Development of Multimodal Applications for Disabled People

Andrey Ronzhin[1]*, Alexey Karpov*[1]*, Milos Zelezny*[2]*, Roman Mesheryakov*[3]*, Ruediger Hoffmann*[4]

[1]St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Russia
{ronzhin,karpov}@iias.spb.su

[2]University of West Bohemia in Pilsen (UWB), Czech Republic
zelezny@kky.zcu.cz

[3]Tomsk State University of the Control system and Radioelectronics (TSUCSR), Russia
mrv@keva.tusur.ru

[4]Institut für Akustik und Sprachkommunikation, Technische Universität Dresden (IAS TUD), Germany
ruediger.hoffmann@ias.et.tu-dresden.de

Abstract

Now European Society pays especial attention to the problems of handicapped persons, which have partial dysfunctions of some organs. There are several kinds of disabilities: voice, vision, hands, legs, ears. They are not able to use standard means to interact with computer, domestic technique, moving objects, etc. It leads to limitation of their social activity and life status. To rehabilitate the lost ability such modern scientific direction as multimodal interfaces is developed. Multimodal interfaces combine speech with other natural modalities (gestures, movements of head, lips, etc.). Multimodality allows choosing an accessible way of interaction in the concrete application for the concrete user. In multimodal systems the information from several video, audio, sensor communication channels is continuously detected and processed to create real and virtual environment, satisfy the wishes of a user and quickly adapt to the current task and applied aspects.

1. Introduction

The aim of the research is development of mathematical models of multimodal interaction and design of applications for groups of disabled people. Fundamental research of human-computer interaction taking into account the needs of a concrete human will have innovative character for designing the pilot systems for end-users. Based on study of persons' needs the flexible architecture of the multimodal interface will be developed. This model will allow replacing inaccessible modalities and carry out the effective fusion and fission of information streams. Three applied areas are chosen for research: (1) speech rehabilitation of oncological patients after full removal of throat; (2) assistive control system for people with disabilities of hands; (3) audio-visual speech recognition system for deaf-and-dumb people. The prototypes of the real applications for disabled people will be developed and tested by end-users.

Restoration of speech at patients with infringement of the speech formation system (the resection of throat, a tumour of tongue, etc), as a rule, is accompanied by the problems demanding active actions as the patient, and the attending physician. Active use of computer facilities at a stage of diagnostics of diseases is obvious, but at stage of treatment with use of a biological feedback it will allow to raise speed and quality of speech rehabilitation of patients.

Many people are unable to operate a standard computer mouse or keyboard because of disabilities of their hands or arms. One alternative is multimodal systems, which allow a person to control a computer without using standard mouse and keyboard, for example: (1) using head movements to control the cursor across the computer screen; (2) using the speech for giving the control commands. Speech and head-based control systems have a great potential to increase the life comfort of disabled people, their social protectability and independence from other people. Thus a hands-free control devices such as hands-free mouse and keyboard for access to PC is one effective application of these technologies. Users who have difficulties using a standard devices could manipulate mouse cursor merely by moving their heads and giving the speech command instead of clicking the buttons.

Fusion of audio and visual information helps the deaf-and-dumb people in interaction with other people or transform the current information into an accessible form. The bimodal audio-visual speech recognition system will use two kinds of information audio and visual (lips movement while pronouncing the speech). Also audio-visual speech recognition is useful in the different conditions: (1) in noisy environment the audio information is not enough and visual information allows to increase the speech recognition accuracy; (2) in poor illumination the audio information can fill the gap of visual information.

For this research the efforts and experience of Russian and European universities will be applied. IAS TUD has wide experience in development of multilingual system for speech synthesis and recognition, as well as creation of large speech databases in framework of INTAS projects [1]. TSUCSR investigates the speech production and perception models for oncological medical applications. Audio-visual speech recognition developed by UWB are adjusted to Russian language and combined with Russian audio speech recognition system, developed by SPIIRAS. As the result of integration of researches of TUD, UWB, TSUCSR and SPIIRAS the bimodal system for audio-visual Russian speech recognition will be created and applied to assist for disabled.

2. Models of speech production for dumb people

TSUCSR constructs serial mathematical models for formation of normal and esophagus speech for investigation of processes of generation vocal sounds [2]. The comparative analysis of the speech signals generated in models with real speech is carried out.

The given researches at association with the biology feedback are actively used at medical researches and speech rehabilitation of oncological patients after operation on full removal of a throat. We developed the technique of an estimation of quality esophagus voices on the basis of the quantitative scales. The techniques and algorithms of biology feedback of parameters esophagus voices are created. These parameters are: (1) Maximal phonation time; (2) Standard deviations and factor of variation of fundamental frequency of fluctuation and intensity; (3) Instability of voice frequency; (4) Stability of peak values of amplitudes; (5) Average spectrum of a voice; (6) Fundamental frequency; (7) Duration of a pause; (8) Melodiously variations; (9) Intensity variations; (10) Duration and parities of words, sounds; (11) Distribution of pauses in a speech stream; (12) Signal and noise parity.

Approbation of the technique was carried out during 2004-2006. Feature of receipt of patients in scientific-research institute of oncology with localization of a cancer tumour in area of throat that patients act very non-uniformly. Average growth of quantity of the periods of the basic tone for training at the first stage of 11,5 %. Results of the second stage are presented in Figure 1. The third stage passes clinical approbation. Growth of fundamental frequency for one training cycle is 3.5 Hz in average.

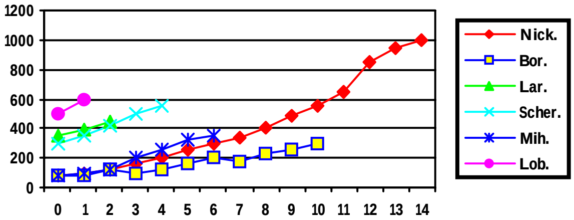

Figure 1: Growth of phonation duration (ms) from quantity of trainings

This data allows saying about efficiency of biology feedback trainings.

Table 1: The complex estimation of gullet voices.

patient	FF estimation	Estimation of FF deviation	Estimation of phonation duration
Bor.	92%	87%	22%
Nick.	100%	100%	100%

Unconditional achievement was that conscious management of speech rehabilitation process with the control of each stage for the first time has been applied and it has already allowed to modify a technique of training.

3. ICanDo: Intellectual Computer AssistaNt for Disabled Operators

A prospective multimodal system ICanDo intended for assistance to persons without hands or with disabilities of their hands or arms in human-computer interaction are developed in SPIIRAS. This system combines the modules for automatic speech recognition and head tracking in one multimodal system [3].

ICanDo system can use the voice commands of a user in two languages: Russian and English. For automatic speech recognition the SIRIUS system (SPIIRAS Interface for Recognition and Integral Understanding of Speech), developed in Speech Informatics Group, is applied. SIRIUS had already used successfully for automatic speech recognition in several multimodal applications [4]. This automatic speech recognition system is mainly intended for recognition of Russian speech and contains several original approaches for processing of Russian speech and language, in particular, the morphemic level of the representation of Russian speech and language.

The list of voice commands for ICANDO contains 40 voice commands for PC control (for instance, “print” or “save”). These commands are similar to keyboard shortcuts. The control of cursor is provided by head (nose) movement. The cursor coordinates are tied with the position of tip of nose and any change of head position produces the cursor movement. At that the recognized speech command are fulfilled taking into account the current cursor position. At allows operating of GUI of the operational system Windows and peripheral devices. In last version of ICANDO system the software method for tracking operator’s head (tip of nose) motions was realized. The system uses standard web-camera, which provides video and audio signal with acceptable quality. It simplifies the usage of the system, since no any additional hardware, like microphone, helmet or reference device unit, is required.

The special approach was developed for control of the mouse cursor, which is able to work in real-time mode. It includes two stages of functioning: calibration and tracking. At first short starting stage the position of face in the video is defined. It is realized by the software module which uses the Haar based object detector to find rectangular regions in the given image that likely can contain face of a human. This region should not be less than 250 per 250 points that allows accelerating video processing. Then taking into account the standard proportions of a human’s face the approximate position of nose is marked by blue point on the image. During several seconds of calibration process a user should combine the tip of his nose with the position of this blue point. Then this point is captured by the system and the tracking algorithm is started. This algorithm uses the iterative Lucas and Kanade technique for optical flow, which is an apparent motion of image brightness. Sometimes the algorithm loses the position of human’s nose that is caused by the lack of light or very quick movements of user’s head. To solve this problem the special voice command “Calibration” was introduced in the system, which runs the process of calibration described above. Thus, the tip of nose defines the position of mouse cursor on the desktop of operating system Windows.

Real work of the multimodal system for hands-free computer control based on speech recognition and head tracking was shown in the main Russian TV channel ("First channel") in the news program ("Vremja") on 6 September 2005. During the demonstration the impaired person successfully worked with a personal computer by ICanDo system (http://www.1tv.ru/owa/win/ort6_main.main?p_news_title_id=82825&p_news_razdel_id=4).

4. Audio-visual speech recognition model

To design a natural and flexible multimodal system it is necessary to fuse different natural modalities. The bimodal audio-visual speech recognition system will use two kinds of information audio and visual (lips movement while pronouncing the speech). For this aim we will develop the techniques for audio-visual speech processing and Russian audio-visual corpus.

The image preprocessing data consists of the static data and dynamic data. First, for all speakers the face is found. Then the region containing only skin colored parts of the face is extracted and stored. The mean value of the skin color of the face is stored separately in a text file. Also, the regions for eyes and nose are extracted and stored in separate image files.

Then each video file is processed as a whole. The head is found using firstly converting the image into CR/CB space and then thresholding. The face region is detected as a convex hull of the thresholding result. At this moment we know the position and roughly the orientation of a head. We can find the region of the eyes and the mouth. The algorithm detecting the position of lips is similar to the one used in [5].

The algorithm continues as follows. In first frame the eyes are found and their distance is measured. The size of the region of interest of lips is then set according to this distance. For each subsequent frame then the center of lips and eyes are measured and used for detecting the region of interest, as shown in Figure 2. Using this information the region of interest can be easily extracted from each video frame during processing. This preprocessing stage was applied for Czech and Russian speech databases.

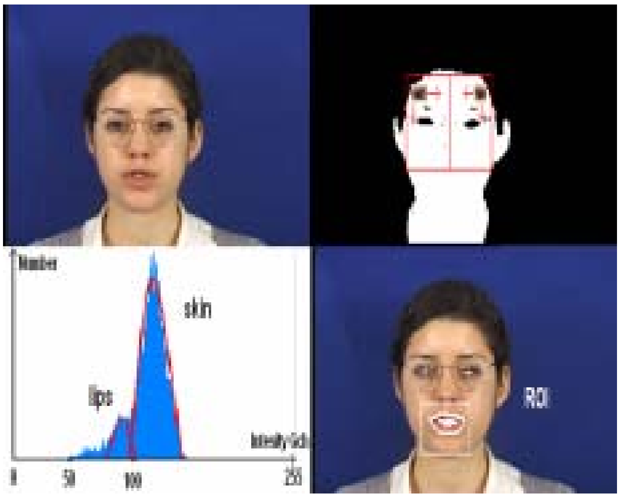

Figure 2: Localization of head, eyes and lips

Shape-based geometric parameterization is based on a description of the shape of lips and the positions of tongue and teeth. Therefore we have to detect outer and inner lip contours for each frame. The process of lip-tracking should be fast because we have to process as many frames per second as possible. It is used simple and fast thresholding method to get a binary image of lips. We work just with the ROI. The position of ROI was estimated during pre-processing of the corpus. The ROI is represented in chromatic colors. We use just the part Gch (green). R, G, B are parts of RBG color space.

We use chromatic color to avoid the influence of illumination. The threshold is estimated by an automatic clustering method based on GMM algorithm. We tried to get the threshold by the analysis of histogram but the method worked only for good conditions. In the ROI there are two main objects (mouth and skin). Therefore the clustering algorithm decides the pixels of ROI to two clusters and gives as a variance and a mean value of these objects.

The multi-stream model of speech recognition was applied for AVSR. This model belongs to the class of state synchronous decision fusion. The multi-stream model was realized by Hidden Markov Model Toolkit. This toolkit allows building and manipulating both single-stream and multi-stream Hidden Markov Models. The main difference between single-stream HMMs, which are used for speech recognition mainly, and multi-stream HMMs consists in diverse calculation of the probability of audio-visual observation vector $o^{(t)}$ in a state *c* of a multi-stream HMM. This probability can be calculated as follows:

$$P[o^{(t)} \mid c] = \prod_{s \in \{A,V\}} [\sum_{j=1}^{J_{sc}} w_{scj} N(o^{(t)}, m_{scj}, v_{scj})]^{\lambda_{sct}} \quad . \qquad (1)$$

Here λ_{sct} is the positive stream exponent which depends on the type of the modality s, HMM's state c and frame of the speech t. These modality weights are global and constant over the entire speech database. J_{sc} is the number of mixture components in the stream, w_{scj} is the weight of the j-th component and $N(o^{(t)}, m_{scj}, v_{scj})$ is a multivariate Gaussian with mean vector m and covariance matrix v that equals:

$$N(o^{(t)}, m_{scj}, v_{scj}) = \frac{1}{\sqrt{(2\pi)^n \mid v \mid}} e^{-\frac{1}{2}(O-m)^T v^{-1} (O-m)} \quad , \qquad (2)$$

where n is the dimensionality of the feature vector O. During the training process the modality weights are tuned manually by minimizing the WER of audio-visual recognition model. All other parameters of HMMs are re-estimated by Baum-Welch procedure.

For parameterization of audio signal 12 MFCC features are used and geometrical shape-based features for video signal. The results of experiments (WER) are presented in the Table 2 [6]. The size of vocabulary in this task is 102 words and the table shows the accuracy of word recognition. It can be seen from the table that the AVSR shows the better results than audio only speech recognizer. The modality weights were manually adjusted for maximal WER and the weight for video stream was 0.2 and the weight for audio stream is 1.8. The signal-to-noise ratio (SNR) for audio signal was 10 db for all the experiments (clean speech).

Table 2: Results of experiments with Russian AVSR

	Audio features	Audio + shape-based features
Accuracy rate	90.1 %	92.3 %

The developed model and audio-visual corpus will be used for creation of special assistive technologies for the deaf-and-dumb people to help them in interaction with other people or transform the current information into the accessible form. Also the model is applicable for design of intellectual applications such as smart phones, tourist information systems, car control, video conference communication etc.

5. Development of flexible architecture for joint multimodal interface

The described models will be applied in integrated multimodal architecture for smart assistance to impaired persons. The main objective of the research is to apply and improve already created models into the united assisting system (Fig. 3). The adaptation module and the interface shell will provide the addition of new modalities during human-computer interaction. In future this model will be used for designing the prototypes of new applications for disabled people and tested by end-users.

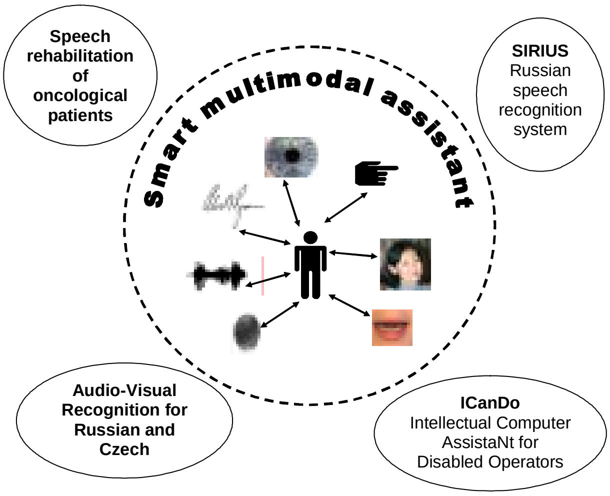
Speech rehabilitation of oncological patients
SIRIUS Russian speech recognition system
Smart multimodal assistant
Audio-Visual Recognition for Russian and Czech
ICanDo Intellectual Computer AssistaNt for Disabled Operators

Figure 3: Smart multimodal assistant for disabled

Needs of users at the interaction with domestic technique and computers will be studied at first. It will allow to estimate the requirements of disabled people and design the prototypes adapted to users. The possible ways for replacement of lacking modalities will be studied for the groups of users with special needs. It is necessary to provide the convenient and effective approaches for choosing accessible modalities for input and output of information. To fusion and fission of input/output modalities it is necessary to create the mathematical models and algorithms.

Investigation of human-computer interaction requires participation of the specialists from the various scientific areas. Especially it is required for introduction of the model into assistive systems. Multidisciplinary character of the research and attraction of people with special needs for testing will allow development of more effective tools to satisfy the necessities of users.

6. Conclusions

This paper presents the results of the cooperative work of 4 scientific organizations: IAS TUD, UWB, TSUCSR, SPIIRAS. The developed models of audio-video processing are intended to improve the life of disabled people. These models solve the problem of computer control by people with arm disabilities; speech rehabilitation of oncological patients after operation of full removal of a throat; interaction the deaf-and-dumb people with other people and transformation of the current information into the accessible form. It is proposed to use the accumulated experience and the tools for creation of the united flexible model for multimodal assistance for disabled people. Such model will allow to choose accessible modalities for interaction and design various applications for special groups of users with special needs.

7. Acknowledgements

This research is supported by the European Community in framework of the SIMILAR NoE “The European taskforce creating human-machine interfaces SIMILAR to human-human communication”, project No. 507609, INTAS Project № 04-77-7404 and the Grant Agency of the Academy of Sciences of the Czech Republic, project No. 1ET101470416.

8. References

[1] Hoffmann, R., Shpilewsky, E., Lobanov, B., Ronzhin, A. Development of multi-voice and multi-language text-to-speech (TTS) and speech-to-text (STT) conversation system (languages: Belorussian, Polish, Russian), 9-th International Conference SPECOM-2004, St. Petersburg: “Anatoliya”, 2004, pp. 657-661.

[2] Bondarenko, V. P., Balatskaya L. N., Choinzonov E. L. “Application of biological feedback in system for reabilitation after full glottal resection”. Sybirian oncology magazine. 2004, Vol. 4(12), pp.17-20 (in Russian).

[3] Ronzhin, A.L., Karpov, A.A. “Assistive multimodal system based on speech recognition and head tracking”. In Proc. of 13-th European Signal Processing Conference EUSIPCO-2005, September, 2005, Antalya, Turkey.

[4] Karpov, A. A, Ronzhin, A. L, Li, I.V., “SIRIUS: A System for Speaker-Independent Recognition of Continuous Russian Speech”. In TRTU Proceedings, № 10, 2005, pp. 44-53 (in Russian).

[5] Císař, P., Železný, M., Krňoul, Z., Kanis, J., Zelinka, J., Müller, L. “Design and recording of Czech speech corpus for audio-visual continuous speech recognition”, In Proc. of the AVSP-2005. Causal Productions, Adelaide, Australia, 2005.

[6] Císař, P., Zelinka, J., Železný M., Karpov A., Ronzhin, A. “Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian)”, In Proc. of 11-th International Conference SPECOM-2006, St. Petersburg: “Anatoliya”, 2006.