J Multimodal User Interfaces (2008) 2: 105–116
DOI 10.1007/s12193-008-0012-2

ORIGINAL PAPER

Multimodal user interface for the communication of the disabled

Savvas Argyropoulos · Konstantinos Moustakas · Alexey A. Karpov · Oya Aran · Dimitrios Tzovaras · Thanos Tsakiris · Giovanna Varni · Byungjun Kwon

Received: 21 December 2007 / Accepted: 5 June 2008 / Published online: 15 July 2008

Abstract In this paper, a novel system is proposed to provide alternative tools and interfaces to blind and deaf-and-mute people and enable their communication and interaction with the computer. Several modules are developed to transform signals into other perceivable forms so that the transmitted message is conveyed despite one's disabilities. The proposed application integrates haptics, audio and visual output, computer vision, sign language analysis and synthesis, speech recognition and synthesis to provide an interactive environment where the blind and deaf-and-mute users can collaborate. All the involved technologies are integrated into a treasure hunting game application that is jointly played by the blind and deaf-and-mute user. The integration of the multimodal interfaces into a game application serves both as an entertainment and a pleasant education tool to the users.

Keywords Multimodal interfaces · Multimodal fusion · Sign language analysis · Gesture recognition · Sign synthesis

1 Introduction

Recent technological advances have improved communication between the disabled people. The emerging artificial intelligence techniques are starting to diminish the barriers for impaired people and change the way individuals with disabilities communicate. A common problem in communication between impaired individuals is that, in general, they do not have access to the same modalities and the perceived communicated message is limited by one's disabilities. A quite challenging task involves the transformation of a signal into another perceivable form to enable or enhance communication. Ideally, a recognition system should combine all incoming modalities of an individual, perform recognition of the transmitted message, and translate it into signals that are more easily understood by impaired individuals.

S. Argyropoulos (✉) · K. Moustakas
Electrical & Computer Engineering Dept., Aristotle University of Thessaloniki, Hellas, Greece
e-mail: savvas@ieee.org

K. Moustakas
e-mail: kmoustak@iti.gr

A.A. Karpov
Speech Informatics Group, SPIIRAS, Russian Academy of Sciences, St. Peterburg, Russia
e-mail: karpov@iias.spb.su

O. Aran
Perceptual Intelligence Lab, Bogazici University, Istanbul, Turkey
e-mail: aranoya@boun.edu.tr

D. Tzovaras · T. Tsakiris
Informatics and Telematics Institute, Centre for Research and Technology, Hellas, Greece

D. Tzovaras
e-mail: Dimitrios.Tzovaras@iti.gr

T. Tsakiris
e-mail: atsakir@iti.gr

G. Varni
InfoMus Lab—Casa Paganini, DIST-University of Genoa, Genoa, Italy
e-mail: giovanna@infomus.dist.unige.it

B. Kwon
Koninklijk Conservatorium, Hague, Netherlands
e-mail: byungjun@gmail.com

Recently, the desire for increased productivity, seamless interaction and immersion, e-inclusion of people with disabilities, along with the progress in fields such as multimedia and multimodal signal analysis and human-computer interaction, have turned multimodal interaction as a very active field of research [1, 2].

Multimodal interfaces are those encompassing more than the traditional keyboard and mouse. Natural input modes are put to use [3, 4], such as voice, gestures and body movement, haptic interaction, facial expressions [5], and more recently physiological signals. As described in [6], multimodal interfaces should follow several guiding principles: multiple modalities that operate in different spaces need to share a common interaction space and to be synchronized; multimodal interaction should be predictable and not unnecessarily complex, and should degrade gracefully, for instance by providing for modality switching; finally multimodal interfaces should adapt to user's needs, abilities, environment.

A key aspect in multimodal interfaces is also the integration of information from several different modalities in order to extract high-level information non-verbally conveyed by users. Such high-level information can be related to the expressive and emotional content that the user wants to communicate. In this framework, gesture has a relevant role as a primary non-verbal conveyor of expressive, emotional information. Research on gesture analysis, processing, and synthesis has received a growing interest from the scientific community in recent years and demonstrated its paramount importance for human machine interaction.

The present work aims to make the first step in the development of efficient tools and interfaces for the generation of an integrated platform for the intercommunication of blind and deaf-mute persons. It is obvious that while multimodal signal processing is essential in such applications, specific issues like modality replacement and enhancement should be addressed in detail.

In the blind user's terminal the major modality to perceive a virtual environment is haptics while audio input is provided as supplementary side information. Force feedback interfaces allow blind and visually impaired users to access not only two-dimensional graphic information, but also information presented in 3D virtual reality environments (VEs) [7]. The greatest potential benefits from virtual environments can be found in applications concerning areas such as education, training, and communication of general ideas and concepts [8–10]. Several research projects have been conducted to assist visually impaired to understand 3D objects, scientific data and mathematical functions, by using force feedback devices [11].

PHANToM is one of the most commonly used force feedback device; it is regarded as one of the best on the market. Due to its hardware design, only one point of contact at a time is supported. This is very different from the way that we usually interact with surroundings and thus, the amount of information that can be transmitted through this haptic channel at a given time is very limited. However, research has shown that this form of exploration, although time consuming, allows users to recognize simple 3D objects. The PHANToM device has the advantage to provide the sense of touch along with the feeling of force feedback at the fingertip. Another device that is often used in such cases is the CyberGrasp. It combines a data glove (CyberGlove) with an exosceletal structure to provide force feedback to each of the fingers of the user (5 Degrees of Freedom (DoF) force feedback, 1 DoF for each finger). In the context of the present work we used the PHANToM desktop device to enable haptic interaction of the blind user with the virtual environment.

Deaf and mute users have visual access to 3D virtual environments; however their immersion is significantly reduced by the lack of audio feedback. Furthermore, effort has been made to provide applications for the training of the hearing impaired. Such applications include the visualization of the hand and body movements performed in order to produce words in sign language as well as applications based on computer vision techniques that aim to recognize such gestures in order to allow natural human machine interaction for the hearing impaired. In the context of the presented framework, the deaf-mute terminal incorporates sign-language analysis and synthesis tools to allow physical interaction of the deaf-mute user and the virtual environment.

The paper is organized as follows: Sect. 2 presents the overall system architecture, and Sect. 3 describes the modality replacement framework. Subsequently, Sects. 4 and 5 present the audio and visual speech recognition modules, respectively. In Sect. 6, the audio and visual multimodal fusion framework is described and the employed algorithms are analytically discussed. In the following, Sect. 7 presents the path sketching module using gesture recognition. Then, Sect. 8 presents the sign language recognition module. Finally, Sect. 9 presents the application scenario and conclusions are drawn in Sect. 10.

2 Overall system description

The basic development concept in multimodal interfaces for the disabled is the idea of modality replacement, which is the use of information originating from various modalities to compensate for the missing input modality of the system or the users.

The main objective of the proposed system is the development of tools, algorithms and interfaces that will utilize modality replacement to enable the communication between blind or visually impaired and deaf-mute users. To achieve the desired result the proposed system combines the use of a set of different modules, such as:

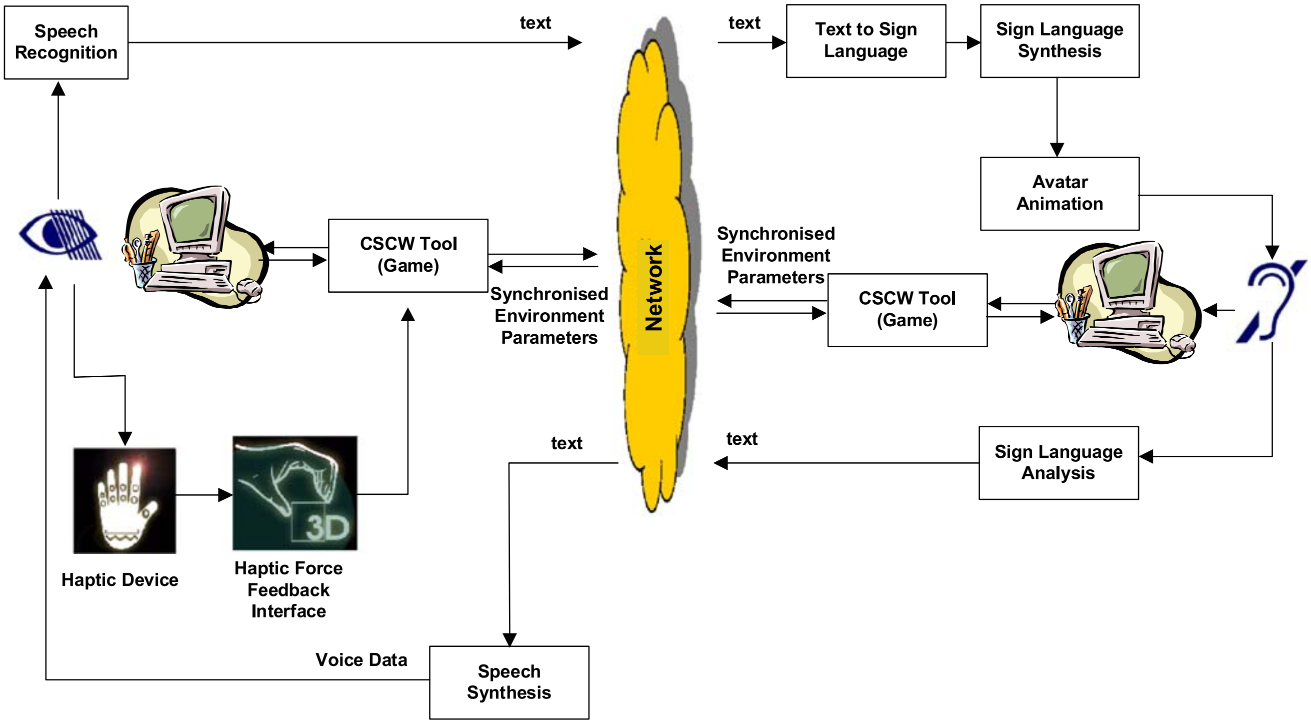

Fig. 1 Architecture of the collaborative treasure hunting game

- gesture recognition
- sign language analysis and synthesis
- speech analysis and synthesis
- haptics

into an innovative multimodal interface available to disabled users. Modality replacement was used in order to enable information transition between the various modalities used and thus enable the communication between the involved users.

Figure 1 presents the architecture of the proposed system, including the communication between the various modules used for the integration of the system as well as intermediate stages used for replacement between the various modalities. The left part of the figure refers to the blind user's terminal, while the right refers to the deaf-mute user's terminal. The different terminals of the treasure hunting game communicate through asynchronous TCP connection using TCP sockets.

The following sockets are implemented in the context of the treasure hunting game. The interested reader is referred to [12] for additional details.

- SeeColor terminal
  Implements a server socket that receives queries for translating color into sound. The code word consists of the following bytes, "b; R; G; B", where b is a boolean flag and R, G, B the color values.
- Blind user terminal
  Implements three sockets; a client socket that connects to the SeeColor terminal; a server socket to receive messages from the deaf-mute user terminal; and a client socket to send messages to the deaf-mute user terminal.
- The deaf-mute user's terminal
  Implements a server socket to receive messages from the blind user terminal and a client socket to send messages to the blind user terminal

Also, file sharing is used to ensure consistency between the data used in the various applications.

3 Modality replacement

The basic architecture of the proposed modality replacement approach is depicted in Fig. 2. The performance of such a system is directly dependent on the efficient multi-modal processing of two or more modalities and the effective exploitation of their complementary nature and their mutual information to achieve accurate recognition of the transmitted content. After the recognition has been performed effectively, either a modality translator can be employed in order to generate a new modality or the output can be utilized to detect and correct possibly erroneous feature vectors that may correspond to different modalities. The latter could be very useful in self-tutoring applications.

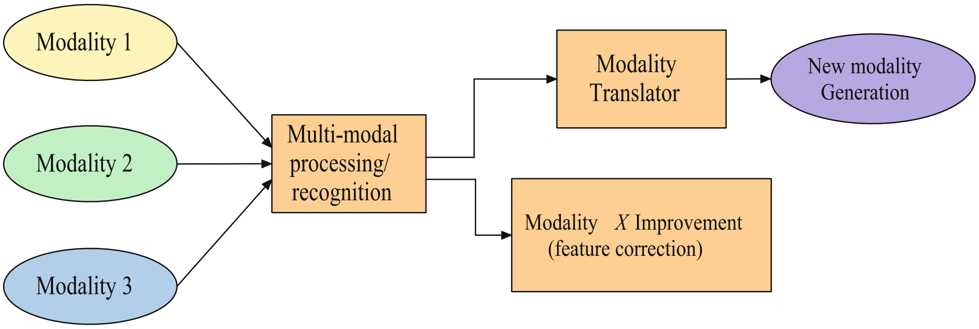
Modality 1
Modality 2
Modality 3
Multi-modal processing/ recognition
Modality Translator
New modality Generation
Modality X Improvement (feature correction)

Fig. 2 The modality replacement concept

The basic idea is to exploit the correlation between modalities in order to enhance the perceivable information by an impaired individual who cannot perceive all incoming modalities. In that sense, a modality, which would not be perceived due to a specific disability, can be employed to improve the information that is conveyed in the perceivable modalities and increase the accuracy rates of recognition. The results obtained by jointly fusing all the modalities outperform those obtained using only the perceived modalities since the inter- dependencies among them are modeled in an efficient manner.

A critical feature of the proposed system is its ability to adaptively assess the reliability of each modality and assign a measure to weight its contribution. There exist different approaches to measure reliability, such as taking into account the noise level of the input signal. The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average [13]. The proposed scheme aims at maximizing the benefit of multimodal fusion so that the error rate of the system becomes less than that of the cases where only the perceivable information is exploited. Modality reliability has also been examined in [14], in the context of multimodal speaker identification. An adaptive cascade rule was proposed and the order of the classifiers was determined based on the reliability of each modality combination.

A modified Coupled Hidden Markov Model (CHMM) is employed to model the complex interaction and inter- dependencies between audio and visual information and combine them efficiently in order to recognize correctly the transmitted words. In this work, modality reliability is regarded as a means of giving priority to single or combined modalities in the fusion process, rather than using it as a numerical weight.

4 Audio speech recognition

Audio speech recognition is one part of the proposed audio-visual speech recognition interface intended for verbal human-computer interaction between a blind person and the computer. 16 voice commands were selected to be pronounced by the blind person. For the demonstration purposes one man was selected to show eyeless human-computer interaction so the automatic recognition system is speaker-dependent. All the voice commands can be divided into two groups: (1) communication with the game process; (2) eyeless interaction with GUI interface of the multimodal system, as illustrated in Table 1.

Table 1 Recognition vocabulary with phonemic transcription

Voice command	Phonemic transcription	Interaction type
Catacombs	k a t a c o m s	game
Click	k l i k k l i	interface
Door	d o r	game
East	i s t i s	game
Enter	e n t e r	game
Exit	e g z i t e g z i	game
Go	g o u	game
Help	h e l p h e l e l	interface
Inscription	i n s k r i p s i o n	game
North	n o r s n o r	game
Open	o p e n	game
Restart	r i s t a r t r i s t a r	interface
South	s a u s s a u	game
Start game	s t a r t g e i m s t a r g e i m	interface
Stop game	s t o p g e i m	interface
West	u e s t u e s	game

HTK 3.4 toolkit [15] was employed to process the audio signal. The signal is captured by the microphone of a web-camera and sampled at 11025 Hz with 16 bits on each sample using a linear scale. Cepstral coefficients are computed for the 25 ms overlapping windows (frames) with 10 ms shift between adjacent frames applying the bank of triangular filters calculated according to the mel-scale frequencies by the equation:

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

Mel-frequency cepstral coefficients (MFCCs) are calculated from the log filter-bank amplitudes mi using the discrete cosine transform according to the equation [15]:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{2}$$

where the number N of filters in the filterbank equals 12. So the audio speech recognizer system calculates 12 MFCCs as well as estimates the first and second order derivatives that forms an observation vector of 36 components.

The acoustical modeling is based on continuous Hidden Markov Models (HMMs) [16], applying mixtures of Gaussian probability density functions that are defined according to the equation:

$$\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{o}-\boldsymbol{\mu})} \tag{3}$$

where $\mathcal{N}$ is a Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and n is the dimensionality of the observable vector $\mathbf{o}$. HMMs of phonemes have three meaningful states and two “hollow” states intended for concatenation of the models of phonemes in the models of words and phrases, as illustrated in Fig. 3. Each word of the vocabulary is obtained by concatenation of context-independent phonemes. The speech decoder uses Viterbi-based token passing algorithm [15]. The input phrase syntax is described in a simple grammar that allows to recognize only one command in a hypothesis.

In order to train the speech recognizer an audio-visual speech corpus was collected in an auditorium room using a USB web-camera. 320 utterances were used for training HMMs of phonemes and 100 utterances for the testing purpose. The wave audio files were extracted from the training video files using the VirtualDub software.

After expert processing of the utterances, it was found that the SNR for audio signal is quite low (15–20 db) because of far position (about 1 meter) of the speaker in front of the microphone and usage of the microphone built in a standard web-camera. Thus some explosive consonants (for instance “t” or “k”) at the beginnings and endings of phrases are not identified in the wave files. In Table 1, some words have several different variants of transcriptions, it is explained by periodical loss of explosive consonants in the speech signal. 30% training utterances were manually labeled on phonemes by the software WaveSurfer, and the remaining data were automatically segmented by the Viterbi forced alignment method [15].

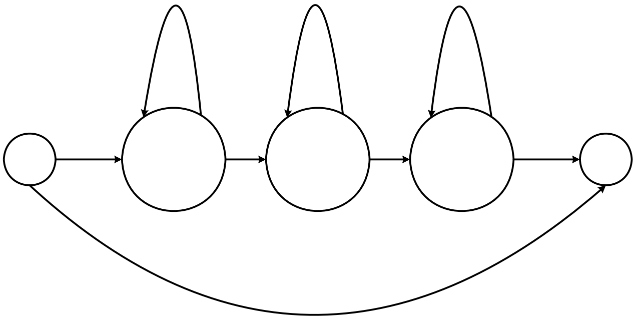

Fig. 3 Topology of the Hidden Markov Model for a phoneme

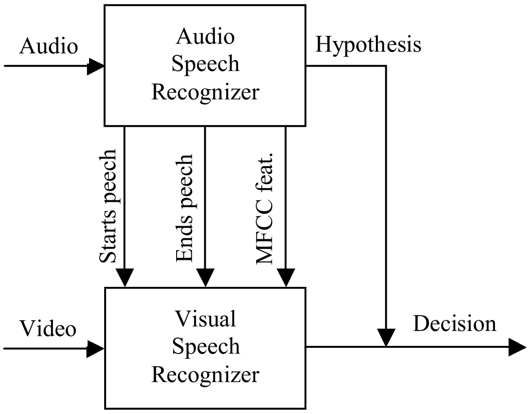

Fig. 4 General data flow in audio-visual speech recognition system

The audio speech recognizer was compiled as dynamic link library ASR.dll, which is used by the main executable module that combines the modules for audio and visual speech recognition (Fig. 4).

The audio speech recognizer can work independently or jointly with the visual speech recognizer. In the on-line operation mode the audio speech recognizer uses an energy-based voice activity detector to find the speech frames in audio signal. When any speech activity is found the module sends the message WMSTARTSPEECH to the window of the main application as well as when speech frames are changed by pause frames the message WMENDSPEECH is sent. After receiving one of the messages the visual recognizer should start or finish the video processing, correspondingly. The audio speech recognizer operates very fast so the result of speech recognition will be available almost immediately after the message WMENDSPEECH. Moreover, the MFCC features, calculated while processing speech, are

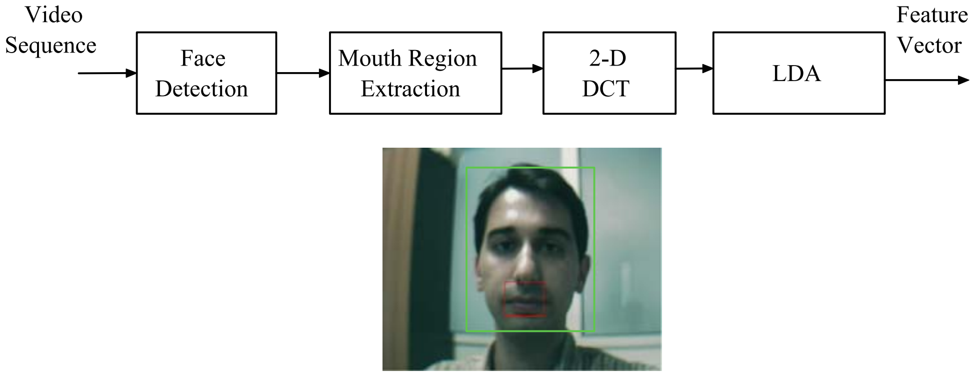

Fig. 5 Lip motion extraction process

stored in an internal buffer and can be transferred to the visual speech recognizer in order to fuse these parameters with visual parameters of the lips region.

5 Visual speech recognition

For the lip shape modality, the robust location of facial features and especially the location of the mouth region is crucial. Then, a discriminant set of visual observation vectors have to be extracted. The process for the extraction of the lip shape is presented in [17], and is described in brief below so that the paper is self-contained.

Initially, the speaker's face is located in the video sequence as illustrated in Fig. 5. Subsequently, the lower half of the detected face is selected as an initial candidate of the mouth region and Linear Discriminant Analysis (LDA) is used to classify pixels into two classes: face and lip. After the lip region segmentation has been performed the contour of the lips is obtained using the binary chain encoding method and a normalized 64 × 64 region is obtained from the mouth region using an affine transform. In the following, this area is split into blocks and the 2D-DCT transform is applied to each of these blocks and the lower frequency coefficients are selected from each block, forming a vector of 32 coefficients. Finally, LDA is applied to the resulting vectors, where the classes correspond to the words considered in the application. A set of 15 coefficients, corresponding to the most significant generalized eigenvalues of the LDA decomposition is used as the lip shape observation vector.

6 Audio-visual speech recognition

6.1 Multimodal fusion

The combination of multiple modalities for inference has proven to be a very powerful way to increase detection and recognition performance. By combining information provided by different models of the modalities, weakly incorrect evidence in one modality can be corrected by another modality. Hidden Markov Models (HMMs) are a popular probabilistic framework for modeling processes that have structure in time. Especially, for the applications that integrate two or more streams of data, Coupled Hidden Markov Models (CHMMs) have been developed.

A CHMM can be considered as a collection of HMMs, one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t-1$ for all the related HMMs. It must be noted that CHMMs are very popular among the audio-visual speech recognition community, since they can model efficiently the endogenous asynchrony between the speech and lip shape modalities. The parameters of a CHMM are described below:

$$\pi_0^c(i) = P(q_t^c = i), \tag{4}$$

$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i), \tag{5}$$

$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^A = j, q_{t-1}^L = k) \tag{6}$$

where q_t^c is the state of the coupled node in the c_{th} stream at time t, $\pi_0^c(i)$ is the initial state probability distribution for state i in c_{th} stream, $\mathbf{O}_t^c$ is the observation of the nodes at time t in the c_{th} stream, $b_t^c(i)$ is the probability of the observation given the i state of the hidden nodes in the c_{th} stream, and $a_{i|j,k,n}^c$ is the state transitional probability to node i in the c_{th} stream, given the state of the nodes at time $t-1$ for all the streams. Figure 6 illustrates the CHMM employed in this work. Square nodes represent the observable nodes whereas circle nodes denote the hidden (backbone) nodes.

One of the most challenging tasks in automatic speech recognition systems is to increase robustness to environmental conditions. Although the stream weights needs to be properly estimated according to noise conditions, they cannot be determined based on the maximum likelihood criterion. Therefore, it is very important to build an efficient stream weight optimization technique to achieve high recognition accuracy.

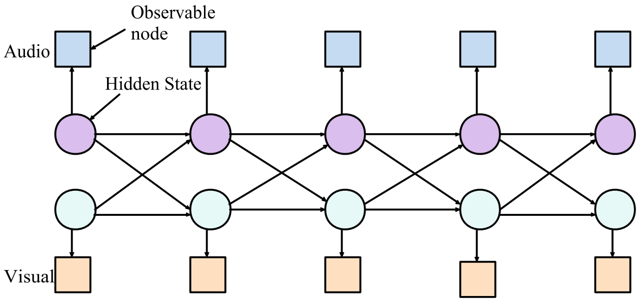

Fig. 6 Coupled HMM for audio and visual information fusion

6.2 Modality reliability

Ideally, the contribution of each modality to the overall output of the recognition system should be weighted according to a reliability measure. This measure denotes how each observation stream should be modified and acts as a weighting factor. In general, it is related to the environmental conditions (e.g., acoustic noise for the speech signal). The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average. Thus, the probability $b_m(\mathbf{O}_t)$ of a feature $\mathbf{O}_t$ for a word m is given by:

$$b_m(\mathbf{O}_t) = w_A \cdot b_A(\mathbf{O}_t^A) + w_L \cdot b_L(\mathbf{O}_t^L) \tag{7}$$

where $b_A(\mathbf{O}_t^A)$, and $b_L(\mathbf{O}_t^L)$ are respectively the likelihoods for an audio feature $\mathbf{O}_t^A$ and a lip shape feature $\mathbf{O}_t^L$. The parameters w_A and w_L are audio and lip shape weights, respectively, and $w_A + w_L = 1$.

In the proposed method, a different approach is employed to determine the weights of each data stream. More specifically, for each modality, word recognition is performed using a HMM for the training sequences. The results of the (unimodal) word recognition indicate the noise levels in each modality and provide an approximation of their reliability. More specifically, when the unimodal HMM classifier fails to identify the transmitted words it means that the observation features for the specific modality are unreliable. On the other hand, a small word error rate using only one modality and the related HMM means that the corresponding feature vector is reliable and should be favored in the CHMM. Let e_A denote the word error rate of the audio recognition system and e_L denote the word error rate of the visual recognition system. We define their ratio λ as:

$$\lambda = \frac{e_A}{e_L}. \tag{8}$$

Based on the value of λ and a set of predefined thresholds determined experimentally, the weight w_A is defined by the following rule:

$$w_A = \begin{cases} 0.1, & \lambda \geq 1.5, \\ 0.2, & 1.2 \leq \lambda < 1.5, \\ 0.4, & 1.1 \leq \lambda < 1.2, \\ 0.55, & 0.7 \leq \lambda < 1.1, \\ 0.75, & 0.5 \leq \lambda < 0.8, \\ 0.9, & 0.1 \leq \lambda < 0.5, \\ 0.95, & \lambda < 0.1. \end{cases} \tag{9}$$

As mentioned before, the value of w_L is computed as $1 - w_A$.

6.3 Word recognition

The word recognition is performed using the Viterbi algorithm, for the parameters of all the word models. It must be emphasized that the influence of each stream is weighted at the recognition process because, in general, the reliability and the information conveyed by each modality is different. Thus, the observation probabilities are modified as:

$$b_t^A(i) = b_t(\mathbf{O}_t^A | q_t^A = i)^{w_A}, \tag{10}$$

$$b_t^L(i) = b_t(\mathbf{O}_t^L | q_t^L = i)^{w_L} \tag{11}$$

where w_A and w_L are respectively the weights for audio and lip shape modalities and $w_A + w_L = 1$. The values of w_A and w_L are obtained using the methodology of Sect. 6.2.

6.4 Experimental evaluation

The audio-visual recognition module was evaluated on the words reported in Table 1. Each word in the vocabulary was repeated 26 times; 18 instances of each word were used for training and eight instances were used for testing. The performance of the audio only, visual only, and audio-visual recognition modules is depicted in Fig. 7 for various levels of environmental acoustic noise. As expected, the recognition accuracy of the audio-visual recognizer is higher than the accuracy of the unimodal recognizer. Moreover, the experimental results indicate that the performance gain becomes larger in low SNR values (high levels of acoustic noise). More specifically, at low SNR levels, the audio-visual recognizer increases by 30% the recognition accuracy.

7 Path sketching

In this revised version of the Treasure Hunting Game, the engagement of deaf-and mute players is improved by path

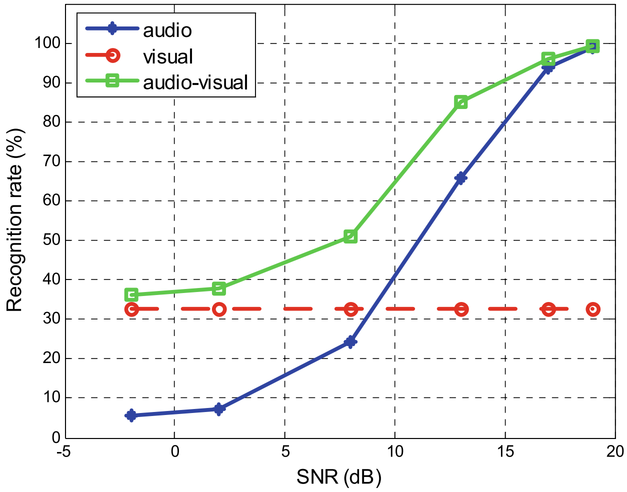

Fig. 7 Word recognition accuracy of the unimodal and the multimodal speech recognition systems at various acoustic noise levels

sketching based on gesture modality. The user can interact with the interface by means of a gesture performed by his/her hand to navigate on the village map and to explore the main areas of this map (e.g., temple, catacombs). This real time navigation process is implemented in the three steps: Hand detection, trajectory extraction, and sketching.

7.1 Hand detection

Detection and tracking was a nontrivial step because occlusions can occur due to overlap of each hand on the other or of the other skin colored regions (e.g., face and harms). To solve this problem and to make the detection easier a blue glove was worn from the player. In this way, we could deal with detection and tracking of hand exploiting techniques based on color blobs.

The colored region is detected via the histogram approach as proposed in [18]. For the glove color, we train a histogram of the components of the color space. HSV color space is preferred because of its robustness to changing illumination conditions. For the training, we collect a number of images that contain the color of the glove. The H, S and V components are divided into bins, and for each pixel in the training images, we update the histogram. At each bin of the histogram, we calculate the number of occurrences of pixels that correspond to that bin. Afterwards, the bins of histogram are normalized such that the maximum is 1. The histogram resembles the pdf of the detected pixel locations but the maximum value is 1 instead of the sum. To segment the hand in an image, we traverse each pixel, find the histogram bin it corresponds and apply thresholding. We set two thresholds, low and high, and choose double thresholding to ensure connectivity, and to avoid spikes in the binary image: A pixel is considered as a hand pixel if its histogram value is higher than the high threshold, or it is higher than the low threshold and the previous pixel was labeled as glove . The final hand region is assumed to be the largest connected component over the detected pixels. Thus we had only one component identified as hand.

7.2 Hand tracking and trajectory extraction

The analysis of hand motion is done by tracking the center of mass (CoM) and calculating the velocity of each segmented hand. However, these hand trajectories are noisy due to the noise introduced at the segmentation step. Thus, we use Kalman filters to smooth the obtained trajectories. The initialization of the Kalman Filter is done when the hand is first detected in the video. The CoM of each hand is given as the measurements to a four state Kalman filter: positions and velocities in x, y coordinates. The Kalman filter tries to estimate the state vector x_k at time k by the following equation:

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \tag{12}$$

with a measurement z_k

$$z_k = Hx_k + v_k. \tag{13}$$

The variables w_k and v_k represent the process and measurement noise respectively and are assumed to be normally distributed with zero mean and with covariances Q and R, respectively. The matrix A relates the state at time $k-1$ to the state at time k. The B matrix relates the optional control input, u. The matrix H relates the state at time k to the measurement.

At each frame, Kalman filter time update equations are calculated to predict the new hand position. The hand position found by the hand segmentation is used as measurements to correct the Kalman Filter parameters. Posterior states of each Kalman filter is defined as feature vectors for x, y coordinates of CoM and velocity. The hand can be lost due to occlusion or bad lighting in some frames. In that case, Kalman Filter prediction is directly used without correcting the Kalman Filter parameters. The hand is assumed to be out of the camera view if no hand can be detected for some number of (i.e. six) consecutive frames.

7.3 Sketching on the map

The extracted trajectory is then superimposed to the map, so that player can sketch directly the path on the map during the whole game.

Hand gestures performed by player were encoded with respect to the elongation of the hand. Positions of the hand were used as drawing controller, when player puts the hand in the vertical position with respect to the ground, drawing is enabled (Fig. 8a) and s/he can start to sketch trajectory on

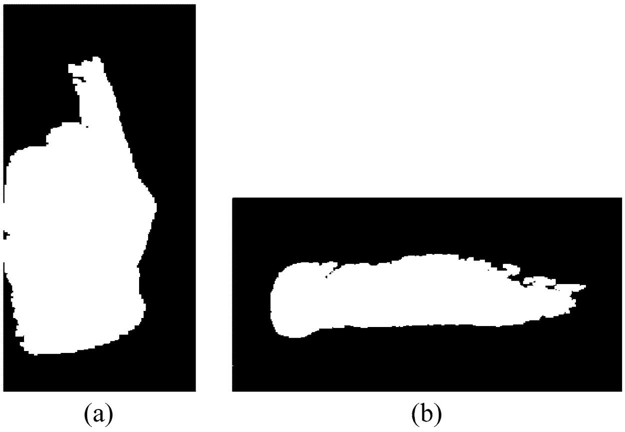

Fig. 8 The detected binary hand. The elongation of the hand sets whether the drawing is (**a**) on or (**b**) off

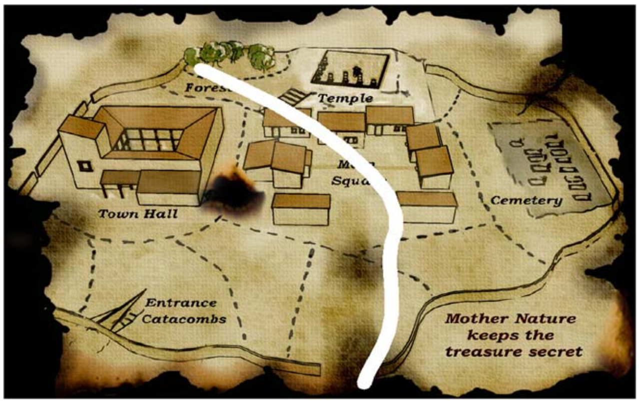

Fig. 9 The sketched trajectory on the map

the map. When the hand is moved to the horizontal position with respect to the ground, the drawing is disabled (Fig. 8b). If the user moves her/his hand to the top left corner of the map, the drawing is deleted and the user may start from the beginning.

The predefined locations on the map are used as the start and stop locations of the path. The drawing only starts when the user is around the starting position and the drawing ends when the path reaches to the stopping position (Fig. 9). The evaluation of the path sketching module is not feasible since it is obvious that there is not only one correct path from the starting point to the end point in the map. Thus, the users are free to sketch by hand any line and there are no ground truth data so that the line which is drawn on the map can be tested with. However, subjective evaluation showed that the trajectory which is drawn on the map matches the trajectory of the hand and user's comments were quite positive on this functionality.

The position of the hand which controls drawing can be easily found given that there is no similar color as the marker color in the camera view. In our tests, when this condition is satisfied, the hand region is found without any problem. If the condition is not satisfied, there are two different scenarios:

- if the hand is not in the camera view, some other region is considered as hand.
- if the hand is in the camera view, the algorithm chooses the largest connected component as the hand.

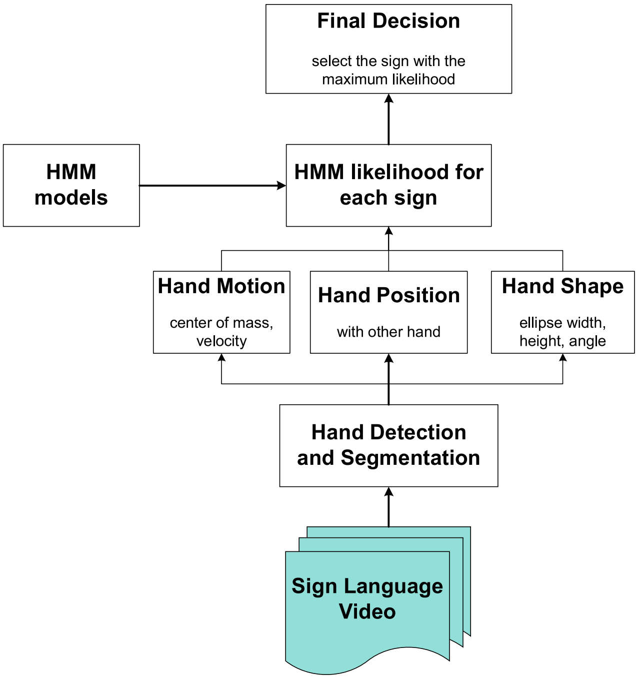
Final Decision
select the sign with the maximum likelihood
HMM models
HMM likelihood for each sign
Hand Motion
center of mass, velocity
Hand Position
with other hand
Hand Shape
ellipse width, height, angle
Hand Detection and Segmentation
Sign Language Video

Fig. 10 Sign language recognition system block diagram

A region can only be considered as hand if its area is greater than the area threshold. Moreover, we see that, if the above condition is met, the elongation of the hand can be correctly found since the segmentation of the hand region is correct.

8 Sign language recognition

Figure 10 depicts the steps in sign recognition. The first step in hand gesture recognition is to detect and track both hands. This is a complex task because the hands may occlude each other and also come in front of other skin colored regions, such as the arms and the face. To make the detection problem easier, we have used colored gloves worn on the hand (see Fig. 11). Once the hands are detected, a complete hand gesture recognition system must be able to extract the hand shape, and the hand motion. We have extracted simple hand shape features and combined them with hand motion and position information to obtain a combined feature vector [19].

Our sign database consists of four ASL signs for directions: north, south, east, and west. For each sign, we

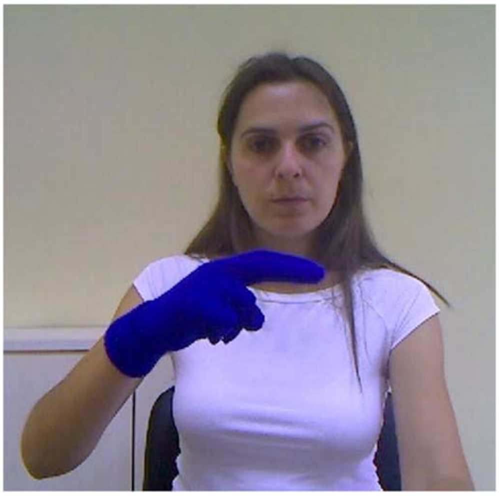

Fig. 11 The user wearing colored gloves

Table 2 Sign recognition accuracy

	% Signer dependent	% Signer independent
Train	100	100
Test	97.5	97.5

recorded 15 repetitions from two subjects. The video resolution is 640 × 480 pixels and the frame rate is 25 *fps*. A left-to-right continuous HMM model of four states and with no state skips, similar to the one in Fig. 3 is trained for each sign in the database. For the final decision, likelihoods of HMM for each sign class, $P(O|\Theta_i)$, are calculated and the sign class with the maximum likelihood is selected as the base decision.

$$\operatorname*{argmax}_i(P(O|\Theta_i)). \tag{14}$$

We tested the recognition performance of the system for the signer-dependent and the signer-independent cases. In the signer-dependent case, we apply a leave-one-out cross validation where at each fold and for each sign, one repetition of each subject is placed into the test set and the remaining ones to the training set. In the signer-independent case, we apply a 2-fold cross validation where at each fold and for each sign, we use examples from one subject into the training set and examples from the other subject to the test set, and vice versa. The average accuracies are given in Table 2. The system gives an average test accuracy of 97.5% and there is no difference between the signer-dependent and signer-independent cases.

9 Application scenario

The aforementioned technologies were integrated in order to create an entertainment scenario. The scenario consists of seven steps. In each step one of the users has to perform one or more actions in order to pass successfully to the next step. The storyboard is about an ancient city that is under attack and citizens of the city try finding the designs in order to create high technology war machines.

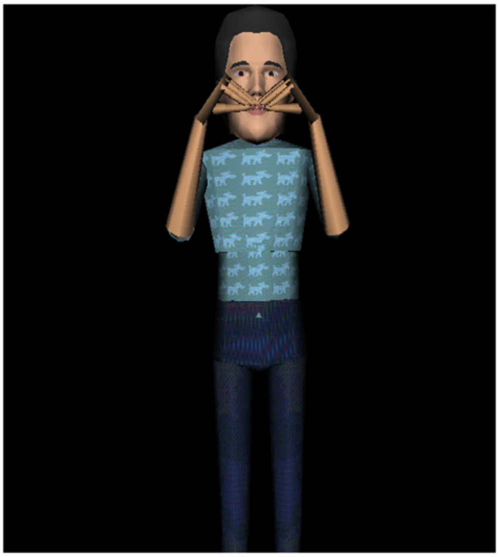

Fig. 12 Sign language synthesis using an avatar

In the first step, the blind user receives an audio message and is instructed to “find a red closet”. Subsequently, the blind user explores the village using the haptic device. It is worth noting that audio modality replaces color modality using the SeeColor module. Thus, the user can select the correct closet and receive further instructions which are transmitted to the other user.

In the second step, the deaf-and-mute person receives the audio message which is converted to text using the speech recognition tool and then to sign language using the sign synthesis tool. Finally, the user receives the message as a gesture through an avatar, as depicted in Fig. 12. The interested readers are referred to [20] for additional details on the implementation of the text-to-sign avatar. This message guides the deaf-and-mute user to the town hall, where the mayor provides the audio message “Go to the temple ruins”.

The third step involves the blind user, who hears the message said by the mayor and goes to the temple ruins. In the temple ruins the blind user has to search for an object that has an inscription written on it. One of the columns has an inscription written on it that states, “The dead will save the city”. The blind user is informed by an audio message whenever he finds this column and the message is sent to the deaf-mute user’s terminal.

The fourth step involves again the deaf and mute user. The user receives the written text in sign language form. The text modality is translated to sign language symbols using the sign synthesis tool. Then the deaf and mute user

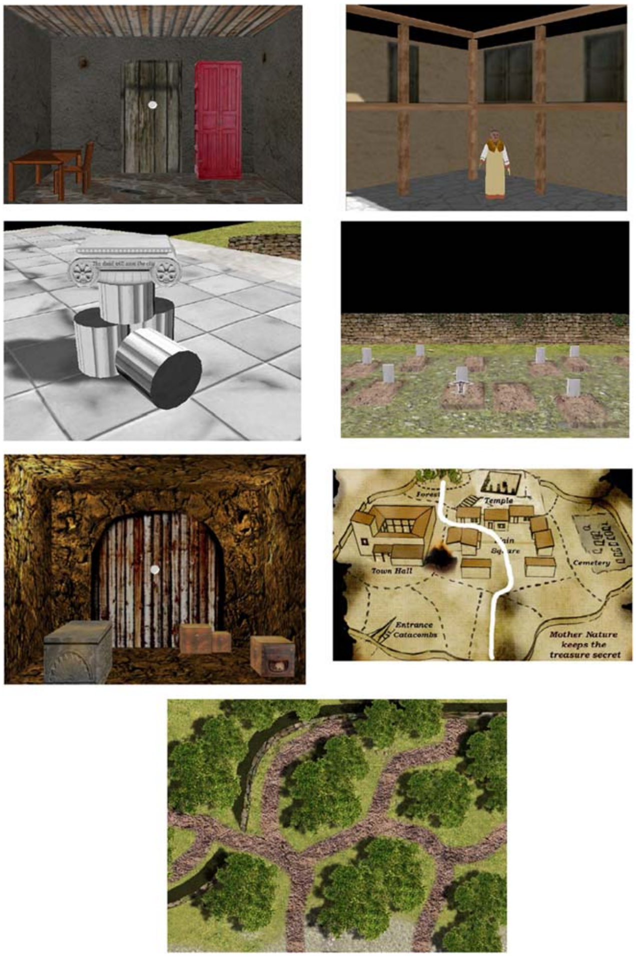

Fig. 13 The seven steps of the virtual game

has to understand the meaning of the inscription “The dead will save the city” and go to the cemetery using the mouse where he/she finds a key with the word “Catacombs” written on it.

In the fifth step, the text-to-speech (TTS) tool is employed to transform the instructions written on the key (“CATACOMBS”) to an audio signal that can be perceived by the blind user. The user has to search for the catacombs enter in them and find the box that contains a map (Fig. 13). The map is then sent to the next level.

In the sixth step, the deaf user receives the map, and has to draw the route to the area where the treasure is hidden. The route is drawn on the map and the map is converted to a grooved line map, which is send to for the last level to the blind user.

In the seventh step, the blind user receives the grooved line map and has to find and follow the way to the forest where the treasure is hidden. Although the map is presented again as a 2D image the blind user can feel the 3D grooved map and follow the route to the forest. The 2D image and the 3D map are registered and this allows us to visualize the route that the blind user actually follows on the 2D image. The blind user is asked to press the key of the PHANToM device while he believes that the PHANTOM cursor lies in the path. Finally, after finding the forest he obtains a new grooved line map where the blind user has to search for the final location of the treasure. After searching in the forest streets the blind user should find the treasure.

The treasure hunt game was played both by disabled and non-disabled people. However, objective evaluation of the game is not feasible since there are not any strict objective performance metrics that can be associated with the developed application. Instead, the overall evaluation of the system can be performed by evaluating its components, as described in the previous sections. Additionally, subjective evaluation of the system based on the comments of the users indicated that the use of speech commands and sign language were the features with the most positive comments. A demonstration of the developed application played by disabled people can be found in [21].

10 Conclusions

In this paper, a novel system for the communication between disabled users and their effective interaction with the computer was presented based on multimodal user interfaces. The main objective was to address the problem caused by the fact that impaired individuals, in general, do not have access to the same modalities. Therefore, the transmitted signals were translated into a perceivable form. The critical point for the automatic translation of information is the accurate recognition of the transmitted content and the effective transformation into another form. Thus, an audio-visual speech recognition system was employed to recognize phonetic commands from the blind user. The translation of the commands for the deaf-mute was performed using a sign synthesis module which produces an animation with an avatar. On the other hand, the deaf-mute user interacts using sign language and gestures. The system incorporates a module which is capable of recognizing user gestures and translate them using text-to-speech applications. As an application scenario, the aforementioned technologies are integrated in a collaborative treasure hunting game which requires the interaction of the users in each level. Future work will focus on the extension of the developed modules in order to support larger vocabularies and enable more natural communication of the users. Furthermore, the structure of the employed modalities should be studied more to reveal their inter-dependencies and exploit their complementary nature more effectively.

Acknowledgements This work was supported by the EU funded SIMILAR Network of Excellence.

References

1. Jaimes A, Sebe N (2007) Multimodal human–computer interaction: a survey. Comput Vis Image Underst 108(1–2):116–134
2. Richter K, Hellenschmidt M (2004) Interacting with the ambience: multimodal interaction and ambient intelligence. In: Proceedings of the W3C workshop on multi-modal interaction, vol 19, Sophia Antipolis, France, July 2004
3. Marsic I, Medl A, Flanagan J (2000) Natural communication with information systems. Proc IEEE 88:1354–1366
4. Lumsden J, Brewster SA (2003) A paradigm shift: alternative interaction techniques for use with mobile and wearable devices. In: Proceedings of the 13th annual IBM centers for advanced studies conference (CASCON 2003), Toronto, Canada, pp 97–100
5. Tangelder JWH, Schouten BAM (2006) Transparent face recognition in an unconstrained environment using a sparse representation from multiple still images. In: ASCI 2006 conference, Lommel, Belgium, June 2006
6. Raman TV (2003) Multimodal interaction design principles for multimodal interaction. In: Proceedings of computer human interaction (CHI 2003), Fort Lauderdale, USA, pp 5–10
7. Luciano C, Banerjee P, Florea L, Dawe G (2005) Design of the ImmersiveTouch™: a high-performance haptic augmented virtual reality system. In: Proceedings of the 11th international conference on human-computer interaction, Las Vegas, Nevada, July 2005
8. Sjostrom C (1999) Touch access for people with disabilities. Licentiate thesis, CERTEC Lund University, Sweden, 1999
9. Nelson B, Ketelhut D, Clarke J, Bowman C, Dede C (2005) Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. Educ Technol 45(1):21–28
10. Lim CP, Nonis D, Hedberg J (2006) Gaming in a 3D multiuser virtual environment: engaging students in Science lessons. Br J Educ Technol 37(2):211–231
11. Scoy V, Kawai I, Darrah S, Rash F (2000) Haptic display of mathematical functions for teaching mathematics to students with vision disabilities. In: Haptic human-computer interaction workshop
12. Moustakas K, Nikolakis G, Tzovaras D, Deville B, Marras I, Pavlek J (2000) Multimodal tools and interfaces for the intercommunication between visually impaired and deaf-and-mute people. In: Proceedings of eNTERFACE 2006, Dubrovnik, Croatia, July 2006
13. Tamura S, Iwano K, Furui S (2005) A stream-weight optimization method for multi-stream HMMS based on likelihood value normalization. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'05), vol 1
14. Erzin E, Yemez Y, Tekalp A (2005) Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. IEEE Trans Multimedia 7(5):840–852
15. Yound S et al (2006) The HTK book, HTK Version 3.4. Cambridge University Engineering Department
16. Rabiner L, Juang B (1993) Fundamentals of speech recognition. Englewood Cliffs, Prentice-Hall
17. Nefian A, Liang L, Pi X, Liu X, Murphy K (2002) Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J Appl Signal Process 2002(11):1274–1288
18. Jayaram S, Schmugge S, Shin MC, Tsap LV (2004) Effect of colorspace transformation, the illuminance component, and color modeling on skin detection. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)
19. Aran O, Akarun L (2006) Recognizing two handed gestures with generative, discriminative and ensemble methods via Fisher kernels. In: International workshop on multimedia content representation, classification and security (MRCS '06), Istanbul, Turkey, September 2006
20. Papadogiorgaki M, Grammalidis N, Tzovaras D, Strintzis MG (2005) Text-to-sign language synthesis tool. In: 13th European signal processing conference (EUSIPCO2005), Antalya, Turkey, September 2005
21. http://avrlab.iti.gr/SIMILAR/GameEvaluation.html