have received a great deal of interest, both in their application to natural human-computer dialogue, and their benefit to the intelligibility of synthesised speech. A model for the realistic synthesis of visual speech animation is described in this paper. Images representing the key visual speech poses (visemes) are pre-recorded and labelled. Transitions between visemes are created by using an image morphing technique based upon the use of Radial Basis Functions. Timing information from the FreeTTS speech synthesis system is used to plan the appropriate transitions to create realistic speech animation. A model of coarticulation is included in the system to improve the realism of articulatory motion.

1. Introduction

Speech as a part of a natural dialogue is a collaboration of both audio and visual stimuli. Whilst the audio component is clearly the most important, visual cues such as the movement of the lips and the visibility of the tongue allow us to disambiguate what is heard. This is most obviously the case with the hearing-impaired, so called lip-readers, who use visual signals to make up for a loss in the ability to perceive audio. However, it is also reported that the visual component of speech can make as much as a +15dB improvement in signal-to-noise ratio [31], and a corresponding increase in the intelligibility of the speech. This, along with interest in making the human-computer dialogue more natural, have led to a great deal of interest in the synthesis of visual speech.

This paper discusses the implementation of an audio-visual speech synthesizer. The system described is an extension of the FreeTTS speech synthesis engine [13], a Java implementation of the Festival synthesizer [4]. The visual component is created by morphing images representing important speech lip poses (visemes) thus creating visually realistic animation.

2. Background and Previous Work

Most research into audio-visual synthesis focuses upon the modelling of facial movement. Many techniques use three-dimensional models of the face, using deformation algorithms to recreate facial expressions. These models typically fall into two main categories [24]: (i) Physically/Anatomically-based techniques; (ii) Terminal analogue techniques. Physical models attempt to model the structure and function of the muscles and skin in the face, whereas terminal analogue methods model facial expression in isolation from its physical-means of production. Networks of masses connected by springs are typically used in physically modelling facial behaviour [28, 20, 15], with forces applied to the network to simulate the action of muscles. Non-physics models use generic animation techniques such as morph-targets [26] and spatial deformations [16, 9, 17] to create facial expressions. An overview of three-dimensional facial animation techniques can be found in [27]. For the remainder of this section we shall focus upon two-dimensional techniques more related to the work in this paper.

Image metamorphosis (morphing) techniques relate to the transition between two digital images. Whilst cross-dissolving (linear colour interpolation) images produces disturbing double exposure effects (fig. 2f), image morphing methods use geometric transformation in coordination with colour interpolation to create fluid transitions. In order to create the geometric transformation, key features are labelled in the source and destination images. The location of pixels are defined in relation to the placement of key features, and thus displacement of the features will produce a warping of the image.

Several different methods have been described for the warping of images. Beier and Neely [3] describe the field warping algorithm, which uses line pairs in the source and destination images to define a coordinate mapping. Lee *et al* [19] describe a morphing algorithm based upon the use of Multilevel Free-Form Deformations (MFFDs). These relate the deformation of a parallelpiped lattice to a transformation of the underlying image. The MFFD model is controlled by a set of feature points which place positional constraints on the MFFD lattice. Finally, several authors [1, 21, 30] have described the use of scattered data interpolation techniques to geometric transformation of an image. These methods use Radial Basis Functions (RBFs) to define a surface which passes through a few scattered feature points. Displacement of the feature points thus leads to a deformed surface and underlying image. An overview of these and similar image metamorphosis techniques can be found in [32].

Liu *et al* [22] describe an alternative method for animating faces in images. Facial expressions from one photograph can be mapped onto another using Expression Ratio Images (ERIs). Each ERI models the variation in illumination due to a variation in facial expression. In combination with a geometric warp an ERI allows the application of an expression from one individual to another.

The synthesis of visual speech has coevolved with methods to synthesise human facial expression. Simple techniques linearly interpolate between static visual speech postures (visemes) [9], not dealing with the underlying complexities of articulatory control. More recent methods incorporate models of coarticulation [7, 29] which blend the visemes together in a physically plausible way. Whilst many three-dimensional models for simulating visual speech movements incorporate coarticulation effects, there is a greater inherent difficulty in simulating this blending for image-based methods where no natural parameterisation exists.

The Miketalk system [10] is an audio-visual speech synthesis system based upon morphing viseme images. The morphing algorithm used relies upon concatenated optical flow, where optical flow describes the mapping between the source and destination images. Optical flow refers to a family of techniques designed to calculate the movement between subsequent frames in an image sequence (see [2]). Unfortunately, given the large displacements between visemes, optical flow may not accurately calculate the required warp function. For this reason Miketalk uses concatenated optical flow which may require several in-between frames to guide the transition between visemes. This is a disadvantage of the technique, requiring the capture of a large number of visemes and transitional-visemes to allow the image morphing algorithm to work.

Bregler *et al* [5] describe a speech synthesis system based upon the concatenation of small video sequences representing triphones (three phonemes in sequence, e.g. /p-aa-p/). This method relies upon an extremely large dataset of speech units, which must be specially recorded for the purpose. Such databases are difficult to capture making it difficult to change the system to control a different talking head.

Brooke and Scott [6] use Hidden Markov Models (HMM) to model triphone sequences for speech synthesis. The output of the system is image sequences encoded using Principal Components Analysis (PCA) for data compression purposes. The quality of synthesis using HMMs is entirely determined by the quantity of data acquired initially to train the system, and cannot work without a sufficient dataset being available.

3. Our Approach

Our implemented system controls the synthesis of visible articulations during speech. The approach taken, similar to [10], involves morphing between static viseme images. Where our approach is distinct is in the method of producing transitions between phonemes. Radial Basis Functions are used to morph between hand-labelled viseme images. This provides a morphing algorithm which does not rely upon computationally expensive optical flow algorithms, or in-between images to guide the transition between visemes. A statistical model of facial shape is used to provide a more efficient parameterisation for visible speech articulation. This parameterisation enables a model of speech coarticulation to be included in the system, leading to more realistic and natural animations. The stages of our approach are as follows:

- **Data Acquisition** - Forty English visemes were captured, directly corresponding to the phonemes used in the FreeTTS synthesis system (see Table 1).
- **Labelling** - Individual visemes are labelled with 22 feature points (a subset of the MPEG-4 feature point set [18]).
- **Model Construction** - A Principal Components (PC) model of visual articulation is created from the labelled feature points. The major components calculated using this technique are retained allowing a certain degree of data compression. Each PC is subsequently a parameter in the system used in the coarticulation of visemes.
- **Audio – Visual Synthesis** - Using output phone timings from the FreeTTS speech synthesis system, visemes are aligned with the output audio. In-between frames are rendered by interpolating the visemes using an image metamorphosis method based upon a scattered data-interpolation technique.

This paper continues with a discussion of the image morphing algorithm (Section 4), a description of the parameterisation of speech articulation (Section 5), an overview of how this is used in the framework of a text-to-visual-speech synthesiser (Section 6), and finally conclusions of this work and discussion of possible future improvements (Section 7).

4. Image Morphing

Image morphing algorithms deal with the transition between two digital images. Smooth transformations cannot be produced by simply fading one image into another, because the features in the two images may be unaligned producing a disturbing double-exposure effect. In order to compensate for this, image morphing algorithms rely upon both a geometric alignment and colour interpolation to produce smooth transitional images.

Given two images, I_0 and I_1, two warping functions are specified, $d_{0\to1} : \mathbb{R}^2 \to \mathbb{R}^2$ and $d_{1\to0} : \mathbb{R}^2 \to \mathbb{R}^2$, which respectively forward warp I_0 to coincide with I_1 and backward warp I_1 to coincide with I_0. Once the geometric mapping between images is determined the pixels can simply be interpolated between the two images, both in colour and location.

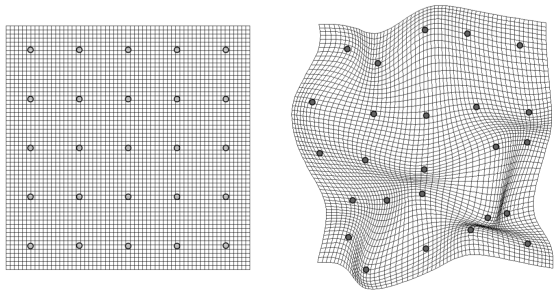
Some pixels will not be filled by such a technique as holes will appear where one pixel maps onto several in the destination image, i.e. dilation of part of the image due to the warping functions. These holes are removed using a simple scanline or bilinear fill algorithm.

The most important feature of image metamorphosis methods is the technique for determining the coordinate mapping between I_0 and I_1. In this work we use an implementation of the technique described in [1, 21, 30], which uses Radial Basis Functions to warp the images to coincide with one another.

4.1. Radial Basis Functions for Image Warping

The most general means of specifying image features relies upon the placement of primitives on the two matched images. These primitives may be points, lines or curves. The warp between the two images can then be defined by finding surfaces that interpolate the feature primitives. Curves and lines can both be point sampled, allowing us to define the surfaces which pass through the point sets using scattered-data interpolation techniques. Radial Basis Functions (RBFs) are one such means of producing a smooth interpolated surface from a set of scattered feature points.

The RBF approach constructs the interpolant as a linear combination of basis functions (the RBFs). Defining a surface which interpolates a number of known points relies upon determining the coefficients α_i from (1).



(a) (b)

Figure 1. Warping a 2D mesh with RBFs: a) original mesh; b) mesh after warping.

$$f(x) = p_m(x) + \sum_{i=1}^{n} \alpha_i \phi_i(\|x - x_i\|) \tag{1}$$

The value of the function ϕ_i depends only upon the distance from its centre x_i and thus is called radial. The weights, α_i, of the basis functions are found by placing the centres back into (1) and solving the resulting set of linear equations.

The polynomial term p_m is included to allow a certain degree of polynomial precision, but may be excluded altogether. If p_m is of degree $m = 1$ then the polynomial term is a simple affine transformation. Where the influence of the RBFs tend to zero, the result of the interpolation will be dominated by the influence of the polynomial term. In this work we use the identity transform for the polynomial term, i.e. $p_m(x) = x$.

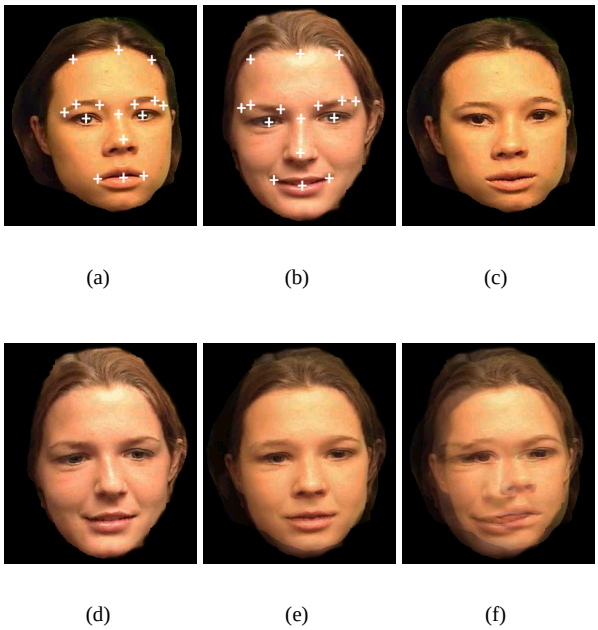
For the purposes of image warping we use the inverse multiquadric (2) for the RBF. These produce C_∞ interpolated surfaces when the functions are global and not locally bounded. Other possible alternatives for ϕ_i include the gaussian and the thin-plate spline.

$$\phi(x) = \frac{1}{\sqrt{x^2 + r^2}} \tag{2}$$

The radii of each function r_i is unique, and is determined as the least distance to its surrounding data points (3).

$$r_i = min_{i \neq j} \|x_i - x_j\| \tag{3}$$

Figure 1 demonstrates the use of RBFs in warping a 2D mesh. The same approach can be used to warp images (fig. 2), as a part of an image metamorphosis method. The surfaces created by RBF interpolation are smooth, making for good image transitions. The disadvantage of RBFs lies in their global nature, since for every pixel in the image *every*



(a) (b) (c)

(d) (e) (f)

Figure 2. Image metamorphasis with RBFs: a) source image I_0; b) destination image I_1; c) forward warping I_0 with $d_{0\to1}$; d) backward warping I_1 with $d_{1\to0}$; e) result of morphing between I_0 and I_1; f) cross-dissolved image.

RBF must be taken into account. An improvement in efficiency, as proposed in [30], can be made by evaluating the RBFs on a lower resolution grid, and interpolating the grid displacements to warp individual pixels.

5. Statistically Modelling Visible Articulation

The morphing technique described in Section 4 relies upon the placement of key feature points on each of the images. This is a rather crude parameterisation of visible articulation when the movement of these points is not independent. A statistical model of visible articulation allows us to define the feature point set in terms of a very few parameters. In this work Principal Component Analysis (Section 5.1) is used to define parameters for facial shape directly from a labelled dataset of visemes.

5.1. Principal Components Analysis

Principal components analysis (PCA) is a classical multivariate statistical technique which provides a generative model for an input population. We define a random vector population, consisting of s samples (4) as having mean vector μ_v (5) and covariance matrix C_v (6).

$$v = \{v_0, ..., v_s\}^T \tag{4}$$

$$\mu_v = E\{v\} \tag{5}$$

$$C_v = E\{(v - \mu_v)(v - \mu_v)^T\} \tag{6}$$

We can define this population using the mean vector and the combination of a set of orthogonal basis vectors e_i weighted by a set of factors b_i (7).

$$v = \mu_v + \sum_{i=1}^{s} e_i b_i \tag{7}$$

As the covariance matrix is symmetric, the orthogonal basis set can be calculated by finding its eigenvectors e_i and eigenvalues λ_i (8).

$$C_v e_i = \lambda_i e_i, i = 1...s \tag{8}$$

$$|C_v - \lambda I| = 0, \tag{9}$$

The eigenvalues λ_i are the solutions to the characteristic equation (9) which can be solved by reducing to tridiagonal form and using the QR algorithm. The product of the orthogonal transformations used in the QR algorithm give the eigenvectors of C_v. Further discussion into the solution of general eigenproblems and a description of the QR algorithm can be found in [11].

The eigenvectors e_i of the covariance matrix C_v are called the principal components (PCs) of the vector population, ordered by the eigenvalues λ_i which hold the variance σ^2 for each PC. The number of PCs will equal the number of dimensions in the sample vectors s. However, in most cases the first n (where $n \ll s$) components will account for enough of the variance within the population for the rest to be discarded. For this reason a truncated principal component model can be used both to extract important relationships, and as a form of data compression. Choosing how many PCs should be kept is a matter of debate - for discussion on this subject and PCA in general refer to [14, 12].

5.2. Formulation of the Model

In order to create a statistical model of visible articulation we consider each set of feature points to be a single sample v_i within the system.

$$v_i = \{x_{i0}, y_{i0}, x_{i1}, ..., x_{i(n-1)}, y_{i(n-1)}\} \tag{10}$$

We perform the PCA upon this dataset which determines the most highly correlated vectors (the eigenvectors e_i), of which we keep those which account for the top 99% of the

Vowels			
AA	odd	UW	two
AE	at	EH	Ed
AH	hut	ER	hurt
AO	ought	EY	ate
AW	cow	IH	it
AY	hide	IY	eat
UH	hood	OW	oat
OY	toy		
Consonants			
B	be	K	key
M	me	D	dee
P	pee	T	tea
CH	cheese	HH	he
JH	gee	L	lee
SH	she	N	knee
ZH	seizure	NG	ping
DH	thee	R	read
TH	theta	S	sea
F	fee	W	we
V	vee	Y	yield
G	green	Z	zee

Table 1. English phone classification used in FreeTTS.

variance (3 PCs). This leads to a parameterisation of the feature points which is subsequently used in the modelling of coarticulation (Section 6.1).

6. Speech Synchronised Animation

The model described works in collaboration with the FreeTTS system to synthesize each new utterance. The FreeTTS synthesizer works by concatenating low level audio units (diphones) to produce the audible speech signal. Phone timing information is extracted from the FreeTTS system, and is used to control the synthesis of the visual signal.

For each audio phone (Table 1) a corresponding visual unit (viseme) is stored. Transitions between visemes are synthesised using the image metamorphosis method described in Section 4. Trivially, linear interpolation is used to determine in-between frames, however, the resulting animations are not physically correct. In order to create a better approximation of speech articulation a model of coarticulation has been included in the model as we shall now describe.

6.1. Modelling Coarticulation

The process of speech production cannot simply described as a linear interpolation between invariant linguistic units (in our case visemes). The movement of the articulators is complex, and the obscuration of the boundaries between visemes should be correctly accounted for.

The blending of visemes can be split two ways: the anticipatory preparation for a future viseme (backward coarticulation), and the inertia of a preceding viseme affecting future articulations (forward coarticulation). Typically coarticulation prevents the articulators from meeting ideal phone targets during speech production, which is why simple linear interpolation models can look unnatural.

Several models for coarticulation have been proposed in the speech production literature (e.g. [25, 23]). These are based upon the use of functions to blend articulatory movements over time. In this work we use a model initially proposed by Cohen and Massaro [7]. Each viseme, s, has a corresponding dominance function, D_{sp}, for each PC parameter p. A negative exponential function (11) is used to define the dominance a segment (viseme) exerts over the utterance.

$$D_{sp}(\tau) = \begin{cases} \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau|^c} \text{ if } |\tau| \geq 0 \\ \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau|^c} \text{ if } |\tau| < 0 \end{cases} \quad (11)$$

In (11), α is the magnitude of the dominance function, τ is the time distance from the segment center, c determines the width of the function, and $-\theta_{\leftarrow sp}$ and $-\theta_{\rightarrow sp}$ define the rate of forward and backward coarticulation respectively. A weighted combination of these dominance functions (12) is used to translate static visemes into the resultant articulatory curve for an utterance. Figure 3 shows the effect of dominance functions on the interpolation of a single speech parameter. As a result of the described coarticulation model the articulatory parameters *may* not pass directly through their target values, thus mimicking the role of context in speech production.

$$val_p = \frac{\sum_{s=0}^{n} D_{sp}(\tau) target_{sp}}{\sum_{s=0}^{n} D_{sp}} \quad (12)$$

In order to apply the result of the dominance functions, a further warp is applied to the result of a linear morph between the source and destination viseme images. This accounts for the difference between a linear interpolation and a more accurate model of speech production. The warp is applied as described in Section 4.1, which requires that a further linear system must be solved every frame.

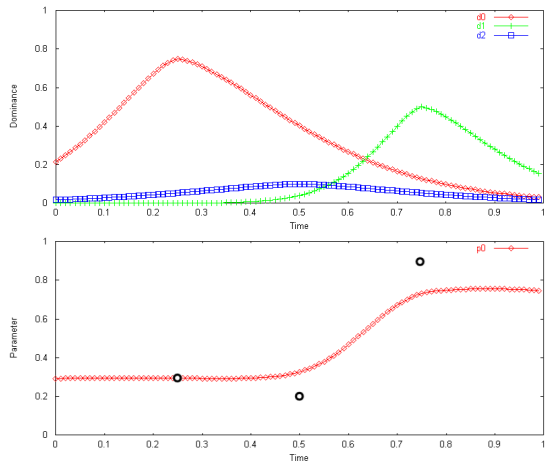


Figure 3. Modelling coarticulation: a) dominance functions for several speech segments; b) output parameter curve (target parameters shown as circles).

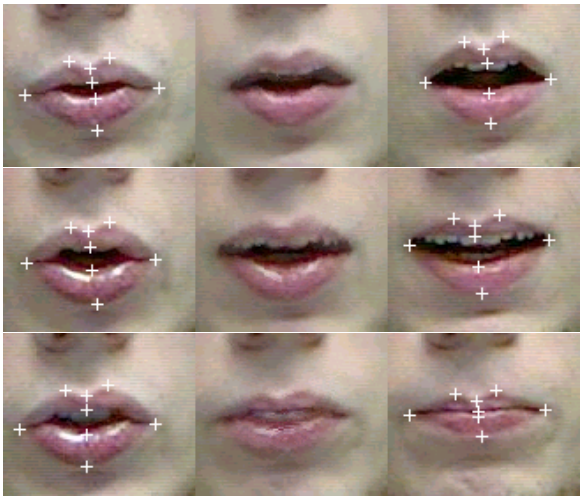


Figure 4. Synthesized viseme transitions. Central column contains transitional frames between the source and destination visemes.

7. Conclusions

The use of RBF-based image morphing produces smooth natural transitions between neighbouring visemes. This is an improvement over other morph-based techniques such as concatenated optical flow, due to the decreased data storage requirements of the method. The central column of fig. 4 demonstrates some of the transitions generated by the system in synthesising speech movements. Unfortunately, the labelling of visemes with feature points is a time consuming task which must be performed before any animations can be created. A possible improvement would be to automatically label visemes with feature points using a method such as the Active Appearance Model described in [8].

One of the major contributions of this paper is the implementation of coarticulation functions for image-based models. Whilst these are increasingly common in three-dimensional talking heads, little research has been carried out into controlling image morphs with dominance functions. Our results lead to less robotic speech movements than reported [10] with simple linear interpolation.

Naturally, image-based methods, such as that described in this paper, have advantages over models which attempt to model the complexities of facial expression and appearance in three-dimensions. As would be expected, the animations output by the system appear photo-realistic. A disadvantage of our current image-based method is the lack of expressive capability. Our talking head cannot smile, or raise its eyebrows as would occur in normal human-human communication. Such capabilities could be incorporated using the currently implemented warping functions. However, issues regarding the blending of visual speech with expressions are still unresolved. Another current problem is that the head cannot be reoriented like a three-dimensional model, and so we cannot see the animation from a different viewpoint. View morphing could be implemented to enable this functionality, although the amount of viseme data required would be greater. For these reasons three-dimensional heads dominate research, even though it could be argued that image-based methods could be used to create realistic looking animation on lightweight platforms such as mobile phones and PDAs.

8. Acknowledgements

The authors would like to thank the EPSRC for providing funding for this research. James Edge would also like to thank Mark Eastlick, Michael Meredith, and Manuel Sánchez Lorenzo for their help and support in producing this report. Images used in fig. 2 are from the CUAVE database (`http://ece.clemson.edu/speech/cuave.htm`).

References

[1] N. Arad, N. Dyn, D. Reisfeld, and Y. Yeshurun. Image warping by radial basis functions: Application to facial expressions. *Computer Vision, Graphics, and Image Processing. Graphical Models and Image Processing*, 56(2):161–172, 1994.

[2] J. Barron, D. Fleet, S. Beauchemin, and T. Burkitt. Performance of optical flow techniques. *CVPR*, 92:236–242, 1992.

[3] T. Beier and S. Neely. Feature-based image metamorphosis. *Computer Graphics*, 26(Annual Conference Series):35–42, 1992.

[4] A. Black, P. Taylor, and R. Caley. The festival speech synthesis system. (`http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`), June 1999.

[5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. *Computer Graphics*, 31(Annual Conference Series):353–360, 1997.

[6] N. Brooke and S. Scott. Two and three-dimensional audio visual speech synthesis. *Proc. AVSP'98*, 1998.

[7] M. Cohen and D. Massaro. Modeling coarticulation in synthetic visual speech. *Computer Animation '93*, 1993.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Lecture Notes in Computer Science*, 1407, 1998.

[9] J. Edge and S. Maddock. Expressive visual speech using geometric muscle functions. *Proc. Eurographics UK*, pages 11–18, 2001.

[10] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. Technical Report AIM-1658, MIT, 1999.

[11] G. H. Golub and C. F. V. Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, 1996.

[12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independant Component Analysis*. John Wiley and Sons Inc., 2001.

[13] S. M. Inc. Freetts programmer's guide. (`http://freetts.sourceforge.net/docs/ProgrammerGuide.html`), 2001.

[14] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[15] K. Kähler, J. Haber, and H.-P. Siedel. Geometry-based muscle modeling for facial animation. *Proc. Graphics Interface 2001*, pages 37–46, 2001.

[16] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. Simulation of facial muscle actions based on rational free form deformations. *Computer Graphics Forum (EUROGRAPHICS '92 Proceedings)*, 11(3):59–69, 1992.

[17] S. King, R. Parent, and B. Olafsky. An anotomically-based 3d parametric lip model to support facial animation and synchronized speech. *Proc. Deform 2000*, pages 7–9, 2000.

[18] R. Koenen. Overview of the mpeg-4 standard. Technical Report ISO/IEC JTC1/SC29/WG11 N2725, Moving Picture Experts Group, March 1999.

[19] S.-Y. Lee, K.-Y. Chwa, and S. Y. Shin. Image metamorphosis using snakes and free-form deformations. *Computer Graphics*, 29(Annual Conference Series):439–448, 1995.

[20] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. *Computer Graphics*, 29(Annual Conference Series):55–62, 1995.

[21] P. Litwinowicz and L. Williams. Animating images with drawings. *Computer Graphics*, 21(Annual Conference Series):409–412, 1994.

[22] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. *Computer Graphics*, 35(Annual Conference Series):271–276, 2001.

[23] A. Löfqvist. Speech as audible gestures. *Speech Production and Speech Modelling*, pages 289–322, 1990.

[24] D. Massaro. *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle.* Bradford Books Series in Cognitive Psychology. MIT Press, 1998.

[25] S. E. G. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310–320, 1967.

[26] F. I. Parke. A parametric model for human faces. Technical report, University of Utah, 1974.

[27] F. I. Parke and K. Waters. *Computer Facial Animation*. A. K. Peters, Ltd., 1996.

[28] S. M. Platt and N. I. Badler. Animating facial expressions. *Computer Graphics*, 15(Annual Conference Series):245–252, 1981.

[29] L. Revéret, G. Bailly, and P. Badin. Mother : A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *6th Int. Conference of Spoken Language Processing, ICSLP'2000*, 2000.

[30] D. Ruprecht, R. Nagel, and H. Müller. Spatial free-form deformation with scattered data interpolation methods. *Computers and Graphics*, 19(1):63–71, 1995.

[31] W. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[32] G. Wolberg. Image morphing: a survey. *The Visual Computer*, 14(8/9):360–372, 1998.