Universität des Saarlandes
Fachrichtung 4.7 – Phonetik
Sommersemester 2002
Seminar: Audiovisuelle Sprache in der Sprachtechnologie
Seminarleitung: Dr. Jacques Koreman

image-based visual synthesis: facial overlay

Video Rewrite and AT&T's 2D photo-realistic talking head

Kurze Ausarbeitung des Vortrags vom 23. Mai 2002

Caren Brinckmann
Scheidter Straße 123
66123 Saarbrücken
e-mail: caren@brinckmann.de

Studienziel: Magister Artium
Studienfächer: Phonetik, Computerlinguistik, Germanistik

CONTENTS

1. Introduction

In the first two sessions of the course "Audiovisuelle Sprache in der Sprachtechnologie", we familiarized ourselves with *model-based* visual synthesis. In model-based systems a face is modelled as a 3D structure, which is then deformed with control parameters using geometric, articulatory, or muscular models. The two visual synthesis systems that are presented in this paper (based on the talk given on May 23, 2002) are not model-based but *image-based*. In image-based systems, new videos are generated by retrieving segments of videos of a human speaker from a database. In a second step those segments are processed so that they fit together, and finally they are concatenated.

This division into model-based and image-based visual synthesis is similar to the division into parametric and concatenative speech synthesis. In parametric systems (i.e. articulatory or formant speech synthesis) speech is generated from a model, whereas in concatenative speech synthesis (e.g. diphone synthesis and non-uniform unit-selection) a database of recordings of a human speaker is used to generate new utterances.

In the following sections of this paper, I will give an overview of the system "Video Rewrite" (Bregler et al., 1997), the motivation behind it and a detailed description of the system's architecture. Then I will compare it to AT&T's 2D photo-realistic talking head (Cosatto & Graf, 1998) and point out the differences between the two systems.[1] Since this is mainly a summary, the interested reader is referred to the two original papers for more detailed information.

2. Video Rewrite

2.1 motivation

Close-ups in dubbed movies are often disturbing because speech and lip motions are not synchronized. On the other hand, the special effects in the movie "Forrest Gump" (where the main character meets John F. Kennedy and talks to him) are so compelling because the original footage is lip synched to the movie's new soundtrack. But there is one severe drawback: a manual modification of the original lip motions is very labor-intensive.

This is where Video Rewrite comes in: In the analysis stage Video Rewrite takes existing video material of a person and a new soundtrack (either spoken by another person or by a TTS system). In the synthesis stage it automatically creates a new video of that person with lip and jaw movements that synchronize to the new audio. The result is a final video of realistic quality without labor-intensive interaction. Possible applications are photo-realistic avatars and a reduction of the data rate for video conferencing.

2.2 analysis

The material Video Rewrite analyzes during the analysis stage is unconstrained, i.e. the speaker does not have to wear bright lipstick to make the automatic recognition of the lips easier, and she can move her head around while speaking.

---

[1] Example video clips of the two systems can be found on the following webpages:
http://www.coli.uni-sb.de/~cabr/vortraege/facialoverlay/FacialOverlay.htm
http://graphics.stanford.edu/~bregler/videorewrite/
http://www.research.att.com/projects/AnimatedHead/

Most of the development of Video Rewrite was done with one subject, Ellen. Eight minutes of video were used, containing 109 sentences of a fairy-tale with 1700 different triphones (of around 19,000 naturally occurring triphones in English). In addition, two minutes of John F. Kennedy talking about the Cuban missle crisis were used, which contained 1157 triphones.

This video material was labelled on two levels: phoneme labelling and visual labelling. First, the audio track was automatically labelled with the spoken phoneme sequence. This phoneme labelling was used to split the video material into overlapping triphones. Then the frames in the triphone video clips were labelled with specific anchor points on the face, which outline the lips, teeth and jaw. During synthesis these anchor points are used to cross-fade the overlapping regions of neighboring triphone video clips and to align the lips with the background face. As shown in figure 1, the output of the analysis stage is an annotated database of triphone video clips of the mouth/jaw region.

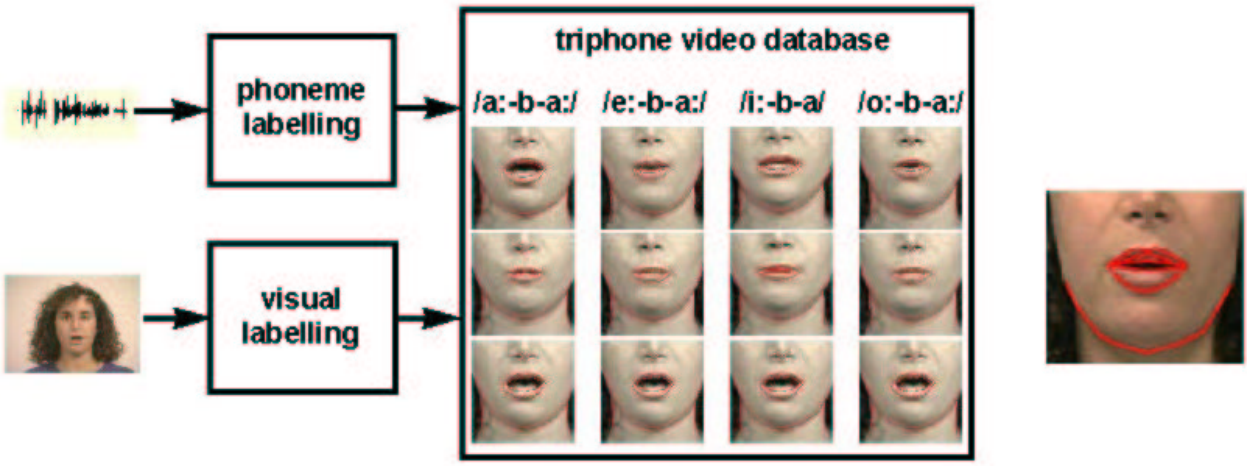

figure 1: analysis stage of Video Rewrite

2.2.1 phoneme labelling

The automatic labelling (alignment) of originally spoken phonemes works as follows: a given (unaligned) graphemic transcription of the spoken utterances is converted into a phonemic transcription by looking up the pronounciation of each word in a pronounciation dictionary.[2] Then a gender-specific system of HMMs, which were trained on the TIMIT database, is used to align this phonemic (or phonetic) transcription with the audio signal by applying forced Viterbi search. The result is a fined-grained transcription of the video's audio track.

Since a /b/[3] followed by a /u/ is very different from a /b/ followed by a /i/, it is very important to capture coarticulation. Therefore, the video signal is split into partly overlapping triphones[4], which are then stored in the triphone video database.

2.2.2 visual labelling

Manual labelling of anchor points that outline the lips, teeth and jaw is error prone and tedious.

[2] However, it remains unclear whether this canonical transcription is kept as it is, or whether it is manually adjusted to the sequence of actually realized phones.
[3] Throughout the paper the SAM Phonetic Alphabet (SAMPA) is used for transcriptions.
[4] For example, an utterance of the German word *Sahne* would consist of the following sequence of triphones: /SIL-z-a:/ + /z-a:-n/ + /a:-n-@/ + /n-@-SIL/, where SIL stands for silence (at the beginning and at the end of an utterance).

Therefore, Video Rewrite uses computer-vision techniques that identify the mouth and its shape to label these anchor points automatically. The two main problems are the low resolution of the images[5] and the low signal-to-noise ratio on the inner lip boundaries. Video Rewrite uses the eigenpoints algorithm (Covell & Bregler, 1996) to solve these problems. As shown in figure 2, the eigenpoints algorithm uses a small set[6] of hand annotated images as training data to find (or "learn") a linear mapping between the brightness of the video signal and the locations of the anchor points on the face. This learned eigenpoint model is then applied to the remaining video frames of the database to label them automatically with the anchor points.

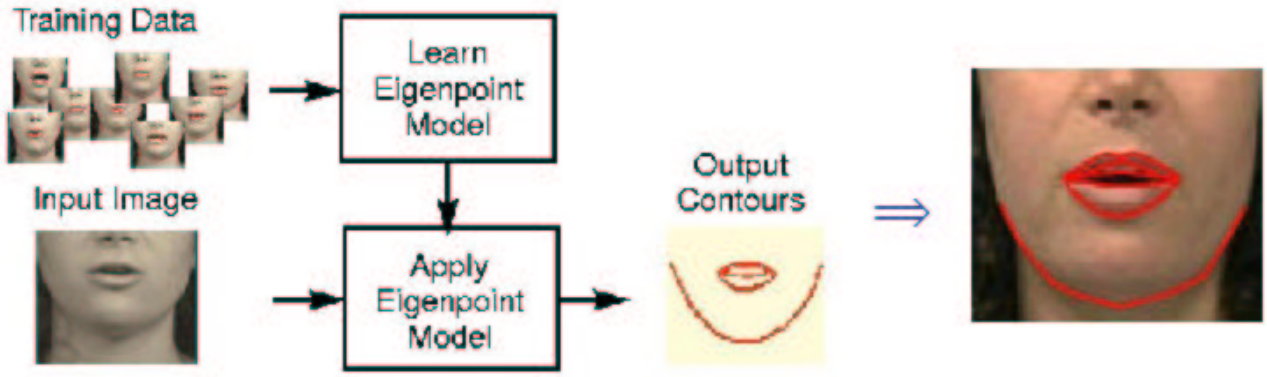

figure 2: eigenpoints algorithm

A major problem of the eigenpoints algorithm is its assumption that the features (mouth or jaw) are undergoing pure translational motion. This leads to poor results in modelling rotations and scale changes, e.g. when the speaker leans toward the camera or turns away from it, or when she tilts her head to the side. Therefore, visual labelling is a three-step method, which first warps each face image to a standard reference plane, then applies the eigenpoint model to label the anchor points, and finally warps these anchor points back onto the original face image.

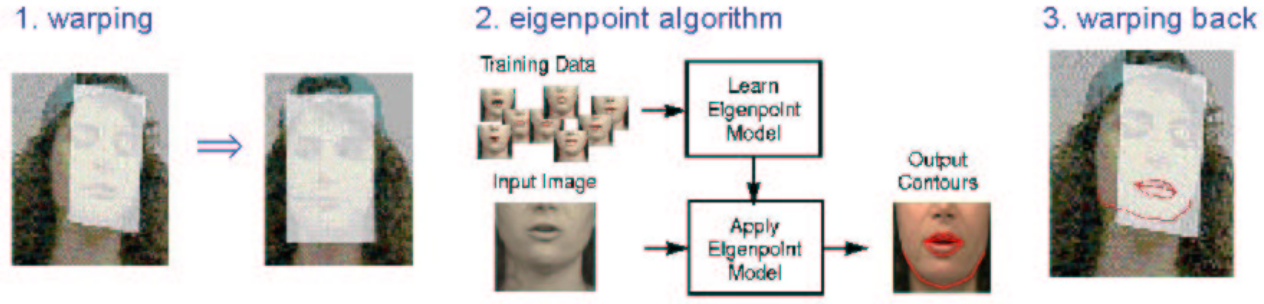

figure 3: visual labelling

2.3 synthesis

As shown in figure 4, during synthesis, first the new soundtrack is automatically labelled with the spoken phoneme sequence using the same method as during analysis. Then, the corresponding triphone video clips are selected from the database. Finally, these triphone videos are stitched into a selected background video using morphing techniques.

The background video is taken from the source video material in the original order (no rearrangement of video frames). It includes most of the speaker's face als well as the scene

[5] width of the lip region in a typical scene: 40 pixels
[6] typically between 20 and 30 images

behind the speaker, thus sets the scene and provides the desired head position and movement (e.g. eyes blink). The triphone video clips show only the motions associated with articulation: mouth, chin, and part of the cheeks. Illumination-matching techniques are used to avoid visible seams between the triphone and the background images.

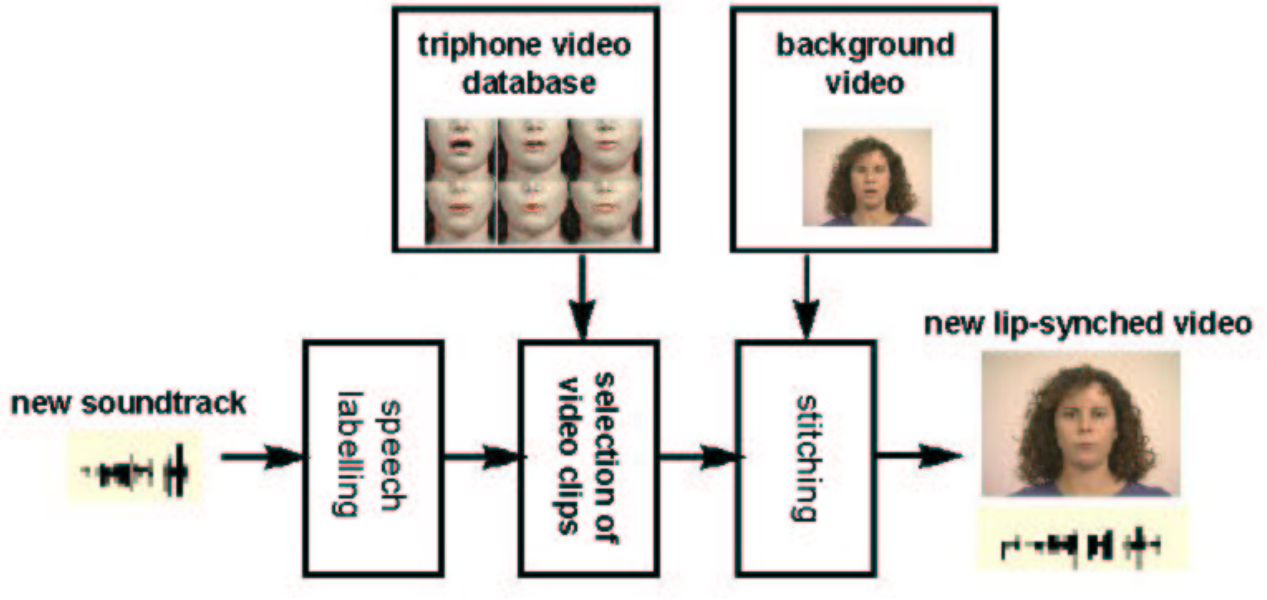

figure 4: synthesis stage of Video Rewrite

Since the first synthesis stage, the speech labelling, works exactly like in the analysis stage, the next two sections will focus on the selection of triphone video clips from the database and the final stitching process.

2.3.1 selection of triphone video clips

For each audio triphone that is needed for the new utterance, it would be ideal to find the corresponding video clip in the database with exactly the transition we need (the matching triphone video). Depending on the size of the database, this is not always possible. For example, for the speaker Ellen in 31% of the cases a non-ideal triphone had to be chosen, whereas for John F. Kennedy even 94% of the triphone videos were not ideal. So in case the ideal triphone is not in the database, the system has to choose a clip that approximates the desired transition. For this reason the system computes a matching distance between the target triphone of the new utterance and each triphone video in the database and then chooses the sequence of triphone videos with the smallest matching distance (using dynamic programming).

The matching distance is computed as follows: $\boxed{\text{total error} = \alpha\, D_P + (1 - \alpha)\, D_S}$, where D_P is the *phoneme-context distance*, D_S is the *lip shape distance*, and α is a constant that trades off D_P and D_S ($0 \le \alpha \le 1$).

D_S measures how closely the mouth contours in overlapping frames of adjacent triphone videos match by computing the Euclidean distance[7] between four-element vectors: overall lip width, overall lip height, inner lip height, and height of visible teeth.

[7] Euclidean distance between two vectors x and y = $\sqrt{\sum_i (x_i - y_i)^2}$

D_p is the weighted sum of phoneme distances between the phonemes of the target triphone and the phonemes of the triphone video in the database, which is based on categorical distances between 26 viseme classes: 15 vowel classes (one class for each vowel), 10 consonant classes[8], and one class for silence. For example, if the target triphone /z-a:-n/ is not in the database, we compute its distance to the existing triphone /d-a:-m/ by a pairwise comparison of the three phoneme-pairs:

- /z/ and /d/ are in the same viseme class → distance is between 0 and 1, e.g. 0.5
- the center phoneme is the same → distance is 0
- /n/ and /m/ in different viseme classes → distance is 1.

These three distance measures are then combined in a weighted sum, where the center phoneme has the largest weight, e.g. 1, and the neighboring phonemes have a smaller weight, e.g. 0.3. In this example we get a total difference measure of 0.3 * 0.5 + 1 * 0 + 0.3 * 1 = 0.45 for the triphone /d-a:-m/. This distance is smaller than for the triphone /z-u:-n/, which would be 0.3 * 0 + 1 * 1 + 0.3 * 0 = 1 (which shows that the center phoneme is the most important one).

This use of viseme classes seems rather suboptimal to us. On the one hand, the distance between phonemes of different viseme classes is always 1, even though the class /p b m/ might be closer to /f v/ than to /k g n l/, which could be reflected in graded differences. On the other hand, the system makes a rather fine-grained difference between different phonemes of the same viseme class. On top of that, some of the viseme classes are rather questionable from a phonetic point of view (e.g. the class /k g n l/).

2.3.2 stitching

Now the sequence of selected triphone videos has to be stitched into the background video. Three steps are necessary: First, adjacent triphone videos are time aligned with one another. This is achieved by choosing a portion of the overlapping triphone videos where the two lip shapes are as similar as possible (by minimizing D_S, the lip shape distance). The result of this first step is a self-consistent sequence of triphone videos.

In a second step, this sequence is time aligned to the new soundtrack by aligning the starting time of the center phone in the triphone video with the starting time of the center phone in the new soundtrack. The triphone videos are then stretched or compressed such that they fit the phoneme durations in the new soundtrack. This gives us a sequence of triphone videos with the correct durations.

Finally, this time aligned sequence of videos of the mouth-jaw region is combined with the background video. Since the lips and head are constantly moving in the triphone and the background video, correctness of facial alignment is crucial: any error in spatial alignment causes the mouth to jitter relative to the face. As shown in figure 5, the combination of the triphone videos with the background video is done in three steps, all of which use the derived anchor points on the face as control points in morphing. First, the overlapping frames of adjacent triphone videos are combined by linear cross-fading. The resulting video sequence is then warped onto the plane of the background video. The final stitching process is a three-way tradeoff in shape and texture among fade-out lip image, fade-in lip image, and background image, e.g. the jaw's shape is a

[8] The ten consonant classes are: /p b m/, /f v/, /w r/, /T D/, /t d s z/, /S Z tS dZ/, /k g n l/, /j/, /N/, and /h/.

combination of the background video's jaw line (near the ears) and the two triphones' jaw lines (near the chin).

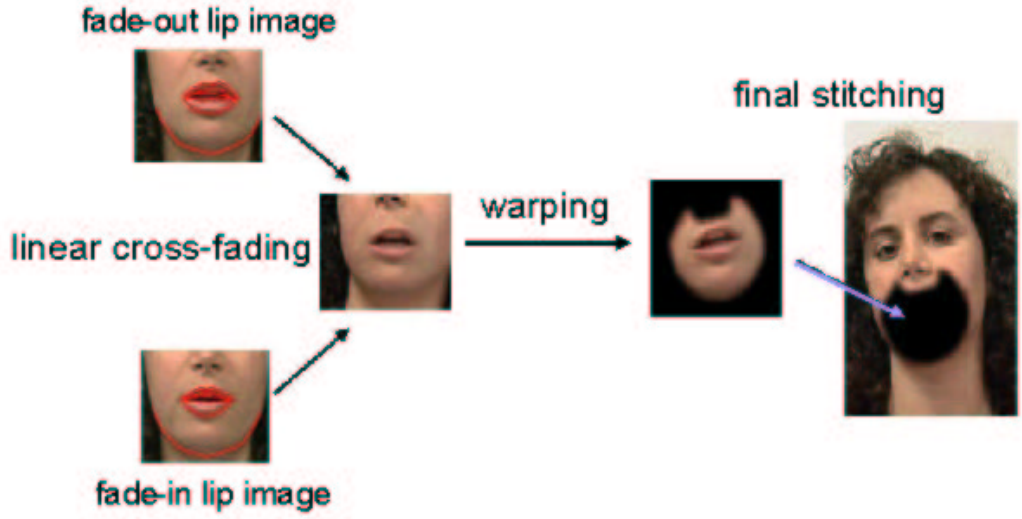

figure 5: combining the lips and the background

The examples of Video Rewrite are of convincing quality, but they are too short to judge whether it would be acceptable in a longer sequence, e.g. reading a whole fairy tale. One severe drawback is that the movements of the head, the eyebrows and the eyes are fixed, because the background video cannot be changed at all, even if the new soundtrack would require different head movements (e.g. for accentuation).

3. AT&T's 2D photo-realistic talking head

AT&T's 2D photo-realistic talking head is an image-based system like Video Rewrite, but there are several differences between the two systems. The main differences on the **analysis** side are the following:

1) the face is separated into six parts
2) visual labelling is achieved with multiple channels of analysis
3) the video database consists of automatically selected, visually distinct video frames (of the six different face parts), which are labelled only with visual parameters (not with phoneme information!)

As a consequence, during **synthesis** these visual parameters have to be predicted from the phoneme information of the new soundtrack, which is synthesized by AT&T's TTS system "Flextalk". According to those predicted visual parameters, the video frames are then selected from the database.

3.1 analysis

3.1.1 parts of the face

The face is separated into six parts: brows, eyes, lips, upper teeth, lower teeth, and jaw. As shown in figure 6, these parts are then combined with a base head. The great advantage of this segmentation is that the parts can be animated independently of the background video (unlike Video Rewrite). As shown in figure 7, these facial expressions can be superimposed over regular speech to make it livelier.

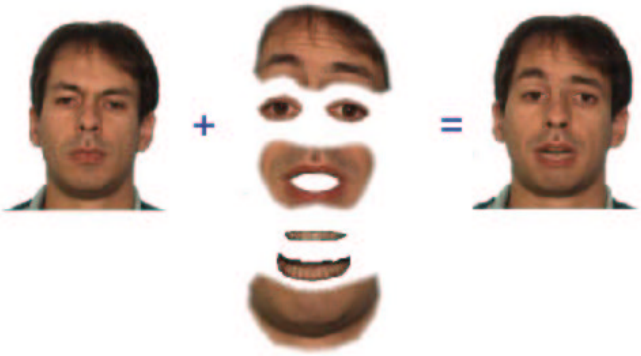

figure 6: parts of the face

figure 7: possible facial expressions

3.1.2 visual labelling: multiple channels of analysis

Instead of the eigenpoints algorithm, several analyses are carried out for each video frame: shape and texture analysis, color analysis, and motion analysis. The features from each channel of analysis are then combined and sent through a classifier (i.e. a machine learning algorithm), which outputs the visual parameters.

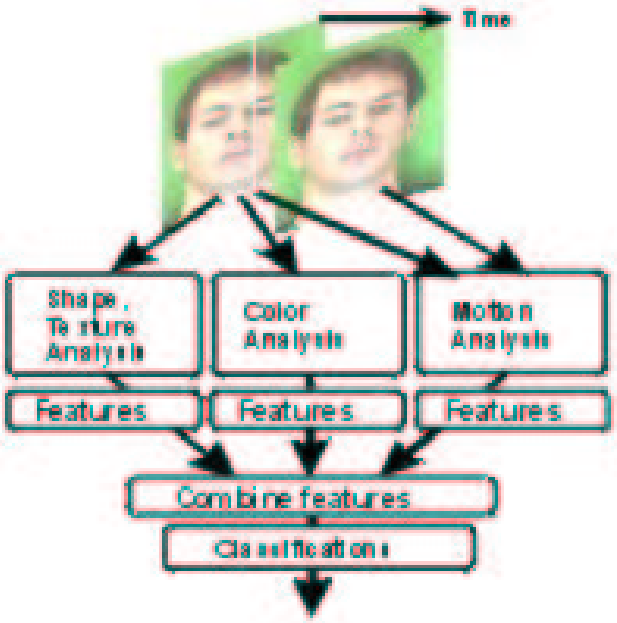

figure 8: visual labelling

3.1.3 visual parameters

The video frames are labelled only with *visual* parameters, e.g. lip width, lip opening, and jaw rotation. By excluding phoneme information, redundant views are eliminated from the video database (see section 3.1.4 below).

As a first classification, the lip shapes are labelled only with the following three parameters: mouth width, position of the lower lip and position of the upper lip. Table 1 shows examples of these parameters for four different phonemes.

sound	mouth width	position lower lip	position upper lip
silence	0.5	0	0
i: (feet)	0.7	0.5	0.5
u: (moon)	0.2	0.5	0.5
f (feet)	1.0	0	0.3

table 1: example of visual parameters for four different phonemes

This representation is very convenient: all possible lip shapes are represented in a compact, low-dimensional space, and it already includes a distance measure between the lip shapes (simply the Euclidean distance of the respective visual parameters). The shapes of the other facial parts (eyes, jaw, and eyebrows) are parametrized in a similar way.

3.1.4 building the database of video frames

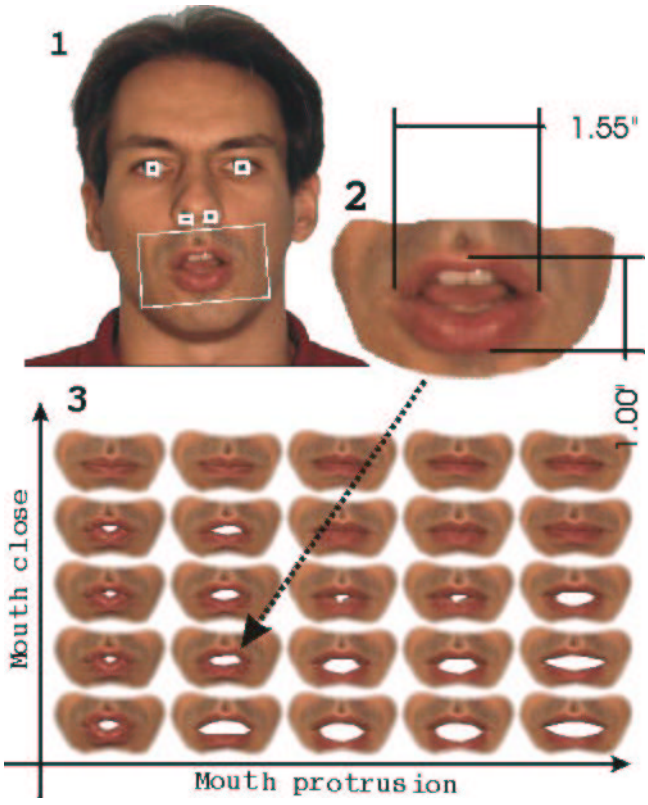

figure 9: populating the parameter space

In order to build the final database, first each video frame is warped to a standard reference plane (just like in Video Rewrite, see section 2.2.2) before the visual parameters are measured automatically. The parameter space forms a hypercube with a different parameter on each axis (in figure 9 a parameter plane for only two selected parameters is shown). In this hypercube, all grid

points that are to be filled with a video frame have to be specified.[9] Then for each grid point a video frame is selected. Since the final database consists only of visually *distinct* video frames, all video frames that are not selected for a specific grid point are not stored in the database. Finally, the database is inspected manually and corrected, if necessary.

3.2 synthesis

AT&T's 2D Talking Head uses the TTS system "Flextalk" to generate the new soundtrack. Figure 10 gives an overview of the whole synthesis stage.

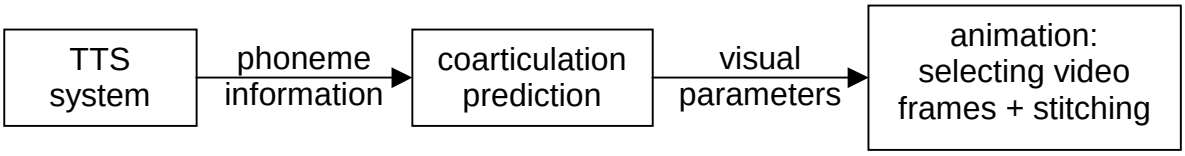

figure 10: overview of the synthesis stage

3.2.1 prediction of visual parameters

Since there is no phoneme information stored in the database, the visual parameters for the desired sequence of video frames have to be *predicted* on the basis of phoneme information. This so-called "coarticulation prediction" can be achieved with two different approaches: In the model-based approach, the visual parameters for each phoneme have to be defined. The visual parameters for a given video frame are then calculated as the weighted sum of the visual parameters of the current, the previous and the following phonemes, with the parameters of each phoneme decaying exponentially over time. In the sample-based approach, the visual parameters of each triphone are looked up in a speaker-specific library instead of defining them beforehand.

3.2.2 animation

During animation, for each facial part the two video frames that are closest to the calculated visual parameters are selected from the database and merged. These video frames are then warped and stitched onto a base head.

The resulting video clips are a little less convincing than Video Rewrite because the lip and jaw movements are too extreme. Probably the model-based approach was used for the coarticulation prediction, and the defined visual parameters for each phoneme are somewhat more pronounced then in natural speech. So the modelling definitely demands more attention in AT&T's system than in Video Rewrite. Nevertheless AT&T's talking head is much more promising since it allows the independent animation of six different facial parts.

References

Bregler, C., M. Covell & M. Slaney (1997): Video Rewrite: Driving Visual Speech with Audio. *Proceedings SIGGRAPH '97,* Los Angeles, CA, pp. 353–360. Retrieved May 20, 2002, from http://graphics.stanford.edu/~bregler/videorewrite/

Cosatto, E. & H. P. Graf (1998): Sample-Based Synthesis of Photo-Realistic Talking Heads. *Computer Animation,* pp. 103–110. Retrieved May 20, 2002, from http://www.research.att.com/projects/AnimatedHead/eric2.htm

Covell, M., & C. Bregler (1996): Eigenpoints. *Proceedings International Conference on Image Processing*, Lausanne, Switzerland, pp. 471–474.

[9] Some parts of the hypercube may be very densely populated to get a high resolution, others may even be left empty (e.g. think of lips that are protruded, while fully open).