# Personalized Unit Selection for an Image-based Facial Animation System

*Axel Weissenfeld, Kang Liu, Sven Klomp and Joern Ostermann, Fellow IEEE*

Institut fuer Theoretische Nachrichtentechnik und Informationsverarbeitung
University of Hanover
Appelstr. 9A, 30167 Hannover, Germany
aweissen, kang, klomp, ostermann@tnt.uni-hannover.de

***Abstract*— This paper describes an image-based facial animation system, which consists of the audiovisual analysis of a human subject and the synthesis of a photo-realistic facial animation. The unit selection algorithm selects for a given audio output the best mouth samples from the database by assigning two costs, the phonetic context and the visual distance between two consecutive samples. Here a novel approach to adapt the unit selection algorithm to an individual human subject is presented, such that a photo-realistic facial animation can be generated.**

***Index Terms*— Facial Animation, Unit Selection, Image-based Rendering, Viterbi, LLE, PCA.**

## I. Introduction

Computer aided modeling of human faces usually requires a lot of manual control to achieve realistic animations and to prevent unrealistic or non-human like results. Humans are very sensitive to any abnormal lineaments, so that facial animation remains a challenging task till today. Facial animation combined with text-to-speech synthesis (TTS), also known as talking head, can be used as a modern human-machine interface. In Fig. 1, a typical application of facial animation is presented. Here an internet-based customer service site integrates a talking head into its web site. Subjective tests showed that Electronic Commerce Web sites with talking heads get a higher ranking than without [1] [2].

Nowadays animation techniques range from animating 3D models to image-based rendering of models. In order to animate a 3D model consisting of a 3D mesh, which defines the geometric shape of the head, vertices of the 3D mesh are moved. The first approaches already began in the early 70's [3]. Since then different animation techniques [4] [5] [6] were developed, which continuously improved the animation. However, animating a 3D model still does not achieve photo-realism. Photo-realism means to generate animations that are undistinguishable from recorded video. Recently, image-based facial animation was introduced [7]. Image-based rendering processes only 2D images, so that new animations are generated by combining different facial parts of recorded image sequences. Hence, a 3D model is not necessary. The system described in [7] can produce videos of people uttering a text, they never said before. Short video clips, each showing three consecutive frames (called tri-phones) are stored as samples, which lead to a large database. Ezzat et al. [8] [9] have demonstrated a sample-based talking head that uses morphing

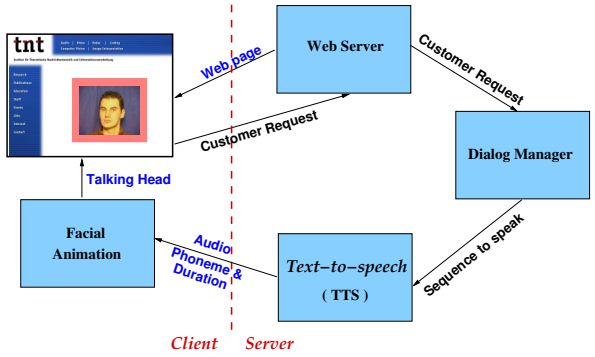


Fig. 1. A web-based information kiosk and a customer service site that integrates a web site with a talking head.

to generate intermediate appearances of mouth shapes from a very small set of mouth samples. Cosatto et al. [10] [11] designed a system, which achieves photo-realistic facial animations and can be currently regarded as the state-of-the-art facial animation engine. The face model mainly consists of a personalized mask and a large database of mouth images and related information. Our system is based on Cosatto's work.

This paper focuses on training the unit selection algorithm, which describes the selection of appropriate mouth samples from a database given an audio file, so that a photo-realistic animation is achieved. The well-known Viterbi search algorithm can be used in order to find the best samples in the large database. Two types of costs are considered: visual differences and phonetic information. Instead of using principal components (PCA) to characterize the visual appearance of mouth images as described by Cosatto [11], we use locally linear embedding (LLE) coordinates which are used to classify nonlinear manifolds of arbitrary dimension, since the relation between mouth appearances and their corresponding image signals is nonlinear. In contrast to Cosatto, who sets the parameters of the unit selection algorithm by experience, we are training our unit selection algorithm on a specific human subject. This assures that the best mouth samples are selected and a high-quality animation, undistinguishable from the specific human subject, is achieved.

In the remainder of this paper, we describe our image-based facial animation system (Section 2). Furthermore is discussed the unit selection algorithm (Section 3) and the training of the unit selection algorithm and examples of some animation

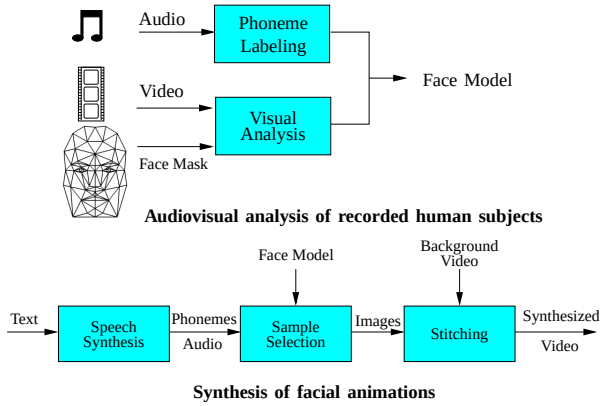


Fig. 2. Overview of analysis and synthesis stage for sample-based face animation.

results (Section 4).

## II. Image-based facial animation system

Our image-based facial animation system consists of two main parts (Fig. 2): Audiovisual analysis of a recorded human subject and synthesis of facial animation.

In the analysis part a database with images of deformable facial parts of the human subject is created. The input of the visual analysis process is a video sequence and a face mask of the recorded human subject. For positioning the face mask to the recorded human subject in the initial frame, facial features such as eye corners and nostrils have to be localized. These facial features, which are independent from local deformations such as a moving yaw or blinking eye, are selected to initially position the face mask. Furthermore, the camera is calibrated so that the intrinsic camera parameters, such as focal length, are known. Thus, only the position and orientation of the mask in the initial image must be reconstructed. This problem is known as the Perspective-n-Point problem in the computer community. We use the reliable and accurate method solving this type of problem as reported in [12]. In order to estimate the pose of the head in each frame, a gradient-based motion estimation algorithm [13] estimates the three rotation and three translation parameters. An accurate pose estimation is required in order to avoid a jerky animation.

After the motion parameters are calculated for each frame, mouth samples are normalized and stored into a database. Normalizing means to compensate for head pose variations. Each mouth sample is characterized by a number of parameters consisting of its phonetic context, original sequence and frame number. Furthermore, each sample is characterized by its visual information, which is required for the selection of samples to create animations. The visual appearance of a sample is described by its first 12 LLE coordinates and geometrical features (mouth height and width). LLE reduces the dimensionality of non-linear structure of data as given by mouth samples, so that each sample can be well characterized by a few coordinates [14]. In Fig. 3 mouth images are represented by their first two coordinates in LLE and PCA space, respectively. The classical technique for dimensionality reduction PCA discovers the true structure of data lying on or

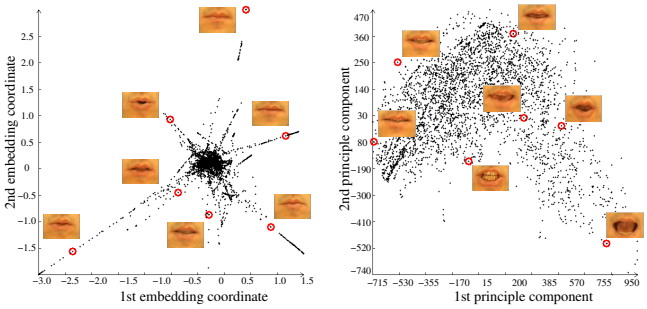


Fig. 3. Data base with mouth samples which are characterized by their first two coordinates in LLE (left image) and PCA (right image) space respectively.

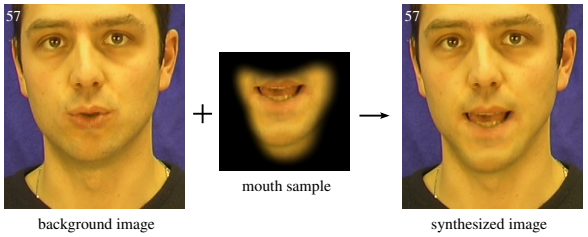


Fig. 4. A synthesized image is generated by stitching a mouth sample into a background image.

near a linear subspace of the high-dimensional input. However, the mouth images contain essential nonlinear structures that cannot be described by PCA, such that image samples are not well clustered in PCA space. Our database consists of approximately 20000 images, which is equal to 10 minutes recording time.

A face is synthesized by first generating the audio from the text using a TTS synthesizer. The TTS synthesizer sends phonemes and their duration to the unit selection engine, which chooses the best samples from the database. Then, image rendering overlay these facial parts corresponding to the generated speech over a background video sequence. Background sequences are recorded video sequences of the human subject with typical short head movements. In order to conceal illumination differences between an image of the background video and the current mouth sample, the mouth samples are blended in the background sequence using alpha-blending (Fig. 4). The mouth sample selection from the database is described in detail in the next section.

## III. Unit selection algorithm

The most important part of an image-based facial animation system is the selection of samples. This algorithm has to define the visual mouth appearance during the animation [15]. The TTS provides the unit selection system with phonemes and their duration, which are transformed into a target feature vector $T_0, ..., T_i, ..., T_N$ with $T_i$ representing a phoneme at frame i. The goal of the unit selection algorithm is to find the most appropriate mouth samples in the database for each target feature vector. The database consists of several thousands images, so that an efficient algorithm is necessary to guarantee a real time animation, which is required by many applications. Therefore, the well-known Viterbi search algorithm is selected,

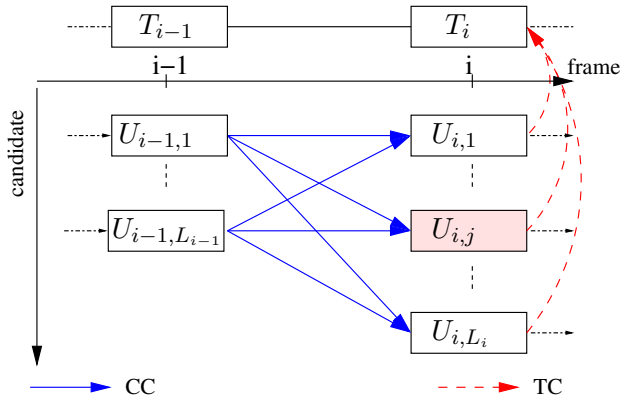


Fig. 5. Selection of mouth samples: To each target feature vector $T_i$ a list with $L_i$ candidate is assigned. To each candidate $U_{i,j}$, one target cost (TC) and all possible concatenation costs (CC) are assigned. The Viterbi search algorithm finds the path with the lowest costs.

which finds the lowest-cost path. Each target $T_i$ obtains a list of candidate images $U_{i,j}$, with i indicating the frame and j the position within the list, which are labeled with the same phoneme as the target (Fig. 5). Target costs $TC_{i,j}$ considering the phonetic context between target and sample, and concatenation costs $CC_{i,j}^{i-1,k}$ with k indicating a candidate from frame i-1 describing the visual distance between two consecutive samples are assigned to each sample $U_{i,j}$.

In order to achieve photo-realistic animations, a synchronization between lip movements and spoken output must be realized. Coarticulation effects describing the influence of consecutive phonemes onto each other do not allow a simple mapping from phoneme to mouth appearance [16]. In order to consider coarticulation effects, a phoneme feature vector: $P_i = (P_{i-n}, ..., P_i, ..., P_{i+n})$ consisting of the phonemes before and after the $i^{th}$ phoneme in the recorded sequence, is assigned to each candidate image $U_{i,j}$. The coarticulation effect lasts for a few hundred milliseconds, so that 2n frames are considered. The target cost $TC_{i,j}$ between the target $T_i$ and phoneme feature vector $P_i$ of a candidate $U_{i,j}$ is calculated as [10]:

$$TC_{i,j} = \frac{1}{\sum_{t=-n}^{n} v_{t+i}} \sum_{t=-n}^{n} v_{t+i} \cdot M(T_{i+t}, P_{i+t}) \quad (1)$$

where $v_i$ is a weight and $M(T_i, P_i)$ is a 42x42 phoneme distance matrix. The matrix is populated using the Euclidian distance in visual feature space, since each phoneme can be described by its mean visual feature vector. In this way the target costs are adapted to the visual pronunciation of each individual human subject.

The weight $v_i$ exponentially decreases with increasing distance to the current phoneme i [10]:

$$v_i = e^{\beta_1 |i-t|}, \quad i\epsilon[t-n, t+n] \quad (2)$$

with the parameter $\beta_1$.

The concatenation costs $CC_{i,j}^{i-1,k}$ describe the visual difference between two consecutive frames. Each mouth sample $U_{i,j}$ is represented by a feature vector $q$ consisting of its first 12 LLE coordinates and 2 geometric features. The difference between two candidate images is determined by calculating the Euclidian distance in visual feature space:

$$f(q_1, q_2) = ||q_1 - q_2||_{L2} \quad (3)$$

The goal of the animation is to use long snipplets of recorded video stored in the database, so that a smooth animation is inherent. Hence, segment transitions are penalized using the visual difference between two consecutive candidates as a determining factor. The cost g($u_1, u_2$) evaluates consecutive images, image repeat, frame skip and different sequences and are calculated as [10]:

$$g = \begin{cases} 0 & : \ |fn(u_1) - fn(u_2)| = 1 \wedge os(u_1) = os(u_2) \\ w_1 & : \ |fn(u_1) - fn(u_2)| = 0 \wedge os(u_1) = os(u_2) \\ w_2 & : \ |fn(u_1) - fn(u_2)| = 2 \wedge os(u_1) = os(u_2) \\ ... & : \\ w_p & : \ |fn(u_1) - fn(u_2)| \geq p \vee os(u_1) \neq os(u_2) \end{cases} \quad (4)$$

with fn and os describing the current frame number and the original sequence number, respectively, and $w_i = e^{\beta_2 i}$.

Finally both parts are combined to calculate the visual distance between two candidates [10]:

$$CC_{i,j}^{i-1,k} = g(u_1, u_2) + f(u_1, u_2) \quad (5)$$

The Viterbi algorithm determines for each candidate the lowest-cost path through the database. The accumulated cost for a candidate $U_{i,j}$ can be iteratively calculated in three steps:

1) Calculate the target cost $TC_{i,j}$.
2) Calculate the concatenation cost between unit $U_{i,j}$ and all former candidates: $CC_{i,j}^{i-1,k} \quad k\epsilon[1, ..., L_{i-1}]$.
3) Calculate the new cost $E_{i,j}^k$ and select the minimum:

$$E_{i,j}^k = E_{i-1}^k + \alpha_1 \cdot TC_{i,j} + \alpha_2 \cdot CC_{i,j}^{i-1,k} \quad k\epsilon[1, ..., L_{i-1}] \quad (6)$$

with $\alpha_1$ and $\alpha_2$ weighting the cost. Finally, the path with the lowest overall costs is selected for the facial animation.

## IV. Training the Unit Selection Algorithm

A complex issue to be addressed is the selection of the parameters $\alpha_1$ and $\alpha_2$, which strongly influence the quality of the animation. $\beta_1$ ,$\beta_2$, $p$ and n have been thoroughly investigated in [10] such that we use these parameters ($\beta_1 = 0.3$, $\beta_2 = 1.0$, $p = 5$ and n=10). However, the impact of target cost (TC) and concatenation cost (CC) on the visual quality needs to be carefully balanced by selecting appropriate values for $\alpha_1$ and $\alpha_2$. This section describes a novel approach for face animation, which has been well known in speech synthesis systems to determine the best parameters for a speech unit selection algorithm [17].

First, an objective measurement must be defined, such that our system can be trained. Since the goal of the animation is a photo-realistic facial animation, one can measure the performance of the selection algorithm by comparing recorded and its corresponding synthesized sequences. The selected recorded sequences for the training are called test sequences.

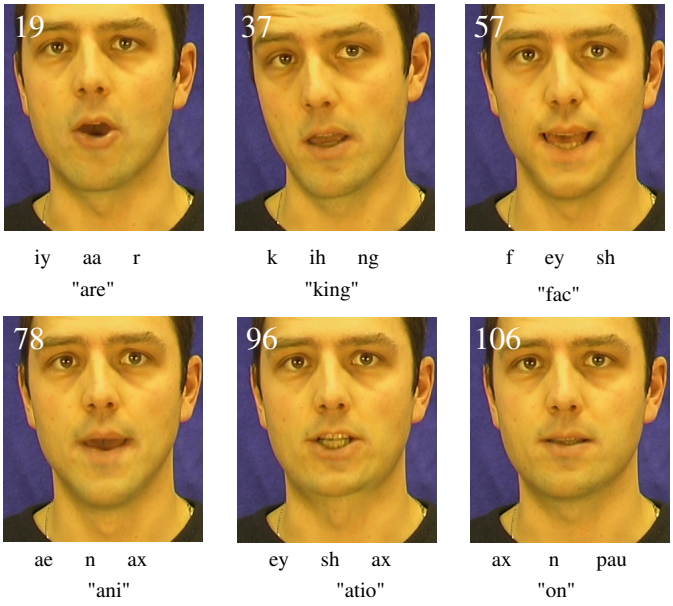


Fig. 6. Snapshots of the animation for the utterance "We are working on facial animation". Each image is labeled with its frame number and its phonetic context. For better understanding also the part of the word belonging to the phonetic context is given.

Each mouth sample in test $S_i^{test}$ and synthetic $S_i^{syn}$ sequence with i indicating the current frame is characterized by its visual feature vector $q$ as described previous. These feature vectors are taken into account to determine the similarity between two samples. The test sequences are phonetically aligned, so that this information can be used as the input to generate an animated sequence. Note that the mouth samples from the test sequences are not stored in the database, so that other samples have to be selected. The quality is objectively measured by determining the Euclidian distance in visual feature space (Eq. 3). The average distance between a test and synthetic sequence with N frames is calculated as:

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} ||q_i^{test} - q_i^{syn}||_{L2} \tag{7}$$

Ten test sequences were considered to determine the quality of a parameter set, such that the average distance in visual feature space is taken into account. Brute force is used to find the best parameters $\alpha_1$ and $\alpha_2$. This approach requires a high computational effort, which is acceptable, since the calculation is only done once for each human subject. This approach guarantees to find the minimum visual distance between test and synthetic sequences. Furthermore, this optimisation is a fully automatic process and can be operated easily by any person nonexperienced in this field.

The simulations showed that the following parameters, $\alpha_1 = 1.18$ and $\alpha_2 = 0.98$, lead to the smallest distance in visual space between test and synthetic sequences. A facial animation generated with these parameters is shown in Fig. 6.

## V. Conclusions

For an image-based facial animation system, we discussed the unit selection algorithm, which enables the synthesis of realistic animations. Two types of costs, target and concatenation costs, are assigned to the samples of the candidate list. The first cost optimizes synchronization between spoken output and mouth appearance, while the second cost ensures a smooth transition between consecutive frames. The unit selection algorithm has to be trained on each human subject. An automatic method to train the parameters of the unit selection algorithm was evident in this paper. The trained selection algorithm chooses the appropriate mouth samples from the candidate list resulting in a photo-realistic face animation. It is also feasible for operation by a common person with less experience in facial animation.

## VI. Acknowledgements

This paper is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

## References

[1] I. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," The Visual Computer, vol. 15, Issue 7/8, 1999.

[2] J. Ostermann, "E-cogent: An electronic convincing agent?" MPEG-4 Facial Animation: The Standard, Implementation and Applications, Igor S. Pandzic (Editor), Robert Forchheimer (Editor), Wiley, Chichester, England, 2002.

[3] F. I. Parke, "Computer generated animation of faces," *Proc. ACM annual conf.*, 1972.

[4] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," *Computer Graphics*, vol. 32, no. Annual Conference Series, pp. 75–84, 1998.

[5] G. A. Kalberer and L. Van Gool, "Face animation based on observed 3D speech dynamics," in *Proceedings of Computer Animation (CA2001)*, November 2001, pp. 20–27.

[6] K. Kähler, J. Haber, H. Yamauchi, and H.-P. Seidel, "Head shop: Generating animated head models with anatomical structure," in *Proceedings of the 2002 ACM SIGGRAPH Symposium on Computer Animation*, S. N. Spencer, Ed., Association of Computing Machinery (ACM). San Antonio, USA: ACM SIGGRAPH, July 2002, pp. 55–64.

[7] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," Proc. ACM SIGGRAPH 97, in Computer Graphics Proceedings, Annual Conference Series, 1997.

[8] T. Ezzat and T. Poggio, "Miketalk: A talking facial display based on morphing visemes," Proc. IEEE Computer Animation, pp. 96-102, 1998.

[9] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," Proc. ACM SIGGRAPH, pp. 388-397, 2002.

[10] E. Cosatto and H. Graf, "Sample-based synthesis of photo-realistic talking heads," Proc. IEEE Computer Animation, pp. 103-110, 1998.

[11] ——, "Photo-realistic talking heads from image samples," IEEE Trans. on Multimedia, vol. 2, no. 3, pp. 152-163, 2000.

[12] D. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," in *European Conference on Computer Vision*, 1992, pp. 335–343.

[13] J. Xiao, T. Kanade, and J. F. Cohn, "Robust full-motion recovery of head by dynamic templates and re-registration techniques," in *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, 2002, p. 163.

[14] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[15] E. Cosatto and H. Graf, "Audio-visual unit selection for the synthesis of photo-realistic talking-heads," ICME 2000, New York, NY, 2000.

[16] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. Magnenat Thalmann and D. Thalmann, Eds. Tokyo: Springer, 1994, pp. 139–156.

[17] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, vol. 1, Atlanta, Georgia, 1996, pp. 373–376.