CONSTRUCTION OF AN INDIVIDUALIZED VISUAL SPEECH-SYNTHESIZER FROM ORTHOGONAL 2D-IMAGES

Arthur Niswar * **Gérard Bailly** ** **Kristian Kroschel** *

* *Institut für Nachrictentechnik, Universität Karlsruhe, Karlsruhe, Germany*
** *Institut de la Communication Parlée - INPG, Université Stendhal, Grenoble, France*

Abstract: In speech communication, the visual information plays an important role. This information, such as lips movement and face expression, is perceived and processed by humans. They can enhance the speech intelligibility, especially in a noisy environment. It is therefore advantageous to incorporate them in a speech synthesis system. This audio-visual speech synthesizer is widely known as talking head. The visual speech synthesizer constructed here is a first step towards creating a complete talking head.
The visual speech-synthesizer constructed in this work is an individualized 3D-face model consisting of 283 3D-points with 6 articulatory parameters, which control the movement of the speech-articulators (jaw, lips, and larynx). This model is created by a non-linear modification of an existing 3D-face model based on a set of the individual's face features. For this purpose, two orthogonal images (i.e. the frontal and profile images) of the person are utilized. The facial features in the frontal and profile images are extracted manually. The texture image of the individual is then mapped to the constructed face-model, to give it a natural look. The articulatory parameters are determined using PCA (Principal Component Analysis) performed on the selected subsets of the 3D-model points.

Keywords: Talking Head, Guided PCA, Dirichlet Free-Form Deformation

1. INTRODUCTION

The construction of a talking head involves the fusion of a TTS (text-to-speech) synthesizer and an articulatory model of a human's head. To create an articulatory head-model, one needs to determine the parameters that control the articulatory movement of the head. These parameters are then incorporated into a head-model to control the required articulation for visual speech synthesis.

The head-model of a person can be created in several ways: through data acquisition of the head (with a 3D-scanner or stereo-reconstruction from several 2D acquisitions) or modification of an existing head-model. In this work an articulatory 3D head-model was first created to serve as a generic model, which can be modified to create another head-model. This model was built using a stereo-reconstruction from the acquired 2D dataset. The articulatory parameters for this model were extracted from the dataset using PCA.

The individualized head-model is then created by modifying this first head-model with the aid of the orthogonal 2D-images of the person, by means of a certain non-linear deformation.

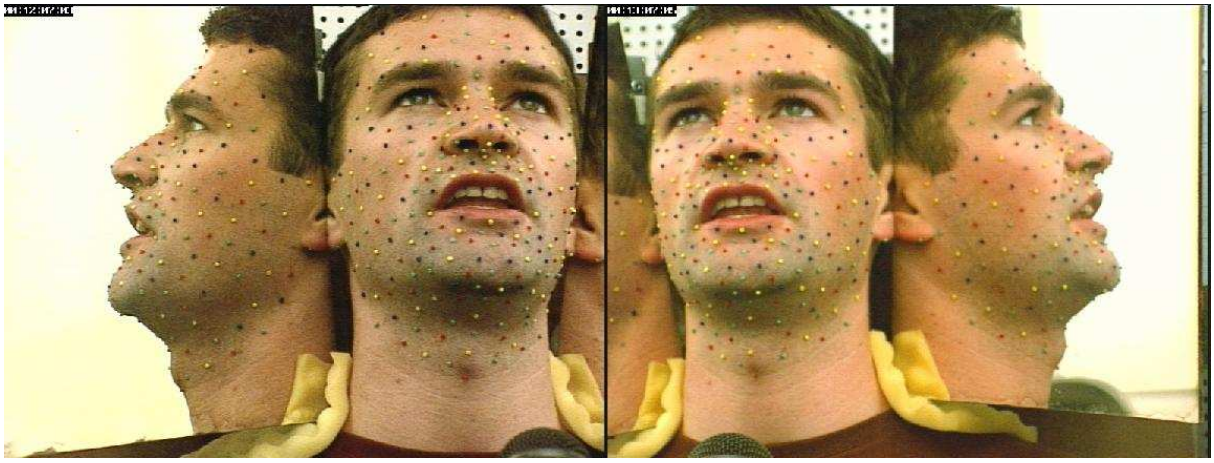

Fig. 1. Sample multi-view frames from the synchronized video-data

2. CONSTRUCTION OF THE FIRST HEAD-MODEL

In order to be able to extract the articulatory parameters, the 2D data was recorded while the subject was pronouncing the expressions from a speech-corpus (see section 2.1). This data-acquisition took place in ICP (*Institut de la Communication* $Parlée$) in Grenoble. Frames from this video-data were extracted and then processed to obtain the head-model and its articulatory parameters.

2.1 Speech-corpus and data acquisition

For the recording, the subject's face had been marked with 243 colored beads (on the salient structures of the face, e.g. jaw line, around the lips, nose and cheek), as shown in Fig. 1. Two synchronous video-cameras were used in this recording. The profile views of the subject were captured at the same time with the aid of 2 mirrors.

The speech-corpus used to construct the model contains a set of German visemes, which consists of sustained hyperarticulated vowels and consonantal closures in context. The vowels and consonants are listed below:

- Vowels : [a], [ɪ], [ʊ], [ɛ], [ɔ], [ʏ], [œ], [aː], [iː], [uː], [eː], [ɛː], [oː], [øː], [yː], [ə], [ɐ]
- Consonants : [b], [p], [d], [t], [g], [k], [f], [v], [s], [ʃ], [h], [ʒ], [j], [ç], [x], [l], [r], [m], [n], [ŋ]

Each of the consonants was uttered in 4 symmetrical vocalic context : /a_a/, /i_i/, /u_u/, /œ_œ/.

After the recording, the frames in the center of each allophone (so-called viseme) were grabbed. The 2D-coordinates of each bead in the frame were determined manually. In addition, 8 points around the eyes, 1 point at the upper teeth and 1 point at the lower teeth were also determined. After the stereo-reconstruction, the 3D-model for each viseme consisted of 283 points, where 30 of them characterized the shape of the lips. These points were obtained by manually fitting a generic 3D-model of lips (Revéret and Benoît, 1998) to all the multi-view images.

2.2 Modeling articulatory movements

The articulatory model results from a statistical analysis of the 3D-data (number of observations $\times$ 283 points $\times$ 3 coordinates). Principal Component Analysis (PCA) is applied successively on particular subsets of the data to generate the linear predictors for the whole dataset. Those subsets represent the speech-articulators. This procedure is also known as *guided PCA* (Elisei, *et al.*, 2001).

The articulatory movements of the facial points $\mathbf{p} = [\ x_1\ y_1\ z_1\ x_2\ y_2\ z_2\ \ldots\ x_{283}\ y_{283}\ z_{283}\ ]^T$ are modeled as follows:

$$\mathbf{p} = \bar{\mathbf{p}} + \mathbf{M} \cdot \boldsymbol{\alpha} \tag{1}$$

where $\bar{\mathbf{p}}$ is the mean value of the data vector $\mathbf{p}$. The matrix $\mathbf{M} = [\mathbf{m}_1\ \mathbf{m}_2\ \ldots\ \mathbf{m}_6]$ is composed of the linear articulatory predictors and the elements of $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_6]^T$ are the articulatory parameters. $\mathbf{M}$ and $\boldsymbol{\alpha}$ describe the movements of the articulators (jaw, lips and larynx), where $\boldsymbol{\alpha}$ is the free parameter. There are 2 parameters for jaw-movement (α_1 and α_5, also called *jaw1* and *jaw2*), 3 for lips-movement (α_2 - α_4, or *lips1-3*) and 1 for larynx-movement (α_6, or *lar1*).

The column vectors of $\mathbf{M}$ (i.e. $\mathbf{m}_i$, $i = 1, \ldots, 6$) are computed successively following this procedure:

- At first α_1 (i.e. *jaw1*) is computed for each viseme from a subset $\mathbf{q}_1$ of the data, which is the facial points on the jaw (see Fig. 2):

$$\alpha_1 = \frac{\mathbf{e}_{1,\mathbf{q}_1}^T \cdot (\mathbf{q}_1 - \bar{\mathbf{q}}_1)}{\sqrt{Var(\mathbf{q}_1)}} \tag{2}$$

where $\mathbf{e}_{1,\mathbf{q}_1}$ is the eigenvector of the covariance matrix $\mathbf{C}_{\mathbf{q}_1}$ corresponding to the greatest eigenvalue $\lambda_{1,\mathbf{q}_1}$, $\mathbf{q}_1$ is the subset of the data vector $\mathbf{p}_1$ (in this case $\mathbf{p}_1 = \mathbf{p}$, the original data), $\bar{\mathbf{q}}_1$ is the mean value of $\mathbf{q}_1$ and $Var(\mathbf{q}_1)$ is the variance of $\mathbf{q}_1$.

- The facial points are then modeled by a multilinear regression model:

$$\hat{\mathbf{p}}_1 = \mathbf{m}_{0,p_1} + \mathbf{m}_1 \cdot \alpha_1 \quad (3)$$

The parameters $\mathbf{m}_{0,p_1}$ and $\mathbf{m}_1$ are obtained by minimizing the mean-square error for all N visemes in the data (indexed by k):

$$S = \sum_{k=1}^{N} (\mathbf{p}_{1,k} - \hat{\mathbf{p}}_{1,k})^T \cdot (\mathbf{p}_{1,k} - \hat{\mathbf{p}}_{1,k}) \quad (4)$$

$\hat{\mathbf{p}}_1$ is substituted from Eq. 3. The parameters are then obtained:

$$\mathbf{m}_{0,p_1} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{p}_{1,k} = \bar{\mathbf{p}}_1 \quad (5)$$

$$\mathbf{m}_1 = \frac{Var(\mathbf{q}_1)}{N\lambda_{1,\mathbf{q}_1}} \sum_{k=1}^{N} \alpha_{1,k} \cdot \mathbf{p}_{1,k} \quad (6)$$

where $\lambda_{1,\mathbf{q}_1}$ is the greatest eigenvalue of the covariance matrix $\mathbf{C}_{\mathbf{q}_1}$.

- Before computing the next parameter (*lips1*), the contribution of $\mathbf{m}_1$ is subtracted from $\mathbf{p}_1$ (for all visemes):

$$\mathbf{p}_2 = \mathbf{p}_1 - \mathbf{m}_1 \cdot \alpha_1 \quad (7)$$

Then α_2 (i.e. *lips1*) is computed for each viseme as in the first step, using the subset $\mathbf{q}_2$ (*lips1*-Points). According to (Elisei, *et al.*, 2001), the articulatory parameters are computed in this order: *jaw1*, *lips1-3*, *jaw2* and *lar1*.

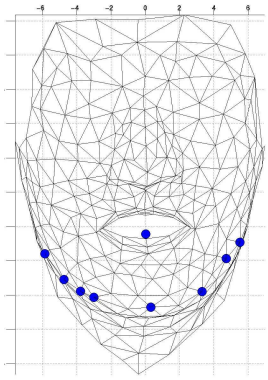

Fig. 2. The *jaw1*-Points

2.3 Visual appearance of the model

The 3D-points of the model are connected in triangles, to form a mesh. The image of the speaker is then mapped to this mesh, to give it a natural look (Fig. 3).

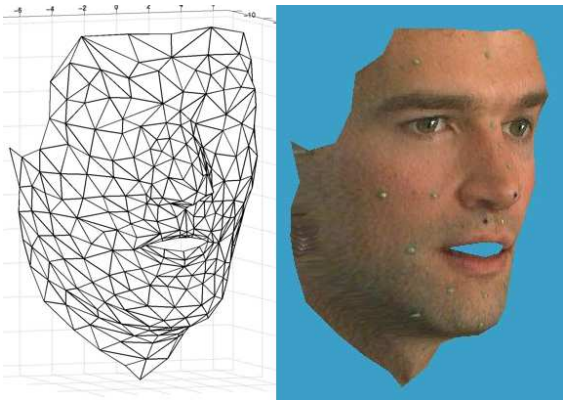

Fig. 3. The 3D head-model

3. CREATING THE INDIVIDUALIZED HEAD-MODEL

To create the individualized head-model from the first model, one needs 2 orthogonal images of the person, i.e. the frontal and profile images. Moreover, the first model has to be "neutralized" first (i.e. to bring it in a "neutral" position, like the head-position of the person in the images). Then this neutral 3D-model is projected into the 2D-space, and one obtains 3 2D-projections: 1 frontal and 2 profile projections (left and right).

Before performing the modification, some reference points in the 2D-projections are chosen. The corresponding points in the images are determined manually. The 2D-projections are then modified using a non-linear deformation called DFFD (*Dirichlet Free-Form Deformation*), to form the individualized head-model.

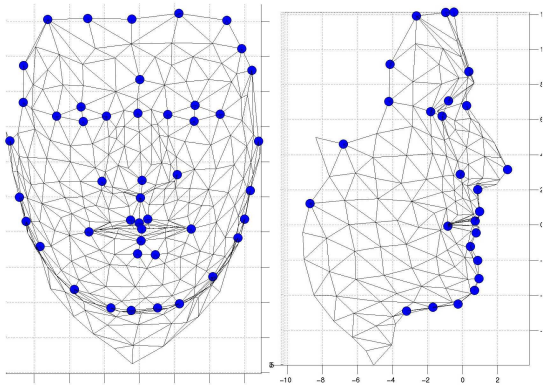

Fig. 4. The reference points in the 2D-projections of the first model

3.1 The reference points

The reference points chosen to aid the modification characterize the structure of the face, e.g. the corners of the eyes, nose-tip and the corner of the lips. In addition, some points on the edge of the 2D-projections are also taken as reference, so that DFFD can work

correctly. These reference points are shown in Figure 4.

The corresponding reference points in the orthogonal images are then determined manually. This is shown in Fig. 5. Those points are then scaled to the unit used in the first model (in cm). Moreover, the point of origin in the images and in the first model are set to the root of the nose.

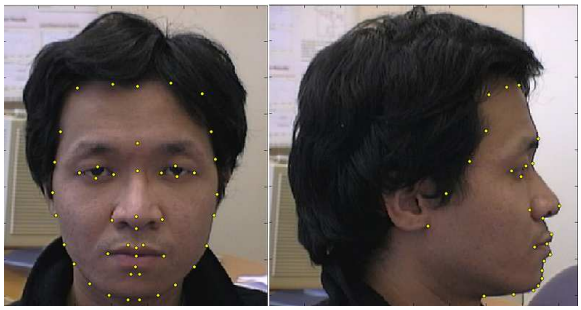

Fig. 5. The reference points in the orthogonal images

3.2 Modification of the model

To construct the new model, one needs to calculate the position of the non-reference points in the orthogonal images. This is performed by determining the displacements from the 2D-model points. For this purpose, DFFD is used. The basic idea of this non-linear deformation is the dependency of displacements: the displacement of one point is dependent on the displacements of the points surrounding it. This dependency is measured by the so-called *natural coordinates* of the point relative to its *natural neighbors*, i.e. the points surrounding it. Calculation of these natural coordinates and natural neighbors involves the determination of the **Voronoi diagram** and the **Delaunay triangulation** (Aurenhammer, 1991) for the point set. This is depicted in the next figure.

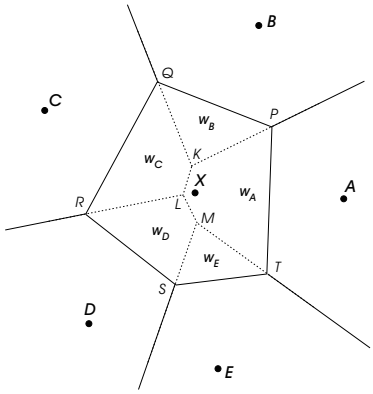

Fig. 6. The natural neighbors ($A - E$) and the natural coordinates ($w_A - w_E$) of X

In Fig. 6, the points $A - E$ are the natural neighbors of X, and $w_A - w_E$ are the corresponding natural coordinates. These coordinates are the ratios between the corresponding partial area and the area of the cell enclosing X.

The displacements of the non-reference points can then be computed iteratively:

- Determine the Voronoi diagram for the existing point set (in the beginning just the reference points).
- Re-determine the Voronoi diagram after the addition of a new point to the point set.
- Calculate the natural coordinates and the natural neighbors of the new point.
- Calculate the displacement of the new point:

$$\Delta X = \sum_i w_i \cdot D_i \tag{8}$$

where D_i is the displacement of the i-th natural neighbor of X and w_i is the corresponding natural coordinate.

The result of this calculation is shown in Fig. 7.

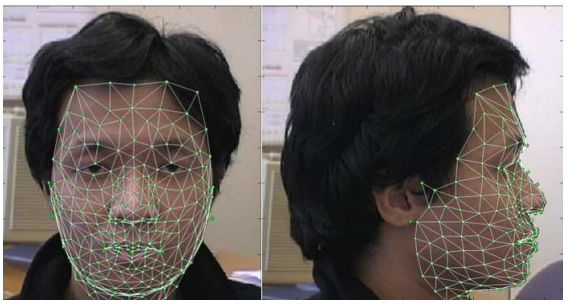

Fig. 7. The new 2D-model points

The new 2D-models are then merged to form the 3D-model. To compute the articulatory parameters for this new model, a new dataset of visemes must be created first. This dataset is created using DFFD from the first dataset. Afterwards, the articulatory parameters are extracted from the new dataset using guided PCA. The image is then mapped to the new model (Fig. 8).

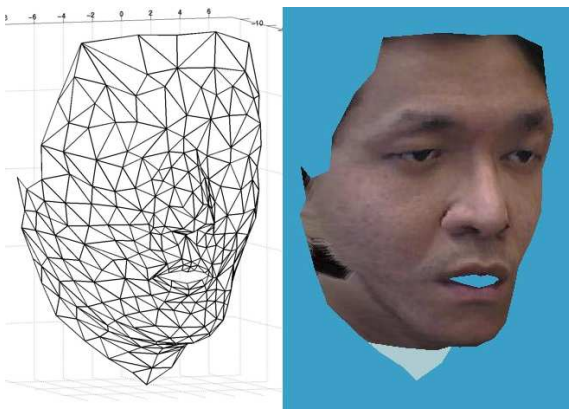

Fig. 8. The new 3D-model

4. SYNTHESIZING VISUAL SPEECH

The synthesis of visual speech involves at first the translation of the phrase to be synthesized into phonetical segments with specific durations. This is accomplished by an existing text-to-speech synthesis system (Möhler, *et al.*, 2001). After that, the visemes for each segment are chosen from the dataset, and the frames in which the visemes occur are determined. The articulatory parameters (α) for each of these visemes are then calculated. The values of α in the remaining frames are currently determined with cubic spline interpolation.

As an example, to synthesize the phrase "**Guten Tag**", it is first translated into phonetical segments (55 frames):

_	g	u:	t	ə	n	t	a:	k	_
1	12	14	20	24	27	30	33	40	44

where [_] symbolizes silence, and the numbers represent the start-frame for each viseme. For this sequence, the visemes chosen from the dataset and their durations (in frames) are:

Viseme	Duration (frames no.)
_	1-8
/u**g**u/	12
u:	16-17
/u**t**u/	20
/i**t**i/	22
ə	25
/œ**n**œ/	27
/a**t**a/	30
a:	35-36
/a**k**a/	40
_	47-55

Then the values of α (*jaw1-2*, *lips1-3*, *lar1*) are computed for each viseme, based on Eq. 1. For silence, α is computed for the neutral position of the head. The values of α in the remaining frames are computed with cubic spline interpolation. This is shown in Fig. 9.

5. CONCLUSIONS

The individualized visual speech-synthesizer constructed in this work is a first step toward creating an individualized talking head. It is created from an existing head-model with only 2 orthogonal 2D-images of the person. With the frontal image of the person mapped on the new 3D-model, the result looks natural, but with some distortions. To improve it, one needs to combine the frontal and profile images into a cylindrical texture image and map it onto the model. Nevertheless, the person is still recognizable.

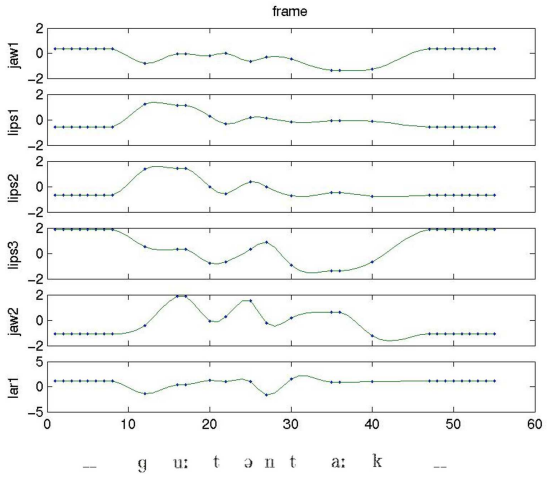

Fig. 9. Generating "Guten Tag" using cubic spline interpolation

The synthesis of visual speech in this work doesn't take into consideration the coarticulation effects between visemes. In the example of "Guten Tag" above, the transition between "n" in "Guten" and "t" in "Tag" in the synthesized visual speech is rather unidentifiable, because there is normally a coarticulation between these 2 visemes which isn't incorporated into this visual speech-synthesizer yet.

This coarticulation models will therefore be investigated in the further work, to improve the intelligibility of the synthesized visual speech. Finally this visual speech-synthesizer must be merged with a text-to-speech synthesizer to create a fully functional talking head.

ACKNOWLEDGEMENT

Thanks to Frédéric Elisei and Christophe Savariaux from ICP-Grenoble for their assistance in gathering and analyzing the data.

REFERENCES

Aurenhammer, F. (1991). Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. In: *ACM Computing Surveys*, Vol. **23**, No. 3, pp. 345–405.

Black, A.W., P. Taylor and R. Caley (1999). The Festival Speech Synthesis System - System Documentation. University of Edinburgh.

Elisei, F., M. Odisio, M., G. Bailly, and P. Badin (2001). Creating and Controlling Video-Realistic Talking Heads. In: *Auditory-Visual Speech Processing Workshop*, Aalborg, Denmark, pp. 90–97.

Ip, H.H.S. and L. Yin (1996). Constructing a 3D individualized head model from two orthogonal views. In: *The Visual Computer*, 12:254-266. Springer Verlag.

Lee, W., E. Lee and N.M. Thalmann (1998). Real Face Communication in a Virtual World. In: *Virtual Worlds, First International Conference, VW98*, Paris, France (J. Heudin, Ed.), pp. 1–13.

Moccozet, L. and N.M. Thalmann (1997). Dirichlet Free-Form Deformations and their Application to Hand Simulation. In: *Proc. Computer Animation*, pp. 93-102. IEEE Computer Society.

Möhler, G., A. Schweitzer and M. Breitenbücher (2001). The IMS German Festival Manual. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Revéret, L. and C. Benoît (1998). A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Recognition. In: *Auditory-Visual Speech Processing Workshop*, Terrigal, Australia, pp. 207-212.