Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 174192, 13 pages
doi:10.1155/2009/174192

Research Article

Optimization of an Image-Based Talking Head System

Kang Liu and Joern Ostermann

Institut für Informationsverarbeitung, Leibniz Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany

Correspondence should be addressed to Kang Liu, kang@tnt.uni-hannover.de

Received 25 February 2009; Accepted 3 July 2009

Recommended by Gérard Bailly

This paper presents an image-based talking head system, which includes two parts: analysis and synthesis. The audiovisual analysis part creates a face model of a recorded human subject, which is composed of a personalized 3D mask as well as a large database of mouth images and their related information. The synthesis part generates natural looking facial animations from phonetic transcripts of text. A critical issue of the synthesis is the unit selection which selects and concatenates these appropriate mouth images from the database such that they match the spoken words of the talking head. Selection is based on lip synchronization and the similarity of consecutive images. The unit selection is refined in this paper, and Pareto optimization is used to train the unit selection. Experimental results of subjective tests show that most people cannot distinguish our facial animations from real videos.

1. Introduction

The development of modern human-computer interfaces [1–3] such as Web-based information services, E-commerce, and E-learning will use facial animation techniques combined with dialog systems extensively in the future. Figure 1 shows a typical application of a talking head for E-commerce. If the E-commerce Website is visited by a user, the talking head will start a conversation with the user. The user is warmly welcomed to experience the Website. The dialog system will answer any questions from the user and send the answer to a TTS (Text-To-Speech Synthesizer). The TTS produces the spoken audio track as well as the phonetic information and their duration which are required by the talking head plug-in embedded in the Website. The talking head plug-in selects appropriate mouth images from the database to generate a video. The talking head will be shown in the Website after the right download and installation of the plug-in and its associated database. Subjective tests [4, 5] show that a realistic talking head embedded in these applications can increase the trust of humans on computer.

Generally, the image-based talking head system [1] includes two parts. One is the offline analysis, the other is the online synthesis. The analysis provides a large database of mouth images and their related information for the synthesis. The quality of synthesized animations depends mainly on the database and the unit selection.

The database contains tens of thousands of mouth images and their associated parameters, such as feature points of mouth images and the motion parameters. If these parameters are not analyzed precisely, the animations look jerky. Instead of template matching-based feature detection in [1], we use Active Appearance Models- (AAM-) based feature point detection [6–8] to locate the facial feature points, which is robust to the illumination change on the face resulted from head and mouth motions. Another contribution of our work in the analysis is to estimate the head motion using gradient-based approach [9] rather than feature point-based approach [1]. Since feature-based motion estimation [10] is very sensitive to the detected feature points, the approach is not stable for the whole sequence.

The training of image-based facial animation system is time consuming and can only find one of the possible optimal parameters [1, 11], such that the facial animation system can only achieve good quality for a limited set of sentences. To better train the facial animation system, an evolutionary algorithm (Pareto optimization) [12, 13] is chosen. Pareto optimization is used to solve a multiobjective problem, which is to search the optimal parameter sets in

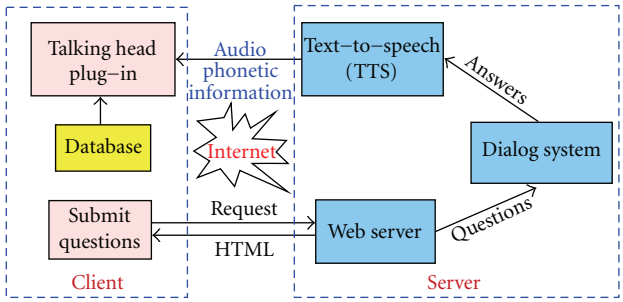


Figure 1: Schematic diagram of Web-based application with talking head for E-commerce.

the parameter space efficiently and to track many optimized targets according to defined objective criteria. In this paper, objective criteria are proposed to train the facial animation system using Pareto optimization approach.

In the remainder of this paper, we compare our approach to other talking head systems in Section 2. Section 3 introduces the overview of the talking head system. Section 4 presents the process of database building. Section 5 refines the unit selection synthesis. The unit selection will be optimized by Pareto optimization approach in Section 6. Experimental results and subjective evaluation are shown in Section 7. Conclusions are given in Section 8.

2. Previous Work

According to the underlying face model, talking heads can be categorized into 3D model-based animation and image-based rendering of models [5]. Image-based facial animation can achieve more realistic animations, while 3D-based approaches are more flexible to render the talking head in any view and under any lighting conditions.

The 3D model-based approach [14] usually requires a mesh of 3D polygons that define the head shape, which can be deformed parametrically to perform facial actions. A texture is mapped over the mesh to render facial parts. Such a facial animation has become a standard defined in ISO/IEC MPEG-4 [15]. A typical shortcoming is that the texture is changed during the animation. Pighin et al. [16] present another 3D model-based facial animation system, which can synthesize facial expressions by morphing static 3D models with textures. A more flexible approach is to model the face by 3D morphable models [17, 18]. Hair is not included in the 3D model and the model building is time consuming. Morphing static facial expressions look surprisingly realistic nowadays, whereas a realistic talking head (animation with synchronized audio) is not possible yet. The physics-based animation [19, 20] has an underlying anatomical structure such that the model allows a deformation of the head in anthropometrically meaningful ways [21]. These techniques allow the creation of subjectively pleasing animations. Due to the complexity of real surfaces, texture, and motion, talking faces are immediately identified as synthetic.

The image-based approaches analyze the recorded image sequences, and animations are synthesized by combining different facial parts. A 3D model is not necessary for animations. Bregler et al. [22] proposed a prototype called video rewrite which used triphones as the element of the database. A new video is synthesized by selecting the most appropriate triphone videos. Ezzat et al. [23] developed a multidimensional morphable model (MMM), which is capable of morphing between various basic mouth shapes. Cosatto et al. [1] described another image-based approach with higher realism and flexibility. A large database is built including all facial parts. A new sequence is rendered by stitching facial part images to the correct position in a previously recorded background sequence. Due to the use of a large number of recorded natural images, this technique has the potential of creating realistic animations. For short sentences, animations without expressions can be indistinguishable from real videos [1].

A talking head can be driven by text or speech. The text-driven talking head consists of TTS and talking head. The TTS synthesizes the audio with phoneme information from the input text. Then the phoneme information drives the talking head. The speech-driven talking head uses phoneme information from original sounds. Text-driven talking head is flexible and can be used in many applications, but the quality of speech is not so good as that of a speech-driven talking head.

The text-driven or speech-driven talking head has an essential problem, lip synchronization. The mouth movement of the talking head has to match the corresponding audio utterance. Lip synchronization is rather complicated due to the coarticulation phenomena [24] which indicate that a particular mouth shape depends not only on its own phoneme but also on its preceding and succeeding phonemes. Generally, the 3D model-based approaches use a coarticulation model with an articulation mapping between a phoneme and the model's action parameters. Image-based approaches implicitly make use of the coarticulation of the recorded speaker when selecting an appropriate sequence of mouth images. Comparing to 3D model-based animations, each frame in the image-based animations looks realistic. However, selecting mouth images, which provides a smooth movement, remains a challenge.

The mouth movement can be derived from the coarticulation property of the vocal tracts. Key-frame-based rendering interpolates the frames between key frames. For example, [25] defined the basic visemes as the key frames and the transition in the animation is based on morphing visemes. A viseme is the basic mouth image corresponding to the speech unit "phoneme", for example, the phonemes "m", "b", "p" correspond to the closure viseme. However, this approach does not take into account the coarticulation models [24, 26]. As preceding and succeeding visemes affect the vocal tracts, the transition between two visemes also gets affected by other neighbor visemes.

Recently, HMMs are used for lip synchronization. Rao et al. [27] presented a Gaussian mixture-based HMM for converting speech features to facial features. The problem is changed to estimate the missing facial feature vectors based

on trained HMMs and given audio feature vectors. Based on the joint speech and facial probability distribution, conditional expectation values of facial features are calculated as the optimal estimates for given speech data. Only the speech features at a given instant in time are considered to estimate the corresponding facial features. Therefore, this model is sensitive to noise in the input speech. Furthermore, coarticulation is disregarded in the approach. Hence, abrupt changes in the estimated facial features occur and the mouth movement appears jerky.

Based on [27], Choi et al. [28] proposed a Baum-Welch HMM Inversion to estimate facial features from speech. The speech-facial HMMs are trained using joint audiovisual observations; optimal facial features are generated directly by Baum-Welch iterations in the Maximum Likelihood (ML) sense. The estimated facial features are used for driving the mouth movement of a 3D face model. In the above two approaches, the facial features are simply parameterized by the mouth width and height. Both lack an explicit and concise articulatory model that simulates the speech production process, resulting in sometimes wrong mouth movements.

In contrast to the above models, Xie and Liu [29] developed a Dynamic Bayesian Network- (DBN)- structured articulatory model, which takes the articulator variables into account which produce the speech. The articulator variables (with discrete values) are defined as voicing (on, off), velum (open, closed), lip rounding (rounded, slightly rounded, mid, wide), tongue show (touching top teeth, near alveolar ridge, touching alveolar, others), and teeth show (on, off). After training the articulatory model parameters, an EM-based conversion algorithm converts audio to facial features in a maximum likelihood sense. The facial features are parameterized by PCA (Principal Component Analysis) [30]. The mouth images are interpolated in PCA space to generate animations. One problem of this approach is that it needs a lot of manual work to determine the value of the articulator variables from the training video clips. Due to the interpolation in PCA space, unnatural images with teeth shining through lips may be generated.

The image-based facial animation system proposed in [31] uses shape and appearance models to create realistic talking head. Each recorded video is mapped to a trajectory in the model space. In the synthesis, synthesis units are the segments extracted from the trajectories. These units are selected and concatenated by matching the phoneme similarity. A sequence of appearance images and 2D feature points are the synthesized trajectory in the model space. The final animations are created by warping the appearance model to the corresponding feature points. But the linear texture modes using PCA are unable to model nonlinear variations of the mouth part. Therefore, the talking head has a rendering problem with mouth blurring, which results in unrealistic animations.

Thus, there exists a significant need to improve coarticulatory model for lip synchronization. The image-based approach selects appropriate mouth images matching the desired values from a large database, in order to maintain the mechanism of mouth movement during speaking. Similar to the unit selection synthesis in the text-to-speech synthesizer, the resulted talking heads could achieve the most naturalness.

3. System Overview of Image-Based Talking Head

The talking head system, also denoted as visual speech synthesis, is depicted in Figure 2. First, a segment of text is sent to a TTS synthesizer. The TTS provides the audio track as well as the sequence of phonemes and their durations, which are sent to the unit selection. Depending on the phoneme information, the unit selection selects mouth images from the database and assembles them in an optimal way to produce the desired animation. The unit selection balances two competing goals: lip synchronization and smoothness of the transition between consecutive images. For each goal a cost function is defined, both of them are functions of the mouth image parameters. The cost function for lip synchronization considers the coarticulation effects by matching the distance between the phonetic context of the synthesized sequence and the phonetic context of the mouth image in the database. The cost function for smoothness reduces the visual distance at the transition of images in the final animation, favoring transitions between consecutively recorded images. Then, an image rendering module stitches these mouth images to the background video sequence. The mouth images are first wrapped onto a personalized 3D face mask and rotated and translated to the correct position on the background images. The wrapped 3D face mask is shown in Figure 3(a). Figure 3(b) shows the projection of the textured 3D mask onto a background image in a correct position and orientation. Background videos are recorded video sequences of a human subject with typical head movements. Finally the facial animation is synchronized with the audio, and a talking head is displayed.

4. Analysis

The goal of the analysis is to build a database for real time visual speech synthesizer. The analysis process is completed in two steps as shown in Figure 4. Step one is to analyze the recorded video and audio to obtain normalized mouth images and related phonetic information. Step two is to parameterize normalized mouth images. The resulted database contains the normalized mouth images and their associated parameters.

4.1. Audio-Visual Analysis. The audio-visual analysis of recorded human subjects results in a database of mouth images and their relevant features suitable for synthesis. The audio and video of a human subject reading texts of a predefined corpus are recorded. As shown in Figure 4(a), the recorded audio and video data are analyzed by motion estimation and aligner.

The recorded audio and the spoken text are processed by speech recognition to recognize and temporally align the phonetic interpretation of the text to the recorded audio data.

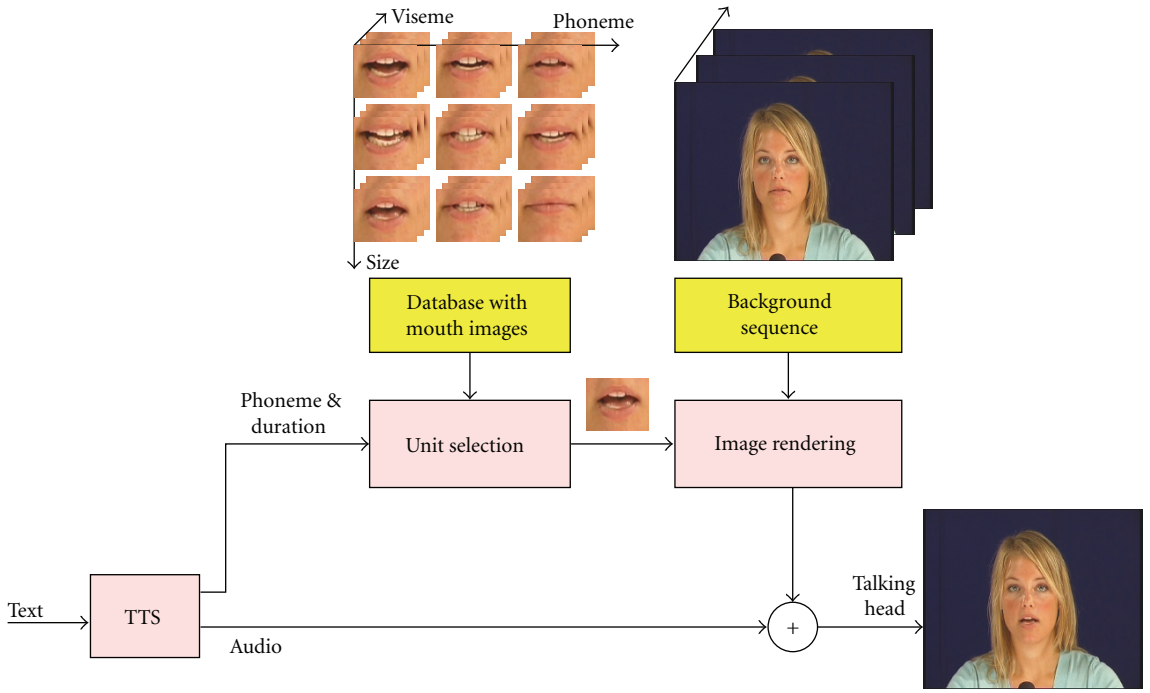
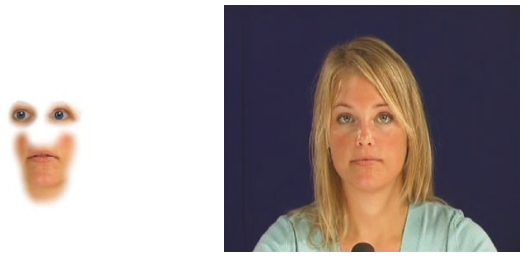


Figure 2: System architecture of image-based talking head system.



(a) (b)

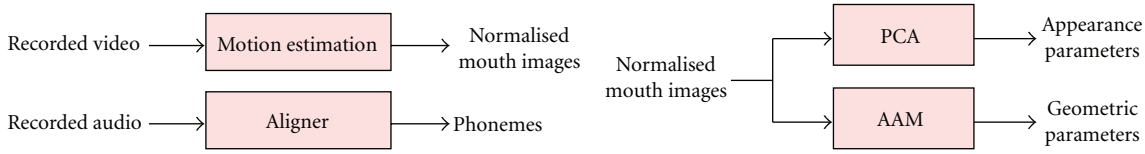
Figure 3: Image-based rendering. (a) The 3D face mask with wrapped mouth and eye textures. (b) A synthesized face by projecting the textured 3D mask onto a background image in a correct position and orientation. Alpha blending is used on the edge of the face mask to combine the 3D face mask with the background seamlessly.

Table 1: Phoneme-viseme mapping of SAPI American English Phoneme Representation. There are 43 phonemes and 22 visemes.

Viseme type no.	Phoneme	Viseme type no.	Phoneme
0	Silence	11	ay
1	ae, ax, ah	12	h, hh
2	aa	13	r
3	ao	14	l
4	ey, eh, uh	15	s, z
5	er	16	sh, ch, jh, zh
6	iy, y, ih, ix	17	th, dh
7	w, uw	18	f, v
8	ow	19	d, t, n
9	aw	20	k, g, ng
10	oy	21	p, b, m

The process is referred to aligner. Finally, the timed sequence of phonemes is aligned up to the sampling rate of the corresponding video. Therefore, for each frame of the recorded video, the corresponding phoneme and phoneme context are known. The phonetic context is required due to the coarticulation, since a particular mouth shape depends not only on its associated phoneme but also on its preceding and succeeding phonemes. Table 1 shows the American English phoneme and viseme inventory that we use to phonetically transcribe the text input. The mapping of phoneme to viseme is based on the similarity of the appearance of the mouth. In our system, we define 22 visemes including 43 phoneme from American English Phoneme Representation of Microsoft Speech API (version SAPI 5.1).

The head motion of the recorded videos is estimated and the mouth images are normalized. A 3D face mask is adapted to the first frame of the video using the calibrated camera parameters and 6 facial feature points (4 eye corners and 2 nostrils). Gradient-based motion estimation approach [9] is carried out to compute the rotation and translation parameters of the head movement in the later frames. These motion parameters are used to compensate head motion such that normalized mouth images can be parameterized by PCA correctly.



(a) Audio-visual analysis of recorded human subjects

(b) Parameterization of the normalized mouth images

FIGURE 4: Database building by analysis of recorded human subject. (a) Analysis of recorded video and audio. (b) Parameterization of the normalized mouth images.

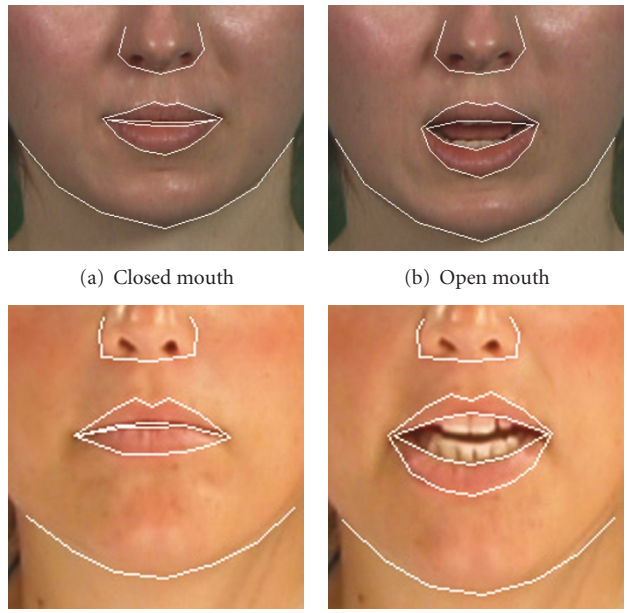
4.2. Parameterization of Normalized Mouth Images. Figure 4(b) shows the parameterization of mouth images. As PCA transforms the mouth image data into principal component space, reflecting the original data structure, we use PCA parameters to measure the distance of the mouth images in the objective criteria for system training. In order to maintain the system consistency, PCA is also used to parameterize the mouth images to describe the texture information.

The geometric parameters, such as mouth corner points and lip position, are obtained by template matching-based approach in the reference system [1]. This method is very sensitive to the illumination change resulted from mouth movement and head motion during speaking, even though the environment lighting is consistent in the studio. Furthermore, the detection error of the mouth corners may be less accurate when the mouth is very wide open. The same problem exists also in the detection of eye corners, which will result in an incorrect motion estimation and normalization.

In order to detect stable and precise feature points, AAM-based feature point detection is proposed in [8]. AAM-based feature detection uses not only the texture but also the shape of the face. AAM models are built from a training set including different appearances. The shape is manually marked. Because the AAM is built in a PCA space, if there are enough training data that can construct the PCA space, the AAM is not sensitive to the illumination change on the face. Typically the training data set consists about 20 mouth images.

The manual landmarked feature points in the training set are also refined by AAM building [8]. The detection error is reduced to 0.2 pixels, which is calculated by measuring the Euclidean distance between the manual marked feature points and detected feature points. Figure 5 shows the AAM-based feature detection used for the test data [32] (Figures 5(a) and 5(b)) and the data from our Institute (Figures 5(c) and 5(d)). We define 20 feature points on the inner and outer lip contours.

All the parameters associated with an image are also saved in the database. Therefore, the database is built with a large number of normalized mouth images. Each image is characterized by geometric parameters (mouth width and height, the visibility of teeth, and tongue), texture parameters (PCA parameters), phonetic context, original sequence, and frame number.



(a) Closed mouth

(b) Open mouth

(c) Closed mouth

(d) Open mouth

FIGURE 5: AAM-based feature detection on normalized mouths of different databases.

5. Synthesis

5.1. Unit Selection. The unit selection selects the mouth images corresponding to the phoneme sequence, using a target cost and a concatenation cost function to balance lip synchronization and smoothness. As shown in Figure 6, the phoneme sequence and audio data are generated by the TTS system. For each frame of the synthesized video a mouth image should be selected from the database for the final animation. The selection is executed as follows.

First, a search graph is built. Each frame is populated with a list of candidate mouth images that belong to the viseme corresponding to the phoneme of the frame. Using a viseme instead of a phoneme increases the number of valid candidates for a given target, given the relatively small database. Each candidate is fully connected to the candidates of the next frame. The connectivity of the candidates builds a search graph as depicted in Figure 6. Target costs are assigned to each candidate and concatenation costs are assigned to each connection. A Viterbi search through the graph finds the optimal path with minimal total cost. Given in Figure 6,

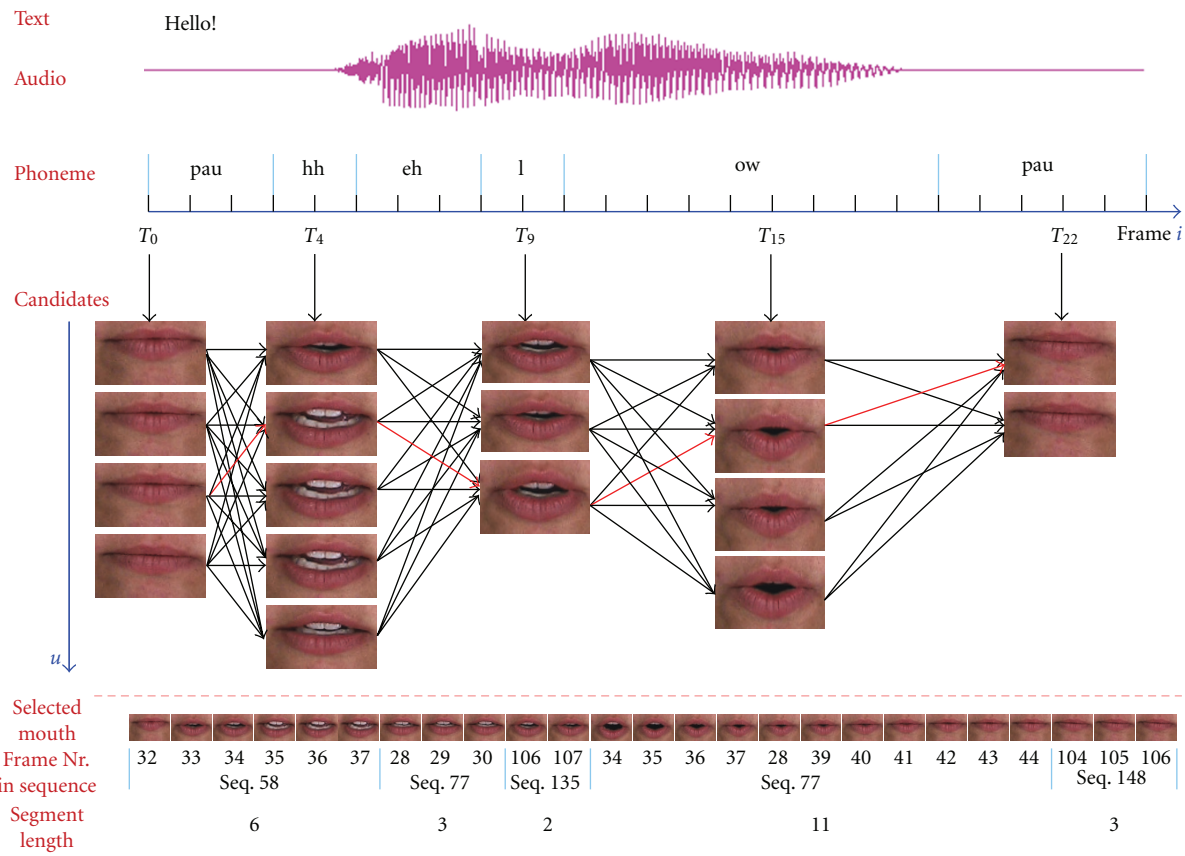


FIGURE 6: Illustration of unit selection algorithm. The text is the input of the TTS synthesizer. The audio and phoneme are the output of the TTS synthesizer. The candidates are from the database and the red path is the optimal animation path with a minimal total cost found by Viterbi search. The selected mouths are composed of several original video segments.

the selected sequence is composed of several segments. The segments are extracted from the recorded sequence. Lip synchronization is achieved by defining target costs that are small for images recorded with the same phonetic context as the current image to be synthesized.

The Target Cost (TC) is a distance measure between the phoneme at frame i and the phoneme of image u in the candidate list:

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

where a target phoneme feature vector

$$\vec{T_i} = (T_{i-n}, \ldots, T_i, \ldots, T_{i+n}) \quad (2)$$

with T_i representing the phoneme at frame i, a candidate phoneme feature vector

$$\vec{P_u} = (P_{u-n}, \ldots, P_u, \ldots, P_{u+n}) \quad (3)$$

consisting of the phonemes before and after the uth phoneme in the recorded sequence and a weight vector

$$\vec{v_i} = (v_{i-n}, \ldots, v_i, \ldots, v_{i+n}) \quad (4)$$

with $v_i = e^{\beta_1 |i-t|}$, $i \in [t-n, t+n]$, n is phoneme context influence length, depending on the speaking speed and the frame rate of the recorded video, we set $n = 10$, if the frame rate is 50 Hz, $n = 5$ at 25 Hz. β_1 is set to -0.3. M is a phoneme distance matrix with size of 43×43, which denotes visual similarities between phoneme pairs. M is computed by weighted Euclidean distance in the PCA space:

$$M\left(\mathrm{Ph}_i, \mathrm{Ph}_j\right) = \frac{\sqrt{\sum_{k=1}^{K} \gamma_k^2 \cdot \left(\overline{\mathrm{PCA}}_{\mathrm{Ph}_i,k} - \overline{\mathrm{PCA}}_{\mathrm{Ph}_j,k}\right)^2}}{\sum_{k=1}^{K} \gamma_k}, \quad (5)$$

where $\overline{\mathrm{PCA}}_{\mathrm{Ph}_i}$ and $\overline{\mathrm{PCA}}_{\mathrm{Ph}_j}$ are the average PCA weights of phoneme i and j, respectively. K is the reduced dimension of the PCA space of mouth images. γ_k is the weight of the kth PCA component, which describes the discrimination of the components, we use exponential factor $\gamma_k = e^{\beta_2 |k-K|}$, $k \in [1, K]$, with $\beta_2 = 0.1$ and $K = 12$.

The Concatenation Cost (CC) is calculated using a visual cost (f) and a skip cost (g) as follows:

$$\mathrm{CC}(u_1, u_2) = \mathrm{wccf} \cdot f(U_1, U_2) + \mathrm{wccg} \cdot g(u_1, u_2) \quad (6)$$

with the weights wccf and wccg. Candidates, u_1 (from frame i) and u_2 (from frame $i-1$), have a feature vector U_1 and U_2 of the mouth image considering the articulator features

including teeth, tongue, lips, appearance, and geometric features.

The visual cost measures the visual difference between two mouth images. A small visual cost indicates that the transition is smooth. The visual cost f is defined as

$$f(U_1, U_2) = \sum_{d=1}^{D} k_d \cdot \left\| U_1^d - U_2^d \right\|_{L2}, \tag{7}$$

where $\|U_1^d - U_2^d\|_{L2}$ measures the Euclidean distance in the articulator feature space with D dimension. Each feature is given a weight k_d which is proportional to its discrimination. For example, the weight for each component of PCA parameters is proportional to its corresponding eigenvalue of PCA analysis.

The skip cost is a penalty given to the path consisting of many video segments. Smooth mouth animations favor long video segments with few skips. The skip cost g is calculated as

$$g(u_1, u_2) = \begin{cases} 0, & |f(u_1) - f(u_2)| = 1 \wedge s(u_1) = s(u_2), \\ w_1, & |f(u_1) - f(u_2)| = 0 \wedge s(u_1) = s(u_2), \\ w_2, & |f(u_1) - f(u_2)| = 2 \wedge s(u_1) = s(u_2), \\ \vdots & \\ w_p, & |f(u_1) - f(u_2)| \geq p \vee s(u_1) \neq s(u_2) \end{cases} \tag{8}$$

with f and s describing the current frame number and the original sequence number that corresponds to a sentence in the corpus, respectively, and $w_i = e^{\beta_3 i}$. We set $\beta_3 = 0.6$ and $p = 5$.

A path $(p_1, p_2, \ldots, p_i, \ldots, p_N)$ through this graph generates the following Path Cost (PC):

$$\mathrm{PC} = \mathrm{wtc} \cdot \sum_{i=1}^{N} \mathrm{TC}\left(i, S_{i,p_i}\right) + \mathrm{wcc} \cdot \sum_{i=1}^{N} \mathrm{CC}\left(S_{i,p_i}, S_{i-1,p_{i-1}}\right) \tag{9}$$

with candidate S_{i,p_i} belonging to the frame i. wtc and wcc are the weights of two costs.

Substituting (6) in (9) yields

$$\mathrm{PC} = \mathrm{wtc} \cdot C1 + \mathrm{wcc} \cdot \mathrm{wccf} \cdot C2 + \mathrm{wcc} \cdot \mathrm{wccg} \cdot C3 \tag{10}$$

with

$$\begin{aligned} C1 &= \sum_{i=1}^{N} \mathrm{TC}\left(i, S_{i,p_i}\right), \\ C2 &= \sum_{i=1}^{N} \left(f\left(S_{i,p_i}, S_{i-1,p_{i-1}}\right)\right), \\ C3 &= \sum_{i=1}^{N} \left(g\left(S_{i,p_i}, S_{i-1,p_{i-1}}\right)\right). \end{aligned} \tag{11}$$

These weights should be trained. In [33] two approaches are proposed to train the weights of the unit selection for a speech synthesizer. In the first approach, weight space search is to search a range of weight sets in the weight space and find the best weight set which minimize the difference between the natural waveform and the synthesized waveform. In the second approach, regression training is used to determine the weights for the target cost and the weights for the concatenation cost separately. Exhaustive comparison of the units in the database and multiple linear regression are involved. Both methods are time consuming and the weights are not globally optimal. An approach similar to weight space search is presented in [11], which uses only one objective measurement to train the weights of the unit selection. However, other objective measurements are not optimized. Therefore, these approaches are only suboptimal for training the unit selection, which has to create a compromise between partially opposing objective quality measures. Considering multiobjective measurements, a novel training method for optimizing the unit selection is presented in the next section.

5.2. Rendering Performance. The performance of visual speech synthesis depends mainly on the TTS synthesizer, the unit selection, and the OpenGL rendering of the animations. We have measured that the TTS synthesizer has about 10 ms latency in a WLAN network. The unit selection is running as a thread, which only delay the program at the first sentence. The unit selection for the second sentence is run when the first sentence is rendered. Therefore, the unit selection is done in real time. The OpenGL rendering takes the main time of the animations, which relies on the graphics card. For our system (CPU: AMD Athlon XP 1.1 GHz, Graphics card: NVIDIA Geforce FX 5200), the rendering needs only 25 ms for each frame of a sequence with CIF format at 25 fps.

6. Unit Selection Training by Pareto Optimization

As discussed in Section 5.1, several weights, influencing TC, CC, and PC, should be trained. Generally, the training set includes several original recorded sentences (as ground truth) which are not included in the database. Using the database, an animation will be generated using the given weights for unit selection. We use objective evaluator functions as Face Image Distance Measure (FIDM). The evaluator functions are average target cost, average segment length, average visual difference between segments. The average target cost indicates the lip synchronization, the average segment length and average visual difference indicate the smoothness.

6.1. Multiobjective Measurements. A mouth sequence $(p_1, p_2, \ldots, p_i, \ldots, p_N)$ with minimal path cost is found by the Viterbi search in the unit selection. Each mouth has a target cost (TC_{p_i}) and a concatenation cost including a visual cost and a skip cost in the selected sequence.

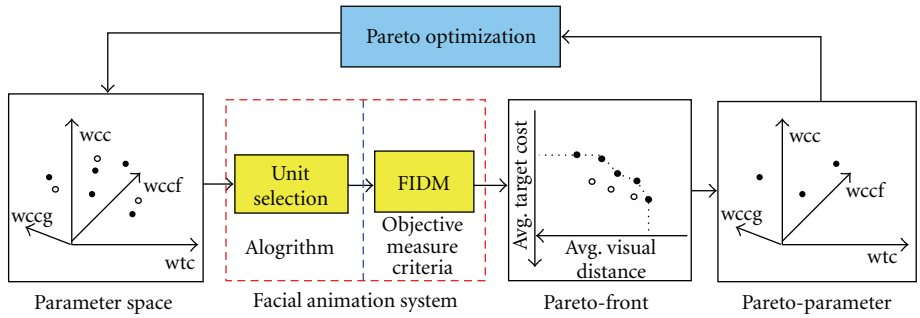


Figure 7: The Pareto optimization for the unit selection.

The average target cost is computed as

$$\mathrm{TC}_{\mathrm{avg.}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{TC}_{p_i}. \tag{12}$$

As mentioned before, the animated sequence is composed of several original video segments. We assume that there are no concatenation costs in the mouth image segment, because they are consecutive frames in a recorded video. The concatenation costs occur only at the joint position of two mouth image segments. When the concatenation costs are high, indicating a large visual difference between two mouth images, this will result in a jerky animation. The average segment length is calculated as

$$\mathrm{SL}_{\mathrm{avg.}} = \frac{1}{L}\sum_{l=1}^{L}(\mathrm{SL}_l), \tag{13}$$

where L is the number of segments in the final animation. For example, the average segment length of the animation in Figure 6 is calculated as $\mathrm{SL}_{\mathrm{avg.}} = (6+3+2+11+3)/5 = 5$.

The Euclidean distance (f_{pca}) between mouth images in the PCA space is used to calculate the average visual difference in the following way:

$$\mathrm{VC}_{\mathrm{avg.}} = \frac{1}{L-1}\sum_{i=1}^{N-1} f_{\mathrm{pca}}(i, i+1), \tag{14}$$

where $f_{\mathrm{pca}}(i, i+1)$ is the visual distance between mouth images at frame i and $i+1$ in the animated sequence. If the mouth image at frame i and $i+1$ is two consecutive frames in a original video segment, the visual distance is set to zero. Otherwise, the visual distance for the joint of the mouth image segments is calculated as

$$f_{\mathrm{pca}}(i, i+1) = \left|\left|\overrightarrow{\mathrm{PCA}_i} - \overrightarrow{\mathrm{PCA}_{i+1}}\right|\right|_{L2}, \tag{15}$$

where PCA_i is the PCA parameter of the mouth image at frame i.

6.2. Pareto Optimization of Unit Selection. Inspired in natural evolution ideas, Pareto optimization evolves a population of candidate solutions (i.e., weights), adapting them to multiobjective evaluator functions (i.e., FIDM). This process takes advantage of evolution mechanisms such as the survival of the fit test and genetic material recombination. The fit test is an evaluation process, which finds the weights that maximize the multiobjective evaluator functions. The Pareto algorithm starts with an initial population. Each individual is a weight vector containing weights to be adjusted. Then, the population is evaluated by the multiobjective evaluator functions (i.e., FIDM). A number of best weight sets are selected to build a new population with the same size as the previous one. The individuals of the new population are recombined in two steps, that is, crossover and mutation. The first step recombines the weight values of two individuals to produce two new children. The children replace their parent in the population. The second step introduces random perturbations to the weights with a given probability. Finally, a new population is obtained to replace the original one, starting the evolutionary cycle again. This process stops when a certain finalization criteria is satisfied.

FIDM is used to evaluate the unit selection and the Pareto optimization accelerates the training process. The Pareto optimization (as shown in Figure 7) begins with thousand combinations of weights of the unit selection in the parameter space, where ten settings were chosen for each of the four weights in our experiments. For each combination, there is a value calculated using the FIDM criteria. The boundary of the optimal FIDM values is called Pareto-front. The boundary indicates the animation with smallest possible target cost given a visual distance between segments. Using the Pareto parameters corresponding to the Pareto-front, the Pareto optimization generates new combinations of the weights for further FIDM values. The optimization process is stopped as soon as the Pareto-front is declared stable.

Once the Pareto-front is obtained, the best weights combination is located on the Pareto-front. The subjective test is the ultimate way to find the best weights combination, but there are many weight combinations performing similar results that subjects cannot distinguish. Therefore, it is necessary to define objective measurements to find the best weight combination automatically and objectively.

The measurable criteria consider the subjective impression of quality. We have performed the following objective evaluations. The similarity of the real sequence and the animated sequence is described by directly comparing the

visual parameters of the animated sequence with the real parameters extracted from the original video. We use the cross-correlation of the two visual parameters as the measure of similarity. The visual parameters are the size of open mouth and the texture parameter.

Appearance similarity is defined as the correlation coefficient (r_{pca}) of the PCA weights trajectory of the animated sequence and the original sequence. If the unit selection finds a mouth sequence, which is similar to the real sequence, the PCA parameters of the corresponding images of the two sequences have a high correlation. Movement similarity is defined as the correlation coefficient (r_h) of the mouth height. If the mouth in the animated sequence moves realistic just as in the real sequence, the coefficient approaches 1. The cross-correlation is calculated as

$$r = \frac{\sum_{i=1}^{N}\left[(x_i - m_x)\cdot\left(y_i - m_y\right)\right]}{\sqrt{\sum_{i=1}^{N}(x_i - m_x)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(y_i - m_y\right)^2}}, \quad (16)$$

where x_i and y_i are the first principal component coefficient of PCA parameter or the mouth height of the mouth image at frame i in the real and animated sequence, respectively. m_x and m_y are the means of the corresponding series, x and y. N is the total number of frames of the sequence.

7. Experimental Results

7.1. Data Collection. In order to test our talking head system, two data sets are used, comprising the data from our Institute (TNT) and the data from LIPS2008 [32].

In our studio a subject is recorded while reading a corpus including about 300 sentences. A lighting system is designed and developed for an audio-visual recording with high image quality [34], which minimizes the shadow on the face of an subject and reduces the change of illumination in the recorded sequences. The capturing is done using an HD camera (Thomson LDK 5490). The video format is originally 1280×720 at 50 fps, which is cropped to 576×720 pixels at 50 fps. The audio signal is sampled at 48 kHz. 148 utterances are selected to build a database to synthesize animations. The database contains 22 762 normalized mouth images with a resolution of 288×304.

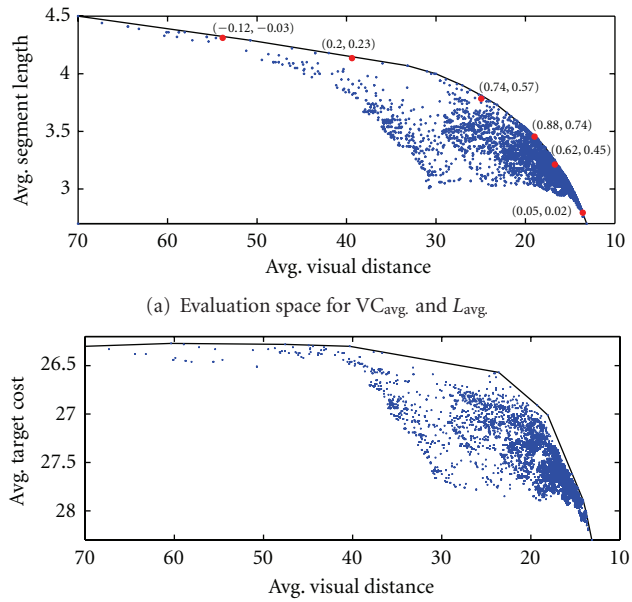
The database from LIPS2008 consists of 279 sentences, supporting the phoneme transcription of the texts. The video format is 576×720 at 50 fps. 180 sentences are selected to build a database for visual speech synthesis. The database contains 36 358 normalized mouth images with a resolution of 288×288.

A snapshot of example images extracted from two databases is shown in Figure 8.

7.2. Unit Selection Optimization. The unit selection is trained by Pareto optimization with 30 sentences. The Pareto-front is calculated and shown in Figure 9. There are many weight combinations satisfying the objective measurement on the Pareto-front, but only one combination of weights is determined as the best set of weights for unit selection. We have tried to generate animations by using several weight

(a) TNT (b) LIPS2008

Figure 8: Snapshot of an example image extracted from recorded videos at TNT and LIPS2008, respectively.



(a) Evaluation space for $\text{VC}_{\text{avg.}}$ and $L_{\text{avg.}}$

(b) Evaluation space for $\text{VC}_{\text{avg.}}$ and $\text{TC}_{\text{avg.}}$

Figure 9: Pareto optimization for unit selection. The curves are the Pareto-front. Several Pareto points on the Pareto-front marked red are selected to generate animations. The cross-correlation coefficients of PCA parameters and mouth height (r_{pca}, r_h) between real and animated sequences are shown for the selected Pareto points.

combinations and find out that they have similar quality subjectively in terms of naturalness, because quite different paths through the graph can produce very similar animations given a quite large database.

To evaluate the Pareto-front automatically, we use the defined objective measurements to find best animations with respect to naturalness. The cross-correlation coefficients of PCA parameter and mouth height between real and animated sequences on the Pareto-front are calculated and shown in Figure 10. The red curve is the cross-correlation of PCA parameter of mouth images between real and animated

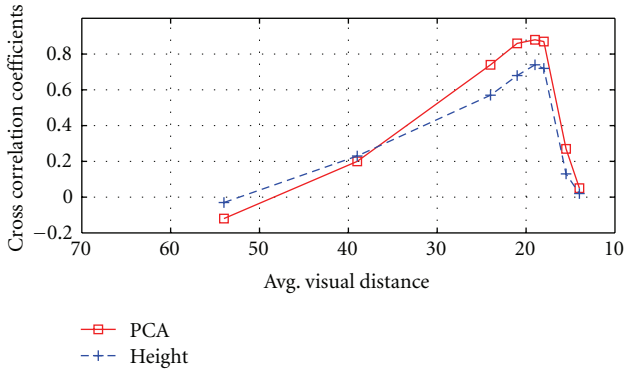
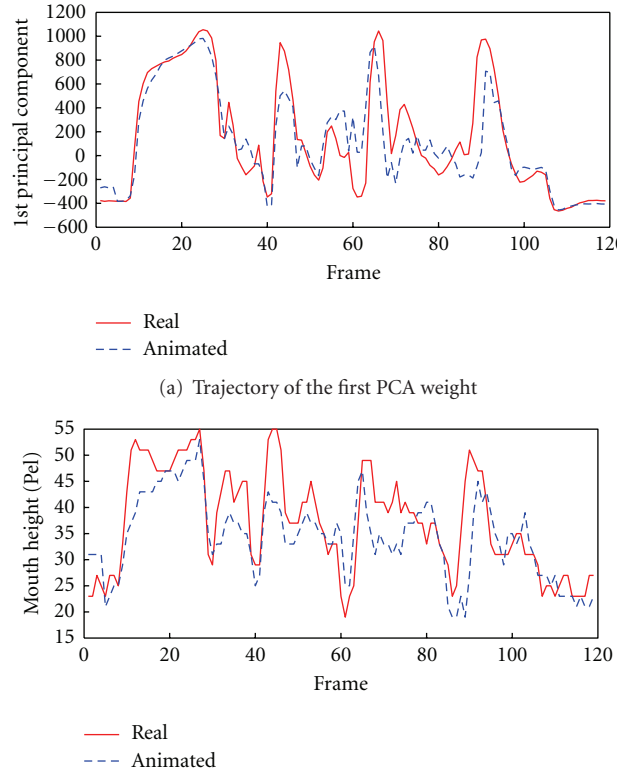


FIGURE 10: Cross-correlation of PCA parameters and mouth height of mouth images between real and animated sequences on the Pareto-front. Red curve is cross-correlation of PCA parameter between real and animated sequences. The blue one is the cross-correlation of mouth height.



(a) Trajectory of the first PCA weight

(b) Trajectory of mouth height of real and animated sequences

FIGURE 11: The similarity measurement for the sentence: I want to divide the talcum powder into two piles. (a) shows the appearance similarity, (b) shows the mouth movement similarity. The red curve is the PCA parameter trajectory and the mouth movement of the real sequence; the blue curve is the PCA parameter trajectory and mouth movement of the animated sequence. The cross-correlation coefficient of PCA parameters between the real and animated sequence is 0.88, the coefficient for mouth height is 0.74. The mouth height is defined as the maximal top to bottom distance of the outer lip contour.

sequences. The blue curve is the cross-correlation of mouth height. The cross-correlation coefficients of several Pareto points on Pareto-front are labeled in Figure 9(a), where the first coefficient is r_{pca}, the second is r_h. Given in Figure 10, the appearance similarity (red curve) and the movement similarity (blue curve) run in a similar way, which reach the maximal cross-correlation coefficients at the same position with the average visual distance of 18.

Figure 11(a) shows the first component of PCA parameters of mouth images in real and animated sequences. The mouth movements of the real and synthesized sequences are shown in Figure 11(b). We have found that the curves in Figure 11 do not match perfectly, but they are highly correlated. The resulting facial animations look realistic compared to the original videos. One of the most important criteria to evaluate the curves is to measure how well the closures match in terms of timing and amplitude. Furthermore, objective criteria and informal subjective tests are consistent to find the best weights in the unit selection. In such a way the optimal weight set is automatically selected by the objective measurements.

The weight set corresponding to the point on the Pareto-front with maximal similarity are used in the unit selection. Animations generated by the optimal facial animation system are used for the following formal subjective tests.

7.3. Subjective Tests. A subjective test is defined and carried out to evaluate the facial animation system. The goal of the subjective test is to assess the naturalness of animations whether they can be distinguished from real videos.

Assessing the quality of a talking head system becomes even more urgent as the animations become more lifelike, since improvements may be more subtle and subjective. A subjective test where observers give feedback is the ultimate measure of quality, although objective measurements used by the Pareto optimization can greatly accelerate the development and also increase the efficiency of subjective tests by focusing them on the important issues. Since a large number of observers is required, preferably from different demographic groups, we designed a Website for subjective tests.

In order to get a fair subjective evaluation, let the viewers focus on the lips and separate the different factors, such as head motions and expressions, influencing the speech perception, we selected a short recorded video with neutral expressions and tiny head movements as the background sequence. The mouth images, which are cropped from a recorded video, are overlaid to the background sequence in a correct position and orientation to generate a new video, named original video. The corresponding real audio is used to generate a synthesized video by the optimized unit selection. Thus a pair of videos, uttering the same sentence, are ready for subjective tests. Overall 5 pairs of original and synthesized videos are collected to build a video database available for subjective tests on our Website. The real videos corresponding to the real audios are not part of the database.

A Turing test was performed to evaluate our talking head system. 30 students and employees of Leibniz University of

Table 2: Results of the subjective tests for talking heads by using TNT database. 5 video pairs were shown to 30 viewers. The number of the viewers, which identified the real and synthesized video correctly (NCI), was counted. The correct identifying rate (CIR) for each video pair was calculated.

Video pair	1	2	3	4	5
NCI	21	16	17	11	21
NTS	30	30	30	30	30
CIR	70%	53%	57 %	37 %	70%

Hanover were invited to take part in the formal subjective tests. All video pairs from the video database were randomly selected and the video pair was itself presented to the participant randomly only once. The participant should decide whether it is an original or a synthesized video immediately after the video pair was displayed.

The results of the subjective tests are summarized in Table 2. The Turing test can be quantified in terms of the Correct Identifying Rate (CIR), which is defined as

$$\text{CIR} = \frac{\text{Number of correctly identified utterances (NCIs)}}{\text{Number of testing utterances (NTSs)}}. \tag{17}$$

Table 2 shows the results of subjective tests. CIR 50% is expected, which means that the animations are as realistic as the real one. From the results of the subjective tests, we can find that the original videos of video pairs 1 and 5 are correctly recognized by 70% of the viewers. The video pairs 2 and 3 are almost indistinguishable to the viewers, where the CIR is approaching 50%. The synthesized video of video pair 4 is decided by most viewers as original video.

Our hypothesis is that original and animated videos are indistinguishable from each other. If the hypothesis is true, the value for NCI is binomially distributed. The probability mass function of binomial distribution is defined in the following way:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{18}$$

with parameters $n = \text{NTS} = 30$, $k = \text{NCI}$, and $P = .5$ for our subjective tests. Figure 12 shows the binomial distribution of the subjective tests. The 95% confidence interval is estimated in the zone between 10 and 20. The video pairs 2, 3, and 4 are kept in the confidence interval, which means that the video pairs are indistinguishable. The video pairs 1 and 5 are outside of the confidence interval, but they are very close to the confidence level. In fact, these video pairs are very difficult to be distinguished according to the feedback of the viewers in the subjective tests.

The generated talking heads using LIPS 2008 database were evaluated on the conference of Interspeech 2008. In comparison to other attended systems [35], our proposed talking head system achieved the most audio-visual consistency in terms of naturalness. The Mean Opinion Score (MOS) of our system was about 3.7 in the subjective test

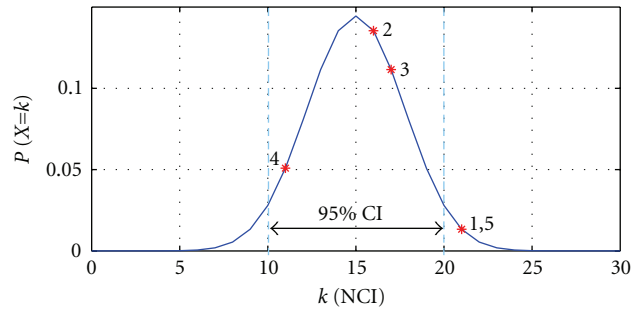


Figure 12: Binomial distribution ($n = 30$, $P = .5$) of the subjective tests. The video pairs are marked with red on the distribution.

evaluated by a 5-point grading scale (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad). The original videos were scored with about 4.7.

The subjective tests carried out in our institute show that the talking head generated by using the database of TNT performs better than the talking head generated by using the database of LIPS2008. A reason for the better animation results is the designed light settings resulting in a high quality recording. All viewers think the videos from TNT look better, since the lighting contrast of the image gives a big impact on the perception of overall quality of talking heads in the subjective tests. Furthermore, the shadow and the illumination changes on the face cause problems in motion estimation, which makes the final animations jerky and blinking. Therefore, talking heads generated by using the database of LIPS2008 do not look as realistic as those heads by using the database of TNT.

Based on the facial animation system, Web-based interactive services such as E-shop and Newsreader were developed. The demos and related Website are available at http://www.tnt.uni-hannover.de/project/facialanimation/demo/. In addition, the video pairs used for the subjective tests can be downloaded from http://www.tnt.uni-hannover.de/project/facialanimation/demo/subtest/.

8. Conclusions

We have presented the optimization of an image-based talking head system. The image-based talking head system consists of an offline audio-visual analysis and an online unit selection synthesis. In the analysis part, Active Appearance Models (AAMs) based facial feature detection is used to find geometric parameters of mouth images instead of color template-based approach that is a reference method. By doing so, the accuracy of facial features is improved to sub-pixel. In the synthesis part, we have refine the unit selection algorithm. Furthermore, optimization of the unit selection synthesis is a difficult problem because the unit selection is a nonlinear system. Pareto optimization algorithm is chosen to train the unit selection so that the visual speech synthesis is stable for arbitrary input texts. The optimization criteria include lip synchronization, visual smoothness, and others. Formal subjective tests show that synthesized animations generated by the optimized talking head system match the

corresponding audio naturally. More encouraging, 3 out of 5 synthesized animations are so realistic that the viewers cannot distinguish them from original videos.

In the future work, we are planning to record additional videos in which the subject is smiling while speaking. We hope to generate expressive talking heads by switching between the smile and the neutral mouth images.

Acknowledgments

This research work was funded by EC within FP6 under Grant 511568 with the acronym 3DTV. The authors acknowledge Holger Blume for his support with the Pareto optimization software. The authors would like to thank Tobias Elbrandt for his helpful comments and suggestions in the evaluation of the subjective tests. The authors also wish to thank all the people involved in the subjective tests.

References

[1] E. Cosatto, J. Ostermann, H. P. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *in Proceedings of the IEEE*, vol. 91, no. 9, pp. 1406–1429, 2003.

[2] K. Liu and J. Ostermann, "Realistic talking head for human-car-entertainment services," in *Proceedings of the Informationssysteme fuer Mobile Anwendungen (IMA '08)*, pp. 108–118, Braunschweig, Germany, September 2008.

[3] J. Beskow, *Talking Heads—Models and Applications for Multimodal Speech Synthesis*, Doctoral thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 2003.

[4] I. S. Pandzic, J. Ostermann, and D. R. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, no. 7-8, pp. 330–340, 1999.

[5] J. Ostermann and A. Weissenfeld, "Talking faces—technologies and applications," in *Proceedings of the International Conference on Pattern Recognition (ICPR '04)*, vol. 3, pp. 826–833, August 2004.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[7] M. B. Stegmann, B. K. Ersbøll, and R. Larsen, "FAME—a flexible appearance modeling environment," *IEEE Transactions on Medical Imaging*, vol. 22, no. 10, pp. 1319–1331, 2003.

[8] K. Liu, A. Weissenfeld, J. Ostermann, and X. Luo, "Robust AAM building for morphing in an image-based facial animation system," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '08)*, pp. 933–936, Hanover, Germany, June 2008.

[9] A. Weissenfeld, O. Urfalioglu, K. Liu, and J. Ostermann, "Robust rigid head motion estimation based on differential evolution," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 225–228, Toronto, Canada, July 2006.

[10] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, 2000.

[11] A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann, "Personalized unit selection for an image-based facial animation system," in *Proceedings of the IEEE 7th Workshop on Multimedia Signal Processing (MMSP '05)*, Shanghai, China, October 2005.

[12] E. Zitzler, M. Laumanns, and S. Bleuler, "A tutorial on evolutionary multiobjective optimization," in *Proceedings of the Multiple Objective Metaheuristics (MOMH '03)*, Springer, Berlin, Germany, 2003.

[13] J. Von Livonius, H. Blume, and T. G. Noll, "Flexible Umgebung zur Pareto-Optimierung von Algorithmen—Anwendungen in der Videosignalverarbeitung," ITG 2007.

[14] Z. Deng and U. Neumann, *Data-Driven 3D Facial Animation*, Springer, 2008.

[15] J. Ostermann, "Animation of synthetic faces in MPEG-4," in *Proceedings of the Computer Animation*, vol. 98, pp. 49–55, Philadelphia, Pa, USA, June 1998.

[16] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proceedings of the 29th ACM annual conference on Computer Graphics (SIGGRAPH'98)*, vol. 3, pp. 75–84, Orlando, Fla, USA, July 1998.

[17] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th ACM Annual Conference on Computer Graphics (SIGGRAPH '99)*, pp. 187–194, Los Angeles, Calif, USA, August 1999.

[18] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," in *Proceedings of the Computer Graphics Forum (Eurographics '03)*, vol. 22, pp. 641–650, Basel, Switzerland, November 2003.

[19] D. Terzopoulos and K. Waters, "Physically-based facial modeling analysis and animation," *Journal of Visualization and Computer Animation*, vol. 1, no. 4, pp. 73–80, 1990.

[20] K. Waters and J. Frisbie, "Coordinated muscle model for speech animation," in *Proceedings of the Graphics Interface Conference*, pp. 163–170, May 1995.

[21] K. Kaehler, J. Haber, H. Yamauchi, and H.-P. Seidel, "Head shop: generating animated head models with anatomical structure," in *Proceedings of the ACM Computer Animation Conference (SIGGRAPH '02)*, pp. 55–63, 2002.

[22] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the ACM Conference on Computer Graphics (SIGGRAPH '97)*, pp. 353–360, Los Angeles, Calif, USA, August 1997.

[23] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the ACM Transactions on Graphics (SIGGRAPH '02)*, vol. 21, no. 3, pp. 388–397, July 2002.

[24] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, M. Magnenat-Thalmann and D. Thalmann, Eds., pp. 139–156, Springer, Tokyo, Japan, 1993.

[25] T. Ezzat and T. Poggio, "MikeTalk: a talking facial display based on morphing visemes," in *Proceedings of the 7th IEEE Eurographics Workshop on Computer Animation*, pp. 96–102, 1998.

[26] N. Hewlett and W. J. Hardcastle, *Coarticulation: Theory, Data and Techniques*, Cambridge University Press, Cambridge, UK, 2000.

[27] R. R. Rao, T. Chen, and R. M. Merserau, "Audio-to-visual conversion for multimedia communication," *IEEE Transaction On Industrial Electronics*, vol. 45, no. 1, pp. 12–22, 1998.

[28] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 29, no. 1-2, pp. 51–61, 2001.

[29] L. Xie and Z.-Q. Liu, "Realistic mouth-synching for speech-driven talking face using articulatory modelling," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 500–510, 2007.

[30] I. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, USA, 1989.

[31] B. J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley, "Near-videorealistic synthetic talking faces: Implementation and evaluation," *Speech Communication*, vol. 44, no. 1–4, pp. 127–140, 2004.

[32] B. Theobald, S. Fagel, G. Bailly, and F. Elsei, "LIPS2008: visual speech synthesis challenge," in *Proceedings of the Interspeech*, pp. 2310–2313, 2008.

[33] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 1, pp. 373–376, 1996.

[34] R. Guenther, *Aufbau eines Mehrkamerastudios fuer audiovisuelle Aufnahmen*, Diplomarbeit, Leibniz University of Hannover, Hannover, Germany, February 2009.

[35] "LIPS2008: Visual Speech Synthesis Challenge," http://www.lips2008.org/.