Realistic Face Animation for a Czech Talking Head

Zdeněk Krňoul and Miloš Železný

Department of Cybernetics, University of West Bohemia,
Univerzitní 8, 306 14 Pilsen, Czech Republic
Email: zdkrnoul@kky.zcu.cz, zelezny@kky.zcu.cz
WWW: http://artin.zcu.cz/projects/visilab

Abstract. This paper is focused on improving visual Czech speech synthesis. Our aim was the design of a highly natural and realistic talking head with a realistic 3D face model, improved co-articulation, and a realistic model of inner articulatory organs (teeth, the tongue and the palate). Besides very good articulation our aim was also expression of the mimic and emotions of the talking head. The intelligibility was verified by the listening test and the results of this test were analysed.
Firstly, the face model reconstruction from real data is presented. A 3D computer vision was employed in order to obtain a model of an arbitrary face. The stereovision technique that is used to reconstruct the model is described in detail in Section 2. Details concerning a visual speech synthesis are discussed in Section 3. Special features of the visual speech synthesis of the Czech language are mentioned, too. Furthermore, the design of a talking head including the solution of co-articulation problem is presented. Next, the modeling of expression and emotions in animation is described. The last part of the paper, Section 4, contains results of the performed listening test.

1 Introduction

Audio-visual synthesis increases the intelligibility of computer speech synthesis. The visual part of synthesis (talking head) with precise articulation can contribute considerably to the intelligibility of speech especially for hearing-impaired people [1] or in the case of significant environment noise or low bandwidth voice transmission. In the latter case, when there is not enough bandwidth for the transmission of accompanying visual signal, the visual part of the synthesis has to be carried out at the "client" side. Such approach can be based on the transmission of the textual information. In another setup the incoming (not necessarily synthesised) voice is recognised and the output of the recogniser drives the visual speech synthesis [2].

The human speech producing organs consist of the breath organ which produces an air stream, larynx which modulates the voice and articulatory organs which create speech. The produced speech is the result of smooth and precise cooperation of all these three parts. The whole cooperation is controlled by the brain. From the outside, we can see only a part of the speech organs. We can see the motion of the jaw, lips and in some cases teeth and tongue. The motion of the other parts of the body belongs to visual speech as well. Gestures indicate word stress, rhythm and phrasing.

The visual speech synthesis comprises geometric parameterisation, morphing between target speech shapes and head animation. Using the presented method of face model

Petr Sojka, Ivan Kopeček, and Karel Pala (Eds.): TSD 2004, LNAI 3206, pp. 603–610, 2004.

reconstruction we get a high resolution 3D static shape of the face and supplementary texture. We are able to parameterise the face model. By supplementing complete articulatory organs and acoustic speech synthesis (TTS – text-to-speech synthesis), we get a highly realistic head model which produces accurate audio-visual speech.

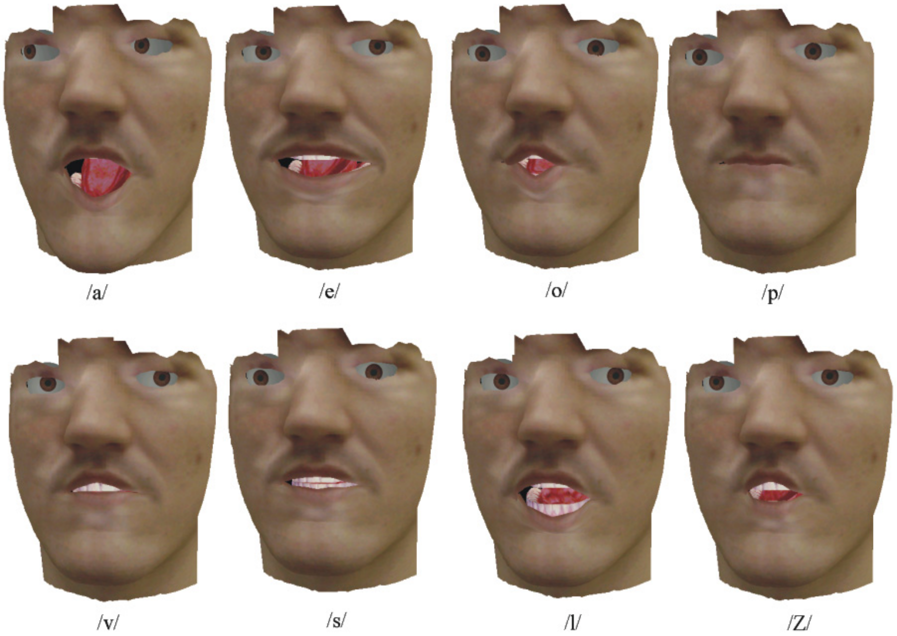


Fig. 1. The visemes: /a/, /e/, /o/, /p/, /v/, /s/, /l/ and /Z/

Our aim is to obtain a correct visual output (visemes) for Czech phonemes. Examples of visemes are shown in Figure 1. The phonemes can be grouped by their external visual perception. Our complete model makes it possible to model internal differences between the phonemes inside these groups. All the visual parameters are stored in a database. Visual speech synthesis can then be carried out by simple concatenating of the segments from the database. However, during speech production it is desirable to model co-articulation, i.e. the influence of adjacent (preceding and following) phonemes.

As already stated, the expression of a face can contribute to overall intelligibility of a whole sentence (or a whole speech). In our model, the emotional expressions of a face may be modeled, too. We present basic techniques for implementing the modeling of emotions.

To verify the intelligibility of our talking head, we carried out a listening test. The results and a detailed analysis of these results are also discussed in this paper.

2 Model Reconstruction

The desired 3D model of a face consists of two parts, a shape and a texture. The task was to obtain information about the shape and texture from a real face, in order to produce thus a real looking face model. The aim was to avoid the use of expensive devices, such as a 3D scanner.

We employed a stereovision-based algorithm [3] for obtaining a 3D geometry of the face. The idea of this algorithm is based on the use of two views of the face from two different points of view. The process of obtaining depth information is based on projective geometry. The off-line calibration step is used for computing a perspective projection matrix from defined points in a space onto a projective plane. The shape of the face is then reconstructed from a sequence of stereo images. For an easy solution of the correspondence problem, a supplementary vertical stripe ray is used in otherwise dark illumination. This ray moves during the image sequence horizontally over the whole face. The last image is used for obtaining the texture. It is acquired using normal (daylight or artificial light) illumination.

The resulting shape model is computed by triangulation. The spatial relationship of the parts of the face is time independent. It remains to set points controlled by parameterisation. The parts adjacent to the control points have to move smoothly according to the movement of the control points. Some interpolation technique should be adopted. The interpolated regions and related control points are depicted in Figure 2. In our case spline interpolation was used. Additionally, models of the internal parts of the mouth (teeth, tongue) and eyes, prepared in advance, were included in the model of the whole face. Using the above described procedure, a complete animated model of the face is obtained. The resulting model looks highly realistic and provides intelligible articulation.

3 Visual Speech Synthesis

From the point of view of lip-reading, i.e. from the point of view of the user of audio-visual speech information, it can be said that the movements of the speech organs are combined into speech images, which are often vague. Sound differences are in such cases produced in the rear parts of the oral cavity (resonance cavities, movement of the tongue behind closed teeth and in the throat). It is thus not possible to visually distinguish all Czech phones [4]. However, from the point of view of synthesis we can define the expression of all organs (including those not visible) for all Czech phones. These can be divided into groups according to their similar speech images: (b-p-m) (v-f) (š-ž-č-ř) (s-z-c) (l-r) (d-t-n) (d'-t'-ň-j) (k-g-ch-h). Vowels (a-e-i-o-u) are produced in a different way. The air flow is not blocked. Each vowel corresponds to a particular speech image. Vowels can thus be easily distinguished.

We defined the parametrical speech image using a modification of our previous parameterisation [5] and in compliance with a MPEG4 standard. Using these parameters we can control the movements of the whole face and the internal mouth organs. This parameterisation makes it possible to express both visual speech and emotions. The fundamental location of the visible parameters for all Czech phonemes is obtained using the Czech audio-visual speech corpus [5]. Methods of 3D computer vision were employed to obtain parameter vectors. At this stage, it would be possible to generate visual speech synthesis by simply concatenating the basic phoneme/viseme models.

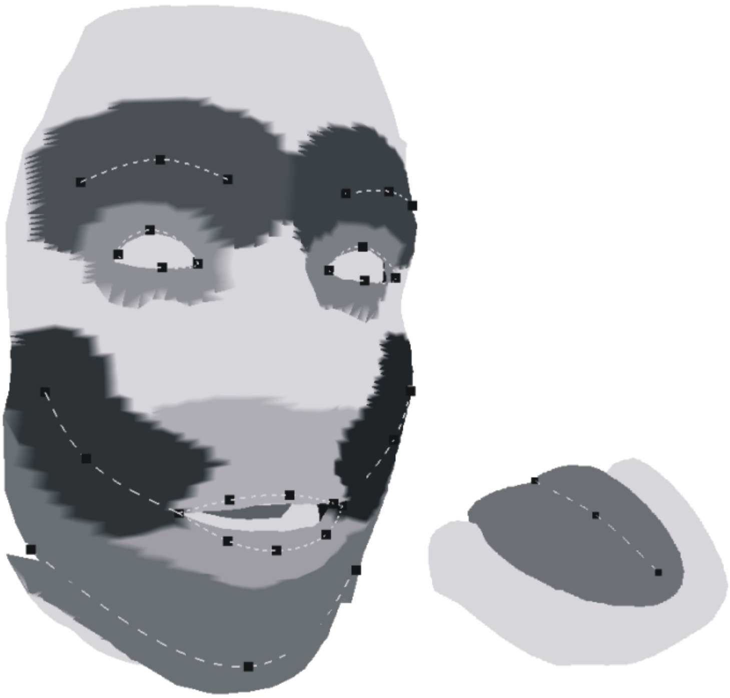


Fig. 2. The influence regions and the place of control points

However, doing so would produce many artefacts at the joints of consequent phonemes. This is caused by the fact that the pattern for each concatenated model in the database was obtained in a context different from the one in which it is used. It is thus necessary to respect co-articulation, i.e. means the influence of adjacent units (phonemes) on each other. Usually, an adjacent consonant and vowel are pronounced jointly. Their joint speech image is different from the individual speech images of the two phonemes. Our approach to the solution to the problem of the co-articulation effect is based on the approach suggested by Cohen and Massaro [6] and already modified by us for the Czech language [7]. This approach is based on the dominance functions, the parameters of which as well as the weights of the speaking images were estimated using the audio-visual speech corpus.

The collected speech database was supplemented by a database of emotional expressions. We selected 8 basic expressions which can enhance the visual speech produced. These emotional expressions are shown in Figure 3. They are depicted on the reconstructed face model of a female speaker. For practical implementation we suggested inserting the expression marks into the text. In this way we can modify the text so as to increase the intelligibility of the whole utterance.

The talking head is a parametrically controlled 3D polygonal (triangular) model. It can be animated synchronously with acoustic speech. For acoustic speech synthesis we use the TTS (text-to-speech) system developed at our department [8]. The module of

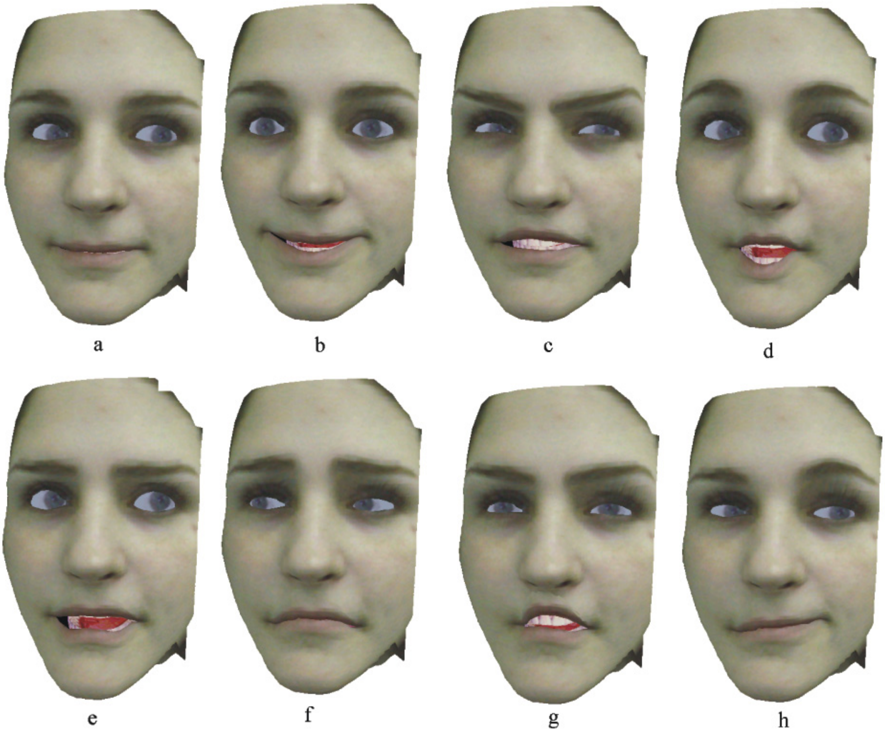


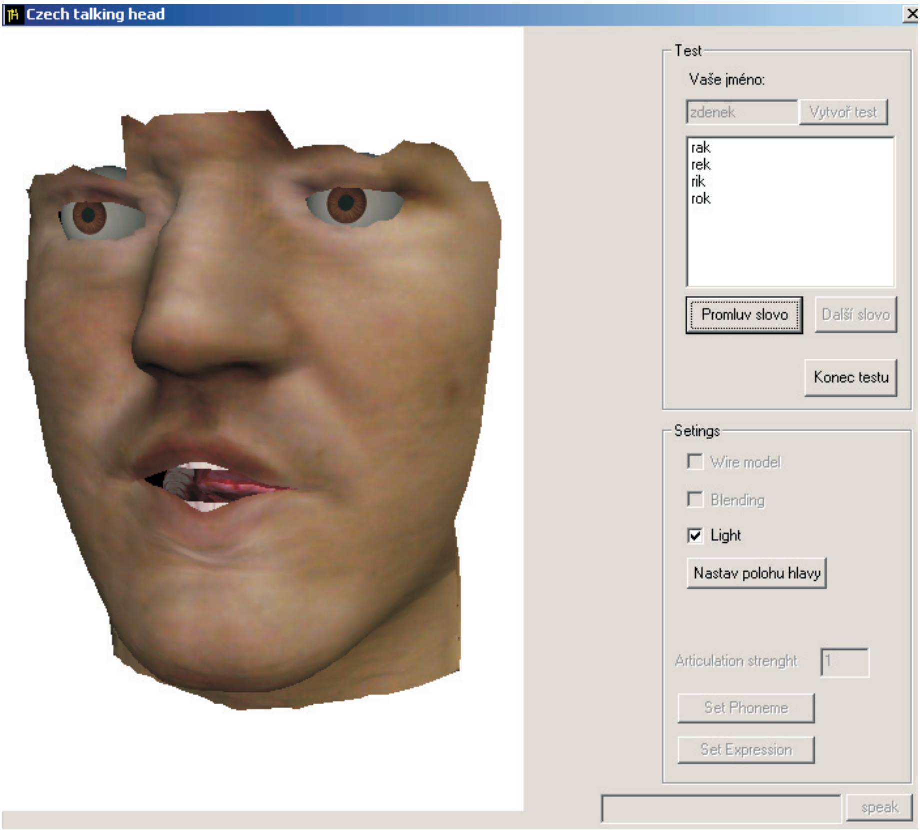
Fig. 3. The Expresions: a) neutral expression, b) happiness, c) anger, d) surprise, e) fear, f) sadness, g) disgust and h) pensive expression

synchronisation with the TTS system uses the notifications produced by this TTS system. These notifications contain the information about boundaries between adjacent phonemes. The face deformations are then controlled by a proposed parameterisation. The face model is rendered synchronously with the acoustic signal produced by the TTS system using the synchronisation information from notifications. For each animated part of the face a set of control points and a set of influence regions is defined. The control points and respective regions are depicted in Figure 2. The vertices of the triangular mesh that belong to a particular region are moved according to the respective control point. The interpolation and smoothing of the point movements is determined by the spline curve in the 3D space. The deformation rules are described by influence weight functions transforming the movement of the curves into individual vertices. For each vertex we can define a deformation equation

$$P'_p = R_p(\Delta S_{pk}(i) w_k(\min_{\forall i \forall k} |S_{pk}(i) - P_p|) + P_p) \tag{1}$$

The new position of vertex P' of part p of mesh is computed from its initial position P. The minimal distance from the control spline function $S_k(i)$ is computed and weighted by a weight w. R_p is the corresponding rotation matrix.

Our implementation allows easy changes of the face model, the change in speech rate, text-driven or iterative change in the expression of emotions and change in the intensity of articulation. Our face animation toolkit is written in the C language. The rendering engine is based on standard 3D graphic libraries.



Vaše jméno: zdenek

rak
rek
rik
rok

☐ Wire model

☐ Blending

☑ Light

Articulation strenght 1

Fig. 4. The testing aplication

4 Listening Test

A listening test was performed with persons with unimpaired hearing. Testing application (depicted in Figure 4) shows the talking head. It allows rotation and change of scale. The approximate height of the head on the 19-inch monitor was 18 cm. The talking head expressed visually the testing word while the sound was off. Simultaneously, several suggestions were shown on the display for each testing word. Only meaningful words were presented. The suggested choices differed in one to three letters. In this way we simulated the state when several words have similar acoustic form, but differ visually. The test was run for 100 words for each listener. The approximate duration of the test was 1 hour per listener.

In the test 20 listeners participated, 18 of whom were males and 2 females. We set a slower speech rate (by 50%) to emphasise the articulation. The test consisted of several parts. The first part was the test of vowels. Vowels are quite easily distinguishable and their count determines the intelligibility. The second part tested words for consonants. The word (choice) list was random. The application awaited a choice from the listener. We computed the overall intelligibility, the CVC test and the McGurk effect test. Results of the test are summarised in Table 1.

Table 1. Results of the listening test

Overall intelligibility	61.6 %
From which	
38 word test of vowels	59.7 %
60 word test of consonants	62.9 %
31 word test of CVC	59.4 %
McGurk effect (ba/da/ga) test	82.4 %

5 Conclusion

The presented talking head indicates a potential of multimodal speech communication. The process of face shape reconstruction generates excellent polygonal representation and extracts a high resolution texture. The face animation of obtained the 3D model supplemented by internal organs (such as tongue, teeth or eyes) performs visual speech synthesis. Running synchronously with the acoustic synthesis, it enhances speech communication by adding a new modality. By including expressions and emotions in face animation we can further extend the range of varieties in multimodal human-computer interaction.

Acknowledgements

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/03/0650 and by the Ministry of Education of the Czech Republic, project No. MSM235200004.

References

1. Krahulcová, B.: Komunikace sluchově postižených – Communication of Hearing Impaired People. Karolinum, Prague, Czech Republic (2002).
2. Angelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Salvi, G., Spens, K.E., Öhman, T.: A synthetic face as a lip-reading support for hearing impaired telephone users – problems and positive results. In: Proceedings of EFAS 1999, Oulo, Finland (1999).

3. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press, Cambridge, UK (2001).
4. Strnadová, V.: Hádej, co říkám aneb Odezírání je nejisté umění. – Guess What I Am Talking or Lip-Reading is Uncertain Art. Ministerstvo zdravotnictví České republiky, Prague, Czech Republic (1998).
5. Železný, M., Císař, P., Krňoul, Z., Novák, J.: Design of an audio-visual speech corpus for the Czech audio-visual speech synthesis. In: Proceedings of ICSLP 2002, Denver, USA (2002) 1941–1944.
6. Cohen, M.M., Massaro, D.W.: Text-to-visual speech synthesis based on parameter generation from hmm. In: Models and Techniques in Computer Animation, Springer-Verlag, Tokyo, Japan (1993) 139–156.
7. Krňoul, Z., Železný, M.: Coarticulation modeling for the Czech audio-visual speech synthesis. In: Proceedings of the ECMS, 6th International Workshop on Electronics, Control, Measurement and Signals 2003, Liberec, Czech Republic (2003).
8. Matoušek, J., Psutka, J.: ARTIC: A new Czech text-to-speech system using statistical approach to speech segment database construction. In: Proceedings of ICSLP2000. Volume IV., Beijing (2000) 612–615.