Video-realistic synthetic speech with a parametric visual speech synthesizer

Sascha Fagel

Institute for Speech and Communication
Technical University Berlin
sascha.fagel@tu-berlin.de

Abstract

The author presents a new face module for MASSY, the Modular Audiovisual Speech SYnthesizer [1]. Within this face module the system combines two approaches of visual speech synthesis. Although the articulation space is parameterized in terms of movements of the articulators, the visual synthesis is image based (video-realistic).

The high-level visual speech synthesis generates a sequence of control commands for the visible articulation. The video synthesis searches an image database for appropriate video frames. If no image with facial properties according to the control commands is found, the missing image is generated by deforming a neutral image. MPEG-4 facial definition parameters (FDPs) [2] and additional points in the mouth opening area and around the lower jaw are defined in the neutral image as feature points. A two-dimensional displacement vector is defined for each feature point. For the image deformation a mesh of triangles connecting the feature points is used. The displacement vector of a point in a triangle is interpolated from the displacement vectors of the vertices.

Hence, the video synthesis algorithm is capable to use either a database of appropriately annotated video frames or a single neutral image with specified feature points and displacement vectors. A simple software tool for marking the feature points in the image was developed. Other well known data (image) based audio-visual speech synthesis systems like MIKETALK [3] and VIDEO REWRITE [4] concatenate pre-recorded video sequences. The presented system demonstrates the compatibility of parametric and data based visual speech synthesis approaches.

1. Introduction

Producing and perceiving speech is the natural way of communication. Man-machine-communication in many cases might be improved by speech, where one important aspect is the speech output of the machine. As human speech communication consists of several information streams, a coherent presentation of audible and visible speech should enhance several quality parameters compared to audio only presentation [5]. E.g. intelligibility is improved by natural or synthetic visual speech when added to audible speech. But research has shown that the possible advantage decreases by growing level of abstraction from natural speech [6][7].

The modularity of MASSY - the system for which the presented face module is developed - allows to exchange modules, to evaluate their specific advantages, and to detect their potential weak points. MASSY is implemented using the module interface specifications defined in [8]. Figure 1 shows a system overview. A plain text serves as system input. The phonetic articulation module creates the phonetic information, which consists of an appropriate phone chain on the one hand and - as prosodic information - phone and pause durations and a fundamental frequency curve on the other hand. From this data, the audio synthesis module generates the audio signal and the visual articulation module generates motion information. This motion information consists of control commands for virtual articulators given by an articulation model. The face module interprets the motion information and adds the audio signal to create the complete audiovisual speech output. MASSY’s previous face module generates an animated three-dimensional synthetic talking head using the Virtual Reality Modeling Language (VRML) [9]. For a detailed description of the system implementation see [10].

The present paper describes the implementation of a new image based face module.

Figure 1: Schematic system overview of MASSY.
2. The new face module

2.1. Facial properties

The control commands for the virtual articulators are converted to a set of facial properties for each frame at a given frame rate (e.g. 25 frames per second). The articulation parameters (virtual articulators) currently used by MASSY are

- the width of the lips \((\text{LipW})\),
- the height of the lower jaw \((\text{LowerJawH})\),
- the height of the lip opening independent from the height of the lower jaw \((\text{LipH})\),
- the retraction of the lower lip \((\text{LowerLipR})\),
- the height of the tongue tip \((\text{TongueTipH})\), and
- the height of the tongue back \((\text{TongueBackH})\).

\(\text{TongueTipH}\) and \(\text{TongueBackH}\) are omitted by the face module and not visualized. \(\text{LipW}\) is adopted without modifications. It describes both the lip narrowing/spreading and the protrusion as there is no phoneme - at least in German and English - that needs to be realized with spread and protruded lips or with narrow and not protruded lips at the same time, respectively. \(\text{LowerJawH}\) and \(\text{LipH}\) are combined to the single facial property \(\text{lip/jaw height}\). Furthermore, as the retraction of the lower lip results in a labiodental constriction, the articulation parameter \(\text{LowerLipR}\) is subtracted from the \(\text{lip/jaw height}\) to visualize this constriction. The three articulation parameters are scaled due to different neutral, minimum and maximum values. The exact form of equation 1 is estimated by subjectively judging the resulting images.

\[
\text{lip/jaw height} = 1.33 \cdot (\text{LowerJawH} - 0.25) - \text{LipH} - 0.5 \cdot \text{LowerLipR}
\]

(1)

The database is searched for an image that has the calculated facial properties with an adjustable accuracy. If no appropriate image is found, a neutral image is deformed by the following method to fit the facial properties.

2.2. Feature points

37 feature points are defined in the neutral image. 27 of them correspond to facial definition parameters (FDPs) standardized in MPEG-4 [2]. Five additional feature points define a surrounding of the lower jaw area to prevent sharp edges when the lower jaw is displaced. Another five feature points mark the upper teeth to save them from being deformed or displaced. Three of these five feature points mark the upper incisors, two point at the intersections of the upper lips and the upper teeth. A simple software tool for the marking procedure by hand was developed. Figure 2 shows the neutral image, the feature points, and the triangle mesh built of the feature points as vertices.

To each feature point two displacement vectors (one per facial property) are assigned and combined before deformation. The software tool can be used to mark the displaced feature points in an image with one facial property different from neutral. The differences of these feature points and those in the neutral image form the displacement vectors. If no displacement vectors are defined, a predefined set of displacement vectors for each facial property is scaled by the width of the lips in the neutral image and then used as preset.

2.3. Finding the triangle

For every pixel inside the bounding box of all feature points that triangle is determined in which the pixel is located. If a triangle is defined circularly, the pixel is inside this triangle if it is on the same side of each edge. On which side of an edge the pixel is located is determined by calculating the scalar product of the orthogonal complement of the edge
vector and the vector connecting one vertex and the pixel. This scalar product is positive if the pixel is on the left side of the edge vector and negative if the pixel is on the right side (figure 3). If all triangles are defined counterclockwise the greater or equal to zero test is sufficient. If some of the triangles are defined clockwise, collapsed triangles cannot be detected and the search order of the triangles becomes important.

The triangles are successively tested which is aborted if the triangle is found. For a better performance the triangle found for the last pixel is tested first for the next pixel and a triangle is tested only if the pixel is inside the bounding box of the triangle.

2.4. Triangular interpolation

The displacement vector of one pixel is composed of the three displacement vectors of the vertices. This triangular interpolation has three constraints:

- In the center of gravity of the triangle all three displacement vectors have to be weighted equally,
- the weight of a displacement vector has to decrease with growing distance from this vertex, and
- if the pixel is located on one edge the displacement vector of the opposite vertex has to be weighted by zero and the position of this vertex must not influence the weights of the other displacement vectors.

The latter constraint is most important for a smooth transition of displacement vectors between the triangles. The displacement vectors for pixels located on an edge shall not depend on which of the adjacent triangles is used for computing the weights. If the pixel is on one of the vertices this results in displacement vectors of both opposite vertices to be weighted by zero.

One reasonably performant method to obtain an appropriate weight per vertex according to the constraints is to calculate the ratio of the height of the pixel above one edge and the height of the opposite vertex above this edge (figure 4). Equation 2 describes the complete interpolation.

\[
\vec{d}_p = \frac{\vec{d}_a \cdot \vec{d}_a + \vec{d}_b \cdot \vec{d}_b + \vec{d}_c \cdot \vec{d}_c}{\vec{d}_a \cdot \vec{d}_a + \vec{d}_b \cdot \vec{d}_b + \vec{d}_c \cdot \vec{d}_c}
\]

with \(\vec{d}_p\) resulting displacement vector of the pixel,

\(\vec{d}_a, \vec{d}_b, \vec{d}_c\) displacement vectors of the vertices A, B, C,

\(h_{pa}, h_{pb}, h_{pc}\) height of the pixel above a, b, c, and

\(h_a, h_b, h_c\) height of the vertices A, B, C above a, b, c.

2.5. Deforming the image

The bounding box of all feature points is used as affected region. Instead of displacing every pixel of the neutral image to its new position the color of every pixel of the resulting (deformed) image is determined by looking where it comes from. This prevents holes in the resulting image arising from stretching an area. Therefore, at first, each feature point is displaced by the sum of its two displacement vectors (lip/jaw height and lip width) weighted by the calculated facial properties. Then for every pixel in the affected region its displacement vector is calculated and subtracted from the position of the pixel. The color of the pixel in the neutral (original) image nearest to this position is written to the pixel in the resulting image.

A scale factor is used at the color lookup operation to allow the resulting image to be of different size than the original. So the original image can be in high resolution and the resulting image can be downsized. This prevents quality loss while stretching an area. Furthermore, the resolution of the resulting image can be chosen with respect to the desired system performance.

2.6. Generating the video

If an image is required as one frame of the video and is not contained in the database it is generated and added to the database for later reuse. The image is written on the hard disk as jpeg file and decompressed to bmp file using the command line options of the image viewer IRFANVIEW [11]. For each frame a hardlink to the corresponding image file (a copy of the file entry in the file table but not of the file itself) is created. The frames are concatenated to an avi movie and the audio file is added with IMP2AVI [12]. Then the video is compressed to a wmv file with the WINDOWS MEDIA 8 ENCODING UTILITY [13]. Figure 5 shows the generated frames for the utterance “Moby” (/m@Ubi/). The algorithm for generating the video is potentially capable of using either the described deforming procedure or a database of marked video frames. The latter or a mixture of both is currently only possible if no head movements occur. The whole system including 3D synthetic and image based face module can be tested at http://fourier.ckw.tu-berlin.de.

3. Conclusions and future work

With the new image based face module the audiovisual speech synthesis system produces video-realistic speech, although the visual articulation itself is parameterized. The 3D synthetic speech visualization of MASSY improves the intelligibility nearly as much as natural visible speech [14].
The quality of the image based visual speech synthesis yet has to be evaluated in terms of intelligibility and naturalness of the generated images and articulation movements. As a further development of the face module the height of the tongue tip shall be added to the facial properties.

A software tool for the extraction of lip shapes was developed. This will be extended to automatically annotate facial properties to images, e.g. frames of a video recording. So an image database can be built without deforming a neutral image to fit facial properties.

Another software tool for automatically fitting once defined feature points to a new neutral image will be developed. This will simplify generating visible speech from a neutral image of any face.

4. Acknowledgements

Parts of the program code were written in the course of a visual speech synthesis seminar at the Technical University Berlin. The author thanks the students for their active participation. Special thanks go to my daughter for serving as "photo model".

5. References


