Translation and Conversion for Czech Sign Speech Synthesis[*]

Zdeněk Krňoul and Miloš Železný

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

Abstract. Recent research progress in developing of Czech Sign Speech synthesizer is presented. The current goal is to improve a system for automatic synthesis to produce accurate synthesis of the Sign Speech. The synthesis system converts written text to an animation of an artificial human model. This includes translation of text to sign phrases and its conversion to the animation of an avatar. The animation is composed of movements and deformations of segments of hands, a head and also a face. The system has been evaluated by two initial perceptual tests. The perceptual tests indicate that the designed synthesis system is capable to produce intelligible Sign Speech.

1 Introduction

For human-computer interaction, speech-based communication systems become more and more widely used. However, people with hearing or speech disorders cannot use such systems. Hence, computer communication systems for aurally or speech disabled people must be based on alternative means, such as Sign Speech. As one part of such communication system, a system for automatic synthesis of the Sign Speech provides speech utterances in these dialogs. A sign of the Sign Speech has two components: the manual one and the non-manual one. The non-manual component is expressed by a gesture of a face, a movement and a position of a head and other parts of the upper half-body. The manual component is expressed by shapes, movements and positions of hands. In principle, there are two ways how to synthesize any sign language utterance by a computer system: a data driven approach or generation of movements from the symbolic formulation (a symbolic system).

The data-driven approach is based on capture of body movements. This synthesis process uses specific movements derived from signing person directly. Second variant uses artificial composition of base movements. The final Sign Speech utterance is achieved by concatenation of relevant isolated signs. There is an analogy with the spoken form of a given language.

Second approach appears to be more general and appropriate. It uses a trajectory generator of the manual component, which employs some symbolic system. However,

[*] Support for this work was provided by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 524–531, 2007.

conversion of symbols to really naturally looking animation is a complicated task. Robotic-like movements are often perceived. As a part of the non-manual component, synchronous synthesis of lip articulation (talking head) is necessary for good overall intelligibility. For this purpose, the goal is not to simulate the musculature of the face or inner mouth organs but to produce sufficient support for lip-reading. Lip articulation should be then controlled by several visual parameters which directly determine required deformations.

For synthesis of the manual component of the Sign Speech we designed solution which is based on the symbolic representation of signs. For this purpose, we have chosen the HamNoSys notation system (Hamburg Notation System for Sign Languages[1]) from several notation systems. This notation system has precise description of signs and is usable for the interpretation in the computer system.

2 Sign Speech Synthesis

Our synthesis system has two parts: the translation and the conversion subsystem (Figure 1). The translation system transfers Czech written text to its textual representation in the Sign Speech (textual sign representation). The conversion system then converts this textual sign representation to animation of the artificial human model (avatar). Resulting animation then represents the corresponding utterance in the Sign Speech.

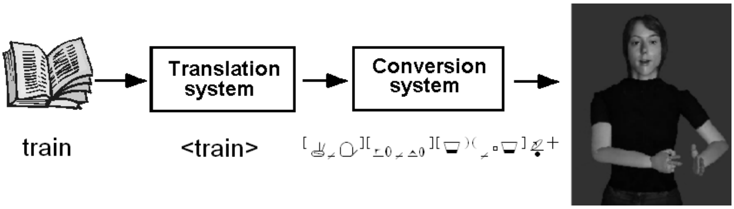

Fig. 1. Schema of the Sign Speech synthesis system

2.1 Translation System

The translation system is an automatic phrase-based translation system. A Czech sentence is divided into phrases and these are then translated into corresponding Sign Speech phrases. The translated words are reordered and rescored using language model at the end of the translation process. In our synthesizer we use our own implementation of the simple monotone phrase-based decoder - **SiMPaD** [1]. Monotonicity means that there is no reordering model. In the decoding process we choose only one alignment with the longest phrase coverage (i.e. if there are three phrases: p_1, p_2, p_3 coverage three words: w_1, w_2, w_3, where $p_1 = w_1 + w_2$, $p_2 = w_3$, $p_3 = w_1 + w_2 + w_3$, we choose the alignment which contains phrase p_3 only). Standard Viterbi algorithm is used for decoding. SiMPaD uses a trigram language model with linear interpolation.

[1] Available at www.sign-lang.uni-hamburg.de/projects/HamNoSys.html

2.2 Conversion System

The conversion system produces two main components: the manual and non-manual component of a sign. The manual component represents movements, orientations and shapes of hands. The non-manual component is composed of remaining movements of the upper half-body, face gestures and also articulation of lips and inner mouth organs. Symbolic representation was designed to solve the animation problem of the manual component and the upper half-body movements. We use HamNoSys 3.0 for this purpose. This notation is deterministic and suitable for processing of the Sign Speech in a computer system. Synthesis of lip articulation is provided by our talking head subsystem.

3 Sign Speech Editor

Methodology of symbolic notation allows precise and extensible description of a sign usable for avatar animation. However, composition of many symbols to a correct string is very difficult. For better coverage of HamNoSys notation features, we improved the SLAPE editor [2]. New functions have been inserted to reach easy transcription. The editor allows notation of two hand movements observed in symmetric signs. Notation of a movement modality is added, too. Modality means a style of a movement (speed, tensity etc.) or a repetition of a movement. The editor allows also notation of a precise location as a contact of the dominant hand with the upper half-body or the second hand. Furthermore, the editor uses avatar animation as a feedback for validation of the string of symbols and easier transcription.

Work with the editor is intuitive. Even those, who are not familiar with a sign language, can use it. Building or correction of signs can be also made in accordance with video records of a signing character. During transcription process, a sign is created by choosing pictures, which represent particular parts of a human body and its basic movements. The final string of HamNoSys symbols is finally saved. For a further synthesis purpose, these strings are stored in the symbol vocabulary.

4 Process of Continuous Sign Speech Synthesis

The synthesis process is based on feature frames and animation trajectories. Firstly, our synthesis system automatically generates sequences of feature frames. These feature frames are then transformed into trajectories. Each trajectory controls relevant rotation of bone joints or deformation of triangular meshes directly. The synthesis process involves analysis of symbols for isolated signs, frame processing and final concatenation of isolated signs. The analysis includes parsing of a symbol string into a tree structure and processing of symbols [2]. Next, the sequences of frames are generated for each isolated sign. The final sequence of frames which describes continuous speech is built using the concatenation technique in accordance with the textual sign representation of input text. The illustration of continuous animation is in Figure 2.

The analysis of symbols allows the computation of trajectories only for hands and upper half-body. Trajectories for lip articulation and face gestures are produced by the ”talking head” subsystem separately.

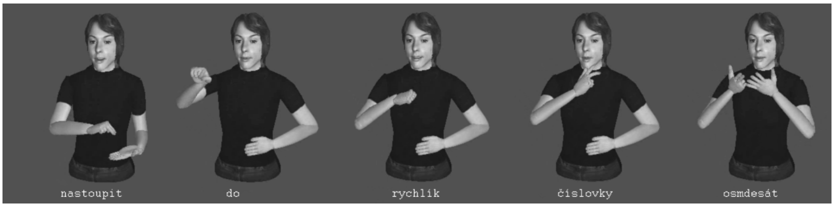

Fig. 2. Example of continuous animation

4.1 Analysis of Symbols and Trajectories for Isolated Sign

The goal of the analysis of symbols is to transform a particular isolated sign into feature frames. The analysis of the HamNoSys symbols leads to the description based on a context-free grammar. We have constructed over 300 parsing rules. If a sign is accepted by a parser, the algorithm creates a parse tree. A path to each leaf node determines types of symbol processing of the particular symbol. Processing is carried out by several tree walks that reduce the number of nodes. In each node of the parse tree, property items of a symbol are joined and blended. For both hands, these items can be divided to three logic parts: a location, motion and a shape. Processing of these items produces several feature frames for each node. As a result of this analysis, we obtain one sequence of frames in the root of the parse tree. It is then transformed into trajectories by the inverse kinematics technique. The frequency of generated frames is implicitly set to 25 frames per second. The non-manual component of body movements can be supplied by computation of relevant trajectories for the upper half-body.

4.2 Talking Head Subsystem

Trajectories for face gestures and also for articulation of lips, a tongue and jaws are created by visual synthesis carried out by the talking head subsystem. This visual synthesis is based on concatenation of phonetic units. Any written text or phrase is represented as a string of successive phones. The lip articulatory trajectories are concatenated by the Visual unit selection method [3]. This synthesis method uses an inventory of phonetic units and the regression tree technique. It allows precise coverage of coarticulation effects. In the inventory of units, several realizations of phoneme are stored. Our synthesis method assumes that the lip and tongue shape is described by a linear model. Realization of phoneme is described by 3 linear components for lip shape and 6 components for tongue shape. The lip components represent linear directions for lip opening, protrusion and upper lip raise. The tongue components consist of jaw height, dorsum raise, tongue body raise, tip raise, tip advance and tongue width. The synthesis algorithm performs a selection of appropriate phoneme candidate according to the context information. This information is built from the triphone context, the occurrence of coarticulation resistant component (of lip or tongue) in adjacent phonemes and also from time duration of neighbored speech segments. Final trajectories are computed by cubic spline interpolation between selected phoneme realizations.

The created trajectories should be time-aligned with the timing of acoustic Sign Speech form. This form is produced by an appropriate TTS system. Synthesis of face gesture trajectories is based on the concatenation and the linear interpolation of the neutral face expression and one of the 6 basic face gestures: happiness, anger, surprise, fear, sadness and disgust. The resulting head animation is in Figure 3.

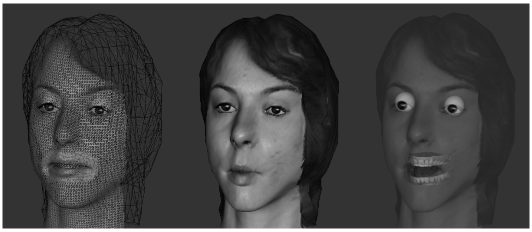

Fig. 3. Animation of talking head

4.3 Synchrony of Manual and Non-manual Components

Synchrony of the manual and non-manual component is crucial in the synthesis of continuous Sign Speech. The asynchronous components cause overall unintelligibility. The asynchrony should be caused by the different speech rate of spoken and the Sign Speech. We designed an effective solution - a synchrony method at the level of words. This method combines basic concatenation technique with time delay processing. Firstly, for each isolated sign, trajectories from the symbol analysis and trajectories from the talking head subsystem are generated. Time delay processing determines duration of all trajectories and selects the longest synthesized variant. The following step of processing evaluates interpolation time which is necessary for concatenation of particular adjacent isolated signs. This interpolation time ensures the fluent shift of body pose. We select the linear interpolation between the frames on the boundaries of concatenated signs. The interpolation of a hand shape and its 3D position is determined by weight average, the finger direction and palm orientation is interpolated by the extension to the quaternion.

5 Animation Model

The animation model tries to produce the Sign Speech by the efficient manner rather than deep understanding to physiological mechanisms. Our animation algorithm employs a 3D geometric animation model of avatar in compliance with H-Anim standard[2]. Our model is composed of 38 joints and body segments. These segments are represented by textured triangular surfaces. The problem of setting the correct shoulder and elbow rotations is solved by the inverse kinematics[3]. There are 7 degrees of freedom for each limb. The rotation of remaining joints and local deformation of the triangular surfaces

[2] www.h-anim.org

[3] Available at cg.cis.upenn.edu/hms/software/ikan/ikan.html

allows to set full avatar poses. The deformation of triangular surfaces is primarly used for animation of a face and a tongue model. The surfaces are deformed according to animation schema which is based on the definition of several control 3D points and splines functions [4]. The rendering of the animation model is determined in C++ code and OpenGL.

6 Perceptual Evaluation

Two tests on the intelligibility of synthesized Sign Speech have been performed. The goal has been to evaluate the quality of our Sign Speech synthesizer. Two participants who are experts in the Sign Speech served as judges. We used vocabulary of about 130 signs for this evaluation purpose. We completed several video records of our animation and also signing person. The video records of signing person are taken from the electronic vocabulary[4]. The capturing of video records of our animation was under two conditions.

6.1 Test A: Isolated Signs

The equivalence test was aimed at the comparison of animation movements of isolated signs with movements of signing person. Video records of 20 pairs of random selected isolated signs were completed. The view on the model of avatar and signing person was from the front. The participants evaluated this equivalence by marks from 1 to 5. The meaning of marks was:

- 1 totally perfect; the animation movements are equivalent to signing person
- 2 the movements are good, the location of hand, shapes or speed of sign are a little different but the sign is intelligible
- 3 the sign is difficultly recognized; the animation includes mistakes
- 4 incorrectly animated movements
- 5 totally incorrect; it is different sign

The results are in top panel of Figure 4. The average mark of participant 1 is 2.25 and of participant 2 is 1.9. The average intelligibility is 70% (marks 1 and 2 indicate the intelligible sign). There was 65% mark agreement between participants. The analysis of signs with lower marks shows that the majority of mistakes is caused by of the symbolic notation rather than inaccuracy in the conversion system. Thus, it is highly important to obtain as accurate symbolic notation of isolated signs as possible.

6.2 Test B: Continuous Speech

We created 20 video animation records of short utterances. The view on the avatar animation was here partially from the side. The participants judged the whole Sign Speech utterance. The subtitles (text representation of each sign) were added to the video records. Thus, the participants knew the meaning of the utterance and determined the overall intelligibility. The participants evaluate the intelligibility by marks from 1 to 5. The meaning of marks was:

[4] Langer, J. et al.: Znaková zásoba českého znakového jazyka. Palacký Univ. Olomouc.

- 1 the animation shows the signs from subtitles
- 2 good intelligible utterance
- 3 hardly intelligible utterance
- 4 almost unintelligible utterance
- 5 total unintelligible utterance

The results are in bottom panel of Figure 4. All utterances were evaluated by mark 1 or 2. In average, the animation of 70% utterances shows the signs from subtitles. The results indicate that the synthesis of continuous speech is intelligible. The concatenation and synchrony method of isolated signs is sufficient.

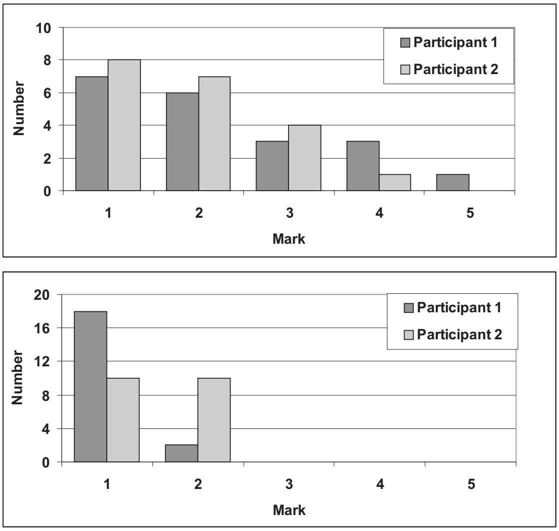

Fig. 4. Perceptual evaluation, on top: isolated signs, below: continuous Sign Speech

7 Summary and Conclusions

The synthesis system automatically converts written text to the animation of avatar. The synthesis system consists of the translation and conversion system. The translation system translates the Czech written text to its text sign representation. The conversion system then transforms this text sign representation into an animation. For purpose of symbolic notation of signs, the SLAPE editor was improved. The SLAPE editor provides online access to creating and editing of signs. Transcription is intuitive and resulting signs are stored in the vocabulary. For a general purpose, each sign is represented by a string of symbols.

The conversion system uses designed synthesis process of continuous Sign Speech. The synthesis process analyses the string of HamNoSys symbols and creates the sequence of feature frames for isolated signs. Designed concatenation and time delay process allows using separately synthesized trajectories from talking head system together with the trajectories generated from symbolic notations. This trajectory representation of signs is depended on the proportion of animation model (our model is composed of 38 joints and segments of body in this time).

The perceptual tests reveal that the synchrony on the level of word preserves the intelligibility for continuous Sign Speech. But the intelligibility of isolated signs highly depends on symbolic notation of particular signs in the vocabulary. Thus, it is necessary to concentrate on acquisition of precise symbolic notation of isolated signs in future work.

References

1. Kanis, J., Müller, L.: Automatic Czech – Sign Speech Translation. In: Proceedings of 10th International Conference on TEXT, SPEECH and DIALOGUE TSD 2007, Springer, Heidelberg (2007)
2. Krňoul, Z., Kanis, J., Železný, M., Müller, L., Císař, P.: 3D Symbol Base Translation and Synthesis of Czech Sign Speech. In: Proceedings of SPECOM. St. Petersburg: Anatolya publisher (2006)
3. 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)
4. Krňoul, Z., Železný, M.: Realistic Face Animation for a Czech Talking Head. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, Springer, Heidelberg (2004)