ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Innovations in Czech Audio-Visual Speech Synthesis for Precise Articulation

Zdeněk Krňoul, Miloš Železný

University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
zdkrnoul@kky.zcu.cz, zelezny@kky.zcu.cz

Abstract

This paper presents new steps toward animation of precise articulation. The acquisition of audio-visual corpus for Czech and new method for parameterization of visual speech was designed to obtain exact speech data. The parameterization method is primarily suitable for training a data driven visual speech synthesis systems. The audio-visual corpus includes also specially designed test part. Furthermore, the paper presents the collection of suitable text material for test of visual speech perception and also the procedure how can be the test performed. The synthesis method based on the selection of visual unit and animation model of talking head is extended. The synthesis system is objectively and subjectively evaluated.

Index Terms: visual speech synthesis, unit selection

1. Introduction

In the previous work, we used to control articulation of lips using data from the audio visual corpus based on stereo records and 3D reconstruction technique [1]. This recording method employs retro-reflex makers glued on speakers face. The method ensures exact lip-tracking but is useable only for shape of outer lip contour. Since the articulation of the inner lip contour is not tracked because markers cannot be placed, visually important bilabial or labiodental occlusions should not be precisely captured. However the shape of inner lip contour is crucial for successful lip-reading. Thus a common video record of visual speech contains the important information for human perception. The front and side view of camera should ensure the equivalent 3D representation. The visual synthesis system can then control the animation model of ”talking head” in a more efficient manner and the applicability of such systems for lip-reading support is rising.

Therefore we designed new audio-visual speech corpus based on audio and video records of continuous speech. To obtain more precise visual synthesis, we composed the corpus from records of speech therapy expert.

The visual speech synthesis from processed corpus data can be based on concatenation of selected lip shapes. We designed synthesis method based on standard regression technique. The binary tree clustering has adequate performance to preserve variable lip shapes in coarticulated continuous speech [2].

No text material for audio visual perception test is available for Czech. Design of new test on audio visual speech perception is generally complicated task. We tried to compose the primary collection of text material and its audio and video records as a part of the designed corpus.

2. Acquisition of audio-visual speech corpus

For data-driven visual synthesis, we collected a new audio-visual speech corpus. The corpus is composed approximately from 2 hours continuous speech of one female speaker. Since the aim has been to get more intelligible articulation data, we captured the audio-visual data from the speech therapy expert. Totally 964 sentences were recorded. Each sentence was recorded only once. The text material of the corpus was divided into two parts: training and test. The text material of 814 sentences was selected for training process to cover as many variabilities of phonetic contexts as possible. Czech newspapers were used as a source of this text material. Length of the sentences differs from short to long. The text material of the test part was selected with different conditions which are described in Section 4.2.

For better quality of audio recordings, we recorded the corpus in a soundproof and reflection-free room using professional audio equipment. The visual part was recorded in calibrated views of 2 digital cameras with resolution 720x576 pixels and frequency 25 frames per second. First camera was directed to the front view and second to the side view at the speaker’s face. The speaker sat in front of the microphone and text data were prompted using a computer screen. The speaker’s head was propped to obviate its turning. The scene was lighted using standard studio lights fixed constant during the whole record stage. The speaker’s lips were not marked by any markers or make-up. The speaker’s head was marked only by auxiliary beads on the forehead and the nose to identify head movements. The recorded data were audio-visually transcribed (real pronunciation was taken down, together with non-speech events such as background noise, loud breathing or lip clicking). Finally the audio and video streams were time synchronized and audio-visual data were stored separately for each sentence.

2.1. Lip-tracking and parameterization

We employed a lip-tracking method based on template matching [3]. The algorithm repeatedly computes the matching with several lip shape templates. This detection technique determines the measure of the similarity between the image and the template by cross correlation score (1).

$$C(u,v) = \frac{\sum_{x,y}(I(u+x,v+y)-\bar{I}_{u,v})\hat{T}_{x,y}}{\sum_{x,y}(I(u+x,v+y)-\bar{I}_{u,v})^2\sum_{x,y}\hat{T}^2_{x,y}} \quad (1)$$

where I is the processed image of a face (video frame), $\bar{I}_{u,v}$ is the local image mean of the pixel location u, v. $\hat{T}_{x,y}$ is difference between template T and its mean $\bar{T}$.

$$\hat{T} = (T(x,y) - \bar{T}) \quad (2)$$

Let us assume that C_i is the best match correlation score for templates T_i and all local positions u, v

$$C_i = \max_{u,v} C(u,v). \tag{3}$$

The algorithm matches N templates and stores C_i for each frame of processed video record. The C_f correlation score of f-th template with the best match is given as maximal value (4).

$$C_f = \max_i C_i \quad i = 1 \ldots N \tag{4}$$

In our approach the repeated template matching algorithm is applied only on video frames from first camera (front view) and the correlation score is calculated in a grayscale representation. The collection of templates was experimentally determined by the selection of more than 200 video frames with different lip shapes and lower teeth placements. We attempted to cover all possible shapes, including mouth opening (from constringed lips through relaxed lips to full opening), different mouth widths, protrusion, several variants of labiodental occlusions as well as several variants of lip rounding. These selected frames of the speaker face were cropped down to an identical image area of 70x120 pixels (lips and surrounding). The cross correlations scores between these templates were calculated and only 83 templates were preserved. In addition, the templates were supplemented with the images from side view camera. The rest more then 150k frames of entire corpus were processed and 99.6% of all frames were detected with correlation score $C_f > 0.9$.

2.2. Parameterization of the template

The lip shape in templates should be described by some arbitrary technique. We used the parameterization based on MPEG-4 FAPs[1]. The outer and inner contour and teeth placement is approximated by several control points, as depicted in Figure 1. The images of the front and side view of the speaker's mouth were manually marked in positions of relevant FAPs. We suppose the symmetric shape of lips, and thus we marked only left half of the mouth in the front view. The teeth placement is parameterized as a vertical coordinate of upper tip of lower teeth. These marked positions provide 2D image points only. The protrusion of the lip corner and the upper and lower lip center is additionally marked in the images of the side view. The calibration of cameras enables to transform these image positions to coordinates in [mm]. The full size of parameter vector is 20. The Figure 1 illustrates the approximation of lips shape and also best matched template.

We apply the principal component analysis (PCA) to all N parameterized templates to reduce the dimension from 20 to 3 main components (PCs). These PCs represent compound lip shape. The meaning of them is: PC1 - lip opening, PC2 - lip width and protrusion, PC3 - upper lip raising. The vertical position of lower teeth is considered as independent and is added as fourth component (PC4). This reduced vector is used for generation of trajectories by the synthesis method.

2.3. Articulatory trajectories

For each j-th video frame, the best match f and correlation $C_f(j)$ is repeatedly computed using (4). The articulatory trajectories can not be directly generated from the parameterization of relevant f templates because it produces jerky movements at borders between templates that do not correspond to the speaker's articulation. The following post-processing is based on the idea that final articulatory trajectories with fluent shifts are obtained by the interpolation of the selected key-frames. The decision which frames will be the key-frames is determined from sequence of C_f scores. This sequence is divided into short segments S of several adjacent frames with the same f templates. If the condition (5) is valid, the j-th frame is selected as key-frame.

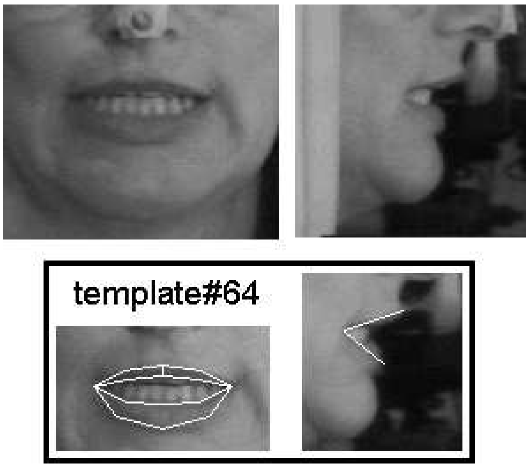

Figure 1: *The approximation of a lip shape by the outer and inner contour, below the best match template with correlation score 0.96.*

$$C_f(j) > (\max_{j \in S} C_f(j) - t_c) \tag{5}$$

Let t_c is correlation threshold. Using the threshold causes that each segment S will be represented by several key-frames instead of only one key-frame target of maximal C_f. Thus the longer segments are more precisely interpolated. The final trajectory is interpolated from all key-frames of all segments S. We used correlation threshold $t_c = 0.001$. For generating final articulatory trajectories, the cubic interpolation method is used.

3. Synthesis system

3.1. Segmentation of visual speech

For the phone time alignment of the corpus, the hidden Markov models (HMMs) and Viterbi forced alignment were employed. The HMMs were trained on compound parameterization vectors that consist of 4 visual PCs and 13 audio and energy LPC parameters. For this purpose, we trained 5-state HMMs for each of 39 Czech phones and non-speech events rather than for equivalent viseme subset. We assume that the time labels of mono-phones are placed to the articulation targets. We extracted these labels as the start time labels of middle HMM state.

3.2. Selection of visual units

In this approach, the visual unit selection method is used to generate articulatory trajectories. The method synthesizes the trajectories by the concatenation of selected lip shapes and interpolation of phone transitions. For each phone in time-aligned input sequence, the appropriate lip shape candidates are selected. This selection is based on the regression tree clustering technique. In this approach, four trees were associated with each

[1] Complete technical information ISO/IEC MPEG-4 Part 2 (Visual)

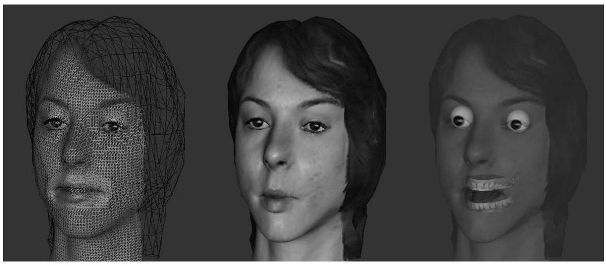

Figure 2: *The animation model "Petra" which was used in perceptual test.*

visual parameter. Trees are constructed for initial clusters collected from all occurrences of particular phone in training part of corpus. The binary division of each cluster the size of which varies from 10 to 4000, is carried out in compliance with specific set of regression questions.

Good choice of questions is crucial for good coverage of coarticulations. We collected following set of questions based on both left and right speech context. The speech context is given in discrete and continuous form. The discrete form determines the phone occurrence and is used either in more general questions as: *"is vowel", "is no speech segment", "is bilabial", "is labiodental", "is fricative"* or very specific question on particular occurrence of closest coarticulatory resistant phone. The set of coarticulatory resistant phones is defined separately for each PC parameter. For example, phones /o/ or /u/ are marked as the lip-rounding resistant (PC2). The continuous form of speech context is given by time distance to preceding (following) phone. In additional, the set of continuous questions is supplemented with the question on energy of acoustic signal in the moment of selection of lip candidate. The target vector of lip candidate is a result of tree-walking of these 4 trees to the terminal nodes. The synthesis process concatenates these target vectors and determines the cubic interpolation according to input time-labels. These trajectories control animation model of "talking head".

3.3. Animation model

We enhanced our animation model of talking head [4]. The animation schema based on spline functions was extended by ability of controlling inner lip contour. 8 control points for approximation of the inner lip contour were additionally added to original 8 points approximating the outer lip contour. The animation of the lip shape together with jaw rotation is thus controlled by 17 points. Since the synthesis method uses reduced parameterization of lips, the animation model is supplemented with linear model which transforms PCs to these control points. The value range of 4 PCs and scale factor of transformation model are manually adjusted on relaxed lip closure and full mouth open.

The animation model "Petra" is illustrated in Figure 2. The model is composed from textured triangular surfaces of face, teeth and tongue. The shape of the face mesh is adjusted using a 3D reconstruction method on the shape and the scale of the specific face [5].

4. Evaluation

4.1. Objective evaluation

Firstly we evaluated synthesis method based on visual unit selection at the trajectory level. The sequences of phones and time labels of test sentence were use for synchronization with the speech rate in the corpus. The trajectories of 160 test sentences were synthesized and compared with original trajectories. The quality was evaluated by root mean square error (RMSE) and correlation scores (CORR). The average RMSE and CORR scores are summarized in Table 1.

Table 1: *Average root mean square error (RMSE) and correlation CORR for test sentences.*

	PC1	PC2	PC3	PC4	average
RMSE [%]	13.1	11.8	9.4	17.4	12.9
CORR	0.6	0.74	0.4	0.48	0.56

4.2. Audio-visual test sentences and perceptual evaluation

We evaluated the benefit of the visual speech synthesis system by calculation of the difference between the auditory and audio-visual percentage intelligibility score. The proposed evaluation process is adjusted for normal-hearing subjects. Three separate visual conditions were used: *audio-alone, audio+synthetic face and audio+natural face*. The result of the audio-alone and audio+natural face conditions were used as the baseline levels of intelligibility for degraded speech. 4 levels of audio degradation combined with visual conditions finally produced 12 audio-visual conditions. Firstly, the suitable text material of test sentences for Czech has to be collected. We chose the methodology of speech reception based on short sentences presented in noise and detection of key words. The sentences of test part of corpus are composed with consideration to the structure of test sentences [6]. Our test is based on 13 sentence lists. Each list consists of 12 sentences. One list is presented in the beginning of each session as a trial to present all conditions to a subject and is not scored.

We assume that each sentence will be presented only once per test subject. Thus the equivalent difficulty of intelligibility or unintelligibility across sentences should be considered. The number of words in test sentences varies from 4 to 6 and constantly 3 words were marked as keywords. The structure of 12 sentences in each list is following: 5 sentences with keywords in form *subject-verb-object*, 5 *subject-verb-adverb*, 1 *subject-verb-complement* and 1 *subject-verb-other*. A lot of such sentences were selected from Prague Dependency Treebank project[2] (PDT). The text source of PDT is Czech newspapers and technical journals. The final set was determined with respect to the neutral predictability and lip-reading difficulty to ensure small differences of score between subjects or conditions. Thus the sentences are composed from familiar words but no sentence should include all keywords with visually distinct phone in the beginning.

White noise was low-pass filtered at 10 kHz and was mixed to audio speech signal of natural voice from the corpus. The intensity was determined by Signal to Noise Ration (SNR). The audio degradation was modeled by four fixed SNR levels: 0dB, -6dB, -12dB and -18dB. For each SNR level, the synchronized video records of natural and synthetic face are supplemented.

[2] PDT 1.0, available at http://ufal.mff.cuni.cz/pdt/

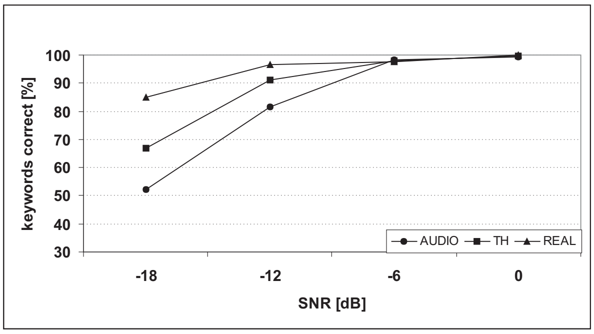

Figure 3: *The result of audio-visual study, percentage average keywords correct of 9 subjects for 4 SNR conditions.*

The video records of natural face were taken from the test part of corpus, the video records of synthetic face were generated by animation model showed in Figure 2. The animation was controlled by synthesized trajectories from objective evaluation 4.1. The animation of tongue and cheeks was not controlled. The resolution of generated video records was 372x480 pixels and frames rate was set to 25 frames per second.

4.3. Test procedure

5 female and 4 male served as subjects. All of them were native Czech speakers with normal hearing and seeing. No subjects were familiar with the text material. Theirs age varied from 20 to 50. The test was arranged in a way that subject sat in font of PC screen with the headset. The responses were recorded on the dictaphone. The size of head on the screen was approximately 12cm. The audio-visual conditions for each sentence list were randomly generated across subjects with all 12 sentences in assigned condition. 144 selected records of these 12 lists were finally presented in random order. For all lists and subjects, the score was computed as percentage of identified key words. Small errors in morphology were ignored.

4.4. Results

The results of audio-visual study are summarized in Figure 3. There are average scores from all subjects. Further the scores were entered into repeated analysis of variance (ANOVA). The presentation of perception test was significant (F(11,96)= 32.41, $p<0.05$). The following pairwise comparison with ($p<0.05$) indicates that the intelligibility score for *audio-alone* condition significantly decreases on SNR -12dB and -18dB. For *audio+natural face*, the significant decreasing was observed only on SNR -18dB. The significant benefit of *audio+synthetic face* against *audio-alone* is observed on SNR -18dB.

5. Summary and conclusions

This paper dealt with the problem of design of precise visual speech synthesis. The acquisition of new audio-visual corpus was one step to achieve it. The corpus is composed from 964 sentences specially selected for training and evaluation of data driven synthesis system. The audio-visual record of speech therapy expert was obtained to obtain precise data. The lip-tracking method was designed to allow visual speech parameterization of these video records. The method employs template matching technique and set of image templates that capture images of mouth in different shapes. The lip shape of the outer and inner contours in the templates are approximated by several control points. The templates with maximal correlation score are selected to parameterize each frame in the video records. The synthesis method based on selection of lip shapes is used to synthesize articulatory trajectories. The lip shapes are given by regression trees. In regression the sequence of questions on phonetic context determines appropriate position of control points to coverage lip coarticulation in fluent speech.

Two measures for objective evaluation were calculated to get to know the similarity of the synthesized articulatory with the trajectories from the corpus. The results indicate that the synthesized trajectories slightly differ. Hence the perceptual evaluation was carried out. As a result, the audio-visual study proved significant benefit of face conditions for higher audio degradations. However the difference in intelligibility of natural face and synthetic face is still significant. The reason for it can be consequence of several factors. The first fact is that the tongue and cheek animation was not included in our synthetic face. The next is that the synthetic face animations generally do not reach intelligibility score of natural face. It can be account synthesis system. In our approach it causes either synthesis method or animation model. Our animation model was semi-automatically reconstructed from human face and with potential errors. Thus the lip shape for particular phones should not be such distinct as in an artificially created face model. We used also automatic time-alignment of phones with the combined parameter vector composed from visual and acoustic parameterization to synthesize test sentences.

6. Acknowledgements

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

7. References

[1] Krňoul, Z., Železný, Císař, P., Holas J., ”Viseme Analysis for Speech-Driven Facial Animation for Czech Audio-Visual Speech Synthesis” In Proceedings of SPECOM’2005; University of Patras, Greece (2005)

[2] Krňoul, Z., Železný, M., Müller, L., Kanis, J., ”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).

[3] Lewis, J. P., ”Fast Normalized Cross-Correlation”, Industrial Light & Magic http://www.idiom.com/ zilla/Papers/nvisionInterface/nip.html

[4] Krňoul, Z., Železný, M., ”Realistic Face Animation for a Czech Talking Head” In Proceedings of 7th International Conference on TEXT, SPEECH and DIALOGUE TSD 2004. Springer-Verlag Berlin Heidelberg (2004).

[5] Krňoul, Z., Železný, M., Císař, P., ”Face Model Reconstruction for Czech Audio-Visual Speech Synthesis” In Proceedings of SPECOM’2004; Saint-Petersburg (2004)

[6] MacLeod, A., Summerfield, A.Q., A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use British Journal of Audiology, 24(1), 29-43.