Input and output modalities used in a sign-language-enabled information kiosk

Marek Hrúz (1), Pavel Campr (1), Alexey Karpov (2),
Pınar Santemiz (3), Oya Aran (3) and Miloš Železný (1)

(1) Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

campr@kky.zcu.cz mhruz@kky.zcu.cz zelezny@kky.zcu.cz

(2) Speech and Multimodal Interfaces Laboratory,

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

karpov@iias.spb.su

(3) Computer Engineering Department, Boğaziçi University, İstanbul, Turkey

aranoya@boun.edu.tr psantemiz@gmail.com

Abstract

This paper presents description and evaluation of input and output modalities used in a sign-language-enabled information kiosk. The kiosk was developed for experiments on interaction between computers and deaf users. The input modalities are automatic computer-vision-based sign language recognition, automatic speech recognition (ASR) and a touchscreen. The output modalities are presented on a screen displaying 3D signing avatar, and on a touchscreen showing special graphical user interface for the Deaf. The kiosk was tested on a dialogue providing information about train connections, but the scenario can be easily changed to e.g. SL tutoring tool, SL dictionary or SL game. This scenario expects that both deaf and hearing people can use the kiosk. This is why both automatic speech recognition and automatic sign language recognition are used as input modalities, and signing avatar and written text as output modalities (in several languages). The human-computer interaction is controlled by a computer-driven dialogue system.

1. Introduction

Deaf and hearing-impaired people have limited possibility of communication both with hearing people and computers [1]. They cannot use speech-based automatic information services. To enable such a service the dialogue system should be designed to be accessible by the deaf users. The first step is to use sign language synthesis to provide feedback to the user who can use standard devices for input, such as mouse, keyboard or touchscreen. Inspired by voice-controlled applications, this setup can be enhanced by adding sign language recognition as a new input modality. For some scenarios (e.g. our examined train information service) it is sufficient to use only touchscreen without sign language recognition, but for other scenarios this input is the primary one (e.g. sign language tutoring tools or sign language games). In this paper we present all modalities which are used in our prototype of sign-language-enabled information kiosk, which is able to communicate with the user in sign language in both directions. Additionally, the kiosk can be used by hearing people with automatic speech recognition as another input modality. The kiosk was tested on train connection information service scenario where the user can query the kiosk for information such as train departures or timetables.

2. System overview

The kiosk is an alone-standing machine equipped with a PC, a touchscreen displaying dialogue status and buttons (fig. 1), large screen presenting sign language output by 3D avatar, microphone and cameras (fig. 2).

The kiosk uses a computer-driven dialogue for communication with the user. When the user comes to the kiosk the face detector triggers the start of the dialogue. The avatar guides the user through the dialogue. When an input from the user is expected, the avatar asks a question and the user can answer either by sign language, speech or by pressing appropriate button on the touchscreen, where all possible answers for current question are listed. When the input is incorrectly recognized or the user wants to change some answer, he can click appropriate button on the touchscreen, where all questions and already entered answers are listed (see left part of fig. 1).
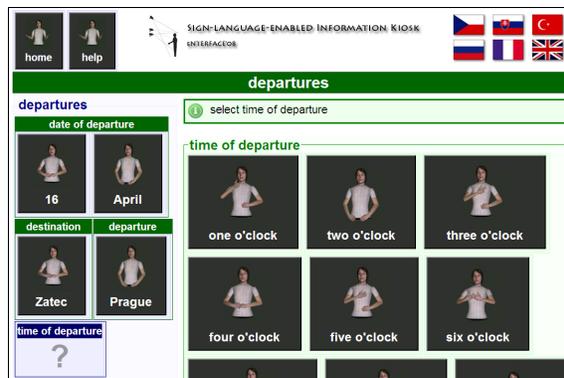

Figure 1: Sample screen of touchscreen graphical user interface. Gray buttons and flags are clickable. Top: main menu. Left: dialogue status with answered and unanswered questions. Right: current question and all possible answers which can be selected by haptic, speech or sign-language modality.

2.1. Dialogue control and graphical user interface

The interaction between the user and the kiosk is managed by a computer-driven dialogue system (fig. 3). The dialogue consists of several scenarios, which are defined as a set of questions, their possible answers and answer generator. The answers can
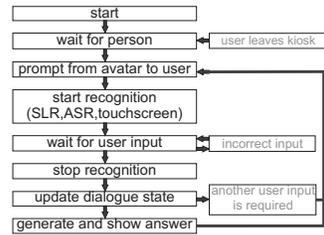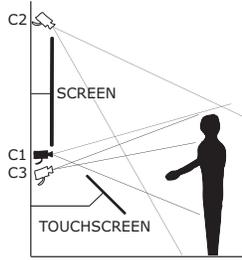
Figure 2: Kiosk arrangement
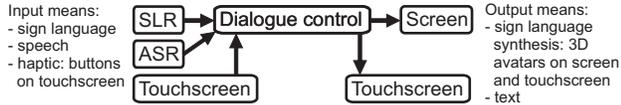
Figure 3: Dialogue flow

Figure 4: Modules overview

be entered as one- or multiple-word sentence (now supported by speech recognizer, planned for SL recognizer in the future). All entered sentences must satisfy a context-free grammar defined by BackusNaur Form (BNF), which allows to generate a list of all possible answers and to validate the user's answer. When all required questions are answered the dialogue system generates an answer for the user's query.

Graphical user interface (GUI) consists of two screens. The first shows a signing avatar which guides the user through the dialogue by asking questions and presenting an answer when all questions are answered. To increase interactivity the avatar turns in the direction to the user. This is achieved by face detection.

The touchscreen presents main menu, current dialogue status and a list of all possible answers for current question.

3. Sign language recognition

For the training purposes a small database was created. In total we recorded 338 video files with one male and one female signer. It contains 50 signs from Signed Czech language. Each sign was repeated three times by the user. We used special recording conditions such as long sleeves, non-skin-colored clothes, uniform background and constant illumination.

3.1. Skin Color Segmentation

Skin color is widely used to aid segmentation in finding parts of human body [2] in images. Our model of skin color was trained on manually segmented images from a training set derived from the recorded database and for better robustness also from database UWB-06-SLR-A [3]. In total, we processed 50 video segments. An example of training images is shown in Fig 5. We model the colors as a Gaussian Mixture Model (GMM) in RGB color space. The segmentation is described in detail in [4] and a result can be seen in Fig 6.

3.2. Hands and head tracking

We use a joint Particle Filter (PF) that calculates a combined likelihood of all objects. This is done by modeling the likelihood of each object with respect to the other objects. This algorithm is described in detail in [5].

The state vector for a single object consists of the position, the velocity and the shape parameters. The shape parameters are the width, the height and the angle of an ellipse surrounding the

Figure 5: Skin-color examples of one signer.

object. Then, the joint particle is a single 21 dimensional vector containing all the objects in the scene:

$$\mathbf{x}_t^n = \{\mathbf{x}_t^{n,f}, \mathbf{x}_t^{n,r}, \mathbf{x}_t^{n,l}\}^T \qquad (1)$$

where f, r, l are indexes of the face, right and left hand. $\mathbf{x}_t^n$ represents the joint particle and $\mathbf{x}_t^{n,i}$ represents the particle or the sub-particle, alternatively.

For each object, the position and the velocity parameters are modeled by a damped velocity model and the shape parameters are modeled by a random walk model. For multiple objects in the joint PF, the dynamic model is applied to each object. We additionally apply mean shift (MS) [6] to the sub-particles of each object independently. The MS algorithm moves the particle centers to the areas with high skin color probability. This allows us to use particles effectively, since the particles with low weights will be less likely. As a result, a PF with MS needs fewer particles than a standard PF.

As observation we use the skin color probability image, which has a positive probability for skin color pixels and zero probability for other colors. To calculate the likelihood of a single object, we make two measurements based on the ellipse that is defined by the state vector of the particle:

- A: The ratio of the skin color pixels to the total number of pixels inside the ellipse.

- B: The ratio of the skin color pixels to the total number of pixels at the ellipse boundary.

These two ratios are considered jointly in order to make sure that our measurement function gives high likelihood to particles that contain the whole hand without containing many non-hand pixels [5]. If we do not take the ellipse boundary into account, smaller ellipses are favored and particles tend to get smaller. We design our measurement function Eq.2 to return a high likelihood when A is as high as possible and B is as low as possible:

$$z_t^{n,i} = \begin{cases} 0 & , A < \Phi_p \\ 0.5 \cdot A + 0.5 \cdot (1 - B) & , otherwise \end{cases} \qquad (2)$$
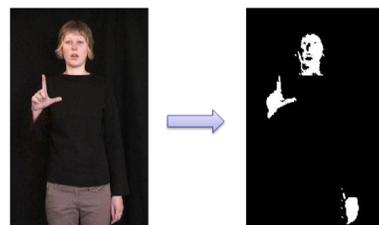

Figure 6: Skin color segmentation example
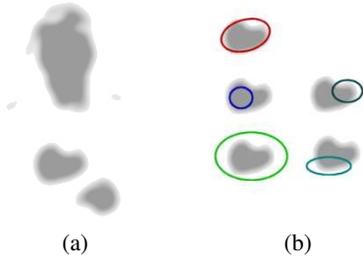
(a)   (b)

Figure 7: (a) The thresholded image that is used in likelihood calculation, (b) hand region and different particles. The likelihood function gives the highest likelihood for the particle at the top.

where $z_t^{n,i}$ denotes the likelihood of a single object, i, for the particle n.

The first line in Eq.2 is required to assign low likelihood to particles that have zero or very few skin color pixels. Otherwise these particles will receive 0.5 likelihood value even if they do not contain any skin color pixels. The equation takes its highest value when there are no skin colored pixels at the boundary ($B = 0$), and when all the inner pixels are skin colored. Figure 7b shows the grey-level hand image and possible particles.

Hand Shape Feature Extraction   In order to recognize the signs, other than the tracking features we also need shape information. For the shape description, we implemented five algorithms and tested their performance over a set of finger alphabet in Signed Czech Language.

The algorithms used were computation of DCT coefficients from the gray scale image [7], Hu moments from the segmented hand image [8], and Fourier descriptors from the contour points of the hand shape [9]. The best method proved to be the DCT coefficients.

Recognition   For every frame we extract 7 tracking features for each object of interest (hands and head) and 14 DCT coefficients for each hand. Together it is 49 features for one frame. For the purpose of sign recognition we use Hidden Markov Model (HMM) with 8 states (Fig. 8) [10].

One model of a sign was trained on 4 out of 6 video sequences. PCA and ICA were tested to reduce the dimensionality of the data and to align the data according to the feature space coordinate system. Both methods failed to improve the recognition rate since there were too few samples available. The remaining two video sequences were used to test the system.

4. Automatic speech recognition

The lexicon of our Automatic Speech Recognition (ASR) contains 101 diverse words for Czech and English (Czech towns, digits, names of months and weekdays, etc.) The audio signal is captured by a microphone of a headset and sampled at 16 KHz
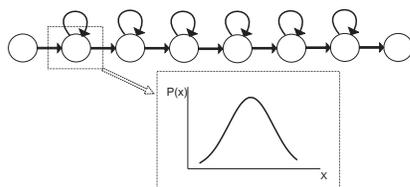

Figure 8: The structure of HMM used for modeling the signs.

with 16 bits on each sample using a linear scale. The system is intended for distant talking and human-computer interaction, so the microphone was selected to be able to capture speech signal at the distance of 2 meters from a speaker with an acceptable SNR.

Feature Extraction   The signal is divided into the frames and cepstral coefficients are computed for the 25 ms overlapping frames with 10 ms shift between adjacent frames applying the bank of triangular filters calculated according to the Mel-scale frequencies. Mel-frequency cepstral coefficients (MFCCs) are calculated from the log filter bank amplitudes using the discrete cosine transform. Hence the audio speech recognizer system calculates 13 MFCCs (including 0-th coefficient) as well as estimates the first and second order derivatives that forms an observation vector of 39 components. The acoustical modeling is based on left-right continuous Hidden Markov Models (HMMs) [11], applying mixtures of Gaussian probability density functions. HMMs of phonemes have three meaningful states and two hollow states intended for concatenation of the models of phonemes in the models of words. As the base technology for realization of speech recognizer Hidden Markov Models Toolkit (HTK) [12] was used.

Training of the recognizer   In order to train the speech recognizer a speech corpus was recorded in office conditions using the distant talking directed microphone. About 1000 utterances of two users were recorded and used for training HMMs of phonemes. These data were labeled semi-automatically in the terms of phoneme sets. Totally 41 different phonemes are used in transcriptions of the lexicon. 20% of training utterances were manually labeled by the WaveSurfer software, the rest of the data were automatically segmented by the Viterbi forced alignment method with the flat start [12]. The speech decoder uses Viterbi-based token passing algorithm [12].

5. Sign language synthesis

3D animation model of the avatar is in compliance with the H-Anim standard. Currently the model covers 38 joints and body segments. Each segment is represented as textured triangular surface, 16 segments for fingers and the palm, one for the arm and one for the forearm. The thorax and the stomach are together represented by one segment. The talking head is composed from seven segments. The body segments are connected by the avatar skeleton. One joint per segment is sufficient for this purpose. Controlling of the skeleton is carried out thought the rotation of segments (3 DOF per joint). The rotation of the shoulder, elbow, and wrist joints are completed by the inverse kinematics from 3D positions of the wrist and shoulder joints.

The animation of the talking head is performed by a local deformation of the triangular surfaces [13]. The triangular surfaces are deformed according to a set of 3D control points. The number of these points is reduced by PCA to 9 animation parameters. The rendering of the animation model is implemented in C++ using OpenGL.

The manual component of signed speech is generated by a trajectory generator. It performs the syntactic analysis of a input HamNoSys string and creates a parse tree structure by using 374 parse rules. The structurally correct string is decomposed to nodes in accordance with the parsing rules. There are two key frames to distinguish between the dominant and the non-dominant hand, both composed of 17 items. The leaf nodes are filled from the symbol descriptors stored in the definition file covering 138 HamNoSys's symbols.

For each rule, one of 39 designed rule actions is added to

transform parse tree to the control trajectories in accordance with the timing of the particular nodes. Finally, the trajectories for both hands are obtained in the root node. The final step is the conversion of the trajectories into the avatar animation and synchronization with talking head trajectories generated by the selection of articulatory targets [13].

6. Evaluation

From all of the mentioned methods of hand shape recognition the best results were achieved with the DCT coefficients. The reason is that letters in Czech finger alphabet are similar in borders (contours) but differ in the texture which the DCT takes into account. The recognition rate was 75%.

The signs were modeled as an 8-state HMM (two of these states are non-emitting). Each state is modeled as one Gaussian. This was due to the relatively small amount of data. With this configuration we achieved a recognition rate of 81.63%. To obtain more presice results more data are needed.

The performance of ASR was evaluated by speech data collected in the same office conditions as the training part. The word recognition rate (WRR) was over 90%. This rate is acceptable for our task since ordinary single microphone is used for distant speech capturing providing quite low SNR. In further research a microphone array and corresponding digital signal processing methods for speaker localization and noise elimination are supposed to be applied [14].

A subject evaluation of the quality of synthesized signed speech has been performed with deaf children. The evaluation was scored on the isolated signs. Two experiments (A,B) have been designed. Five deaf pupils (5-6 years old) and 6 deaf pupils (11-13 years old) participated in the experiments. Both experiments contained synthesized animations of 15 isolated signs (5 non-scored for demonstration and 10 for test) which were presented on a wall in the classroom by a data projector. The experiment A was a multiple-choice test with three options (one was correct). In the experiment B there were no options and the pupils were asked to write down the correct answer.

The score for a correct answer was one point and zero for wrong. The evaluation was done by one-sample and one-sided t-test. The results show significantly better understanding of the signed speech than a chance (the chance level 33.3%). The younger children achieved a 82.5% success in the experiment A. The older achieved 95%. The experiment B was attended only by the older pupils and they achieved 80% success rate. The results have proved that the sign language sythesis is understandable by hearing impaired people.

Because the kiosk should be usable by all hearing-impaired people (even who cannot read) we designed the GUI in a way that all important text labels are accompanied by an animation of signing avatar with corresponding sign. This graphical user interface component can be used as a clickable button. This component and whole GUI layout is designed as a XHTML web page, so that the whole application could be used online (without ASR and SLR input components).

Currently, the biggest usability problem is the SL recognition where the user has to start and finish the performed sign in the initial position. This condition must be explained to the user in the early beginning of the dialogue.

The other parts of the kiosk proved that they can be used without any major usability difficulties. This is achieved by using similar concept of the dialogue that is used in hypertext browsing which is well-known and the users don't have to think how to control the kiosk.

8. References

[1] J. Langer, "Přínos elektronických výukových pomůcek a slovníků znakového jazyka," *In Vzdělávání sluchově postižných. Praha: MŠMT, 2006.*, 2006.

[2] V. Sazonov V. Vezhnevets and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Graphicon*, 2003, pp. 85–92.

[3] P. Campr, M. Hrúz, and M. Železný, "Design and Recording of Czech Sign Language Corpus for Automatic Sign Language Recognition," *Proceedings of the Interspeech 2007, Antwerp, Belgium*, 2007.

[4] Oya Aran, Ismail Ari, Pavel Campr, Erinc Dikici, Marek Hruz, Deniz Kahramaner, Siddika Parlak, Lale Akarun, and Murat Saraclar, "Speech and Sliding Text Aided Sign Retrieval from Hearing Impaired Sign News Videos," *eNTERFACE 2007 Proceedings*, 2007.

[5] O. Aran and L. Akarun, "A particle filter based algorithm for robust tracking of hands and face under occlusion," in *IEEE 16th Signal Processing and Communications Applications (SIU 2008)*, 2008.

[6] E. Maggio and A. Cavallaro., "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[7] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Prentice-Hall, 2001.

[8] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, pp. 179187, 1962.

[9] D. S. Zhang and G. Lu, "A comparative study of three region shape descriptors," in *Proc. of the Sixth Digital Image Computing Techniques and Applications (DICTA02)*, 2002, pp. 86–91.

[10] J. Trmal, M. Hruz, J. Zelinka, P. Campr, and L. Muller, "Feature Space Transforms for Czech Sign-Language Recognition," *Proceedings of the Interspeech 2008, Brisbain, Australia*, 2008.

[11] R. Rabiner and B. Juang, "Fundamentals of speech recognition," in *New Jersey: Prentice-Hall, Englewood Cliffs, USA*, 1993.

[12] S. Young et al., "The htk book," in *HTK Version 3.4, Cambridge University Engineering Department*, 2006.

[13] Zdeněk Krňoul and Miloš Železný, "A development of Czech talking head," in *Proceedings of ICSPL 2008*, in press, 2008.

[14] S. Brandstein and D. Ward, "Microphone arrays," in *Springer Verlag*, 2000.