DESIGN OF AN AUDIO-VISUAL SPEECH CORPUS
FOR THE CZECH AUDIO-VISUAL SPEECH SYNTHESIS

Miloš Železný, Petr Čísař, Zdeněk Krňoul, Jan Novák

University of West Bohemia in Pilsen
Department of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic

ABSTRACT

Our long-term goal is to design a system for the Czech visual synthesis, that means an animated synthetic face (often called talking head) imitating pronouncing of a speech by a human being. In this paper we present techniques used for acquiring data and building the audio-visual speech corpus, especially its visual part. This process involves the recording of stereoscopic video data and solving of related problems as synchronization. Apart from that, we present simple method of utilization of such corpus using stereo vision principles and modelling shape of the lips by simple triangular mesh.

1. INTRODUCTION

People use both auditory and visual information for perception of a speech. This is even more important in the case when speech is disturbed by a noise or when listening person is hearing-impaired. In such cases the visual perception (lip-reading) helps people in correct understanding of a speech being heard [1]. For these reasons it is suitable to supplement a system of (acoustic) speech synthesis with a visual synthesis, that means with an animated synthetic face that emulates pronouncing of a speech by people. Another use of visual synthesis is in the case when only the speech is available as a supplement of an unknown speech being heard. The use of such approach is for example in noisy environment or as a communication aid for hearing-impaired people. Similar aid is used also for language and speech teaching of deaf children [2]. In this case it is needed to show with the incoming speech also the imitation of a face. It is needed to either recognize the speech or at least estimate type of the phoneme. As the phonetic and acoustic parameters and face shape and expression parameters differ among various languages [3, 4], the specifics of the Czech language have to be taken into account.

As in the case of acoustic synthesis, the method of concatenation can be used for visual synthesis. For this purpose the database of basic speech units must be created. In this paper we will focus on the technique of acquiring data and building such database.

2. CORPUS PREPARATION

The first task of preparation for corpus recording is the selection of sentences [5]. Assuming the fact that the same text is used both for training of acoustic speech synthesizer and visual speech synthesizer, the text must conform to requirements of both parts of audio-visual synthesis. For acoustic synthesis the ARTIC text-to-speech system [6] will be used.

Next problem is the choice of parameterization. To design a synthetic 3D model of a face, either fully artificial model can be created or 3D information can be estimated from real visual data. The design of a parametric model can be divided into two tasks: design of a static model of a face (or a whole head) and design of a parametrically controllable model of moving parts.

Fig. 1. An illustration of the layout of the control points.

We can vary the complexity of a task by specifying more parts of a face as moveable and parametrically controllable. In our task we decided to assume only lips and chin as moveable and rest of the face as static. Furthermore we assume to get only 3D face model and not the model of the whole head. We chose 11 points that control the contour of lips and position of a chin. Layout of the points for the first version of synthetic face is shown on figure 1. These points refer to the reference point (marked as R) that is placed at the root of the nose. This reference point is regarded as the origin of the coordinate system.

3. RECORDING SETUP

To estimate 3D parameters of parametric parts of a model, the whole corpus must be recorded using stereoscopic view. Usual

This research was supported by the Ministry of Education of the Czech Republic, project No. MSM235200004.
setup for stereovision is illustrated in figure 2. Two cameras are needed for acquiring visual data from both angles. In addition to the fact that the video stream must be synchronized with an audio stream, both video streams have to be synchronized with each other. Besides it, both cameras must operate synchronously to ensure that both frames of stereo pair are acquired at the same time. It causes difficulties of the project solution and requirements of very special hardware.

![Fig. 2. Usual setup for stereovision.](image1)

Our solution is to record both stereo images using only one camera and system of 4 mirrors. Proposed setup is illustrated on figure 3. The whole stereo image pair is acquired using only one camera. Our setup is more complicated than that one used in [7], but retains the same distance between camera and head in both stereoscopic frames. By this way we avoid the problem of synchronization of both cameras with each other. It is obvious that this advantage is at the expense of image resolution.

Next we solved the problem of synchronization of the video data with acoustic data. The acoustic data for training the acoustic speech synthesis module are acquired separately. It is needed for achievement of required quality. Data for acoustic synthesis are acquired by a high quality microphone and second channel is acquired by an electroglottograph. Synchronization is made using standard clapperboard used in cinematography. Using this simple approach the moment when the clapperboard clicks can be easily identified both in video (the first frame when both parts of clapperboard touch each other) and audio stream (a pulse produced by this clicking).

![Fig. 3. Proposed setup for stereovision using only one camera.](image2)

The corpus. This is done within the training of acoustic models of speech segments.

### 4. DATA PROCESSING

Processing of acquired data is probably the most important task within the visual database design. Methods used for solving this task are discussed in this section. Two main problems can be seen in this task. Firstly we must detect interest points in each frame or, to be precise, in both parts of each frame. Then we must compute 3D coordinates of each interest point from the two corresponding points from the two parts of a frame.

Also, we need the information about beginnings and endings of phones in the recorded text. This segmentation information is provided by the output of training process on acoustic part of the corpus. This is done within the training of acoustic models of speech segments.

#### 4.1. Image segmentation

As mentioned in section 2, we need to detect 12 points in both parts of each frame, i.e. 24 2D points. For simplification of points detection, we enhanced speaker’s lips by red-colored lipstick. Furthermore, we enhanced both important points outside lips, i.e. reference point $R$ and chin point $\delta$ by a blue marker. As the whole recording process was accomplished in one day, there was no problem retaining these points at the same place.

This preparation allows easy segmentation. For detection of the reference point and the point at chin the thresholding method used on image converted to grayscale is sufficient. For detection of lip contours the image has to be converted to IHS (intensity, hue, saturation) color representation model. After that the method of color thresholding (thresholding on the hue component of IHS model) was applied.

After segmentation the points are to be detected. Points $I$ and $II$ are detected as points of segmented object with the minimal and maximal value of $y$ coordinate. Point $4$ is detected as local minimum between two local maxima of the upper contour of lips (in the $x$ direction) and point $7$ as minimum in $x$ direction at lower lip contour. Remaining points are detected by the edge detection at the midpoints between detected points.
4.2. Stereo processing

Once we have detected interest points, we can compute the 3D coordinates for each point from a couple of corresponding 2D points. Usually, the stereo vision is based on the assumption of collinearity of camera axes [8]. This method cannot be used in our case. Particular formulae can be derived from the sketch at the figure 5.

\[
x = z' \sin \left( \beta - \arcsin \frac{d_R}{z'} \right)
\]

(2)

\[
z = \sqrt{(z')^2 - x^2}
\]

(3)

Let us note that this computing is done in XZ plane and does not affect the y coordinate.

5. RESULTS

Resulting audio-visual speech corpus consists of important speech material needed to model speech units and to create both acoustical and visual speech segment database. Two levels of visual corpus data are presented in this paper. Firstly the visual data and secondly the method of speech units modelling and preparation of the speech segment database. The corpus parameters are summarized in table 1.

<table>
<thead>
<tr>
<th>Common parameters</th>
<th>600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td></td>
</tr>
<tr>
<td>Total length</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acoustical part</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel 1</td>
<td>Microphone audio</td>
</tr>
<tr>
<td>Channel 2</td>
<td>Electroglossograph signal</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>44 kHz, 16 bits</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visual part</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
<td>720x576, 25 fps</td>
</tr>
<tr>
<td>Codec</td>
<td>Indeo video 5.10</td>
</tr>
</tbody>
</table>

Table 1. Parameters of the speech corpus.

Visual corpus data can be used for generating desired speech segment database. However, to proof this ability we derived simple 3D visual speech database from this corpus. We used parameterization mentioned in section 2 and using image segmentation and stereo processing we created the database consisting of 3D models of lips. Example of resulting model for the phone a is on figure 6, example of the model for the phone p is on the figure 7. Design of a 3D model of whole face using the visual data acquired is a subject of future work.

Proposed parametric 3D models of lips serve as control parameters for the process of visual synthesis. First of all, the process of acoustic synthesis is performed. Its output is a wave file and phoneme notifications. These notifications control length of individual phonemes in visual synthesis. In the online mode the synthesized sound is playing and the visual synthesis module generates corresponding frames using the information about the phoneme being played and using the database of parametric models.

6. CONCLUSION AND FUTURE WORK

We proposed a method of data acquisition for the visual speech corpus creation. Proposed approach of acquiring stereo image data allows to derive 3D models of the face from corpus data. It also preserves obtaining the whole stereo pair at the same time using only one camera and the system of 4 mirrors. The visual data collected have to be linked to the acoustic data, so that both parts produce the audio-visual speech corpus. Acoustic part of the corpus is used for training of the acoustic text-to-speech system ARTIC [9].
developed at our department. Our work tries to extend this system by a “text-to-face” system. It can be used also in standalone mode using unknown speech and methods of speech recognition for obtaining corresponding text.

We also propose simple method of generating 3D models for lips movement modelling. Lips are modelled by simple triangular mesh using ten points depicted in figure 1. The resulting 3D model of lips shown in figures 6 and 7 will be part of the 3D model of a whole face. Lip model represents the moving and controllable part of this model. Results show the possibility of derivation 3D models from collected corpus.

Future work includes the completion of the design of 3D model of the whole synthetic face. Further optimization of all parts of the process is needed. We will also develop better method for extraction of speech units from the input data. Lastly we will design a speech database based on the context dependent speech units, triphones.

7. REFERENCES


