A Development of Czech Talking Head

Zdeněk Krňoul, Miloš Železný

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia, Pilsen, Czech Republic
zdkrnoul@kky.zcu.cz, zelezny@kky.zcu.cz

Abstract

This paper presents a research on the Czech talking head system. It gives an overview of methods used for visual speech animation, parameterization of a human face and a tongue, necessary data sources and a synthesis method. A 3D animation model is used for a pseudo-muscular animation schema to create such animation of visual speech which is usable for a lipreading. An extension of animation schema is presented to reach more precise deformations mainly in a lip area. Furthermore, a problem of forming articulatory trajectories is formulated to solve labial coarticulation effects. It is used for the synthesis method based on a selection of articulatory targets and interpolation technique.

Index Terms: synthesis of visual speech, talking head, facial animation

1. Introduction

A research in an area of a speech synthesis attempts to design methods that produce a speech output recognizable by humans. The methods are implemented in systems whose common feature is communication between people and computers. The applications of visual speech synthesis are in the areas of speech production or training of lipreading ability. Such systems help Deaf or hearing impaired people. The other applications should be considered in the future because the computer systems increasingly attach a lot of human activities. In that connection, we can find several synthesis methods and talking head systems designed for particular languages and particular purposes. In the majority of cases, audiovisual speech synthesis methods convert a text form of speech to audiovisual speech. The text form expresses a phrase given by a sequence of phonemes and audiovisual speech is then represented as simultaneously articulated acoustic and visual cues.

The research on a synthesis of Czech speech is in progress for two last decades and the synthesis of visual speech the last seven years. At present, our talking head system automatically converts input text to audiovisual speech in the form of an animation of 3D human head model. An input sequence of phonemes is transformed into a continuous stream of lip and tongue movements. We can change an appearance or a parameterization of the 3D model as well as a type of the synthesis method. In previous work, the audiovisual corpora of Czech speech were recorded. We have concentrated on a study of labial coarticulation and a creation of such speech output which can be used as a lipreading support.

This paper presents advancements in previous research on Czech talking head system. The section 2 makes an overview of our talking head system. Required components of the system and accompanying research activities are described there. In the section 3, the formerly designed animation schema is recapitulated and some new extensions are introduced. An acquisition of visual data and collections of audiovisual speech corpora are summarized in the section 4 and the section 5 makes a formulation of our synthesis method.

2. Talking head system overview

The Czech talking head system provides the conversion of text to audiovisual Czech speech. The audiovisual speech consists from acoustic part in the form of a waveform and a visual part in the form of a rendered animation. A synthesis method for the acoustic part is designed separately. For this purpose, we have employed text-to-speech system (TTS) ARTIC [1]. This TTS converts an input text to high quality Czech voice and provides a phonetic transcription and a time labeling. However, the talking head system can potentially employ another TTS system providing mentioned features.

Figure 1 illustrates talking head system as a block schema. *An animation module* combines the acoustic part with the visual one and ensures final rendering with the animation model. *A 3D face reconstruction technique* can be used to determine a new shape of face of the model. For a control of the animation model, articulatory trajectories are used. These trajectories are completed by a block of a *synthesis*. The phonetic transcription and time labeling from *TTS module* are used for necessary synchronization. We consider a phoneme as base units which are stored by *a database of units*.

3. Animation schema

The animation schema is based on a geometric representation of a human face and other necessary parts of head in 3D space by polygonal meshes. We consider a computation of defor-

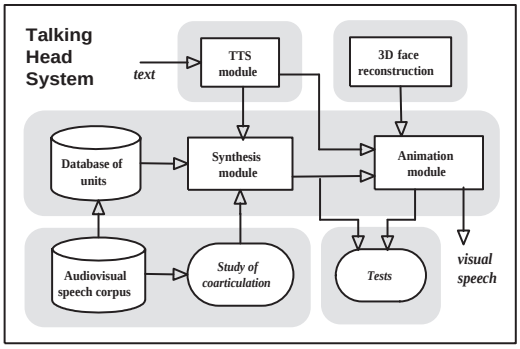

Figure 1: A block schema of Czech talking head system.

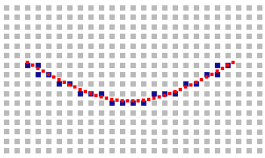

Figure 2: The example of *a curve-print*. The big light squares are vertices of the mesh and the small squares are the approximation points of the curve. The big black squares show the vertices of the curve-print.

mations on the outermost layer of the face skin (epidermis) or tongue without subsurface layers. An application of the animation schema is mainly to provide synthesized visual speech as a lipreading support.

The principle of the animation schema is based on feature points and deformation zones [2, 3]. One deformation zone is considered as a area on the 3D face surface given by a set of vetices and the feature point as one 3D point in this area. A shift of its position causes changes in whole deformation zone. It uses one polygonal mesh only often formed in neutral face pose in comparison with designs used interpolation of multiple face poses. The low requirements on the shape of the polygonal mesh caused it very flexible, too. However, for better approximation of face shape around lips, deformations commutated around a curve appear to be more suitable [4]. Therefore, we have designed an animation schema which accounts both mentioned cases [5]. The animation schema is based on 3D cubic spline curves each defined by several control points. We can found similarity with the controlling of the tongue model defined in 2D space [6]. On the other hand, an advantage of a 3D cubic spline curve is that it uses feature points directly matched with animation parameters of a 3D head model.

Two examples of an influence zone affected by our animation schema are shown in Fig. 3. There is a simple mesh and the deformation commutated by one feature point and by one spline curve created from four feature points. In general, the curves can be open, as shown in the figure, or can be closed when first and the last feature point have same position. In advance, these closed splines are used for the approximation of the areas around eyes or the *Orbicilaris Oris* muscle. The cubic curve is fitted by several approximation points equidistant align in (x, y), (x, z) and (y, z) planes. The Euclidean distance of two adjacent approximation points determines a density of the curve. This distance could be bigger than length of edges of mesh polygons in the affected influence zone to avoid unnatural displacements of the mesh.

3.1. Computation of deformation

The animation model is represented by several separated polygonal meshes describing the shapes of relevant parts of head. A definition of the model includes a set of the feature points, sizes of influence zones and the cubic spline curves. One feature point should be used for a determination of more than one curve. However one curve should be assigned to one mesh only. The 3D positions of feature points are manually placed at moment of new model creation.

Firstly we describe animation schema with non-overlapping influence zones. S_{kp} is $k-th$ curve defined on the mesh p. Its shape is given by one or more feature points. The influence zone enclosing the curve is composed from a subset of the vertices $V_p(i)$. The subset is given by *a propagation process* which starts from *a curve-print* and ends at a boundary of the influence zone. The curve-print of $k-th$ curve is given by the positions of such vertices which satisfy condition (1).

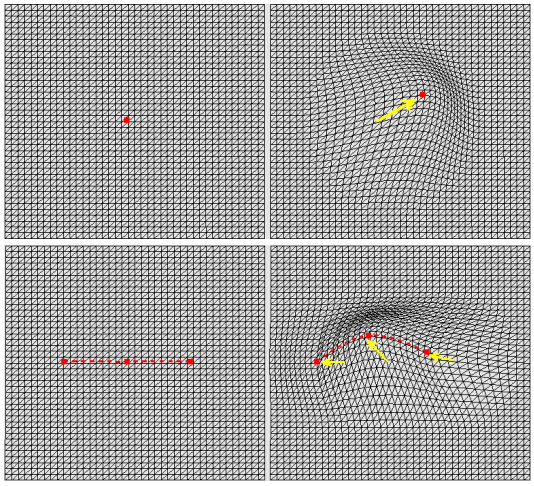

Figure 3: An examples of curve definitions, influence zones and deformations.

$$min_{\forall i}(|S_{kp}(j) - V_p(i)|) \qquad i = 1..N, j = 1..M \quad (1)$$

$S_{kp}(j)$ are approximation points given by the interpolation between feature points. M is determined in accordance with the density of the polygonal mesh determined the influence zone. Figure 2 shows an example of a curve-print.

By the propagation process, weights for all vertices in the influence zone are determined. The weights are based on the Euclidean distance between processed vertex $V_p(i)$ and the closest approximation point, Eq. (2). The examples of acceptable weight functions are summarized by Eq. (3).

$$D_p(i) = min_{\forall j}(|S_{kp}(j) - V_p(i)|) \qquad j = 1..M \quad (2)$$

$$\begin{aligned} w_\alpha(d) &= 0.5(cos(d) + 1) \\ w_\beta(d) &= (0.5(cos(d) + 1))^2 \\ w_\gamma(d) &= cos(d)/2 \end{aligned} \quad (3)$$

These weight functions have value 1 for distance $d = 0$ and 0 for distance $d = 1$ reached on the boundary of the influence zone. The Eq. (4) is used to determine transformation of all vertices in the influence zone.

$$V_p^{'}(i) = R_p(\Delta S_{kp}(j) w_k(D_p(i)) + V_p(i)) \quad (4)$$

New position $V_p^{'}(i)$ is determined from its neutral position $V_p(i)$, the shift of relevant approximation point $\Delta S_{kp}(j)$ weighted by w_k and a rotation of the mesh R_p.

3.2. Enhancement of animation schema

The animation schema described in the previous section has a limitation caused by overlapped influence zones. An area of the mesh in an intersection of two or more influence zones is divided in assumption the nearest distance. It causes unnatural displacements. Therefore we have extend it about new feature providing a computation of deformation in the overlap zones. The curve-print and propagation process is carried out for each spline curve separately in compliance with the formulas (1)

and (2). However the propagation process determines distance $D_{kp}(i)$ for all S_{kp}. Thus each vertex can be influenced more than one curve. The transformation of V_p is given by Eq. (5).

$$V_p^{'} = R_p(\frac{\sum_{\forall k}(w_{bk}(D_{kp}(i))w_o(D_{kp}(i))\Delta S_{kp}(j))}{\sum_{\forall k} w_{bk}(D_{kp}(i))} + V_p) \tag{5}$$

$V_p^{'}$ is computed as weighted sum of the shifts $\Delta S_{kp}(j)$ from all curves k defined on the polygonal mesh p. The weight function w_{bk} should be one of (3) and the weight function w_o is given by (6).

$$w_o(d) = \frac{1}{1 + exp(c_1 d - c_2)} \tag{6}$$

c_1 and c_2 are well defined constants to determine the relative redistribution of w_{bk} in the overlapped area.

3.3. Parameterization of visual speech

The parameterization of a lip or a tongue is given by a layout of the feature points on the polygonal meshes. The layout is not fixed but some standard can by used with advantage. We use a low level parameterization defined in MPEG4 standard [7].

We have done a connection of the feature points to several spline curves to cover the deformations caused by muscles under epidermis layers. The lips are parameterized by 16 feature points and two closed curves. One curve connects eight points from FAP group 8 except the point 8.1 to approximate the outer lip conture. The inner lip conture is created from FAP group 2. The deformation of dermis caused jaw rotation is approximated by one open curve constructed from FAPs 2.1, 2.13 and 2.14. The centers of cheeks are controlled by two isolated feature points FAPs 5.1 and 5.2. The tongue model is controlled by two curves defined in *sagital* and *transversal* planes. The first one is defined by five feature points. The tongue tip and tongue dorsum correspond to FAPs 6.1 and 6.2. The second one controls the width of tongue body and is constructed from five feature points. The sides of tongue body are controlled by FAPs 6.3 and 6.4 and the feature point on the tongue tip is shared with the first curve.

A data analysis of measured FAPs can produce a high level parameterization of visual speech. We have used the principal component analysis (PCA) to reduct a feature point space to four dimensions. Currently we have these parameters: *lip opening*, *lip protrusion*, *upper lip raising* and *jaw rotation.*

4. Data acquisition and audiovisual corpora

For Czech speech, audiovisual corpus suitable for synthesis was not available. We have designed methods for a capturing of visual speech as well as a reconstruction of 3D shape of face. For reconstruction, we designed scanning process to get a polygonal mesh approximating static shape of speakers face [8]. This 3D data are used for the animation schema described in the section 3. For a capturing of visual speech, we designed two methods to get articulatory trajectories of lip movements. Both are suitable for the training process describe in the section 5.1. First one is based on an optical tracing and retro-reflex marks glued on speakers face [9]. A infrared illumination, video records and a stereo vision technique are used to create time sequence (trajectory) of 3D positions of these marks. We collected one corpus composed from 318 sentences and three speakers (one female and two male). In addition, the corpus contains CVC and VCV words artificially constructed from the Czech vowels and consonants.

The second method computes articulatory trajectories by using the template matching technique [10]. The standard colored video records of front view on speaker face are parametrized by several templates. The predefined templates capture images of speaker mouth (lip and teeth). We collect more than 80 templates represented the key lip shapes and labiodental configurations. The templates are manually parametrized and data then reducted by PCA. Processing of all frames in the the video records produces the required articulatory trajectories. We completed one corpus consisted from 974 sentences and one female speaker who is a speech therapy expert. Both corpora are time labeled to phoneme speech segments using HMMs and the Viterbi algorithm.

5. Synthesis method

Currently we use two synthesis methods which are able to form articulation trajectories. Both take account labial coarticulation effects in fluent speech. First one is well known Cohen-Massaro coarticulation model [11], the second one is our method using the selection of articulation targets [9].

5.1. Selection of articulatory targets

The synthesis method is based on the regression tree technique (CART) [12]. The method predicts the articulatory targets as one continuous value given for a phoneme and articulatory parameter separately. Let $d(x)$ is a predictor returning one value z derived from attributes stored in $x \in X$. x describes the processed speech segment and its phonetic context. x can be composed from real or categorical value. For one phoneme and one parameter, the predictor $d(x)$ is constructed from the set $\mathcal{L}$ composed from pairs $\mathcal{L} = (x_1, y_1), \ldots, (x_N, y_N)$ completed from the training part of audiovisual corpus. y_n are articulatory targets obtained from the centers of speech segments and N is number of all occurrences. The synthesis problem can be decomposed to:

- a computation of the predicted value in leaf nodes
- a cut of one node to two children nodes
- a end condition determining leaf nodes

The predicted value is based on the computation of the least squared error:

$$R(d) = \frac{1}{N}\sum_{n=1}^{N}(y_n - d(x_n))^2, \tag{7}$$

where $R(d)$ is a prediction error of the predictor d.

Let T_{sp} is a binary tree for speech segment s and articulatory parameter p, b is its subtree determined by certain pairs (x_n, y_n). With respecting (7), the predicted value z_{sp} is obtained as a optimal value:

$$\bar{z}_{sp}(b) = \frac{1}{N(b)}\sum_{n \in b} y_n, \tag{8}$$

where $N(b)$ is number of pairs (x_n, y_n) in b. For the cut of b, the cross validation is used to estimate the the prediction error $R^{CV}(b)$. $R^{CV}(b)$ is computed from 10-fold randomly divided train set $\mathcal{L}$ to subset $\mathcal{L}_1, .., \mathcal{L}_v$, Eq. (9). Ten predictors $d^v(x)$ are trained on the sets $\mathcal{L} - \mathcal{L}_v$.

$$R^{CV}(b) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(x_n, y_n) \in \mathcal{L}_v}(y_n - d^v(x_n))^2, \tag{9}$$

For a cut σ of b to a left subtree b_L and a right subtree b_R, a decrease of the prediction error is given by the formula (10).

$$\Delta R(\sigma, b) = R^{CV}(b) - R^{CV}(b_L) - R^{CV}(b_R) \quad (10)$$

An optimal cut σ^* from all $\mathcal{S}$ is:

$$\Delta R(\sigma^*, b) = \max_{\sigma \in \mathcal{S}} \Delta R(\sigma, b) \quad (11)$$

Let $T_{sp\,max}$ is final tree computed by the recursive application of Eq. (11) and with end condition $N(b) \leq N_{min}$. Furthermore a pruning of $T_{sp\,max}$ is used to reduct its size. The prediction error (9) is determined for a sequence of trees T_k created by a sequential pruning of the branches of $T_{sp\,max}$. The optimal tree T_{k_0} is selected according to Eq. (12).

$$R^{CV}(T_{k_0}) = \min_k R^{CV}(T_k) \quad (12)$$

The last step of the designed method is a formulation of regression questions to define the space X. The type of questions is crucial. We have formulated following questions specially collected for covering of the labial coarticulation:

- left/right phonetic context (triphone),
- is left/right phoneme vowel,
- is left/right phoneme no speech segment,
- is left/right phoneme bilabial stop,
- is left/right phoneme labiodental segment,
- is left/right phoneme fricative segment,
- the nearest left/right articulatory resistant phoneme.

The articulatory resistant phoneme is such phoneme which is dominant in particular articulatory parameter, for example phoneme /u/ for lip protrusion. The continuous values of x are rather used for prosodic properties of speech segments:

- time duration of the speech segment of the processed target,
- time duration of the preceding/following speech segment,
- the energy of the acoustic signal at moment of selection.

The articulatory trajectories are formed for each articulatory parameter separately. For each speech segment from the synthesized trajectory, the synthesis method selects articulation targets z_{sp} derived from relevant T_{k_0}. The articulatory targets are placed to the centers of the speech segments and a interpolation is used to create the continuous trajectory (e.g. 25 frames per seconds).

6. Conclusions

The Czech talking head system has been evaluated [10]. Significant benefits of the intelligibility of synthesized visual speech were observed. In this article, we made an outline of our talking head system. The summarization of new features is presented. The extended animation schema allows more precise approximation of inner and outer lip contour as well as deformations of the tongue mesh. The parameterization of visual speech is based on MPEG4 standard. The acquisition of necessary data and audiovisual corpora is mentioned and included articulatory trajectories are suitable for experiments with data driven synthesis methods.

The mathematical formulation of our synthesis method was introduced. This synthesis method based on the selection of articulatory targets converts phonetic input into the articulatory trajectories as well as precisely coverages of labial coarticulation effects observed in visual speech. At the present, the talking head system is applied in synthesis of Czech signed speech. An perception study with deaf children will be carried out to extend the system about new features.

7. Acknowledgements

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 and by the Ministry of Education of the Czech Republic, project No. ME08106.

8. References

[1] J. Matoušek, J. Romportl, and D. Tihelka, "Current state of Czech text-to-speech system ARTIC," in *TSD 2006*. Pilsen, Czech republic: Springer-Verlag, Berlin, Heidelberg, 2007, pp. 439–446.

[2] S. Kshirsagar, S. Garchery, and N. Magnenat-Thalmann, "Feature point based mesh deformation applied to MPEG-4 facial animation," in *Deform'2000*. Kluwer Academic Publishers, 2000, pp. 23–34.

[3] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi, "Modelling an Italian talking head," in *AVSP 2001*, Aalborg, Denmark, September 2001.

[4] L. Revret and C. Benot, "A new 3D lip model for analysis and synthesis of lip motion in speech production," in *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

[5] Z. Krňoul and M. Železný, "Realistic face animation for a Czech Talking Head," in *Proceedings of TEXT, SPEECH and DIALOGUE, TSD 2004*, Brno, Czech republic, 2004.

[6] M. M. Cohen, J. Beskow, and D. W. Massaro, "Recent developments in facial animation: an inside view," in *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

[7] I. S. Pandzic and R. Forchheimer, "The origins of the MPEG-4 facial animation standard," in *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, I. Pandzic and R. Forchheime, Eds. John Wiley & Sons, 2002.

[8] Z. Krňoul, P. Císař, and M. Železný, "Face model reconstruction for czech audio-visual speech synthesis," in *SPECOM 2004*, St. Petersburg, Russian Federation, 2004.

[9] Z. Krňoul, M. Železný, L. Müller, and J. Kanis, "Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis," in *Proceedings of INTERSPEECH 2006 - ICSLP*, Bonn, 2006.

[10] Z. Krňoul and M. Železný, "Innovations in czech audio-visual speech synthesis for precise articulation," in *Proceedings of AVSP 2007*, Hilvarenbeek, Netherlands, 2007.

[11] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. M. T. . D. Thalmann, Ed. Tokyo: Springer-Verlag, 1993.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, Kalifornia, USA: Wadsworth and Brooks, 1984.