The Fifth International Conference on Neural Networks and Artificial Intelligence
May 27-30, Minsk, Belarus

Audio-Visual Russian Speech Recognition and Synthesis for a Multimodal Information Kiosk

Alexey Karpov 1, Boris Lobanov 2, Andrey Ronzhin 1, Liliya Tsirulnik 2
1) St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
199178, SPIIRAS, 39, 14-th line, St. Petersburg, Russia
{karpov, ronzhin}@iias.spb.su, http://www.spiiras.nw.ru/speech
2) United Institute of Informatics Problems, National Academy of Sciences, Surganov str., 6, 220012, Minsk, Belarus
{lobanov, l.tsirulnik}@newman.bas-net.by

Abstract *- This paper describes the process of development of the models of audio-visual speech recognition and audio-visual speech synthesis (talking head) which are the key components of the multimodal information kiosk. Audio-visual speech recognition of Russian was implemented as a state synchronous decision fusion model and compared with an audio-based speech recognizer by the parameter of phrase recognition rate. The results obtained have demonstrated the importance of the visual information for automatic speech recognition. The paper presents also the technology of creation of natural personalized talking heads for realization of the avatars in smart kiosks.*

Keywords *– multimodal interfaces, automatic speech recognition, talking head, intelligent information kiosk*

I. INTRODUCTION

Information queuing systems such as automatic teller machines or enquiry automatons have increasingly being embedded in many application domains. At present there exist variety of information kiosks equipped with touchscreens, they can work in the 24x7 mode and often replace human-staff.

However in order to communicate effectively with these devices new ergonomic and natural means of human-computer interaction are needed. Contemporary computer systems have to provide a user with intelligent control systems, their functions must be easily accessible in intuitive and clear manner. Clients of the kiosks should use these devices without knowledge of principles of their work as well as without necessity to learn any special commands or operations. It is important also that intelligent control systems could operate without own errors and be robust to mistakes in users' actions.

Last decade to solve the problem of organization of an effective human-computer interaction the developers and scientists investigate new models based on simultaneous usage of several information input channels (modalities) that are natural for human being such as speech, gestures by hands and body, gaze, facial expressions, lip movements, etc. Such multimodal interfaces combine several recognition technologies in one system and provide natural interaction of a human being with diverse automated computer devices in a manner similar to human-human communication.

At present both in Europe and in USA automatic information kiosks with multimodal interfaces are actively studied. Such devices are called "intelligent multimodal kiosks" or "smart kiosks". These are information enquiry automatons which can automatically detect presence of a user in a working area and verbally communicate with clients. An information input can be organized simultaneously by touchscreen, keyboard as well as by voice and manual gestures. Among successful researches of multimodal kiosks the following systems should be mentioned [1,2]: Touch'n'Speak system developed by the Tampere University (Finland); Memphis Intelligent Kiosk Initiative (MIKI) from the Memphis University (USA); French system Multimodal-Multimedia Automated Service Kiosk (MASK); Multimodal Access To City Help Kiosk (MATCHKiosk) produced by AT&T company.

In SPIIRAS in order to study the peculiarities of multimodal human-machine interaction a prototype of an intelligent information automaton with natural user-friendly interface was developed. This computer device is able to detect a user in front of the kiosk as well as communicate with detected client by voice.

The main attention in the given paper is paid to the models of audio-visual recognition and synthesis of continuous Russian speech for the multimodal kiosk, which are studied and developed by SPIIRAS and UIIP teams in collaboration since 2007.

II. ARCHITECTURE OF THE MULTIMODAL KIOSK

The general architecture of software-hardware complex of the proposed information kiosk is presented in Fig. 1. The kiosk contains several hardware components and software technologies which have to be synchronized. The main of these modules are [3]: (1) video processing using a technology of computer vision in order to detect position of human's body, face and some facial organs; (2) speaker-independent system of audio-visual recognition of continuous Russian speech that uses a microphone array to eliminate acoustical

noises and to localize a source of a useful voice signal at distant speech recording, as well as a video camera and a computer vision technology for visual recognition of speech by lip reading additionally to audio speech recognition in order to increase the recognition accuracy and robustness; (3) a module for Russian speech synthesis to be applied for realization of a virtual character – avatar; (4) an interactive graphical user interface based on a touchscreen; (5) a model of dialogue and a dialogue manager that include a database of an applied domain and a system for dialogue strategy control.

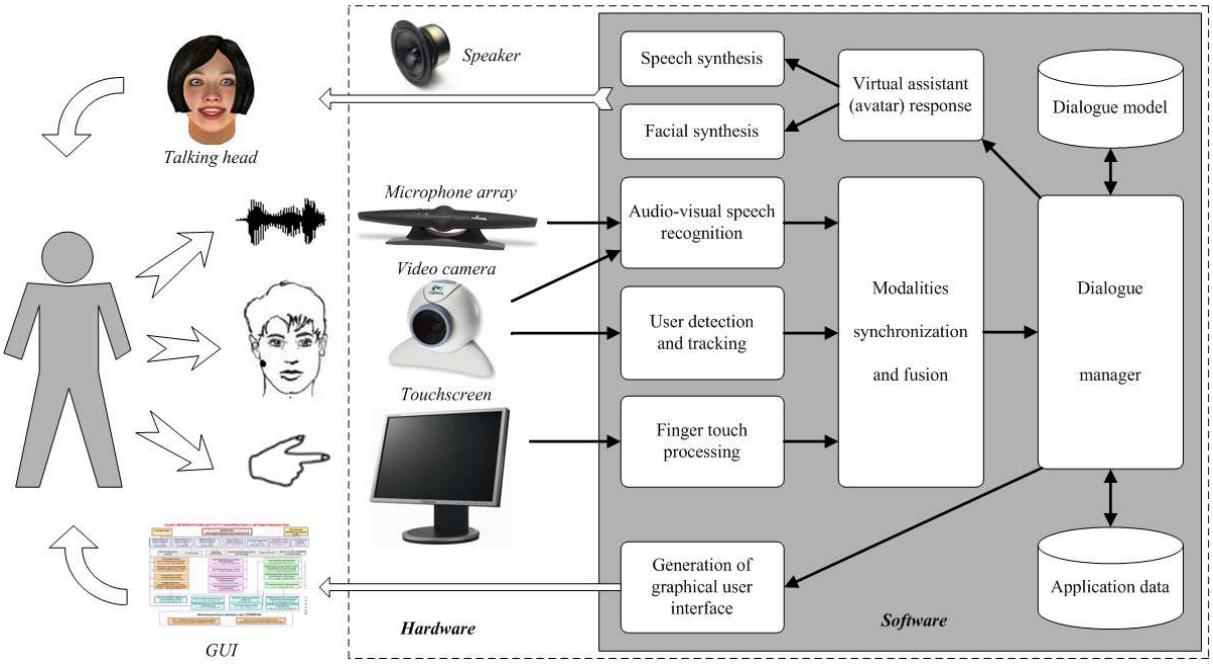

Fig. 1. General architecture of the multimodal information kiosk

For debugging of the hardware and software parts of the multimodal kiosk an application domain was chosen which has quite small amount of information data. The prototype of the intelligent automaton contains the data about the structure and staff of laboratories, contacts and phone numbers of departments, current events of the SPIIRAS as well as the navigation information helpful both for visitors and workers of the institute. In order to access to the interactive diagram of SPIIRAS structure a user may use either touchscreen or voice queries.

A subsystem of dialogue manager controls the kiosk functionality, generates its behavior and replies users' requests by graphics and synthesized voice. The kiosk has been designed to operate in two modes: an entertainment mode (attraction of clients) and an interaction mode. When there are not any users in operation area in front of the kiosk, it works in the entertainment mode and demonstrates continuously presentations and videos about the institute. If a human being shows an interest to the kiosk it switches to the interaction mode. In this mode an avatar welcomes clients and the start page with the interactive diagram is displayed on the screen after that the client can make queries to search an information needed. Finally when a user has completed his/her work with the terminal it says goodbye and switches back to the entertainment mode.

III. AUDIO-VISUAL SPEECH SYNTHESIS

A. Talking head in intelligent kiosks

Along with the interactive graphical user interface one of the main components of the information output subsystem of the multimodal kiosk is an animated virtual character (avatar). Avatar is a 3D model of human's face that is able to move eyes, mouth and facial muscles. The model of avatar can speak, synchronizing movements of its mouth, lips and teeth with synthesized speech. The synchronization of lips motion with human's speech creates an expressive illusion of "an alive talking head". Now diverse talking heads are available for the most of the European languages [4], however up to date there are not such systems for the Russian language.

Avatar in an information kiosk has two main purposes. At first it serves for attraction of the clients, analyzing the information about user detection and tracking process from the computer vision subsystem. The kiosk tracks location and position of potential users and the avatar is able to turn to this direction and look at coming visitor that attracts clients to the kiosk. When a client comes rather close to the automaton, the avatar welcomes him/her. The second purpose of the avatar is assistance to

clients. It instructs about operation with the terminal, as well as answers user's requests and gives useful information to the client in a verbal form.

There exist two common techniques for designing of avatar animation. Firstly a parametrical approach, when 2D or 3D face model is created using advantages of MPEG-4 format, can be applied [5]. Face parameters are tuned to express facial mimics, movements of lips and articulation organs. The second way is a compilation approach where the talking head is made by search, selection and concatenation of video frames from the visual database [6]. When using the first technique the volume of the database needed for visual speech synthesis is significantly lower rather than in the second approach. Besides it does not require collection of a video corpus and its tedious labeling using visemes (images of the face and shapes of the mouth while pronouncing diverse phonemes). Nevertheless the first way has a disadvantage concerning with inevitable sketchiness of movements of the facial organs that can not allow to use this approach for natural personalized audio-visual text-to-speech synthesis (TTS). Another disadvantage of this method is high computational complexity of its realization. Thus compilation audio-visual TTS is considered as more adequate for development of natural personalized talking heads.

B. Audio-visual Russian speech synthesis

As it is known, phonemes in the speech flow are realized in the form of allophones, or otherwise, in the form of positional and combinatory variants of phonemes. Experience of creating Russian-speaking TTS has shown, that synthesized speech of sufficiently good quality can be reached under conditions of generating the necessary set of positional and combinatory allophones.

For a set of allophones of vowels 3 types of positional allophones are created: stressed - (0), pre-stressed - (1) and post-stressed - (2). To present the left context the following combinatory allophones of vowels are created: after a phrase pause - (0), after hard consonants labial - (1), alveolar - (2), velar - (3), and after soft consonants - (4). Totally there are 5 types of left contexts. For the right context there are the following combinatory allophones of vowels: before a phrase pause - (0), before hard consonants labial - (1), alveolar and velar - (2), and before soft consonants - (3). For 6 Russian vowels /*a, e, i, o, u, y*/ we have $Nv = 3*5*4*6 = 360$ allophones. Allophones of consonants are created only with regard to the right context: before a pause - (0), before unvoiced - (1) and voiced - (2) consonants, before unstressed- (3) and stressed - (4) vowels. In total for 36 Russian consonants we obtain $Nc = 5*36 = 180$ allophones. Overall, in a set of Russian phonemes we have $360+180 = 540$ allophones. For the designation of allophones the symbols of corresponding phonemes (in Latin letters) with 3 digital indexes are used, where the first index designates the positional type of a phoneme, the second index is the type of the left context, and the third index is the right context. For example, A_{023} is an allophone of phoneme /a/ in stressed position (0), after a hard alveolar consonant (2) and before a soft consonant (3).

At first the sets of allophonic Natural Speech Waves (NSW) and Natural Face Movements (NFM) for audio-visual TTS system should be created. The process of creation of speaker dependent DBs of NSWs and NFMs includes the following operations:

- formation of the phonetically representative text corpus;
- audio-visual recordings of several speakers, corresponding to the text corpus;
- processing of the created audio-visual recordings including phonemic segmentation, allophonic marking of audio-visual segments and preservation of the obtained sets in NSWs and NFMs DBs.

The text corpus is created on the basis of especially selected mini-set of words that covers all allophones. Automatic creation of the NSW set is realized by data driven voice "cloning" technology [7]. The set of NFMs for audio-visual TTS synthesis is created manually. The general structure of the multi-voice and multi-face audio-visual TTS-synthesizer is shown in Fig. 2.

The incoming orthographic text undergoes a number of successive operations carried out with the help of specialized processors: textual, phonemic, prosodic, acoustical and visual.

The textual processor is devised to transform the incoming orthographic text into a prosodically marked text. The processor carries out the following tasks:

- dividing an orthographic text into utterances;
- transforming numbers, abbreviations, etc;
- dividing an utterance into phrases;
- placing word's accents (weak and strong);
- dividing phrases into accentual units (AU);
- marking the intonation type of the phrases.

The prosodically marked text is then sent to the phonemic processor that performs the following tasks:

- phonemic transcription of the text;
- transforming the phonemic text into allophonic.

The prosodic parameters processor performs the following tasks:

- splitting AU into the elements of accentual units (EAU): pre-nuclear, nuclear and post-nuclear parts;
- generating the fundamental frequency (F_0) contour as well as the amplitude (A) and phoneme duration (T) assigning values according to the accent unit portrait for each accent unit.

The acoustical processor uses the information incoming from the phonemic and prosodic processors in order to provide:

- prosodic parameters modification of NSWs;

- concatenation of NSWs to the appropriate sequence.

The visual processor also uses the information incoming from the phonemic processor to concatenate NFMs in the appropriate sequence of face movements.

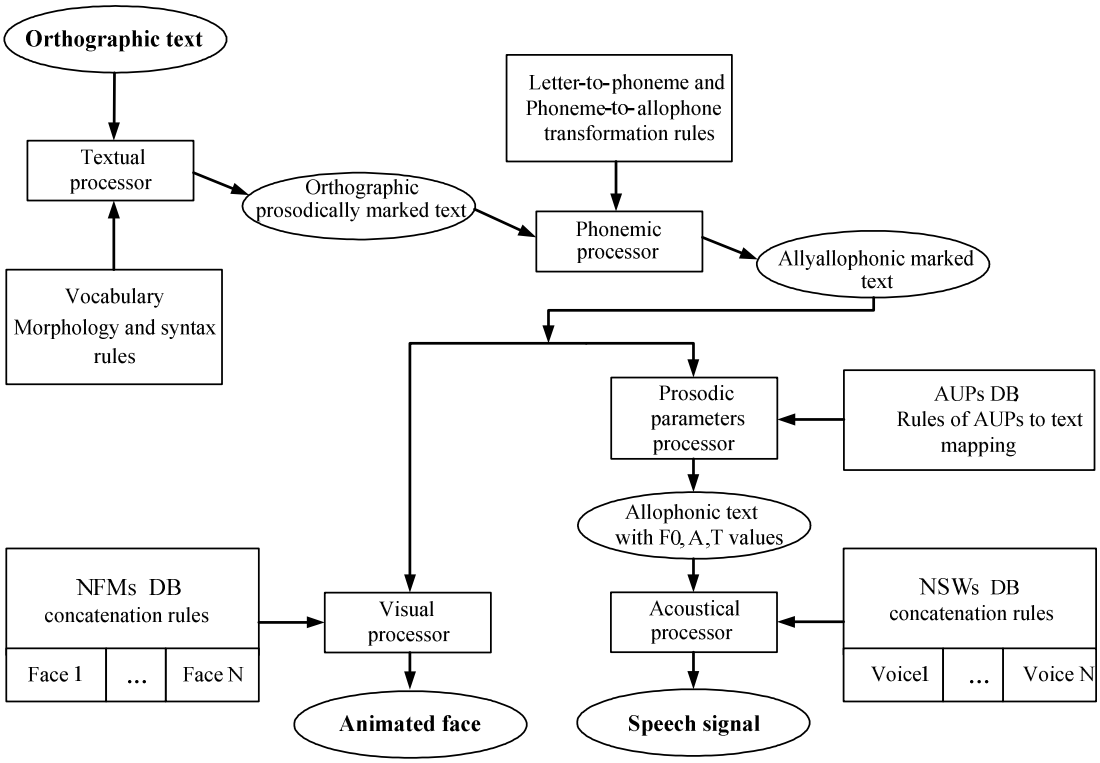

Fig. 2. General structure of the multi-voice and multi-face audio-visual TTS-synthesizer

IV. AUDIO-VISUAL SPEECH RECOGNITION

The core component of the multimodal kiosk is the speech recognition module based on joint information from audio waves of an utterance and visual information of lips movements that are made while uttering any words by a human being. Audio and visual modalities can supplement each other under diverse conditions, that improves quality of audio-visual speech recognition (AVSR). In conditions with low level of audio signal, noisy environments or unwanted sounds the standard audio-based speech recognition systems can not provide the required accuracy. As opposed to this under poor illumination environments the audio information can fill the gap of visual information. Thus it is clear to increase robustness of speech recognition process it is necessary to use visual information in addition to audio signal.

A. Collection of bimodal audio-visual database

A bimodal corpus of speech is required for the purpose of training of the speech recognition system. The database should consist of both visual and acoustic parts, and although we could record both data streams by a single device, either acoustic or visual part would not be recorded in a sufficient quality. To obtain highest possible quality of the recordings, it is best to record the two streams by different devices. Since the recording of the parts is made by different means, the data streams have to be synchronized. Especial software aimed for audio and video recordings has been developed using Direct Show facilities. Then it was required to choose between several types of video cameras with respect to both availability and compatibility with the device that is supposed to be used in the target application. Thus the visual part of the corpus has been recorded by a good quality consumer market digital camcorder (MiniDV camera Sony DCR-PC1000E). Using good quality recordings we can easily perform experiments on lower quality of video which can be obtained by various processes of degradation, such as blurring, noise contamination, and color distortion, reduction in spatial and temporal resolution. While recording the visual data were stored on a medium (MiniDV tape) of the camcorder and acquired later offline using the standard IEEE1394 (FireWire) interface.

The parameters of the visual part of the corpus are: frontal view, portrait orientation, constant illumination, image resolution of 720x576 pixels, 25 frames per second, Sony DV codec. The visual data were stored in video files (.avi format) with the digital video codec. Using this codec, the visual data for one speaker occupy approximately 4 GB of disk space. All speakers are asked not to move with the head during the recording process,

thus head position can be considered as static in all recordings. A uniform color background is used for easy segmentation of the head using a technique such as chromakey (Fig. 3).

Fig. 3. A fragment of bimodal corpus of speech

The acoustical part of the corpus was independently recorded using a dedicated computer equipped with high quality digital signal processing sound card. The parameters of the acoustical data are: headset microphone Sony DR-50, PCM audio format, 16 kHz sampling rate, 16 bits per sample, mono signal.

Ten native Russian speakers (both male and female) were recorded in the office conditions. The average age of the speakers is 20 years old. Each speaker had to read 200 phrases (voice queries) with the length up to 5 words. Recording time for a speaker was approximately 15 minutes. This database was divided into two parts: 80 percents of utterances of each speaker were used for the training purpose and other rest of the data for the model testing [8].

B. PCA pixel-based parameterization

The principal component analysis (PCA) was applied for extraction of features of the visual part of the bimodal corpus. In this parameterization method the obtained images of mouth region (ROI) are normalized to 32×32 in size, the gray level pixels in the mouth region are mapped into a 32-dimensional feature vector using PCA [9]. The PCA projection was computed from a set of two thousands of training mouth region images from the training database.

For calculation of PCA projection the following data were used: $U = (u^1, u^2, ..., u^N)^T$ is a *N*-dimensional vector containing the pixels of an image with the size $W \times H$. Having a set of *M* input vectors $\{U_1, U_2, ..., U_m\}$, the mean vector μ and the covariance matrix *C* are defined as:

$$\mu = \frac{1}{M}\sum_{k=1}^{m} U^k \quad (1)$$

$$C = \frac{1}{M}\sum_{k=1}^{M}(U_k - \mu)(U_k - \mu)^T = \frac{1}{M}\sum_{k=1}^{M} U_k \cdot U_k{}^T - \mu\mu^T \quad (2)$$

Also the sum vector $\tilde{\mu}$ and the partial covariance matrix $\tilde{C}$ are calculated according to the formulas:

$$\tilde{\mu} = \sum_{k=1}^{M} U^k \quad (3)$$

$$\tilde{C} = \sum_{k=1}^{m} U_k U_k{}^T \quad (4)$$

The first *p* largest eigenvalues and the corresponding eigenvectors of PCA are $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_p\}$ and $V = (V_1, V_2, ..., V_p)^T$, hence the projection of input vector *u* in the p-dimensional subspace $Y = (y_1, y_2, ..., y_p)^T$ is calculated by the following formula:

$$Y = V(U - \mu) \quad (5)$$

The result vector is normalized using eigenvalues:

$$\hat{y}_i = \frac{y_i}{\sqrt{\lambda_i}} \quad (6)$$

The final PCA feature vector is:

$$\hat{Y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_p)^T \quad (7)$$

These visual features were applied for AVSR of the Russian language and the results of experiments are presented below.

C. Multi-stream recognition model

The multi-stream model of speech recognition was applied for AVSR of Russian [10]. This model belongs to the class of state synchronous decision fusion models [11]. The main difference between single-stream HMMs, which are used for audio-based speech recognition mainly, and multi-stream HMMs consists in different calculation of the probability of an audio-visual observation vector $o^{(t)}$ in a state *c* of a multi-stream HMM. This probability is calculated by the following formula:

$$P[o^{(t)} \mid c] = \prod_{s \in \{A,V\}} [\sum_{j=1}^{J_{sc}} w_{scj} N(o^{(t)}, m_{scj}, v_{scj})]^{\lambda_{sct}} \quad (8)$$

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 weights of modalities 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:

$$N(o^{(t)}, m_{scj}, v_{scj}) = \frac{1}{\sqrt{(2\pi)^n \mid v \mid}} e^{-\frac{1}{2}(O-m)^T v^{-1}(O-m)} \qquad (9)$$

where *n* is the dimensionality of the feature vector *O*.

During the training process the modality weights are tuned manually by minimizing the word error rates of speech recognition. All other parameters of HMMs are re-estimated by the standard Baum-Welch procedure.

D. *Experiments with AVSR of Russian*

The multi-stream model was adopted for audio-visual recognition of continuous Russian speech. For parameterization of the audio signal 12 Mel-frequency cepstral coefficients (MFCC) were used and pixel-based PCA features were applied for the video signal. Table 2 presents the list of Russian visemes and viseme-to-phoneme mapping of AVSR [12].

TABLE I
THE LIST OF RUSSIAN VISEMES FOR AVSR

Visemes	Russian phonemes
s	silence
a	а, а!, е, е!
i	и, и!, ы, ы!
o	о!, у, у!
v	ф, ф', в, в'
z	з, з', с, с', ц, ч
p	м, м', б, п, п'
t	т, т', д, д', н, н', к, к', г, г'
l	л, л', р, р'
j	ж, щ, х, ш, й

The results of the experiments (phrases recognition rate) are presented in Table 2. The size of vocabulary of this application domain is above 100 words and the table shows the accuracy of phrases recognition by audio-based speech recognizer and by bimodal speech recognition models. The modality weights were manually adjusted for minimizing WER and the weight for the video stream was 0.2 and the weight for audio stream equals 1.8. The signal-to-noise ratio (SNR) for the audio signal in the experiments was about 20 db. It can be seen from the table that AVSR model outperforms audio speech recognizer that confirms a hypothesis of effectiveness of visual information for recognition and understanding of speech utterances. In further research the experiments with diverse SNRs will be made.

TABLE II
THE RESULTS OF EXPERIMENTS WITH RUSSIAN AVSR

	Audio-based Russian speech recognition	Audio-visual Russian speech recognition
Phrase recognition rate	90.1 %	92.3 %

V. CONCLUSIONS

Application of the speech and multimodal recognition technologies for controlling the computer systems as well as usage of the virtual assistants allow to design effective, natural and ergonomic interfaces, where in the human-computer communication a human being plays the role of master. Multimodal input and output interfaces allow also people with special needs (for instance, hard of hearing people) to use computer without any restrictions. Such natural interfaces break psychological barriers of unexperienced persons (for instance, elderly people) in their interaction with computer systems.

The proposed multimodal information kiosk will serve as a prototype for manufacturing the devices that will be the kernel of diverse information enquiry systems for automatic services located in airports, hotels, museums, shops, business centers.

ACKNOWLEDGMENT

This work has been supported by the Russian Foundation for Basic Research (project № 07-07-00073) as well as by the Russian Science Support Foundation.

REFERENCES

[1] D. Andrew, C.L. Avery, and B.L. Avery, “Digital smart kiosk project,” in Proc. of the SIGCHI conference on human factors in computing systems, pp. 155-162, 1998.
[2] L. McCauley, and S. D’Mello, “MIKI: a speech enabled intelligent kiosk,” Intelligent virtual agents. Lecture Notes in Computer Science, vol. 4133, pp. 132-144, 2006.
[3] A. Ronzhin, A. Karpov, An. Leontyeva, and B. Kostuchenko, “Development of the multimodal information kiosk,” SPIIRAS Proceedings, issue 5, vol. 1, pp. 227-245, 2007.
[4] Animated natural talking heads for the Italian language, http://www.pd.istc.cnr.it/LUCIA/home/default.htm
[5] A.M. Tekalp, and J. Ostermann, “Face and 2-D mesh animation in MPEG-4,” Signal Processing: Image Communication, Special Issue on MPEG-4, vol. 15, pp. 387-421, 2000.
[6] E. Cosatto, and H.P. Graf, “Photo-realistic talking-heads from image samples,” IEEE Transactions on Multimedia, vol. 2, pp. 152-163, 2000.
[7] L. Tsirulnik, “Automated System for Individual Phonetic-Acoustical Speech Peculiarities Cloning,” Informatics, issue 2, vol. 10, pp. 47-56, 2006.
[8] M. Železný, P. Císar, Z. Krnoul, A. Ronzhin, I. Li, and A. Karpov, “Design of Russian Audio-Visual Speech Corpus for Bimodal Speech Recognition,” in Proc. of the 10-th International Conference on Speech and Computer SPECOM’2005, pp. 397-400, 2005.
[9] Intel Technology of Audio-Visual Speech Recognition, www.intel.com/technology/computing/applications/avcsr.htm
[10] Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk
[11] C. Neti, G. Potamianos, J. Luettin, et al., “Audio-visual speech recognition,” Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, 2000.
[12] P. Cisar, J. Zelinka, M. Zelezny, A. Karpov, and A. Ronzhin, “Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian),” in Proc. of the 11-th International Conference on Speech and Computer SPECOM’2006, pp. 493-498, 2006.

View publication stats