Evaluation of Synthesized Sign and Visual Speech by Deaf

Zdeněk Krňoul, Patrik Roštík, Miloš Železný

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

Abstract

This paper is focused on an evaluation of quality of synthesized sign speech and a comparison of sign and visual speech. The evaluation has been performed with the Czech sign speech synthesis system. The system produces a manual component as well as a non-manual component given by the lip articulation. The perception test by deaf children from primary school is scored on the isolated signs. Two studies are designed for 5-6 and 11-13 years old pupils. It uses the multiple-choice test composed from picture of the signs to get an intelligibility score. The results confirm intelligibility of the synthesized sign speech as well as visual speech and indicate also statistically a significant difference between perception of sign and visual speech.

Index Terms: perception study, synthesis of visual speech, talking head

1. Introduction

Modern age brings many new possibilities in the field of integration of disabled people into society, because people are increasingly aware that to have handicap does not make them marginalized. New technologies are being developed, which allow them easier integration. In the area of research dealing with the problems of hearing impaired people, we can found tools for the transfer speech into sign speech. There is the question why such systems are designed. It is necessary to mention the fact that for example Czech is not for deaf people natural language. Deaf understand the words which the interpreter signs better than words which they will see written. From this point of view, Czech is their second language.

The use of the above-mentioned tools is advantageous. All information systems, such as for example information boards with departures at railway stations, could be extended with the screen, where the written text or loudspeaker messages are translated into sign speech.

Since July 1st 2004 Department of Cybernetics has been working on the project MUSSLAP[1], which is focused on problems of the sign speech synthesis. One of the main aims is to create applications for the full replacement of a sign language interpreter by a computer. The computer would thus facilitate two-way communication between the deaf and hearing people. The part of project is a research on the technology for the transfer of speech in the Czech sign speech – synthesis of sign speech. The use of the synthesis has another practical applications than mentioned information systems. Applications of synthesis suitable for the other users are language learning tools or sign language dictionaries.

The contribution to the understanding of the visual components of the speech given by lip articulation is known [1, 2, 3]. The study [4] with older normal hearing students shows that the lipreading score achieved for isolated words is 66% for speaker face condition and 52% in the case talking head "Baldi". The study [5, 6] with younger children with hearing disorder are aimed at training the speech production in terms of visual speech given by lip articulation only. The design of training tool for sign speech production is introduced in [7].

The potential contribution and limitation of synthesized sign speech are known but no practical experiences were obtained by now. Hence, this paper presents the evaluation of the current state of sign speech synthesis system. The perception tests are used to find out how hearing impaired people perceive the animation.

2. System Description

Our sign speech synthesis system is composed of the two following parts: translation system and synthesis system. The first part is used to convert the text form of Czech language to sign language expressed as a sequence of signs in symbolic form [8]. Thereafter the synthesis system converts the symbolic form to an animation of human character. Translation subsystem is based on technique of automatic translation of phrases. This means that the sentence in Czech is divided into phrases which are then translated into the relevant phrases in sign language.

The synthesis system produces the manual component of speech as well as the non-manual component. For manual component, we have designed the rule based synthesizer [8] which uses the lexicon based on the symbolic notation HamNoSys[2]. For the non-manual component, we have employed the talking head system [9]. The control of a face animation includes coarticulation model based on the selection of visual targets. Controlling the movement of the head, the direction of eye view and nonverbal facial gestures are not currently included in the system.

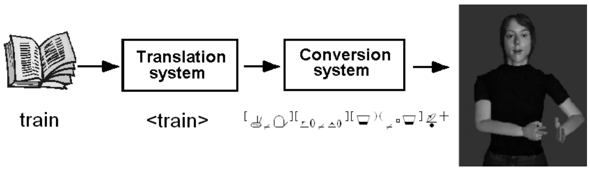

Figure 1: *A schema of the sign speech synthesis system.*

[1] http://musslap.zcu.cz/en/

[2] http://www.sign-lang.uni-hamburg.de/projects/HamNoSys.html

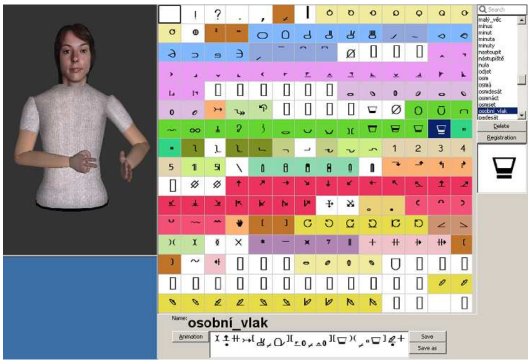

Figure 2: *The screenshot of the editor used for symbolic notation. The right window has been used for creating or editing of a symbolic string. The left window provides the animation of the edited sign.*

2.1. Notation Process

The main structure of HamNoSys is composed from three types of blocks which are notated with about 200 symbols. There are a symmetry block, block of starting position and block of actions. The starting position determines the shape of the hand, hand and palm orientation and location in space. The block of actions is used to write hand movements. Because all movements are implicitly included for the right hand, which is regarded as dominant here, the notation for the left hand can be solved by the symmetry block or these movements can be notated separately. All three above mentioned blocks are optional and can therefore be omitted.

2.2. Editor for Symbolic Notation

An editor for symbolic notation was developed from the need to facilitate the process of creating a symbolic representation of signs. The editor allows annotators to verify the sequence of inserted symbols. The editor consists of the table of symbols, the edit box, the list box of stored symbolic strings and the feedback animation. The feedback animation is given by our synthesis system. The synthesis system checks symbolic string of notated sign and converts properly composed string to the animation of manual component. By using feedback animation, annotators can repeatedly change for example different shapes by hand, the intensity of motion or different points of contact. The animation model can be rotated or zoomed in and created animation can be shown frame by frame forward or backward. A screenshot of the editor is shown in Fig 2.

3. Evaluation Study

The synthesis of sign language has not yet been tested for the understandability with deaf subjects. The first study was conducted with hearing subjects [8]. The study compared the animation of isolated signs with video records of sign language speaker as well as the synthesis of animation for continuous sign speech with text in the subtitles. The signed speech was presented using only manual component.

The aim of this evaluation study is to score quality of synthesized sign speech and also to compare the perception of sign speech and visual speech. For the study, deaf children from primary school have been selected. The study is aimed at the understanding isolated signs. It consists of two experiments and either is composed from two tests. The second test followed the first test after three weeks. Five deaf pupils were chosen from the preliminary class and the first class (5-6 years) as participants for the first experiment and six deaf pupils from the sixth and seventh class (11-13 years) for the second experiment.

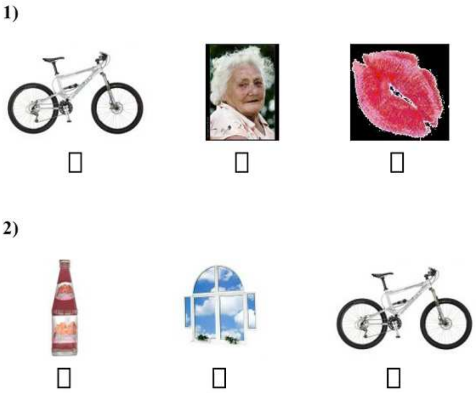

Figure 3: *An example of two choices from the sheet used for filling answers in multiple-choice test. Three choices are offered in form of three illustrations.*

3.1. Test Material

Test material is composed of the synthesized animation and isolated signs only. The video records of a sign language speaker are not used here. Since, the evaluation study is meant for deaf children, we selected such signs, which are familiar for them. We collected 15 signs from videotapes used in the curriculum of the preliminary class. These signs ought to know all participated pupils. The signs are new for pupils from the preliminary class and pupils from the first class should have it adopted. Pupils in sixth and seventh class should have to know these signs very well.

The symbolic notation of these signs was determined before the evaluation study. The editor was used with accordance Czech sign speech vocabulary to create the manual component of sign speech. The phonetic transcription of the word form of the signs and the choice of a sufficient speech rate was used to set the non-manual component. Non-manual component tested in this study is expressed by lip articulation only. This is the signed Czech variant when face gestures or non-verbal expressions are not included. The signs are articulated as well in the written form.

The created signs were captured into the video records. Two types of the records were prepared. The first one captures the entire animated character. The animation includes both manual component as well as simultaneously expressed non-manual component. The second type of the records captures detail on the head of animation model. The manual component is not controlled here. All records contains only the visual information without a sound track. The resolution of video records were 372x480 pixels, 25 frame per second and as compression is used XVid MPEG4 codec. The video records were checked by the teacher before the study and the errors were corrected.

Further, we have prepared sheets for the multiple-choice test with three response options (one correct response) composed from randomly arranged pictures of the tested words, Fig. 3.

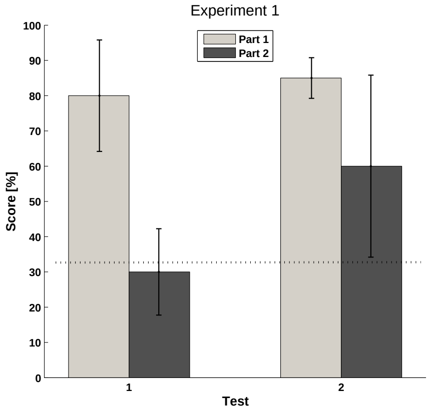

Figure 4: *The result scores of the experiment 1.*

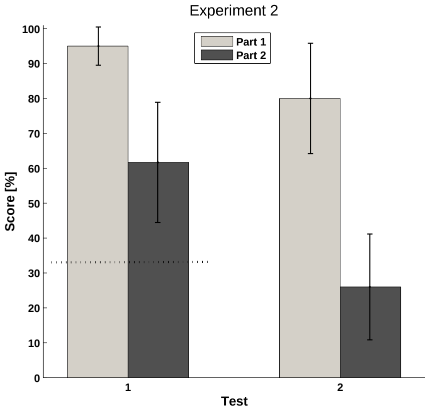

Figure 5: *The result scores of the experiment 2.*

3.2. Procedure

Tests used in the first and second experiment are composed of two immediately consecutive parts. The signs expressed by first type of record are shown to pupils in the first part. In the second part, the same signs are shown in a different order but expressed only the non-manual component (second type of the records). The second part of the test followed the first part after a short break. The first part tests the overall perception of sign language while in the second, the ability of lip-reading is tested only. It is simplified because the same signs as those in the first part are presented.

The presentation of the records consists in a sequential projection of the tested words on the wall in the classroom by the data projector, Fig 7. Firstly, five extra non-scored words were presented at the beginning of the experiments for the presentation of the various options of the study. Next, we projected ten video records capturing complete sign speech. The picture of signed character on the wall was approximately 30 cm high. In the second part of the test, the size of projected talking head was 30 cm too. The presentation have made through ProRec tool[3].

The pupils were not familiar with the tested words before the experiments. Also the scribing was prevented. The procedure of first and second experiment was the same. The only difference was in the second experiment when the pupils did not use the sheets of multiple-choice test. This step was taken because older students already achieved in the first test good results. Therefore in this test, pupils recognize the signs without a multiple-choice chance.

4. Results

Tests have been assessed as follows. Pupil got for each correct answer one point and for wrong none point. The similarity of some signs was not taken into account. The mean and standard error of achieved scores are shown in the graphs in Fig. 4 and Fig. 5.

There are two evaluation of the results. The first evaluation tested the hypothesis that pupils filled the multiple-choice test by a chance. It is used the one-sample and one-sided t-test. The planned comparisons are carried out for both tests of the first experiment and for the first test of the second experiment, $(\alpha = 0.01)$. The results show significantly better understanding of the signed speech than a chance (three options, the chance level 33.3%, p <0.01). The average 80% success was achieved in the first test of first experiment (t (4) = 6.6, p = 0.0014) and 85% for the second test (t (3) = 17.91, p $<$ 0.001). Better results are achieved in the second experiment with older pupils. There are for the first test on average 95% correct answers (t (5) = 27.59, p $<$ 0.001) and the retesting without the possibility of choosing, on average 80% correct answers.

The evaluation in the same assumptions for the second parts of the tests for lip reading shows significantly better understanding for the first test of the second experiment, the mean score is 62% (t (5)= 4.03, p= 0.005). The mean score of the re-testing without the choice is only 26%. In the first experiment can not refute that pupils completed this test by chance (p $>$ 0.01).

The second evaluation was testing hypothesis that the removal of the non-manual component causes a significant decrease in understanding. Testing is at the same level of the significance, $\alpha = 0.01$. It used the one-sided and paired t-test to verify that the mean scores achieved for lip reading are at least identical against variant that the results are worse. The significant decrease of understanding was observed for the first test of the first experiment and for both tests of the second experiment (p $<$ 0.01).

The decrease of the score is for the first test of the first experiment 50%, (t (3) = 5.98, p = 0,002). For the second test of the first experiment, it is not possible to determine a significant decrease, (t (3) = 1.89, p = 0.0776). The decrease of 33% is observed in the first test of the second experiment, (t (5) = 5.0, p = 0.0021). The second test from the second experiment has the largest decline, 54% (t (4) = 4.32, p = 0.0062).

In comparison with the first hypothesis, it may be noted, that despite the significant decrease of understandability, the significant level of understanding in the first test of the second experiment remains for the lipreading condition.

5. Summary and Conclusions

The perception study of sign language and providing the necessary evaluation for further development of synthesis system is the main objective of this paper. The design of sign language synthesis developed at the Department of Cybernetics in Pilsen is mentioned in the first half of this paper. In the second part, the evaluation of the synthesis system is performed. It tests the degree of understanding of isolated signs and explores the decline

[3] available at http://www.phon.ucl.ac.uk/resource/prorec/

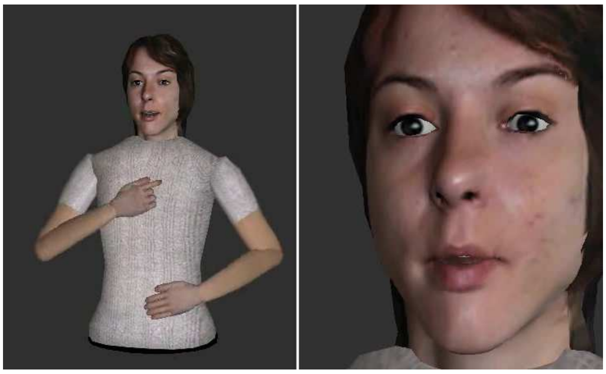

Figure 6: *An example of the video records used in the perception study. The animation of sign speech from the first part of test is on the left and the animation of talking head is on the right.*

of understandability, if the sings are only lipread.

The achieved scores for individual pupils differ very much. This can be counted to the fact that every child in the class is individual with a different diagnosis. For example participants of first experiment, only two of the five deaf children are from deaf families. Boy after polio, autism boy, boy prelingually deaf also participated in the first experiment are from the hearing family. Every of them has prerequisites for another results of the tests. Children from deaf families would have better control of sign language unlike those of hearing families, who should have rather excellence in lipreading.

The first result shows that younger and older pupils are able to understand the animated sign language given by synthesis system. Younger pupils did not understand signs with the removal of manual components while older pupils reached for lipreading the significantly better results than a chance. This can be assigned to a fact that the test includes for them the well-known signs and they have also more experience with lipreading.

The second result shows the significant decline in understandability for removal of the manual component. The highest decrease is observed on the second test of the second experiment when choice sheets are not provided. In this case, understanding the sign language remains in comparison with other tests still high (80%). However, for lipreading, the score is only 26%. There is a considerable difference compared to the score 62% which is reached with the multiple-choice test.

The lipreading was for pupils very difficult when they perceived the signs without the possibility of a choice. It can be considered that the current animation of visual speech does not achieve the accuracy of the speaker. However, if students have the option to choose, then they derive the correct answer from the articulation more easily. For example, in the choices (mísa, máma, limonáda) (eng. bowl, mother, lemonade) with the third correct sign, the dominant lip rounding of the phoneme /o/ and the length of the word lead to the successful selection of the correct answer.

In addition to the tests, we have received comments from deaf pupils. The critical comments were propounded not only for animated sign speech but also for talking head. The low understandability was observed in some consonants. This can be accounted to the setting of the articulatory parameters in the coarticulation model. However, we can found also the reason that the control of tongue was not included in the animation of talking head. The animation of manual component is understandable but still appears to be robotic like. More natural movements of hands would contribute to a better perception of the animation.

Figure 7: *A photo of the testing procedure in the first experiment performed in the preliminary and first class.*

6. Acknowledgements

We would like to thank the primary school for hearing impaired in Pilsen for the possibility of carrying out the study. This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

7. References

[1] W. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise,” *J. Acoustical Society America*, vol. 26, pp. 212–215, 1954.

[2] A. MacLeod and Q. Summerfield, “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*, 1990.

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

[4] S. Ouni, M. M. Cohen, I. Hope, and D. W. Massaro, “Visual contribution to speech perception: Measuring the intelligibility of animated talking heads,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2007.

[5] D. W. Massaro and J. Light, “Using visible speech for training perception and production of speech for hard of hearing individuals,” *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 2, pp. 304–320, 2004.

[6] K. Grauwinkel and S. Fagel, “Visualization of internal articulator dynamics for use in speech therapy for children with sigmatismus interdentalis,” in *AVSP 2007*, Hilvarenbeek, Netherlands, 2007.

[7] O. Aran, I. Ari, A. Benoit, P. Campr, A. H. Carrillo, F. Fanard, L. Akarun, A. Caplier, and B. Sankur, “Signtutor: An interactive system for sign language tutoring,” *IEEE Multimedia*, 2008.

[8] Z. Krňoul, J. Kanis, M. Železný, and L. Müller, “Czech text-to-sign speech synthesizer,” *Machine Learning for Multimodal Interaction, Series Lecture Notes in Computer Science*, vol. 4892, pp. 180–191, 2008.

[9] Z. Krňoul and M. Železný, “A development of czech talking head,” in *Proceedings of ICSPL 2008*, in press, 2008.