

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016,  
9-12 May 2016, Yogyakarta, Indonesia

## Automatic Technologies for Processing Spoken Sign Languages

Alexey Karpov <sup>a\*</sup>, Irina Kipyatkova <sup>a</sup>, Milos Zelezny <sup>b</sup>

<sup>a</sup>St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Russian Federation

<sup>b</sup>University of West Bohemia (UWB), Pilsen, Czech Republic

---

### Abstract

Sign languages are known as a natural means for verbal communication of the deaf and hard of hearing people. There is no universal sign language, and almost each country has its own national sign language and fingerspelling alphabet. Sign languages use visual-kinetic clues for human-to-human communication combining hand gestures with lips articulation and facial mimics. They also possess a special grammar that is quite different from that of speech-based spoken languages. Sign languages are spoken (silently) by a hundred million deaf people all over the world and the most popular are American (ASL), Chinese, Brazilian, Russian, and British Sign Languages; there are almost 140 such languages according to the Ethnologue. They do not have a natural written form, and there is a huge lack of electronic resources for them, in particular, vocabularies, audio-visual databases, automatic recognition and synthesis systems, etc. Thus, sign languages may be considered as non-written under-resourced spoken languages. In this paper, we present a computer system for text-to-sign language synthesis for the Russian and Czech Sign Languages.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the Organizing Committee of SLTU 2016

**Keywords:** Sign language; communication of deaf people; unwritten languages; multi-modal synthesis system; under-resourced languages

---

### 1. Introduction

At present, sign languages (SLs) are well known as a natural means for verbal communication of the deaf, hard-of-hearing people, and people who have some speaking difficulties. There is no universal sign language, and almost

---

\* Corresponding author. Tel.: +7-812-328-0421; fax: +7-812-328-7081.

E-mail address: [karpov@iiias.spb.su](mailto:karpov@iiias.spb.su)

each country has its own national sign language and fingerspelling alphabet. All the sign languages use visual-kinetic clues for human-to-human communication combining manual gestures with lips articulation and facial mimics. They also possess a specific and simplified grammar that is quite different from that of acoustic-based spoken languages.

Sign languages are spoken (silently) by a hundred million deaf people all over the world. In total, there are at least 138 living sign languages according to the Ethnologue catalogue, and many of them are national (state) or official languages of human communication in some countries like the USA, Finland, the Czech Republic, France, the Russian Federation (since 2013), etc. According to the statistics of medical organizations, about 0.1% of the population of any country is absolutely deaf and the most of such people communicate only by sign languages; many people, who were born deaf, even are not able to read. Additionally to conversational sign languages there are also fingerspelling alphabets, which are used to spell words (names, rare words, unknown signs, etc.) letter-by-letter. A fingerspelling system directly depends on the national alphabet of a country; also, there are both one-handed fingerspelling alphabets (e.g., in Russia, France, USA) and two-handed ones (e.g., in the Czech Republic, the UK, Turkey).

The most popular SLs with approximate numbers of native signers/speakers, who use a sign language as the first language of communication, according to the Ethnologue catalogue (<http://www.ethnologue.com/subgroups/deaf-sign-language-0>) and information from Wikipedia ([https://en.wikipedia.org/wiki/List\\_of\\_sign\\_languages\\_by\\_number\\_of\\_native\\_signers](https://en.wikipedia.org/wiki/List_of_sign_languages_by_number_of_native_signers)), are:

- Chinese SL – about 20M signers (there is no exact information in catalogues, it is an estimation)
- Brazilian SL – about 3M signers;
- Indo-Pakistani SL – at least 2.7M signers (up to 10M according to some Internet sources);
- American SL (ASL) – 500K signers, mainly in the USA;
- Hungarian SL – 350K signers;
- Kenyan SL – 340K signers;
- Japanese SL – 320K signers;
- Ecuadorian SL – 188K signers;
- Norwegian-Malagasy SL – 185K signers;
- British SL (BSL) – 125K signers;
- Russian SL – 121K signers in the Russian Federation and up to 100K in other countries;
- French SL – 100K signers in France, plus about 50K in other French-speaking countries, etc.

Russian SL (RSL) is a native language for the deaf in Russia, Belarus, Ukraine, Kazakhstan, Moldova, also partly in Bulgaria, Latvia, Estonia, and Lithuania; there also exist some regiolects of Russian SL (RSL differs even in Moscow and St. Petersburg regions)<sup>1</sup>. Czech SL (CSL) is used by almost 10K deaf people in Czech Republic.

Family relationships of SLs are not well established because of the lack of linguistic research, but it is known that French SL family joins over 50 different SL in Europe (including Russian and Czech SLs), North and Latin America, and Africa; British and Arabic SL families include from 4 to 10 different SLs each.

All the sign languages do not have a natural written form, and there is a huge lack of electronic resources for them, in particular, vocabularies, multimedia and video databases, automatic recognition and synthesis systems, etc. Thus, sign languages may be considered as non-written under-resourced spoken languages. However, in the second half of the 20th century, some useful ways to represent sign languages in a written form have been proposed, which are called sign notation systems, for instance:

- Stokoe notation proposed by W. Stokoe<sup>2</sup>
- SignWriting developed by V. Sutton (<http://www.signwriting.org>)
- Hamburg Notation System (HamNoSys)<sup>3</sup>
- ASL-phabet developed by S. Supalla (<http://www.aslphabet.com>)
- Movement-Hold (M-H) notation proposed by S. Liddell and R. Johnson<sup>4</sup>
- Dimskis notation proposed by L. Dimskis<sup>5</sup>
- Si5s writing system proposed by R. Arnold (<http://www.si5s.org>)

- Sign Language International Phonetic Alphabet (SLIPA, <http://dedalvs.com/slipa.html>), etc.

The last one has been proposed as a phonetic transcription system for signing. However, there is still no official acceptance of a writing system for sign languages. At present, HamNoSys and SignWriting are the most widely used by linguists and developers of information (surdo) systems for the deaf, since these sign notation systems propose both a computer-friendly and user-friendly alphabet (in Unicode) and system fonts.

Animated 3D virtual characters (avatars) are very convenient for computer synthesis of sign language and fingerspelling. There are some researches on 3D signing avatars and SL machine translation systems both in Europe and the USA, for example:

- SIGNSPEAK EU project ([www.signspeak.eu/en](http://www.signspeak.eu/en))
- Dicta-Sign EU project ([www.dictasign.eu](http://www.dictasign.eu))
- SignCom EU project<sup>6</sup>
- Italian SL project<sup>7</sup>
- MUSSLAP project (<http://musslap.zcu.cz/en/sign-language-synthesis/>)
- ViSiCAST with Visia avatar ([www.visicast.co.uk](http://www.visicast.co.uk))
- eSIGN with Virtual vGuido avatar ([www.sign-lang.uni-hamburg.de/esign](http://www.sign-lang.uni-hamburg.de/esign))
- DePaul ASL Synthesizer (<http://asl.cs.depaul.edu>)
- Sign Smith, Sign4Me avatars from Vcom3D ([www.vcom3d.com](http://www.vcom3d.com))
- iCommunicator for ASL ([www.icommunicator.com](http://www.icommunicator.com))
- SiSi (Say It Sign It) system of IBM for BSL (<http://mqtt.org/projects/sisi>), etc.

There is a quite large scientific community aiming at computer processing of sign languages in Europe organized around several European projects, but there is a lack of computer systems for Russian SL processing, including sign language analysis and synthesis. In this paper, we present a computer system for text-to-sign language synthesis for RSL and CSL.

## 2. Text-to-sign language synthesis system

UWB and SPIIRAS teams have developed a multi-modal (audio-visual) text-to-sign language & speech synthesizer for Czech and Russian<sup>8</sup>. Originally, the text-to-sign language system was a part of the Czech project MUSSLAP (<http://musslap.zcu.cz/en/sign-language-synthesis/>).

### 2.1. Architecture of the synthesis system

The proposed synthesizer takes text as an input and translates it into sign language gestures and audio-visual speech. The multi-modal synthesizer system in general consists of the following components:

- text processor that takes text as an input to generate transcriptions of phonemes and visemes (a visual equivalent of phoneme), and control selection of HamNoSys codes intended for hand description<sup>9</sup>;
- TTS (text-to-speech) systems that generate the acoustic speech signal with time labeling corresponding to the entered text for Czech<sup>10</sup> and Russian<sup>11</sup>;
- virtual 3D model of human's head with controlled lips articulation, mimics and facial expressions<sup>12</sup>;
- control unit for the audio-visual talking head that synchronizes and integrates lips movements with the synthesized auditory speech signal<sup>13,14</sup>;
- virtual 3D model of human's upper body; we use a skeletal model of the signing avatar, which is controlled via HamNoSys codes by OpenGL<sup>15</sup>;
- multi-modal user interface, which integrates all the components for automatic generation of SL gestures, auditory and visual speech (articulation and facial expressions) in the signing avatar<sup>8</sup>.

Input text phrases are given to the input of the system and firstly analyzed by a word-processor. Clauses, words

(for audio speech synthesis and video synthesis of lips articulation by the talking head, as well as for sign language output by the avatar) and letters (for fingerspelling synthesis) are found out and automatically processed to the symbols of HamNoSys. On this basis, the signing avatar outputs manual gestures of SL decoding HamNoSys codes.

Thus the visual output is available for deaf and hard of hearing people, who can use SL and/or lip-reading (many deaf people are able to read speech by lips even without manual gestures); the audio output is oriented towards visually impaired people; the audio-visual part of the interface is intended for non-disabled people.

## 2.2. Bimodal talking head

We have implemented a 3D realistic talking head model, which is a text-driven system, and the visual processing part of which is controlled by taking into consideration the results of input text processing and audio TTS with the help of a modality asynchrony model.

The talking head is based on a parametrically controllable 3D model of a human head. Movable parts are animated by a set of control points. The synthesis is based on concatenative principles, i.e. the descriptions of the visemes (in the form of a set of control points) are concatenated to produce continuous stream of visual parameters<sup>12</sup>.

3D head model is represented by a set of points-vertices of a virtual space, which are connected by edges to build up dense triangular surfaces (Figure 1a). The obtained 3D data are processed, completed by adding manually created other facial parts and kept in a file in a virtual reality modeling language (VRML) as a set of vertex coordinates of triangular planes and textures of the face. The head model is described by tens of thousands of vertices (Figure 1b), however, only few of them are active, which can be controlled by the software, simulating movements of facial muscles; control of those allows displaying visemes.

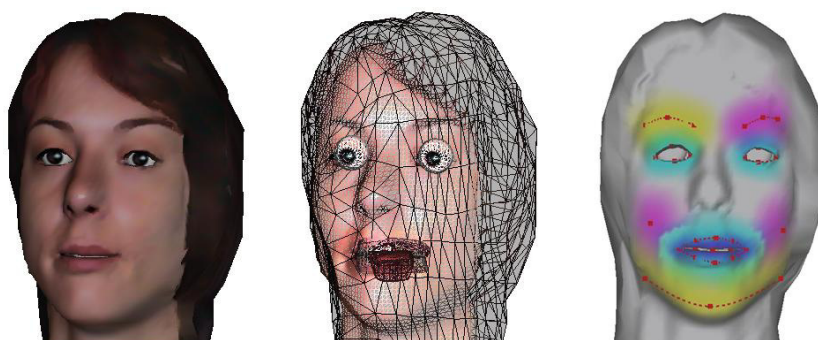


Fig. 1. Talking head: (a) view of 3D model; (b) view of wired model; (c) model influence zones.

For smooth movements the vertices surrounding the control points are interpolated; animation is smoothed using the influence zones approach (Figure 1c). Each influence zone is attached to one set of control points connected to a 3D spline curve. The shape of the head model was created using a 3D scanning system from a real face<sup>12</sup>.

In our audio-visual synthesizer, not only the general model of a head is used, but also additional models of its parametrically controlled facial components: eyes, tongue, upper and bottom jaws, and some internal speech organs. They are created on the knowledge base of anthropological physiology. Each model is controlled by the software.

In the system, an incoming orthographic text to be transformed into audio-visual speech signal undergoes a number of successive operations carried out by specialized processors: textual, phonemic, prosodic, acoustic and visual. The textual processor divides an orthographic text into utterances, transforms numbers and abbreviations into textual form, divides an utterance into phrases, places word stress (weak and strong), divides phrases into accentual units, and finally marks intonation types.

Synchronization of face/lip movements with the synthesized acoustical signal is based on timestamps of allophones in the synthesized speech flow. Synchronization of lips movements with the synthesized speech signal is

realized on the basis of information known about positions of the beginning and end boundaries of each context-dependent phoneme (allophone) in the speech flow. Duration of every allophone is set by the auditory TTS system based on allophone average length and required speech tempo. In order to model natural asynchrony between the audio and visual speech cues and to take into account different speech rates, some context-dependent timing rules for transitions between displaying visemes are applied<sup>13</sup>. This method allows increasing naturalness and intelligibility of generated audio-visual speech<sup>16</sup>.

### 2.3. Formal description and animation of manual gestures

The goal of an automatic sign language synthesizer is a simulation of human behavior during signing. Sign language synthesis is implemented in several steps. First, the input utterance is translated into the corresponding sequence of signs. Then the relevant signs are concatenated to form a continuous utterance.

The synthesis module incorporates conversion algorithm for Hamburg Notation System<sup>9</sup> (HamNoSys) to create necessary SL and fingerspelling gestures<sup>15</sup>. HamNoSys allows describing any manual gesture using only four components for both hands: (1) hand shape; (2) hand/palm orientation; (3) location of the hand; (4) motion type. For each of these components there is its own set of written iconic symbols. Fig. 2 presents an example of formal description of one hand gesture by HamNoSys (Russian letter “C” used in fingerspelling).

The algorithm automatically converts the HamNoSys codes to control trajectories and accepts most of the valid combinations of symbols. Final animation frames are the input to animation model. Time sequences of values determine trajectories controlling the joints, the control points or weights of the morph targets. Besides, it is also important that the virtual avatar simulates the manner of gesticulation as “humanly” as possible.



| Demonstrator   | Sign notation in HamNoSys: $\text{C} \text{ r} \text{ o} \text{ } \text{C}$ |                              | 3D signing avatar   |
|--|---|------------------------------|---|
|  | <b>Hand shape</b>   | $\text{C}$                   |  |
|  | <b>Hand orientation</b>   | $\text{r} \text{ o}$         |   |
|  | <b>Location</b>   | $\text{C} \text{ } \text{C}$ |   |
|  | <b>Motion type</b>  |                              |   |

Fig. 2. Formalization of manual gestures by the HamNoSys notation (on example of the Russian letter “C”).

Signing avatar displays 3D animation of the upper half of a human’s body (a full body model is not required as there are no gestures displayed below the belt level). The baseline system incorporates 3D articulatory model approximating skin surface of the human body by polygonal meshes. The meshes are divided into body segments describing arms, forearms, palm, knuckle-bones plus the parts of the talking head model. The full animation model is designed to express both manual and non-manual components of sign languages. The manual component is fully expressed by rotations of the body segments. The body segments are connected by joints and hierarchically composed into a tree structure (an approximation of body skeleton). Every joint is attached to at least one body segment. Thus the rotation of one body segment causes rotations of other body segments in lower hierarchy<sup>15</sup>.

In addition, synthesis of the non-manual component employs second control through the control points of the talking head model or more general morph targets. Thus the joint connections ensure movements of shoulders, neck, skull, eyeballs (eye gaze) and jaw. The control points and the morph targets allow us to change the local shape of polygonal meshes describing the face, lips, or tongue.

We organized a qualitative user evaluation of the system with the help of some teachers from the deaf people

society, and they positively estimated intelligibility and naturalness of lips articulation of the talking head and intelligibility of manual gestures of the signing avatar.

In future research the signing avatar can be applied in various assistive technologies for human-computer interaction<sup>17</sup>, in multimodal information kiosks<sup>18</sup>, interactive dialogue systems, etc.

### 3. Conclusions

In the paper, we have presented computer-based research and developments in the area of spoken sign languages. Sign languages are considered as non-written under-resourced spoken languages, and at present there are 138 living sign languages. Sign languages do not have a natural written form and there is a huge lack of electronic resources such as vocabularies, audiovisual databases, automatic recognition and synthesis systems, etc. We described several existing sign notation systems used for formal description of manual and non-manual gestures.

We have also presented the multi-modal audio-visual computer system for text-to-sign language and speech synthesis for Russian and Czech, including the general architecture of the synthesis system, bimodal talking head and animation of manual gestures and articulation. The developed system is not only aimed at deaf and hard-of-hearing people, but is useful for both ordinary and visually impaired people as well. Acoustic-based spoken language is a natural modality for communication with hearing-able people. Avatar's lips articulation synchronized with audio signal allows improving both intelligibility and naturalness of speech. The multi-modal system can be applied in various dialogue systems, multi-modal embodied communication agents, and learning systems.

The given paper does not deal with automatic recognition of sign language and fingerspelling gestures, but it is also a very important and dynamic direction for research, for example<sup>19,20,21</sup>.

### Acknowledgements

This research is partially supported by the Council for Grants of the President of Russia (projects MD-3035.2015.8 and MK-5209.2015.8), by the Russian Foundation for Basic Research (projects 15-07-04415 and 15-07-04322), as well as by the Czech Ministry of Education, Youth and Sports (project LO1506).

### References

1. Karpov A, Zelezny M. Towards Russian sign language synthesizer: Lexical level. In: *Proc. 5th International Workshop on Representation and Processing of Sign Languages*, Istanbul, Turkey; 2012. p. 83-86.
2. Stokoe WC. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Studies in Linguistics: Occasional papers* 1960;**8**:78 p.
3. Prillwitz S, et al. *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide*. Hamburg: Signum; 1989.
4. Liddell S, Johnson R. American Sign Language: the phonological base. *Sign Language Studies* 1989;**64**:195-278.
5. Dimskis LS. *Learning sign language*. Moscow: Academia; 2002. 128 p. (in Russian).
6. Gibet S, Courty N, Duarte K, Naour T. The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. *ACM Transactions on Interactive Intelligent Systems* 2011;**1**(1):Article No. 6.
7. Borgotallo R, Marino C, Piccolo E, Prinetto P, Tiotto G, Rossini M. A multi-language database for supporting sign language translation and synthesis. In: *Proc. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta; 2010. p. 23-26.
8. Karpov A, Krnoul Z, Zelezny M, Ronzhin A. Multimodal Synthesizer for Russian and Czech Sign Languages and Audio-Visual Speech. In: *Proc. 15th International Conference on Human-Computer Interaction HCI International-2013*, Las Vegas, USA, Lecture Notes in Computer Science 2013;**8009**:520-529.
9. Hanke T. HamNoSys - representing sign language data in language resources and language processing contexts. In: *Proc. International Conference on Language Resources and Evaluation LREC-2004*, Lisbon, Portugal; 2004. p. 1-6.
10. Tihelka D, Kala J, Matoušek J. Enhancements of Viterbi Search for Fast Unit Selection Synthesis. In: *Proc. International Conference INTERSPEECH-2010*, Makuhari, Japan; 2010. p. 174-177.
11. Hoffmann R, Jokisch O, Lobanov B, Tsirulnik L, Shpilevsky E, Piurkowska B, Ronzhin A, Karpov A. Slavonic TTS and SST Conversion for Let's Fly Dialogue System. In: *Proc. 12th International Conference on Speech and Computer SPECOM-2007*, Moscow, Russia; 2007. p. 729-733.



12. Železný M, Krňoul Z, Cisar P, Matousek J. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing* 2006;**86**(12):3657-3673.
13. Karpov A, Tsurulnik L, Krňoul Z, Ronzhin A, Lobanov B, Železný M. Audio-Visual Speech Asynchrony Modeling in a Talking Head. In: *Proc. INTERSPEECH-2009*, Brighton, UK; 2009. p. 2911-2914.
14. Krňoul Z, Železný M, Müller L. Training of Coarticulation Models using Dominance Functions and Visual Unit Selection Methods for Audio-Visual Speech Synthesis. In: *Proc. International Conference INTERSPEECH-2006*, Pittsburgh, USA; 2006. p. 585-588.
15. Krňoul Z, Kanis J, Železný M, Müller L. Czech Text-to-Sign Speech Synthesizer. In: *Proc. 4th International Workshop on Machine Learning for Multimodal Interaction MLMI-2007*, Brno, Czech Republic, Lecture Notes in Computer Science 2007;**4892**:180-191.
16. Karpov A, Ronzhin A, Kipyatkova I, Železný M. Influence of Phone-viseme Temporal Correlations on Audiovisual STT and TTS Performance. In: *Proc. 17th International Congress of Phonetic Sciences ICPHS-2011*, Hong Kong, China; 2011. p. 1030-1033.
17. Karpov A, Ronzhin A, Kipyatkova I. An Assistive Bi-modal User Interface Integrating Multi-channel Speech Recognition and Computer Vision. In: *Proc. 14th International Conference on Human-Computer Interaction HCII-2011*, Orlando, USA, Lecture Notes in Computer Science 2011;**6762**:454-463.
18. Karpov A, Ronzhin A. Information Enquiry Kiosk with Multimodal User Interface. *Pattern Recognition and Image Analysis* 2009;**19**(3):546-558.
19. Cooper H, Ong EJ, Pugeault N, Bowden R. Sign language recognition using sub-units. *Journal of Machine Learning Research* 2012;**13**:2205-2231.
20. Koller O, Forster J, Ney H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 2015;**141**:108-125.
21. Kindiroglu A, Yalcin H, Aran O, Hruz M, Campr P, Akarun L, Karpov A. Automatic Recognition of Fingerspelling Gestures in Multiple Languages for a Communication Interface for the Disabled. *Pattern Recognition and Image Analysis* 2012;**22**(4):527-536.