Effect of Visual Speech in Sign Speech Synthesis

Zdeněk Krňoul

Abstract—This article investigates a contribution of synthesized visual speech. Synthesis of visual speech expressed by a computer consists in an animation in particular movements of lips. Visual speech is also necessary part of the non-manual component of a sign language. Appropriate methodology is proposed to determine the quality and the accuracy of synthesized visual speech. Proposed methodology is inspected on Czech speech. Hence, this article presents a procedure of recording of speech data in order to set a synthesis system as well as to evaluate synthesized speech. Furthermore, one option of the evaluation process is elaborated in the form of a perceptual test. This test procedure is verified on the measured data with two settings of the synthesis system. The results of the perceptual test are presented as a statistically significant increase of intelligibility evoked by real and synthesized visual speech. Now, the aim is to show one part of evaluation process which leads to more comprehensive evaluation of the sign speech synthesis system.

Keywords—Perception test, Sign speech synthesis, Talking head, Visual speech.

I. Introduction

THE CURRENT trend is to have information in an electronic form. Sign language synthesis including translation from a spoken language to a sign language is the important element of information technologies. Deaf users have problems with these technologies because they could have a difficulty with reading. The direct correlation between the ability to speak and read and very strong correlation between reading skills and the degree of hearing loss is observed [1]. In particular, the Czech grammar is for deaf people much larger problem because the Czech language is inflected and has the freer word order. They can understand all individual words in an utterance but it does not give them sense [2].

In the scope of this paper, we distinguish the Czech Sign Language (CSE) and Signed Czech (SC). The CSE is a natural and adequate communication form and a primary communication tool of the hearing-impaired an deaf people in the Czech Republic. It is composed of the specific visual-spatial resources, i.e. hand shapes and movements (the manual component), facial expressions, head and upper part of the body positions (the non-manual component). CSE is not derived from or based on any spoken language because has its lexical and grammatical structure. On the other hand, SC was introduced as an artificial language system derived from the spoken Czech language to facilitate communication between deaf and hearing people. SC uses grammatical and lexical resources of the Czech language. During the SC production, visual speech is audibly or inaudibly articulated and the CSE signs of all individual words of the sentence are simultaneously signed.

Zdeněk Krňoul is with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic e-mail: zdkrnoul@kky.zcu.cz

Generally, an increase of the intelligibility of synthesized sign speech is expected when the non-manual component of sign speech is supplemented. In the case of SC, visual speech given by the lip articulation is very important element of the speech perception in the process of human-computer interaction. Therefore, the good interpolation of lip parameters is crucial for the overall performance of the sign speech synthesizer used for example in the sign-language-enabled information kiosk [3]. Evaluation of intelligibility of perceived visual speech is one way how to improve this complex problem. This paper is structured as follows. Section II introduces the issue of audiovisual speech databases necessary for both the development and the evaluation of a sign speech synthesis system. In Section III, one evaluation methodology is proposed and this procedure is verified by a perceptual test. The summary is in Section IV.

II. Audiovisual Databases

A. Data for Synthesis Process

Various data sources are required to create automatic 3D synthesis of sign speech (an avatar animation). In order to obtain synthetic lip movements as realistic as possible and apply it in the sign speech synthesis process, lip tracking during articulation of a real speaker has to be considered.

To estimate 3D parameters of parametric parts of the animation model (shape model), a record of an audiovisual database is recommended. Usual setup is using stereoscopic view. For stereovision, we use two cameras acquiring visual data from two different angles. In this setup, these video streams have to be synchronized with each other to get correct 3D correspondences. Besides the fact, that is difficult to ensure that both frames of stereo pair are acquired at the same time, requirement of two cameras makes the considered investigation more expensive.

The setup for stereovision using only one camera was designed [4]. This solution captures stereo images using only one camera and system of 4 mirrors. Proposed setup is illustrated in Fig. 1. The stereo image pair is acquired using only one camera to avoid the mentioned problem of synchronization. This setup takes advantage from the fact that while a face is rather portrait, a video frame is rather landscape and thus putting two images of a head does not decrease the image resolution dramatically. Acoustic data are acquired at the same time by the high quality microphone and by the electroglottograph. Synchronization of video, audio and electroglottograph signals is made using short clapperboard sequence recorded at the beginning of each recording session.

To obtain the 3D parameters from visual data of a real person with high precision, an enhanced technique using reflexive markers in the infrared spectrum can be used. This

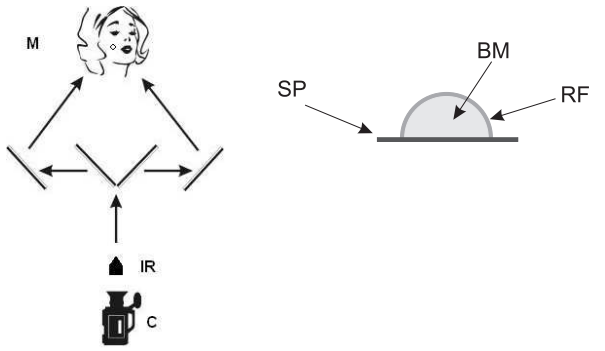

Fig. 1. On left: recording of the THC1 audiovisual database ((C) video camera, (IR) infrared light source, (M) speaker's head); on right: the schema of the reflexive markers detected in the infrared spectrum ((BM) marker, (RF) retro-reflexive film, (SP) sticking strip).

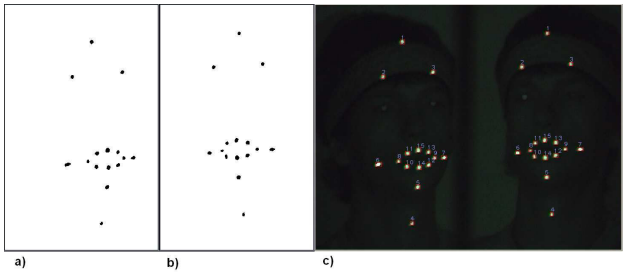

Fig. 2. An example of the left (a) and right (b) stereo view that are preprocessed by image segmentation; (c) identification of the markers in the video files.

way simplifies the image segmentation that produces more accurate model parameters. This proposal considers a system composed of one camera, infrared (IR) source of light, four mirrors with the console and few small passive markers. The mirrors combine two stereo views into one image. The standard camera calibration is processed at the beginning of each record session and is used for the reconstruction of 3D positions of the markers. The principle of tracking is based on the detection of the markers which reflect IR light source, Fig. 2.

Twelve markers were glued to the speaker's chin, cheeks, laryngeal prominence and lips. Three markers were fixed on the forehead for the detection of head movements. The markers have half-spherical shape with reflexive material on the surface, see Fig. 1. The diameter was 5 mm for three markers on the forehead and 3 mm for the rest of markers. The lip contour was approximated by 8 markers. The placement of these points ensures the good lip shape representation and good visibility by camera. The marker on the chin approximates the rotation of the jaw and the remaining markers are used for the articulation of the cheeks and the neck. These markers capture specific movements especially of phonemes /p/, /b/ or /g/. Complete parameterization describes important articulation movement of the face.

The video signal was recorded with 25 interlaced frames per second. The acoustic signal was recorded simultaneously by the external microphone and by the electroglottograph, see Table I.

TABLE II
THE SYNTACTIC STRUCTURE OF SENTENCES INCLUDED IN THE TEST LIST.

Number of Sentences	Types of Sentences
5	Subject – Verb – Adverb
5	Subject – Verb – Object
1	Subject – Verb – Complement
1	Subject – Verb – Another

This method capturing visual speech was experimentally used for the record of the audiovisual speech database (THC1). The text material consists of 318 Czech sentences. A controlled selection of the sentences has to be considered here to get the phonetically balanced text material. The selected sentences contain at least three words and maximally 15 words [5]. The text material does not contain foreign words, numbers or abbreviations. In addition, symmetrical CVC and VCV combinations were incorporated to the text material, where the V group is formed by vowels /a/, /e/, /i/, /o/, /u/ and the C group by consonants /f/, /p/, /s/ , /d/, /D/, /S/, /l/, /r/, /j/, /k/, /h/, see the table of all Czech phoneme units [6]. Finally, isolated pronunciations of the phonemes is recorded as well. Totally, THC1 includes mentioned audiovisual speech data of three speakers (two male speakers (SM1, SM2) and one professional female speaker (SF1).

B. Data for Perceptual Test

In many cases, it is not possible audiovisual data collected for a training process to use for a perceptual test. This is also case of the THC1 database that is conformed to the optical lip tracking under the conditions specified in Section II-A. People percept the visual speech from image of the speaker's face which is naturally illuminated and without auxiliary marks. Therefore, new video records have to be captured to get appropriate speech material.

For new speech database, another suitable text material has to be collected as well. For this purpose, the methodology of visual speech perception based on short sentences presented with acoustic noise and key word detection can be considered [7]. However for this proposal, we have to collect Czech sentences. The number of words is restricted to four, five or six but three words are marked as keywords in all cases. Next, the selection process is driven by the syntactic structure of the sentences, see Table II. Thus, the equivalent intelligibility or unintelligibility across sentences should be considered.

Firstly, a lot of appropriate sentences has to be selected from adequate source. Prague Dependency Treebank (PDT 1.0 is available at http://ufal.mff.cuni.cz/pdt/) is used here [8]. Text material of PDT consists of these sources:

- Lidové noviny (Daily newspapers), 1991, 1994, 1995
- Mladá fronta Dnes (Daily newspapers), 1992
- Českomoravský Profit (Business weekly), 1994
- Vesmír (Scientific magazine), Academia Publishers, 1992, 1993

The final set of the test sentences should be determined with regard on the neutral predictability and lip-reading difficulty to ensure small differences of score between participants and considered conditions. Thus for this proposal, the test

TABLE I
PARAMETERS OF THE AUDIOVISUAL DATABASES, THC1.

Type of Data	Recording Device	Resolution	Data Compression
Video	Video camera, Sony TRV740E	720x576, 25fps	Indeo video 5.11
Audio 1	Microphone, Sennheiser ME65	16bit, 44kHz	PCM
Audio 2	Electroglottograph, EG2-PC	16bit, 44kHz	PCM

sentences are composed from familiar words but any sentence does not include all keywords with visually distinct phoneme at the beginning.

A soundproof room and professional sound recording equipments are premise to get acoustically clear audio traces. For this proposal, visual speech is captured by two digital cameras for the front and the side view. The professional speaker, SF1, was selected and white lighting of constant intensity throughout the recording session was kept. Finally, the audiovisual records were annotated and time-synchronized using the same procedure of the THC1 database. The synchronized records was divided to separate data files. Image processing would be involved to make a compression of video stream and to crop and resize video frames on suitable size.

C. Audiovisual Speech Segmentation

Segmentation of visual speech recorded in audiovisual speech databases provides decomposition of measured articulatory trajectories to appropriate sub-segments. Automatic speech segmentation is necessary condition for training of a control model since manual segmentation of large amounts of speech data to the phoneme units is very time consuming process. Firstly, phonetic transcription and manual identification of sentence borders is determined. There are used the 47 phonemes that are using by Text To Speech synthesis (TTS) and Automatic Speech Recognition (ASR) systems [6]. On principle, both visual and acoustic signal are available for the considered segmentation process. In the case of visual data, the trajectories describing the movement of the lips are directly obtained from the database. In this proposal, the principle of recording of the THC1 database establishes particular type of parameterization of the lip shape. For audiovisual database, THT, parameterization of visual speech is not directly available because we have only raw image data.

In the case of acoustic data, the process of segmentation requires a parameterization of the acoustic components of audiovisual speech. This proposal does not inspect a problem of acoustic speech parameterization. The acoustic speech parameterization is addressed research of TTS as well as ASR systems. Well-known MFCC parameterization is considered here. MFCC acoustic parameter vector is determined with a ten millisecond (msec) window and four msec off-set.

However, the visual parameterization is obtained in principle with a different off-set than the acoustic speech parameterization. This is due to a fixed frame rate of video recording equipment which is in this case 25 fps (40 ms). To ensure the synchronization of these components and to use both acoustic and video signal for segmentation, the articulatory trajectories for visual speech have to be interpolated, for example by the cubic spline interpolation.

In this proposal, the automatic segmentation of continuous speech stored in the THC1 audiovisual database is made from acoustic components only. The tri-phone units are selected and well know Hidden Markov Models (HMMs) and HTK toolkit are suitable software tools (HTK is a toolkit for modeling HMM and it is available on http://htk.eng.cam.ac.uk/). There are five state HMMs and the segmentation process can be performed for each speaker separately. The segmentation process is carried out in two stages. The first stage estimates HMM parameters by the Baum-Welch restimation algorithm. For this proposal, initial speech segments are determined using the Czech ASR system. The second stage carries out an allocation of the segments to individual HMM states using Vitherbi algorithm. The same segmentation procedure is considered the in the case of test sentences from the database, THT. For the synthesis process as well as a perceptual test, the position and duration of each speech segment is crucial for proper synchronization of the visual component with the acoustic component.

III. EFFECT OF VISUAL SPEECH

One possibility, how investigate the benefit of the visual speech automatically generated by a sign speech synthesis system, is an estimation of percentage differences between the auditory and audio-visual intelligibility scores. The proposed evaluation process is adjusted for normal-hearing people because considered perception of the acoustic component could be influenced by the hearing impairment. For this purpose, three visual conditions of presentation are considered: *audio-alone*, *synthetic-face* and *natural-face*. The audio-alone and the natural-face condition are baseline levels of intelligibility.

A. Formulation of Problem

This proposal involves perceptual test considering both the synthesized component of visual speech and real visual speech of a speaker. The acoustic component of speech is used for all conditions of the presentation. The acoustic component has to be collected from clean voice records that will be artificially supplemented by acoustic noise. The proposal uses white noise low-pass filtered at 10 kHz. The acoustic component of speech synthesized by TTS is not appropriate for this purpose because the perceptual test could be influenced by this factor. On the other hand, the proposed test has to enable a comparison of different approaches (different settings of the shape or control model).

For this investigation, the proposal specifically aims to compare two different control models. First one is implementation of the control model using Cohen-Massararo coarticulation (CM) model [9], and second one is the method of selection of articulatory targets, SAT model [10]. Both CM and SAT model

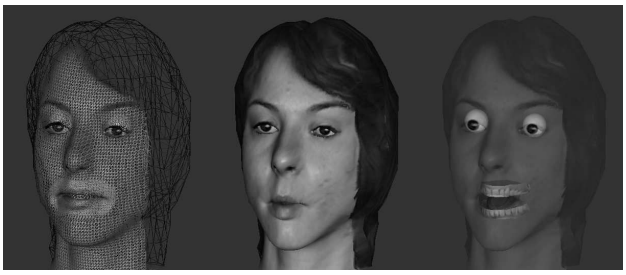

Fig. 3. ”Petra” animation model used in the proposal.

have been set on the same speech data. Continuous visual speech of the speaker, SF1, and the audiovisual database, THC1, are considered. For this purpose, it is used the 270 training sentences. The training process of CM model is carried out in accordance with the Gauss-Newton minimization procedure [11]. SAT model is set in accordance with the classification and regression tree techniques, CART [10], [12].

B. Shape Model

For proposed methodology, an identical shape model of talking head is employed [13]. The shape model uses original design with spline functions to control the shape of the outer lip contour. The shape model totally uses 11 3D control points located around outer lip contour and jaw. Since the control model is designed for a reduced parameterization (PCs), the shape model has to be supplemented about a transformation model which transforms PCs to mentioned 3D control points of shape model [14]. Value ranges of 4 PCs and a scale factor of the transformation model are manually adjusted on relaxed lip closure and full mouth opening.

The animation model ”Petra” is illustrated in Fig. 3. The model is composed from textured triangular surfaces of face, teeth and tongue. For this proposal, the triangular mesh of face is adjusted by 3D reconstruction method on the shape and scale the SF1 speaker [15].

C. Test Procedure

Ten normal hearing and seeing participants served as participants. Nine participants are native Czech speakers and one participant knows Czech language very well. None participant was familiar with the test material before the perceptual test. The participants were divided into two groups of five participants, the group A and B. The first group, A, was tested with test records incorporating the talking heads animation controlled by the CM model; the second group, B, was tested with the records incorporating the SAT model. The group, A, was comprised of three female and two male with an average age of 36 years. The group, B, was formed one female and four male, average age 39.8 years. The perceptual test was arranged so that participants set in font of a PC screen and use the headset. Their responses were recorded by a computer and stored for later evaluation. The size of speaking head on the screen was approximately 120 mm. For perceptual test, the 19 inches LCD monitor is used. The duration of one test session (one participant) is approximately 30 minutes.

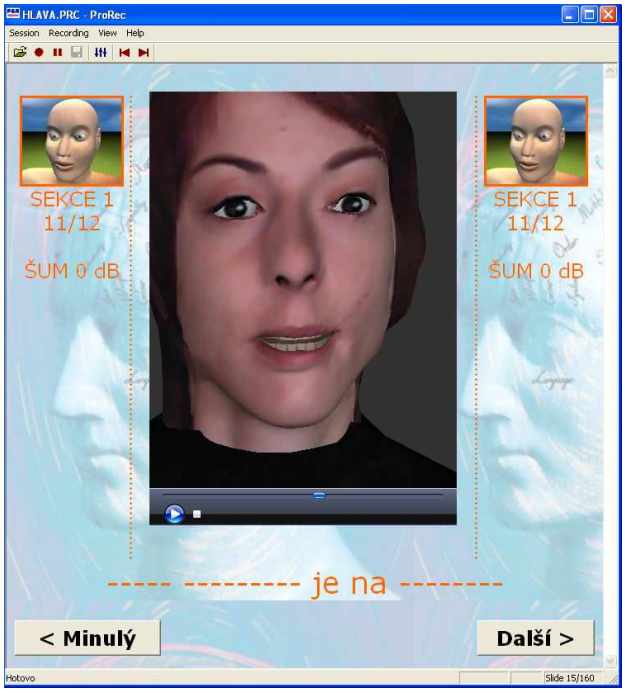

Fig. 4. The screenshot of the test application used for the audiovisual perceptual test.

Intensity of noise mixed to acoustic speech signal was controlled by a signal to noise ration (SNR) on four fixed levels: 0 dB, -6 dB, -12 dB and -18 dB. Four levels of this audio degradation combined with the mentioned visual conditions produce 12 audio-visual conditions. For each SNR level, the synchronized video records of synthesized and natural face are supplemented. The video records of natural face were taken from the THT database, the video records of synthesized face were generated by shape model, see Section III-B.

Furthermore, the talking head system must be modified to render the animation to video files. The format of the video files should be set to the same parameters as the records of the test database. Thus, size of the synthesized face and the speaker’s head is approximately the same. Following requirement has to be considered as well. Talking head application has to expect in the input the test sentences represented by a sequence of phonemes and time labels. The concept of the segmentation process is described in Section II-B. This modification enables the precise synchronization artificially generated visual components with the real acoustic component of speech. The animation of tongue and cheeks was not included here. The resolution of the video records is 372x480 pixels and the frame rate is set to 25 frames per second.

The perceptual test is based on 12 sentence lists each consisted from 12 sentences, totally 144 test sentences. One extra non-scored list is presented at the beginning of each test session as trial to demonstrate the participant all conditions of the perceptual test. 12 audio-visual conditions are randomly assigned to 12 sentence lists across participants. 144 collected records are finally presented in random order. For all lists and all participants, the score was computed as percentage