Multimodal Human Computer Interaction with MIDAS Intelligent Infokiosk

Alexey Karpov, Andrey Ronzhin,
Irina Kipyatkova, Alexander Ronzhin
St. Petersburg Institute for Informatics and
Automation of RAS (SPIIRAS), Russia
e-mail: {karpov, ronzhin, kipyatkova}@iias.spb.su

Lale Akarun
Department of Computer Engineering,
Boğaziçi University
İstanbul, Turkey
e-mail: akarun@boun.edu.tr

Abstract—**In this paper, we present an intelligent information kiosk called MIDAS (Multimodal Interactive-Dialogue Automaton for Self-service), including its hardware and software architecture, stages of deployment of speech recognition and synthesis technologies. MIDAS uses the methodology Wizard of Oz (WOZ) that allows an expert to correct speech recognition results and control the dialogue flow. User statistics of the multimodal human computer interaction (HCI) have been analyzed for the operation of the kiosk in the automatic and automated modes. The infokiosk offers information about the structure and staff of laboratories, the location and phones of departments and employees of the institution. The multimodal user interface is provided with a touchscreen, natural speech input and head and manual gestures, both for ordinary and physically handicapped users.**

Keywords-multimodal user interfaces; human-computer interaction; automatic speech recognition; speech synthesis; artificial intelligence, infokiosk

I. INTRODUCTION

In recent years, smart information kiosks employing speech and multimodal user interfaces are being actively developed. The following systems can be mentioned as examples: Touch'n'Speak developed by Tampere University (Finland); Memphis Intelligent Kiosk Initiative (MIKI) [1] from Memphis University (USA); French system Multimodal-Multimedia Automated Service Kiosk (MASK); and Multimodal Access to City Help Kiosk (MATCHKiosk) [2] manufactured by AT&T company. In a smart kiosk, information can be input by a touchscreen/keyboard, by voice, or by manual or body gestures.

A prototype of an infokiosk called MIDAS has been developed and installed in the SPIIRAS hall. The infokiosk is able to detect a human being inside the working zone with the Haar-based object detector [3], to track his/her movements and demonstrate awareness by a 3D avatar, which tracks clients by rotation of her talking head [4]. The infokiosk realizes a mixed initiative dialogue strategy, starting from the system's initiative and giving initiative for the query to a user after a verbal welcome.

The architecture of MIDAS is presented in Figure 1. It contains a lot of hardware and software technologies, which work simultaneously and synchronously. Most important of these modules are: (1) video processing with two non-stereo video-cameras and a technology of computer vision in order to detect the human's position, face and some facial organs; (2) speaker-independent system of automatic recognition of continuous speech that uses an array of 4 microphones with the T-shape geometry to eliminate acoustical noise and to localize the source of a relevant acoustic signal for distant speech processing; (3) modules for audio-visual speech synthesis to be applied for a talking avatar; (4) an interactive graphical user interface with a touchscreen; (5) a dialogue and data manager that accesses an application database, generates multimodal output and synchronizes input modalities fusion and output modalities fission.

The kiosk has been designed for multimodal HCI by users with special needs as well. It includes a module for contactless control of the mouse cursor by head movements; which is helpful for hand-disabled people. It is based on a Lucas-Kanade feature tracker for optical flow and uses two video cameras. We are working also on the task of equipping the infokiosk with technologies for sign language analysis and synthesis; [5] extending the 3D talking head and automatic recognition of manual gestures of sign language [6] available to deaf people.

The proposed kiosk provides information about the structure and staff of laboratories, location and phones of departments and employees, current events as well as contact information needed both for visitors and employees of the institute. To access the interactive diagram of SPIIRAS structure, users may apply both touchscreen and/or voice requests.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.941

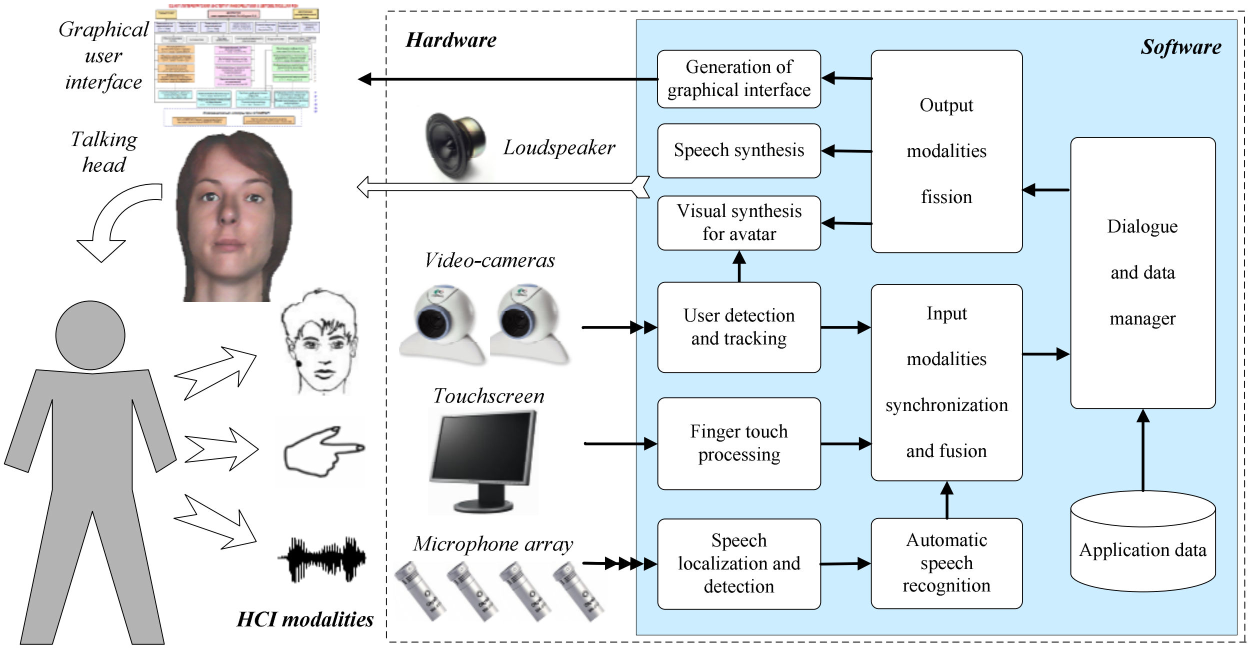
Graphical user interface
Talking head
Hardware
Software
Loudspeaker
Video-cameras
Touchscreen
Microphone array
HCI modalities
Generation of graphical interface
Speech synthesis
Visual synthesis for avatar
Output modalities fission
Dialogue and data manager
User detection and tracking
Finger touch processing
Input modalities synchronization and fusion
Speech localization and detection
Automatic speech recognition
Application data

Figure 1. Hardware/software architecture of MIDAS intelligent information kiosk.

II. Deployment of Speech Technologies

Deployment of computer technologies for automatic speech recognition (ASR) and synthesis in the multimodal kiosk was made in three stages (Figure 2), by applying the methodology Wizard of Oz [7]. At the first stage, a hidden human-operator handles clients' voice queries without any automated speech-to-text processing, and simulating real interaction between a user and the kiosk that is needed in order to attract potential clients to the system and to collect users' requests, phrases and dialogues. These data are required for training initial acoustical models of ASR and a baseline phonetic vocabulary as well as for creation of a grammar and a dialogue model for HCI. At the second stage, an intermediate version of ASR is employed and speech is processed in the automated mode while a human-corrector with a remote PC may edit the mistakes committed by ASR. When the editor is not available (for example, at night time), the system works in the automatic mode. At this stage, on-line and semi-automatic update of the phonetic vocabulary and grammar is made. It is the main phase for collecting speech data of real interaction in a real environment. Current speaker-independent ASR technologies are rather error-prone because of the variability of voices, noises, and ambiguous input. However, continuously collected speech corpus allows time-to-time retraining of acoustical and language models to obtain higher recognition accuracy. This step also allows developers to renew application data adding information, which is often asked by clients. The third stage is completely automatic, where the device uses prepared ASR, TTS and dialogue manager. Developers have to check the system's log once in a while to analyze interaction, recognition mistakes and uncompleted dialogues. On-line automatic adaptation of ASR to new speech data is performed that improves the word recognition rate.

SIRIUS ASR system [8] has been embedded into the automaton in order to recognize and interpret continuous Russian speech detected by the microphone array. The captured signal is digitized with the sampling rate of 16 kHz; then, samples are combined in the segments (100 segments per second) and Mel-frequency cepstral coefficients are extracted. Acoustical models are based on Hidden Markov Models (HMMs) with mixture Gaussian probability density functions. HMMs of triphones have 3 meaningful states and 2 additional states intended for concatenation of context-dependent phonemes (triphones) in the word models.

The input phrase syntax is described in a grammar that allows recognizing one voice message in each hypothesis. Fragment of the finite-state grammar (translation to English), used by ASR to understand users' queries concerning employees and departments of the institute, is shown in Figure 3. Some examples of acceptable questions are: “Where is a laboratory of Professor Boris Vladimirovich Sokolov?”, “Phone of Doctor Smirnov”, “Please, Ronzhin”, etc.

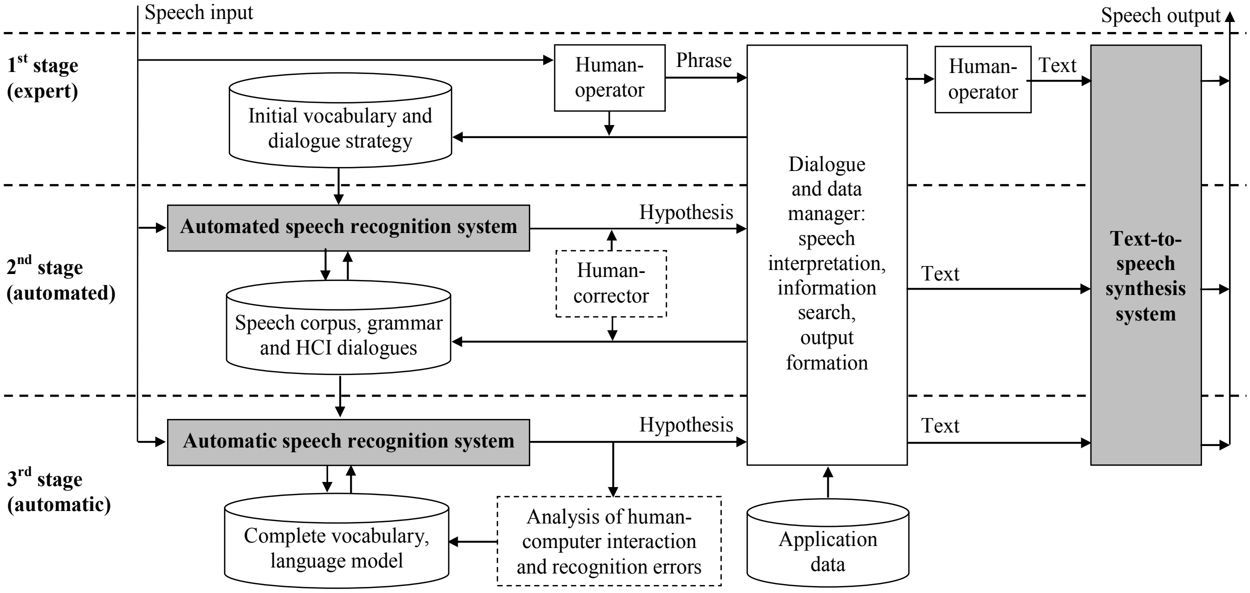
Speech input
Speech output
1st stage (expert)
2nd stage (automated)
3rd stage (automatic)
Human-operator
Phrase
Initial vocabulary and dialogue strategy
Human-operator
Text
Dialogue and data manager: speech interpretation, information search, output formation
Automated speech recognition system
Hypothesis
Human-corrector
Speech corpus, grammar and HCI dialogues
Text
Text-to-speech synthesis system
Automatic speech recognition system
Hypothesis
Text
Complete vocabulary, language model
Analysis of human-computer interaction and recognition errors
Application data

Figure 2. Stages of speech technologies deployment in the intelligent information automaton.

Totally there are more than 400 keywords in the recognition vocabulary (small-sized lexicon), including the list of permanent employees, names of departments (laboratories and groups) with possible rephrasings and abbreviations. To implement a robust ASR a garbage model (that models any out-of-vocabulary (OOV) items or a sound sequence) and a silence model (pause) were introduced in ASR to discard possible uninformative words and their acoustical models were trained by the collected speech data. Viterbi-based token passing method was realized for ASR and recognition results are presented as N-best lists of hypotheses with highest likelihoods.

III. An analysis of multimodal interaction

The paper presents results of two initial stages of deployment of speech technologies in MIDAS kiosk and an analysis of transmission from the 2nd to the 3rd stage. The infokiosk was designed during 2008, first stage of deployment with WOZ technology was started since early 2009 and the second stage was launched in June 2009. Since the end of 2009, it operates in the automatic mode with minor expert support. Statistics of HCI are based on analysis of system's logs, voice dialogues and a gallery of photos of users within the last five months of 2009 in the automated mode with WOZ (in the automatic mode in off hours). Altogether, there were 668 sessions of HCI with MIDAS (in 15 % of the sessions, a group of two or more people worked with the infokiosk). In the other 372 cases, the kiosk initiated communication by a welcoming phrase after detecting users inside the working zone; but clients did not keep the dialogue and went away without any actions. Users made 1202 meaningful voice queries and produced 652 non-informative messages or phrases addressed to other people. All users could easily observe video-cameras and microphones of the kiosk but they were not informed about on-line audio-visual recordings.

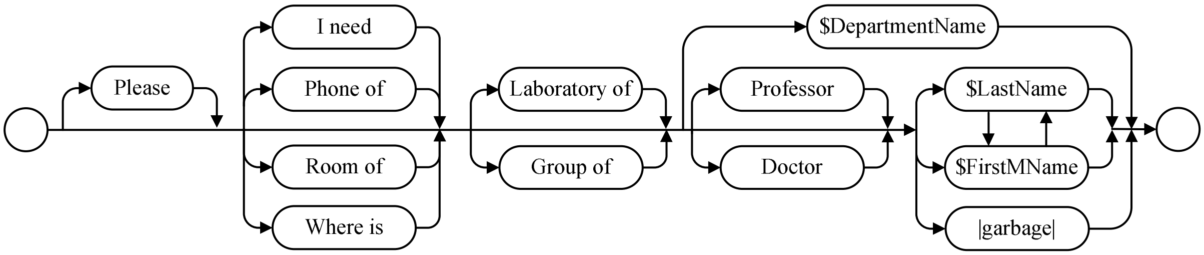

Figure 3. Grammar for the automatic speech recognition system in the multimodal infokiosk.

One month after installation, people actively started to use the kiosk. There were 266 sessions per month. During the next month, we had 108 sessions only due to this being a season of summer vacations, and in September, the number of clients has increased to 122, while in the next two months, 105 and 67 sessions were completed. So there is a tendency of a decrease in kiosk usage caused by the effect of habituation. In the starting phase, people often tried to play with the kiosk asking many different questions or even showing printed 2D face images to examine the face detection module. 84% of the sessions were initiated from 9 AM to 6 PM (working hours). But surprisingly, 10 sessions were made at night time by office-cleaners, who also tested the new device. Table 1 summarizes main statistics of multimodal HCI.

TABLE I. STATISTICS OF MULTIMODAL INTERACTION

Quantitative Indicator	Value
Avg. sessions per day	4.20
Unique users (based on face analysis)	145
Avg. speech inputs per session	2.77
Avg. informative voice queries / session	1.80
Sentence recognition rate by ASR	55%
Dialogue completion rate with WOZ	96%
Number of requests by the touchscreen	1018

It was discovered that the most frequent inquiries were related to information about "Director" and "Library", constituting almost 20% of all the voice questions. Often, users asked contacts of deputy directors and heads of departments. However, names of departments were rarely pronounced (less than 5%) Instead, users asked questions like "Sokolov's laboratory". It is quite difficult to differentiate user sessions that were caused by real needs and sessions with the aim to play with new technologies. For example, in some questions, users asked the names of D. Medvedev (Russian President) or V. Putin who have no relations to SPIIRAS, or other out-of-vocabulary (OOV) words and uninformative phrases. The percentage of the total number of OOV words pronounced was approximately 15%. They were misrecognized by ASR, but the human-corrector modified kiosk's answers in most of such cases. Besides, at many speech dialogues clients asked the surname "Ivanov", which is the most frequent Russian surname. It is clear that in some cases, users did not mean a real person working in the institute. After analysis of verbal interaction, information on a few new persons was added to the application database and the grammar was extended.

IV. CONCLUSIONS

The infokiosk is able to detect clients and support verbal HCI as described. It combines both the standard means for information input/output (touchscreen and loudspeaker) and the devices for contactless HCI (video-cameras and microphones). Usability results and the analysis of real exploitation have demonstrated that people prefer to communicate in a multimodal way; more than 60% of the user requests were made by natural speech and the rest by touch. In the automated mode, with WOZ speech dialogue completion, recognition rate was 96%. Afterwards, MIDAS is launched in the automatic mode, results of which will be reported later on.

ACKNOWLEDGMENTS

This work is supported by RFBR and TÜBİTAK foundations in the framework of the bilateral Russian-Turkish project (# 09-07-91220 / 108E113), and by the Grant of the President of Russia (# MK-64898.2010.8).

REFERENCES

[1] L. McCauley, and S. D'Mello, "MIKI: a speech enabled intelligent kiosk. Intelligent virtual agents," LNCS, Springer Verlag, vol. 4133, 2006, pp. 132–144.

[2] M. Johnston, and S. Bangalore, "MATCHkiosk: A Multimodal Interactive City Guide," Proc. Association of Computational Linguistics ACL-2004, Barcelona, Spain, 2004, pp. 223–226.

[3] R. Lienhart, and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection," Proc. IEEE International Conference on Image Processing ICIP-2002, 2002, pp. 900–903.

[4] A. Karpov, L. Tsirulnik, Z. Krnoul, A. Ronzhin, B. Lobanov, and M. Zelezny, "Audio-Visual Speech Asynchrony Modeling in a Talking Head," Proc. 10-th International Conference Interspeech-2009, Brighton, UK, 2009, pp. 2911–2914.

[5] M. Hruz, P. Campr, A. Karpov, P. Santemiz, O. Aran, and M. Zelezny, "Input and Output Modalities Used in a Sign-Language-Enabled Information Kiosk," Proc. 13-th International Conference on Speech and Computer SPECOM-2009, Saint-Petersburg, Russia, 2009, pp. 113–116.

[6] O. Aran, and L. Akarun, "A Multi-class Classification Strategy for Fisher Scores: Application to Signer Independent Sign Language Recognition," Pattern Recognition, vol. 43(5), 2010, pp. 1776-1788.

[7] N. Dahlback, A. Jonsson, and L. Ahrenberg, "Wizard of Oz Studies - Why and How," Knowledge Based Systems, Elsevier, vol. 6(4), 1993, pp. 258–266.

[8] A. Ronzhin, and A. Karpov, "Russian Voice Interface," Pattern Recognition and Image Analysis, MAIK, vol. 17(2), 2007, pp. 321–336.