A MULTIMODAL FRAMEWORK FOR THE COMMUNICATION OF THE DISABLED

Savvas Argyropoulos [1], *Konstantinos Moustakas* [1], *Alexey A. Karpov* [2], *Oya Aran* [3], *Dimitrios Tzovaras* [1], *Thanos Tsakiris* [1], *Giovanna Varni* [4], *Byungjun Kwon* [5]

[1] Informatics and Telematics Institute (ITI), Aristotle University of Thessaloniki, Hellas, Greece
[2] Saint-Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, Russian Federation
[3] Perceptual Intelligence Lab, Boğaziçi University, İstanbul, Turkey
[4] InfoMus Lab - Casa Paganini, DIST, University of Genoa, Italy
[5] Koninklijk Conservatorium, The Hague, Netherlands

ABSTRACT

In this paper, a novel system, which aims to provide alternative tools and interfaces to blind and deaf-and-mute people and enable their communication and interaction with the computer is presented. 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. The proposed application integrates haptics, audio, visual output and computer vision, sign language analysis and synthesis, speech recognition and synthesis, in order to provide an interactive environment where the blind and deaf-and-mute users can collaborate to play the treasure hunting game.

KEYWORDS

Multimodal interfaces – Multimodal fusion – Sign language analysis – Sign language synthesis – Speech recognition

1. INTRODUCTION

The widespread deployment of novel human-computer interaction methods has changed the way individuals communicate with computers. Since Sutherland's SketchPad in 1961 or Xerox' alto in 1973, computer users have long been acquainted with more than the traditional keyboard to interact with a system. More recently, with the desire of increased productivity, seamless interaction and immersion, e-inclusion of people with disabilities, and with the progress in fields such as multimedia/multimodal signal analysis and human-computer interaction, multimodal interaction has emerged as a very active field of research [1].

Multimodal interfaces are those encompassing more than the traditional keyboard and mouse. Natural input modes are employed [2], [3], such as voice, gestures and body movement, haptic interaction, facial expressions, and physiological signals. As described in [4], 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. Also, 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, and the 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 expressive, emotional content 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 twodimensional graphic information, but also information presented in 3D virtual reality environments (VEs) [5]. 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 [6].

Several research projects have been conducted to assist visually impaired to understand 3D objects, scientific data and mathematical functions, by using force feedback devices [7]. PHANToM™ is one of the most commonly used force feedback device. Due its hardware design, only one point of contact at a time is supported. This is very different from the way that people 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.

Deaf and mute users have visual access to 3D virtual environments; however their immersion is significantly reduced by the luck of audio feedback. Furthermore effort has been done to provide applications for the training of 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 so as to

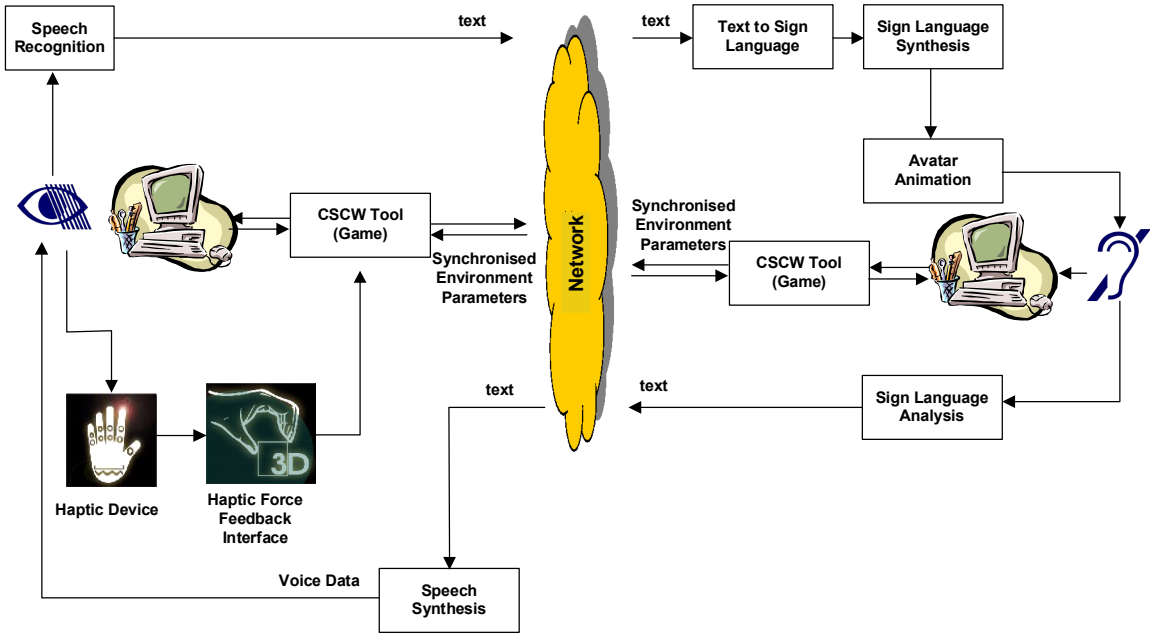

Figure 1: *Architecture of the collaborative treasure hunting game.*

allow physical interaction of the deafmute user and the virtual environment.

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

2. OVERALL SYSTEM DESCRIPTION

The basic development concept in multimodal interfaces for the disabled is the idea of *modality replacement*, which is defined as *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 so as to allow 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

- 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 interested reader is referred to [8] for additional details.

The following sockets are implemented in the context of the treasure hunting game:

- 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.
  - 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.
- 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

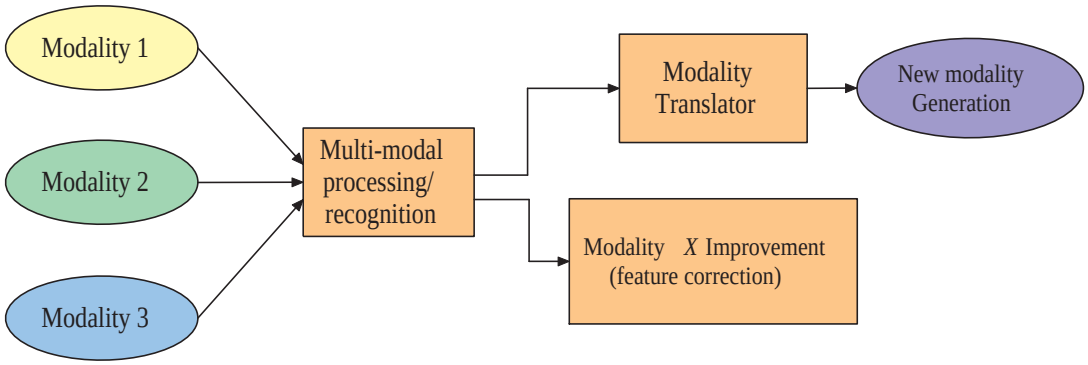

Figure 2: *The modality replacement concept.*

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. For example, if an individual practices sign language, the automatic recognition algorithm could detect incorrect hand shapes (based on audio and visual information) and indicate them so that the user can identify the wrong gestures and practice more on them.

The basic idea is to exploit the correlation between modalities in order to enhance the perceivable information by an impaired individual who can not 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 modelled 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 [9]. The proposed scheme aims at maximizing the benefit of multimodal fusion so that the error rate of the system becomes than that of the cases where only the perceivable information is exploited. Modality reliability has also been examined in [10], 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 interdependencies among the modalities and combine them efficiently in order to recognize correctly the transmitted message. 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. These issues are discussed in detail in the following sections and the unimodal speech recognition modules, based on audio and visual information, are described to illustrate how they can be combined.

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.

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

Table 1: *Recognition vocabulary with phonemic Transcription*

HTK 3.4 toolkit [11] was employed to process speech signal. The audio signal is captured by microphone of webcamera Philips SPC900 and sampled at 11025 Hz with 16 bits on

each sample using a linear scale. Mel-frequency cepstral coefficients (MFCC) are computed for the 25 ms overlapping windows (frames) with 10 ms shift between adjacent frames. Audio speech recognizer system uses 12 MFCCs as well as an estimation of the first and second order derivatives that forms a vector of 36 components.

The acoustical modeling is based on Hidden Markov Models (HMMs) with mixture Gaussian probability density functions [12]. HMMs of phonemes have 3 meaningful states and 2 "hollow" states intended for concatenation of the models (Fig. 3). Each word of the vocabulary is obtained by concatenation of context-independent phonemes. The speech decoder uses Viterbi-based token passing algorithm [11]. The input phrase syntax is described in a simple grammar that allows recognizing only one command each time.

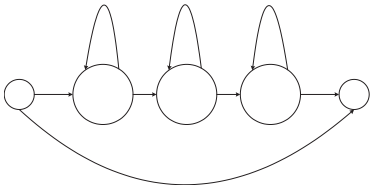

Figure 3: *Topology of the Hidden Markov Model for a phoneme.*

In order to train the speech recognizer an audio-visual speech corpus was collected in an auditorium room using USB webcamera Philips SPC900. 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 avi 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 webcamera. 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 [11].

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 `WM_STARTSPEECH` to the window of the main application as well as when speech frames are changed by pause frames the message `WM_ENDSPEECH` 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 `WM_ENDSPEECH`. Moreover, the MFCC features, calculated while processing speech, are 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.

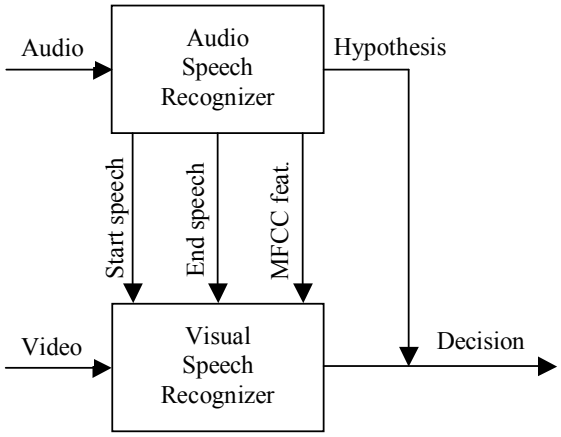

Figure 4: *General data flow in audio-visual speech recognition system.*

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 [13], 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 64x64 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.

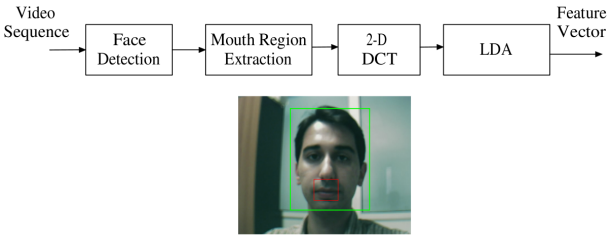

Figure 5: *Lip motion extraction process.*

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 modelling 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:

$$\pi_0^c(i) = P(q_t^c = i) \quad (1)$$

$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i) \quad (2)$$

$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^A = j, q_{t-1}^L = k) \quad (3)$$

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. The distribution of the observation probability is usually defined as a continuous Gaussian Mixture. Fig. 6 illustrates the CHMM employed in this work. Square nodes represent the observable nodes whereas circle nodes denote the hidden (backbone) nodes.

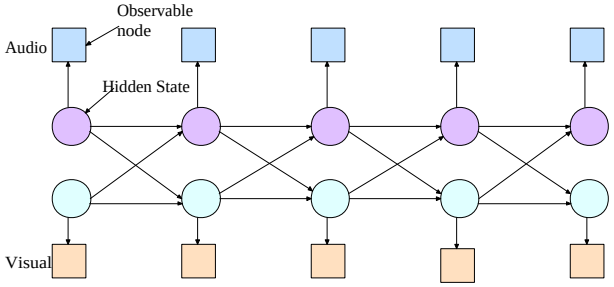

Figure 6: *Coupled HMM for audio and visual information fusion.*

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 can not 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.

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:

$$b_m(\mathbf{O}_t) = w_A \cdot b_A(\mathbf{O}_t^A) + w_L \cdot b_L(\mathbf{O}_t^L)) \quad (4)$$

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 favoured in the CHMM.

6.3. Word Recognition

The word recognition is performed using the Viterbi algorithm, described above, 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:

$$b_t^A(i) = b_t(\mathbf{O}_t^A | q_t^A = i)^{w_A} \quad (5)$$

$$b_t^L(i) = b_t(\mathbf{O}_t^L | q_t^L = i)^{w_L} \quad (6)$$

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 section 6.2.

7. PATH SKETCHING

In this revised version of the Treasure Hunting Game, the engagement of deaf-and mute players is improved by path 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 non trivial 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 [14]. Double thresholding is used to ensure connectivity, and to avoid spikes in the binary image. The scheme is composed of training the histogram and threshold values for future use. To cancel the noise, we selected the largest connected component of the detected regions into consideration. Thus we had only one component identified as hand.

7.2. Hand Tracking and Trajectory extraction

The analysis of hand motion is performed 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. 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. 7a) and s/he can start to sketch trajectory on the map. When the hand is moved to the horizontal position with respect to the ground, the drawing is disabled (Fig. 7b). 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.

(a) (b)

Figure 7: *The detected binary hand. The elongation of the hand sets whether the drawing is (a) on or (b) off.*

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. 8).

8. SIGN LANGUAGE RECOGNITION

Figure 9 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. 10). Once the hands are detected, a complete hand gesture recognition system

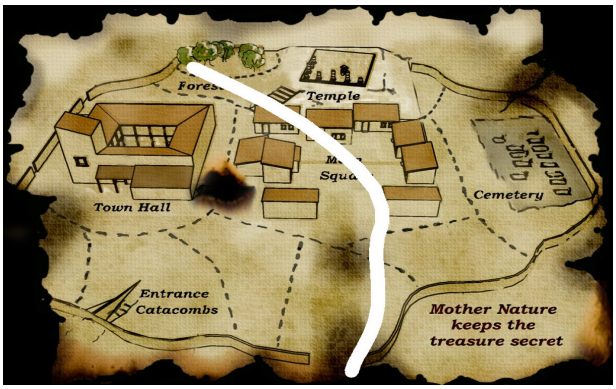

Figure 8: *The sketched trajectory on the map.*

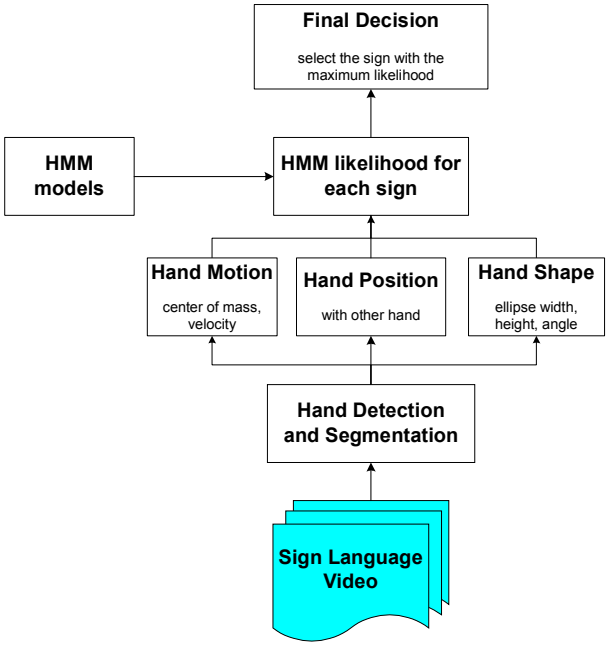

Figure 9: *Sign language recognition system block diagram.*

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 [15].

Our sign database consists of four ASL signs for directions: *north, south, east, and west*. For each sign, we 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 with no state skips is trained for each sign in the database. For the final decision, likelihoods of HMM for each sign class are calculated and the sign class with the maximum likelihood is selected as the base decision.

9. APPLICATION SCENARIO

The aforementioned technologies where 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

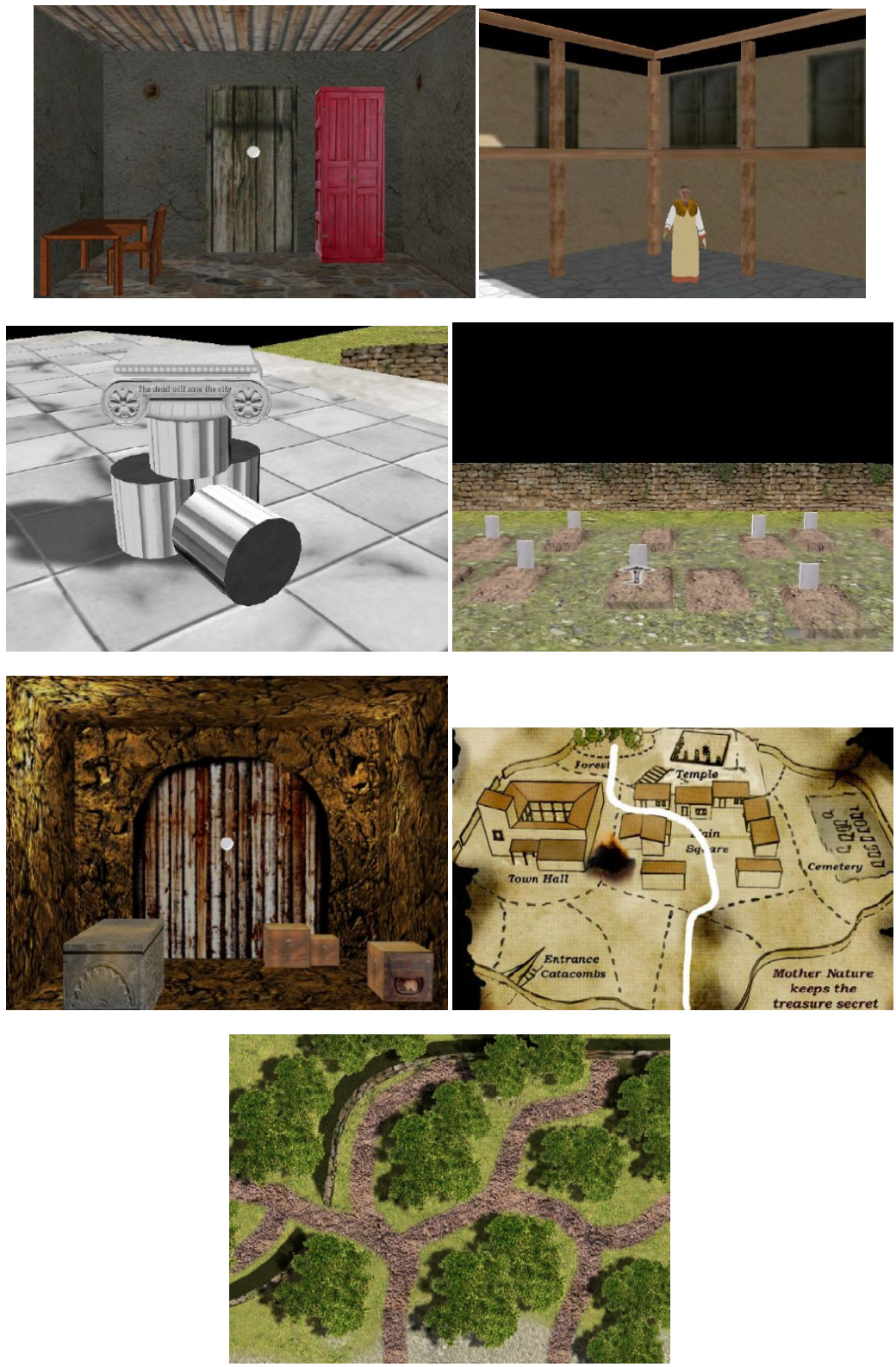

Figure 11: *The seven steps of the virtual game.*

Figure 10: *The user wearing colored gloves.*

war machines.

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.

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. 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 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. 11). 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.

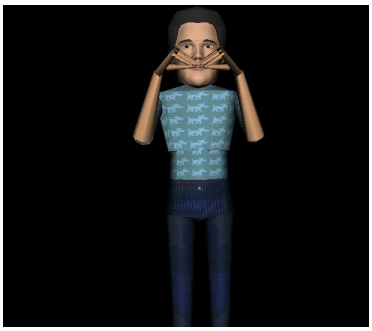

Figure 12: *Sign language synthesis using an avatar.*

10. CONCLUSIONS

In this paper, a novel system for the communication between disabled used 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.

11. ACKNOWLEDGEMENTS

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

12. REFERENCES

[1] "W3C Workshop on Multimodal Interaction", July 2004. http://www.w3.org/2004/02/mmi-workshop-cfp.html. 27

[2] I. Marsic, A. Medl, and J. Flanagan, "Natural communication with information systems", *Proc. of the IEEE*, vol. 88, pp. 1354–1366, August 2000. 27

[3] J. Lumsden and S. A. Brewster, "A paradigm shift: Alternative interaction techniques for use with mobile and wearable devices", in *Proc. 13th Annual IBM Centers for Advanced Studies Conference (CASCON 2003)*, (Toronto, Canada), pp. 97–100, 2003. 27

[4] T. V. Raman, "Multimodal Interaction Design Principles For Multimodal Interaction", in *Proc. of Computer Human Interaction (CHI 2003)*, (Fort Lauderdale, USA), pp. 5–10, 2003. 27

[5] C. Colwell, H. Petrie, D. Kornbrot, A. Hardwick, and S. Furner, "Haptic Virtual Reality for Blind Computer Users", in *Proc. of Annual ACM Conference on Assistive Technologies (ASSETS 1998)*, pp. 92–99, 1998. 27

[6] C. Sjostrom, "Touch Access for People With Disabilities". Licentiate Thesis, CERTEC Lund University, Sweden, 1999. 27

[7] V. Scoy, I. Kawai, S. Darrah, and F. Rash, "Haptic Display of Mathematical Functions for Teaching Mathematics to Students with Vision Disabilities", in *Haptic Human-Computer Interaction Workshop*, 2000. 27

[8] K. Moustakas, G. Nikolakis, D. Tzovaras, B. Deville, I. Marras, and J. Pavlek, "Multimodal tools and interafces for the intercommunication between visually impaired and deaf-and-mute people", in *Proc. of eNTERFACE 2006*, (Dubrovnik, Croatia), July 2006. 28

[9] S. Tamura, K. Iwano, and S. Furui, "A Stream-Weight Optimization Method for Multi-Stream HMMS Based on Likelihood Value Normalization", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP'05)*, vol. 1, pp. 469–472, 2005. 29

[10] E. Erzin, Y. Yemez, and A. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability", *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, 2005. 29

[11] S. Yound *et al.*, *The HTK Book, HTK Version 3.4*. Cambridge University Engineering Department, 2006. 29, 30

[12] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, USA: New Jersey: Prentice-Hall, 1993. 30

[13] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002. 30

[14] S. Jayaram, S. Schmugge, M. C. Shin, and L. V. Tsap, "Effect of Colorspace Transformation, the Illuminance Component, and Color Modeling on Skin Detection", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004. 31

[15] O. Aran and L. Akarun, "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. 32

13. BIOGRAPHIES

Savvas Argyropoulos (S'04) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Hellas, in 2004, where he is currently pursuing the Ph.D. degree. He holds teaching and research assistantship positions at AUTH. He is also a graduate Research Associate with the Informatics and Telematics Institute, Centre for Research and Technology Hellas. Since 2005, he has participated in several research projects funded by the EC. His research interests include distributed source coding, video coding/transmission, multimodal signal processing, and biometric recognition.
Email: savvas@ieee.org

Konstantinos Moustakas received the Diploma degree in electrical and computer engineering and the PhD degree in virtual reality from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2007 respectively. Currently, he is a postdoctoral research fellow in the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include virtual reality, collision detection, haptics, deformable object modeling and simulation, 3D content-based search, computer vision, and stereoscopic image processing. During the last three years, he has been the (co)author of more than 40 papers in refereed journals, edited books, and international conferences. Dr. Moustakas serves as a regular reviewer for various international scientific journals. He has also been involved in many projects funded by the EC and the Greek secretariat of Research and Technology. He is a member of the IEEE and the Technical Chamber of Greece.
Email: moustak@iti.gr

Alexey A. Karpov received the M.S. Diploma from St. Petersburg State University of Airspace Instrumentation and Ph.D. degree in computer science from St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), in 2002 and 2007, respectively. Currently he is a researcher of Speech Informatics Group of SPIIRAS. His main research interests include automatic speech and speaker recognition, multimodal interfaces based on speech and gestures, computer vision techniques. He has been the (co)author of more than 50 papers in refereed journals and international conferences. He has also been involved in SIMILAR Network of Excellence funded by the EC as well as several research projects funded by EU INTAS association and Russian scientific foundations. He is a member of organizing committee of International Conferences "Speech and Computer" SPECOM.
Email: karpov@iias.spb.su

Oya Aran received the BS and MS degrees in Computer Engineering from Boğaziçi University, İstanbul, Turkey in 2000 and 2002, respectively. She is currently a PhD candidate at Boğaziçi University working on dynamic hand gesture and sign language recognition. Her research interests include computer vision, pattern recognition and machine learning.
Email: aranoya@boun.edu.tr

Dimitrios Tzovaras received the Diploma degree in electrical engineering and the Ph.D. degree in 2-D and 3-D image compression from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1992 and 1997, respectively. He is currently a Senior Researcher Grade B in the Informatics and Telematics Institute of the Centre for Research and Technology Hellas in Thessaloniki, Greece. Prior to his current position, he was a Senior Researcher on 3-D imaging at the Aristotle University of Thessaloniki. His main research interests include virtual reality, assistive technologies, 3-D data processing, haptics and semantics in virtual environments. His involvement with those research areas has led to the co-authoring of more than 50 papers in refereed journals and more than 150 publications in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Since 1992, he has been involved in more than 100 projects in Greece, funded by the EC and the Greek Ministry of Research and Technology. Dr. Tzovaras is a member of the Technical Chamber of Greece.
Email: dimitrios.tzovaras@iti.gr

Thanos Tsakiris received the BSc in Computer Science from the Aristotle University of Thessaloniki in 2000 and the MSc in Computer Games Technology from the University of Abertay Dundee in 2001. He is currently working in ITI/CERTH as a research associate in the fields of 3D Graphics,VR and HCI.
Email: atsakir@iti.gr

Giovanna Varni was born in Genoa (Italy), in 1979. She received summa cum laude her master's degree in Biomedical Engineering at the University of Genoa in 2005. She is currently PhD Student at InfoMus Lab (DIST, University of Genoa) where she is working on multimodal feedback and multimodal streams data analysis.
Email: giovanna@infomus.dist.unige.it

Byungjun Kwon was born in Seoul, Korea. He received his bachelor's degree in French language and literature from Seoul National University (1996) and master's in Art- Science from Royal Conservatory, the Hague, Netherlands (2008). He started his musical career in early 90's as a singer/songwriter and has released 7 albums ranging from alternative rock to minimal house. He creates music for records, sound tracks, fashion collections, contemporary dance, theatre plays and interdisciplinary events. Recent works and performances have been presented in many international venues. Now he lives and works in Amsterdam.
Email: byungjun@gmail.com