Multimodal Synthesizer for Russian and Czech Sign Languages and Audio-Visual Speech

Alexey Karpov[1,2], Zdenek Krnoul[3], Milos Zelezny[3], and Andrey Ronzhin[2]

[1] St. Petersburg State University, St. Petersburg, Russia
[2] St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Russia
[3] University of West Bohemia (UWB), Pilsen, Czech Republic
{karpov,ronzhin}@iias.spb.su,
{zdkrnoul,zelezny}@kky.zcu.cz

Abstract. This paper presents a model of a computer-animated avatar for the Russian and Czech sign languages. Basic principles of sign language(s) and their implementation in a computer model are briefly sketched. Particular attention is paid to animation principles of the "talking head", which allows for maximum expansion of the functions of the program, making it suitable not only for deaf and hard-of-hearing people, but for blind and non-disabled people too, so the universal audio-visual synthesizer is proposed.

Keywords: Signing Avatar, Speech Synthesis, Talking Head, Assistive Technology, Multimodal User Interface, Universal Access.

1 Introduction

Nowadays world society pays much attention to the problems of disabled persons with partial or full dysfunctions of their organs and sensory systems such as hearing, vision, speech impairments etc.

Animated 3D virtual characters (avatars) are very convenient for sign language (SL) and fingerspelling (FS) synthesis tasks. There is a lot of recent research on 3D signing avatars and SL machine translation systems both in the USA and Europe, for instance: DePaul ASL Synthesizer [1], EU projects Dicta-Sign [2,3], SIGNSPEAK [4,5], SignCom [6], Italian SL [7], ViSiCAST with Visia avatar [8], eSign with vGuido avatar [9], as well as Sign Smith and Sign4Me avatars of Vcom3D [10], SiSi project of IBM [11], iCommunicator system [12], etc. Recent studies were also made on machine learning and motor control in relation to visual speech synthesis. There are experimental works on computer synthesis of lips and tongue movements interpreted by talking head systems [13,14].

The paper is organized as follows: Section 2 describes specificity and general features of the Russian and Czech sign languages; Section 3 presents structure of the multimodal synthesizer, 3D signing avatar as well as information fusion and user interface; conclusions are given in the end of the paper.

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part I, LNCS 8009, pp. 520–529, 2013.

2 Specificity of Sign Languages

It is generally known that there is no such linguistic phenomenon as The Universal International Sign Language (SL). Quite the contrary, a lot of SLs (at least 130, according to the Ethnologue[1] report for sign languages of deaf people) have appeared in various deaf communities and changed with time having gotten new lexical items and grammar structures, just like natural languages (NL) spoken by us all over the world. Sign languages have gained legal recognition as means of communication in the USA, Canada, Australia, states of EU and some other countries. Since 1998 the law regulates use of Czech sign language and signed Czech as communication means of deaf in Czech Republic, and since the 30th of December 2012, the Russian SL has been officially recognized in its home country too[2].

Elaborating of human-computer interfaces, which make it possible for deaf people to interact with the "normal" world, would help SLs to expand their sphere of influence. One of the most challenging problems in this realm is NL-to-SL, SL-to-NL and SL-to-SL translations. Differences in SLs as well as between NLs and SLs cause still troubles at human and machine translation and interpretation due to general lack of knowledge of SL grammars. There are however essential similarities owing to their natural iconicity and visual interpretation.

The sign languages are mainly preferred by the deaf. They are natural forms developed in communities of deaf people. Signed modes of natural languages (manually coded languages) are a bridge between natural and sign languages. Likewise, they use signs, but the grammar and sentence structure is adopted from spoken languages.

2.1 Russian Sign Language and Fingerspelling

One of the most important problems connected with the Russian sign language (RSL), which is used by several hundred deaf people (approximately a half million) in Russia and some Commonwealth of Independent States (CIS) like Ukraine, Belarus, Kazakhstan, etc., consists of a geographical vast of the country and existence of various dialects of RSL, such as Moscow, St. Petersburg, Novosibirsk, Vladivostok, Minsk dialects and others. An expert analysis of electronic and published dictionaries of RSL shows that only 30-40% gestures are similar in the Moscow and Petersburg dialects nowadays.

Fingerspelling is a representation of letters or numerals using only the hands. In contrast to the British and Czech fingerspelling alphabets, all signs in the Russian fingerspelling system are performed by one (right) hand only. Moreover, seven signs of the 33 Russian letters (cheremes) are dynamic ones. Thus the cardinal differences between Russian or American (on the one hand) and British or Czech (on the other hand) fingerspelling signs are: all the Russian and American gestures for letters are one-handed, while most gestures in British and some signs in Czech fingerspelling are two-handed; quite a few Russian letters are dynamic, while all the British or Czech letters are static; besides that, finger shapes in the RSL are more complicated.

[1] www.ethnologue.com

[2] www.voginfo.ru

2.2 Czech Sign Language and Fingerspelling

It is quite difficult to estimate number of people who are deaf and use Czech sign language. In the Czech Republic the number is about 7500 of deaf signers (0.07% of the population, and about 500000 (4.7%) hard of hearing. Furthermore the Czech sign language can be a primary communication means of the hearing-impaired people. It is composed of the specific visual-spatial resources, i.e. hand shapes (manual signals), movements, facial expressions, head and upper part of the body positions (non-manual signals). It is not derived from or based on any spoken language.

On the other hand the signed Czech was introduced as an artificial language system derived from the spoken Czech language to facilitate communication between deaf and hearing people. Signed Czech uses grammatical and lexical resources of the Czech language. During the production, the Czech sentence is audibly or inaudibly articulated and all individual words of the sentence are simultaneously signed. Czech sign language and signed Czech use fingerspelling less than the sign languages in Russia and CIS countries. The fingerspelling is primarily incorporated in the sentence context of lexical signs to express shortcuts, new words etc.

3 Structure of the Multimodal Synthesizer

In the last years, UWB and SPIIRAS have been developing a universal multimodal (audio-visual) text-to-SL & speech synthesizer for Czech and Russian. The synthesizer takes text as an input and translates it into audio-visual speech and SL.

Thus the visual output is available for deaf and hard of hearing people, who can use SL and/or lip-reading; the audio output is oriented for visually impaired people; the audio-visual part of the interface is intended for non-disabled people. The universal multimodal synthesizer consists of several components (Figure 1):

- especially designed text processor that takes text as an input to generate phoneme (a minimal acoustic element of speech) and viseme (a visual equivalent of phoneme) transcriptions, and control selection of HamNoSys codes primarily intended for hand description of fingerspelling [15];
- TTS (text-to-speech) systems that generate auditory speech signal with time labeling corresponding to the entered text [16, 17];
- virtual 3D model of human's head with controlled lips articulation, mimics and facial expressions [18];
- control unit for the audio-visual talking head that synchronizes and integrates lips movements with synthesized auditory speech signal [19];
- virtual 3D model of human's upper body; we employ own skeletal model of the signing avatar, which can be controlled by HamNoSys codes [20];
- audio-visual multimodal user interface that synchronizes output audio and visual speech and gesture modalities, integrates all the components for automatic generation of auditory speech, visual speech (articulation and facial expressions) and avatar's gestures of signed Czech/Russian language and outputs audio and visual signals by PC loudspeakers and a display; the sign language and visual

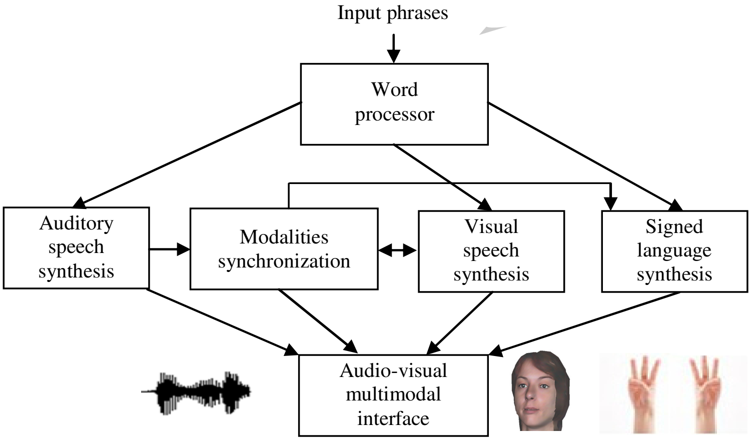
Input phrases
Word processor
Auditory speech synthesis
Modalities synchronization
Visual speech synthesis
Signed language synthesis
Audio-visual multimodal interface

Fig. 1. General structure of the multimodal synthesizer

speech interface is available for deaf and hard-of-hearing people, while audio-speech speech interface is focused on blind and visually impaired people, and the audio-visual talking head provides interaction with ordinary non-disabled users; so the proposed audio-visual synthesizer is a universal HCI interface.

Input phrases of speaking text 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 lip articulation by the talking head, as well as for signed language output by the full avatar) and letters (intended for fingerspelling synthesis) are found out and automatically processed to the symbols of HamNoSys sign notation. On this basis, the signing avatar outputs manual gestures of sign language/fingerspelling decoding HamNoSys notation symbols.

3.1 Architecture of the Audio-Visual Talking Head

We have implemented a 3D realistic talking head model for both the Czech and Russian languages. The talking head 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 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 the sets of control points) are concatenated to produce continuous stream of visual parameters. In the concatenative approach the co-articulation problem has to be solved to avoid unpleasant or unintelligible (unnatural) visual artifacts. In our case the co-articulation of the lips is modeled by method of selection of articulatory targets [21].

The animation technique uses virtual control points predefined on the face or tongue surface to move vertices of the 3D shape of the model. For smooth movements the vertices surrounding the control points are interpolated; animation is smoothed using the influence zones approach (see Figure 2c). 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 3D scanning system from a real face. The scanning system is based on a digital camera, projection of structured light on the face (moving a vertical thin stripe of white light) and system of four mirrors to obtain stereoscopy while avoiding the stereo correspondence problem [18].

The head model as such is represented by a set of points - vertices of a virtual space which are connected by edges to build up dense triangular surfaces and form a three-dimensional model (Figure 2a). The obtained 3D data are processed, completed by adding manually created other face parts and kept in a file in a virtual reality format (VRML) as a set of vertex coordinates triangular planes and textures of the face of the speaker. The full head model is described by tens of thousands of vertices (Figure 2b), however, only few of them are active, i.e. can be commanded by the program, simulating movements of facial muscles; control of those allows displaying visemes.

The talking head model is also employed for creation of a viseme set in a setup phase of the system, specified by the parameters of the model for given language.

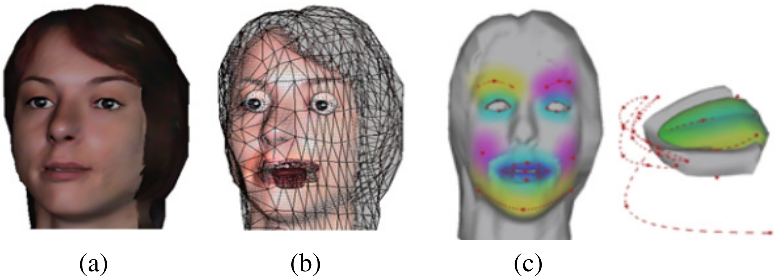

Fig. 2. Talking head: a) view of 3D model; b) view of wired model; c) model influence zones

In the given synthesizer, not only the general model of a head is used, but also additional models of its parametrically controlled facial components: the eyes, tongue, upper and bottom jaws, and internal speech organs. They are created on the knowledge base of anthropological physiology. Each model can be controlled by the program irrespective of the head. One more advantage of creation of separate models of "talking head" elements is possibility to create "an illusion of a live head" (eye-blinking/movements, jaw side-move, breathe movements, etc.).

For Russian language we apply a TTS synthesizer based on allophones and multi-allophones natural waves (ANWs and MANWs) concatenation [16]. The speech prosody synthesis uses an original Accentual Units Portrait (AUP) model for a stylization of tonal, rhythmical and dynamic contours of a phrase. Fusion of these modules allows synthesizing speech with a high degree of intelligibility and naturalness. In the system an incoming orthographic text to be transformed into speech signal undergoes a number of successive operations carried out by specialized processors: textual, phonemic, prosodic and acoustical. 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 (AU), and finally marks the intonation type of the input phrase.

Synchronization of face/lip movements with synthesized acoustical signal is based on timestamps of allophones in the synthesized speech flow. Duration of every allophone is based on allophone's average length and desired speech tempo. Synchronization of virtual face and lip movements with synthesized acoustical signal is realized on the basis of information known about positions of 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 audio and visual speech cues and to take into account different speech rates, 16 context-dependent timing rules for transitions between displaying visemes are applied [17]; this method allows increasing naturalness and intelligibility of generated audio-visual speech [22].

3.2 Architecture of the Signing Avatar

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

The architecture of the signing avatar incorporates a baseline sign language translation module experimentally designed for the signed Czech. In general, the sign language translation module uses an automatic phrase-based translation system. Sentences are divided into phrases and these are then translated into corresponding sign speech phrases. The translated words are reordered and rescored using a language model at the end of the translation process. In the architecture, we use own implementation of a simple monotone phrase-based decoder.

The main resource for the statistical machine translation is a parallel corpus which contains parallel texts of a source and a target language. The acquisition of such corpus is complicated by the absence of an official written form of both the sign and signed languages. Therefore we experimentally collected own Czech - signed Czech parallel corpus for baseline setup of the translation system. The decoder and performance of the system is described in more details in [23]. In general, created sign speech phrases determine sign order, dependences and form.

Signing avatar synthesis system creates 3D animation of the upper half of a human figure (Figure 3). 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. Joint connection incorporates rotation limits to prevent non-anatomic poses of the animation model.

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.

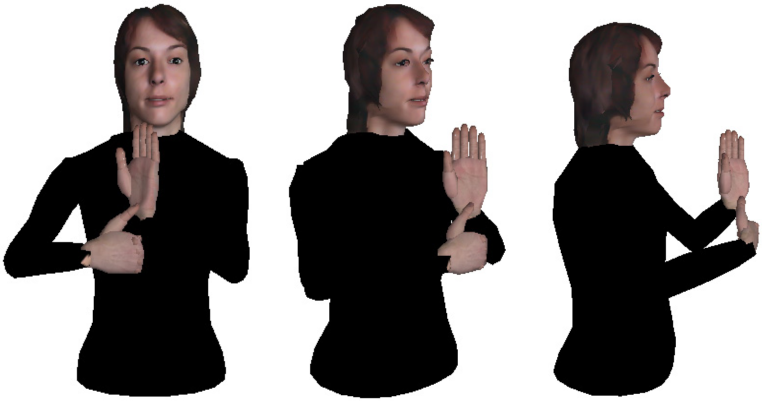

Fig. 3. 3D views of the signing avatar (it shows numeral "16" of the Russian sign language)

The synthesis module incorporates conversion algorithm for Hamburg notation system (HamNoSys, [15]) to create necessary signs and fingerspelling gestures [20]. An 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.

The algorithm allows modeling almost any configuration and hand movements in the presence of corresponding visual means. It is important too, that the virtual avatar simulates the manner of gesticulation as "humanly" as possible.

3.3 Audio-Visual Multimodal Interface

Synchronization of audio-visual speech with visual gestures is controlled by time stamps of start and end of spoken words generated by the audio speech synthesizer. Since natural speech has a higher tempo than the corresponding signs, then the avatar pronounces and articulates isolated spoken words and waits for the following word until completion of the current gesticulation (if there is no sign for a word in the vocabulary then it is spelled by finger signs with the avatar's right hand in the case of Russian or both hands in the case of Czech sign language). The proposed SL interface has a lot of advantages:

- it allows a user to see generated visual data from different sides and viewing angles that results in better understanding of spatial information, e.g. distance between the hands and the body or hands each from other;

- it is possible to add new tokens into the sign vocabulary quite easily thanks to the usage of the animated virtual avatar, so there is no requirement to record one human SL speaker (e.g. identity with same dress, haircut and make-up with similar lighting conditions and equipment, in contrast to video recordings);
- it can generate a continuous stream of visual signs without transitions through a neutral position of hands and there are no seen borders between adjacent signs;
- it is possible to change one 3D avatar with another one and to create new models of human beings or characters;
- synthesized signed phrases can be conformed to type of HCI with any required speed, slowing down or speeding up video stream;
- sign language translation module can be conformed to provide next algorithms in the future, which solving translation problems linked with specificity of sign language grammars

The developed multimodal avatar is convenient for deaf and hard-of-hearing people, as well as for blind and non-disabled people, so it is the universal synthesizer designed especially for HCI, interactive dialogue systems, and communication agents [24, 25]. It generates signals of auditory and visual output modalities and fuses dynamic speech gestures with the avatar's hands, auditory speech and lips articulation/facial expressions. Multimedia demonstrations of the proposed synthesizer for signed languages and fingerspelling (one-handed for Russian and two-handed for Czech) are available in Internet [26]. Initial intelligibility test have been performed with sign language synthesizer and signed Czech [27]. For this purpose, 20 videos of short utterances were synthesized. Hearing participants (sign language experts) evaluate each utterance as a whole. Subtitles were added to the videos to show text transcription to the signs and to give the participants original meaning of the utterance. In accordance to the evaluative criteria, the animation of 14 of 20 utterances showed the signs from subtitles. Qualitative user evaluation of the 3D signing avatar for signed Russian and fingerspelling was made with the help of some representatives of the All-Russian society of the deaf in St. Petersburg. They said on novelty and urgency of the system and positively estimated intelligibility and naturalness of lips articulation of the talking head and recognizability of manual gestures of the virtual avatar.

4 Conclusion

The multimodal system is aimed not only for deaf, hard-of-hearing, and hearing impaired people, but is useful for hearing people as well. It is a universal multimodal computer system that is currently tested for synthesis both Russian spoken language (audio-visual modality) and the sign language (visual modality). Generated audio-visual signed Russian speech and language is a fusion of dynamic gestures shown by the avatar's both hands (or only by the right hand in the case of Russian fingerspelling), lip movements articulating spoken words and acoustic speech. Many deaf people are able to lip-read and to understand phrases even without manual gestures. Acoustic spoken language is a natural speech modality for communication with hearing-able people. Avatar's lips articulation synchronized with audio stream helps to improve both intelligibility and naturalness of generated speech. The proposed universal synthesizer can be applied in various dialogue systems, multimodal embodied communication agents as well as in learning systems.

Acknowledgements. This research is supported by the St. Petersburg State University (project No. 31.37.103.2011), by the Ministry of Education and Science of Russia (contract No. 11.519.11.4020), by the grant of the President of Russia (project No. MK-1880.2012.8), as well as by the Technology Agency of the Czech Republic (project No. TA01011264).

References

1. DePaul ASL Synthesizer, `http://asl.cs.depaul.edu`
2. Efthimiou, E., et al.: Sign Language technologies and resources of the Dicta-Sign project. In: Proc. 5th Workshop on the Representation and Processing of Sign Languages, Istanbul, Turkey, pp. 37–44 (2012)
3. Dicta-Sign Project, `http://www.dictasign.eu`
4. Caminero, J., Rodríguez-Gancedo, M., Hernández-Trapote, A., López-Mencía, B.: SIGNSPEAK Project Tools: A way to improve the communication bridge between signer and hearing communities. In: Proc. 5th Workshop on the Representation and Processing of Sign Languages, Istanbul, Turkey, pp. 1–6 (2012)
5. SIGNSPEAK Project, `http://www.signspeak.eu/en`
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 1(1) (2011)
7. Borgotallo, R., et al.: 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, pp. 23–26 (2010)
8. ViSiCAST Project, `http://www.visicast.co.uk`
9. eSign Project, `http://www.sign-lang.uni-hamburg.de/esign`
10. Vcom3D Company, `http://www.vcom3d.com`
11. SiSi Project, `http://www-03.ibm.com/press/us/en/pressrelease/22316.wss`
12. iCommunicator project, `http://www.icommunicator.com`
13. Beskow, J.: Trainable articulatory control models for visual speech synthesis. Journal of Speech Technology 4(7), 335–349 (2004)
14. Youssef, A., Hueber, T., Badin, P., Bailly, G.: Toward a Multi-Speaker Visual Articulatory Feedback System. In: Proc. International Conference INTERSPEECH-2011, Florence, Italy, pp. 589–592 (2011)
15. 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, pp. 1–6 (2004)
16. Hoffmann, R., Jokisch, O., Lobanov, B., Tsirulnik, L., Shpilewsky, 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, pp. 729–733 (2007)

17. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi Search for Fast Unit Selection Synthesis. In: Proc. International Conference INTERSPEECH-2010, Makuhari, Japan, pp. 174–177 (2010)
18. Železný, M., Krňoul, Z., Cisar, P., Matousek, J.: Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. Signal Processing 86(12), 3657–3673 (2006)
19. Karpov, A., Tsirulnik, L., Krňoul, Z., Ronzhin, A., Lobanov, B., Železný, M.: Audio-Visual Speech Asynchrony Modeling in a Talking Head. In: Proc. International Conference INTERSPEECH-2009, Brighton, UK, pp. 2911–2914 (2009)
20. Krňoul, Z., Kanis, J., Železný, M., Müller, L.: Czech Text-to-Sign Speech Synthesizer. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 180–191. Springer, Heidelberg (2007)
21. 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. 9th International Conference on Spoken Language Processing INTERSPEECH-2006, Pittsburgh, PA, pp. 585–588 (2006)
22. 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, pp. 1030–1033 (2011)
23. Kanis, J., Zahradil, J., Jurčíček, F., Müller, L.: Czech-Sign Speech corpus for semantic based machine translation. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 613–620. Springer, Heidelberg (2006)
24. Hrúz, M., Campr, P., Dikici, E., Kindirouglu, A., Krňoul, Z., Ronzhin, A., Sak, H., Schorno, D., Akarun, L., Aran, O., Karpov, A., Saraclar, M., Železný, M.: Automatic Fingersign to Speech Translation System. Journal on Multimodal User Interfaces 4(2), 61–79 (2011)
25. Krňoul, Z.: Web-based sign language synthesis and animation for on-line assistive technologies. In: Proc. 13th International ACM SIGACCESS Conference on Computers and Accessibility ASSETS-2011, Dundee, Scotland, UK, pp. 307–308 (2011)
26. Audio-visual demonstration of the universal multimodal synthesizer for Russian, `http://www.spiiras.nw.ru/speech/demo/daktilrus.avi`
27. Krňoul, Z., Železný, M.: Translation and conversion for Czech Sign Speech synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 524–531. Springer, Heidelberg (2007)