Multimodal Interface for Effective Man Machine Interaction

N.S. Sreekanth, Nobby Varghese, C.H. Pradeepkumar, Pal Vaishali, R. Ganga Prasad, N. Pal Supriya, and N.K. Narayanan

1 Introduction

Computers are used extensively in day-to-day life; hence more focus is expected to make human computer interaction as natural as possible (Sreekanth, Supriya, Thomas, Hassan, & Narayanan, 2009). It is still a dream to interact with your electronic gadgets like how you interact with your friends. Providing intelligence to machines is being a research area for past few decades. There have been tremendous achievements in this area. The significant progress in the areas of automatic speech recognition, natural language processing and computer vision, facilitate the man-machine interaction environment more intelligent. In the past few decades there have been lots of initiatives for improving human computer interaction. As more powerful and complex computer systems emerged, efforts to make computer user interfaces more simple and natural become important. The effort behind all these works has been to make the interaction between computer and human as natural as the way human beings communicate with each other (Thomas, Hassan, Sreekanth, & Supriya, 2008).

Bringing the research outcomes to practical applications requires massive effort. If we review the progress in the individual threads of machine intelligence like speech recognition, language processing and computer vision, quite a good amount of performance is guaranteed. Human beings are accustomed to convey ideas through various modalities. The five modalities namely speech, hearing, vision, taste, smell and touch are involved in human–human interaction. If you consider

N.S. Sreekanth (✉) • N. Varghese • C.H. Pradeepkumar • P. Vaishali • R. Ganga Prasad • N.P. Supriya
Center for Development of Advanced Computing (C-DAC), #68 Electronics City, Bangalore 560100, Karnataka, India
e-mail: nssreekanth@gmail.com

N.K. Narayanan
Department of Information Technology, Kannur University, Kannur, Kerala, India

A. Lugmayr, C. Dal Zotto (eds.), *Media Convergence Handbook - Vol. 2*, Media Business and Innovation, DOI 10.1007/978-3-642-54487-3_14

human as a machine, it has two output mechanisms and five input mechanisms to send and receive various forms of communication signals. Speech and Gestures are two output mechanisms and hearing, vision, olfaction, taste and haptic (touch) are the five input receptors.

As we discussed when human beings communicate with each other we use various modalities like speech, gestures, text, and images in various combinations. Human cognitive systems are capable of recognizing the combination of various modalities and they can synchronize and understand it. Building user interfaces by mimicking the human way of communication, lead to thinking about multimodal interface. Multimodal interaction is a type of Human Computer Interaction, which combines multiple modalities or different modes of communication like speech, gestures, text and various other combinations. The most common multimodal interface combines a visual modality (e.g. a display, keyboard, and mouse) with a voice modality (speech recognition for input, speech synthesis and recorded audio for output). These devices have grown to be familiar but tend to restrict the information and command flow between the user and the computer system. However other modalities, such as haptic and olfactory can also be combined with the previous ones this limitation has become even more apparent with the emergence of novel display technologies such as virtual reality and wearable computers. Thus, in recent years, there has been a tremendous interest in introducing new modalities into HCI that will potentially resolve this interaction bottleneck.

Multimodal systems are sometimes designed based on one main modality, with the other modalities simply added on top. As handling several modalities together may result in cognitive overload and reduced usability, especially in the demanding usage situations that arise in mobile use. Providing the logical synchronization between the various signals such as speech, haptic, gesture, olfaction seems to be really challenging and this is where human cognition is still a black box to Multimodal researchers.

In this chapter we discuss about the convergence of various modalities to make human machine communication efficient and easier and the best available practices for designing a user friendly and effective multimodal interface.

2 Literature Review

Multimodal interfaces emerged approximately three decades ago within the field of human/computer interaction with Richard Bolt’s “Put-That-There” application. First multimodal systems sought ways to go beyond the standard interaction mode at this time, which were graphical interfaces with keyboards and mice. Bolt’s “Put-that-there” processed spoken commands linked to a pointing gesture using an armrest-mounted touchpad to move and change shapes displayed on a screen in front of the user (Bolt, 1980; Dumas, Lalanne, & Oviat, 2009).

Another interesting study, which has been done at Pennsylvania State University. “A Real-Time Framework for Natural Multimodal Interaction with Large Screen displays” in which they discussed about a framework, which uses speech and gesture to create a natural interface (Krahnstoever, Kettebekov, Yeasin, &

Sharma, 2002). The system is designed to accommodate the use natural gestures and speech commands of an experienced as well as an inexperienced user to increase the usability of the system in domains where user training is not feasible. Another important aspect is the use of a large screen display to provide appropriate feedback to the user. Large screen displays are a natural choice for many applications, especially interaction with spatial/geocentric data, immersive virtual reality environments and collaborative systems that allow interaction with multiple users simultaneously.

"Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios". This paper describes an intelligent system that they developed to support affective multimodal human–computer interaction (AMM-HCI) where the user's actions and emotions are modeled and then used to adapt the HCI and support the user in his or her activity (Maat & Pantic, 2007). The proposed system, which they named Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how). It integrates a number of natural human communicative modalities including speech, eye gaze direction, face and facial expression. To attain a system that can be educated, that can improve its knowledge and decision making through experience. To support concepts of concurrency, persistency, and mobility, Gaze-X has been built as an agent-based system where different agents are responsible for different parts of the processing. A usability study conducted in an office scenario with a number of users indicates that Gaze-X is perceived as effective, easy to use, useful, and affectively qualitative.

UI on the Fly is a system that dynamically presents coordinated multimodal content through natural language and a small-screen graphical user interface (Reitter, Panttaja, & Cummins, 2004). It adapts to the user's preferences and situation. Multimodal Functional Unification Grammar (MUG) is a unification-based formalism that uses rules to generate content that is coordinated across several communication modes. Faithful variants are scored with a heuristic function.

Another interesting work in the category of assistive technology in Kanagawa Rehabilitation Center, Japan "Multi-modal Interface with Voice and Head Tracking for Multiple Home Appliances" addresses a multi-modal interface that allows use of voice and gesture commands for controlling distributed home appliances used by people with disabilities (Ito, 2001). The main objective of this study is combined with nonverbal and verbal interface for intuitive and efficient control that uses hands-free operation. The pointing gesture by facing as nonverbal interface represents selecting one of the home appliances. The voice commands as verbal interface represent button operation of the remote controller such as the power on/off, the channel select and the volume up/down. The prototype system can provide a hands-free remote controller for people with quadriplegia who do not have to send verbal commands for selecting home appliances.

Researchers at AT&T labs are addressing this challenge by developing technologies to support truly multimodal interaction. Various products and prototypes from this lab brought a new dimension in the area of multimodal interaction. The prototypes include MATCH (Multimodal Access to City Help), Multimodal IPTV and Multimodal presentation dash board. Building these systems involves significant advances in the areas of multimodal integration, understanding, multimodal dialog management, and multimodal generation of sentences. These multimodal interface technologies have been applied to a broad range of different application areas, including local search, corporate directory access and messaging, medical informatics, accessing and controlling presentations, and searching and browsing for Internet Protocol television (IPTV) content such as movies-on-demand (http://www2.research.att.com/~johnston/).

3 Multimodal Interface- Methodology and Approach

Multimodal interface provides a very natural way for humans to perform tasks on a machine, using direct manipulation and speech interaction methods similar to those used daily in human-to-human communication. However, despite the availability of high accuracy speech recognizers and the available haptic and gesture-based devices such as gaze trackers, touch screens, and gesture trackers, very few applications take advantage of these technologies. One reason for this may be that the cost in time of implementing a multimodal interface is prohibitive (Flippo, Kerbs, & Marsic, 2003). Multimodal interaction can have many benefits compared to unimodal interaction. It may bring more bandwidth to the communication and provide alternative modalities for the same tasks, for example in the case of disabled users it provides speech based and haptic alternatives for graphical elements. Unfortunately, multimodal systems are sometimes designed based on one main modality, with the other modalities simply added on top. As handling several modalities together, may result in cognitive overload and reduced usability, especially in demanding usage situations that arise in mobile use (Turunen, Hakulinen, Kainulainen, Melto, & Hurtig, 2007).

The speech and visual modes are the most commonly used communication methods in information dissemination and perception process of human–human interaction. Addition of new modalities not only increases the bandwidth of communication, but also resolves the ambiguity in the primarily communicated message. The resolution of ambiguity in one mode of signal can be complemented by the other mode of signal. The best examples are using visual information to understand ambiguous speech (lip tracking for improving the accuracy of speech recognition).

Multimodal systems represent a new class of user-machine interfaces, different from standard WIMP interfaces. WIMP—"Windows Menu, Icon and pointing device"—is a style of Human Computer Interaction. The primary benefit of this style of system is to improve the HCI by enabling better usability for non-technical people, both novice and power users. The Multimodal system differs from WIMP by emphasizing the use of richer, natural ways of communication. Hence, the objectives of multimodal interfaces are to support and accommodate user's perceptual and communicative capabilities; and also to integrate computational skills of computers in the real world, by offering more natural ways of interaction to humans (Dumas, Lalanne, & Oviatt, 2009). The evolution of speech technologies and computer vision (gesture) technologies provides the way to implement naturalness in man–machine interaction. The component of a typical multimodal system is given in Fig. 1.

A typical Multimodal message has candidate elements from various modalities, which is defined as the dimensionality of multimodal signal. Consider speech, gesture, olfaction, taste, haptic and input via conventional input devices (keyboard, mouse) as various input modalities and candidate elements of these modality sets are

Speech {any spoken meaning full units}
Gestures {certain visual patterns generated by the human/ external object}
Haptic {touch input} e.g. Touch at a coordinate location 300,250
Olfaction {any smell}
Taste {any taste}
Conventional Input {input from Keyboard, mouse, joystick etc.}

Comprehension of a multimodal signal will be an appropriate synchronization of various elements drawn from the above said modality set. Dimensionality of a multimodal signal is defined as the number of the participating modality set. The following example gives detailed explanation of multimodal signal.

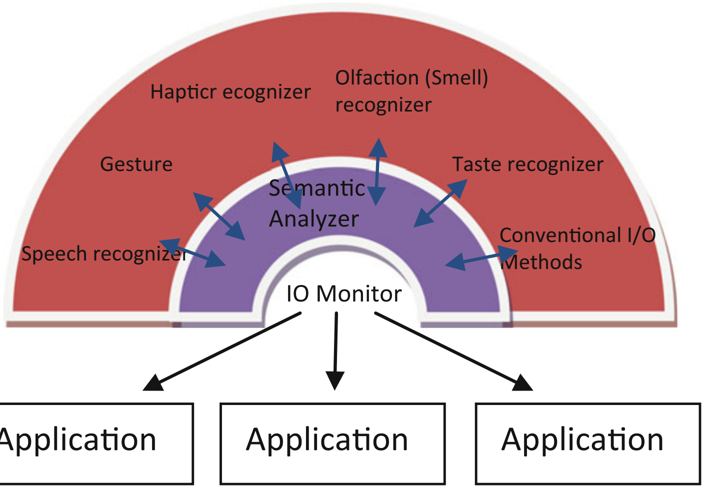


Fig. 1 Component diagram of multimodal interface system

3.1 A Sample Use Case: Multimodal Interface for Interacting with Desktop

A sample multimodal use case is simulated for interacting with the desktop. The user can communicate with the system through speech and hand gestures. Simple file operations like copy, delete etc. are considered as use cases here. The user issues a command to the system through speech "*Copy this file to that folder*" and gestures for the source and destination corresponds to the deictic "*this*" and "*that*" in the utterance. Here the user points twice, where the first pointing gesture is for the source and second one is for the destination. The generated Multimodal message signal has candidate elements of speech and gestures. The sequencing diagram for the above discussed scenario is shown in Fig. 2. The number of candidate modality sets in the communicated multimodal signal defines the input dimensionality of multimodal signal.

In the above example there are two input modalities involved so the dimension of input modality is two. The horizontal axis represents the time and the vertical axis represents the various input modalities involved. Here the speech signals started at time T_0 and ended at T_1. In between the speech event the gesture events *e1* and *e2* also took place. Pointing the source file event e1 happened in the interval T_a and T_b and the gesture which corresponds to pointing the destination folder happened in the interval T_c and T_d.

Here both gestures are identical in structure i.e. first one may be Pointer (200,175) the icon location of the source file to be copied and Pointer(100,230) the location of the destination folder where the file has to be placed. But the chronological ordering of these two gestures is mandatory as they specify the source and destination folders sequentially.

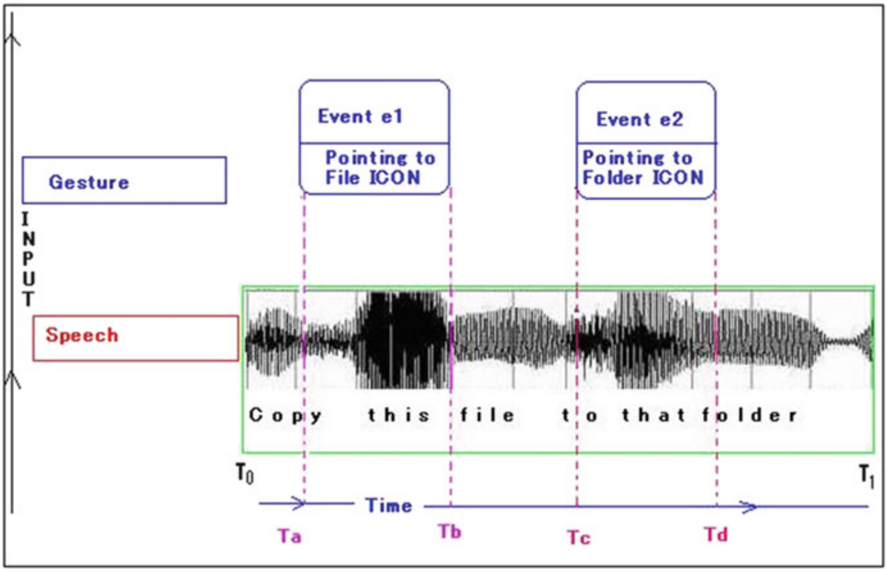


Fig. 2 Sequence diagram with occurrence of speech and gesture

4 Components of Multimodal Interface

Multimodal interface is composed of input modules, output modules and interaction manager (W3C, 2003). The input module is responsible for collecting input from the user and forwarding it to the interaction manager for processing. The processed input signal will be given to the user via output module. Speech, gestures (human body gestures, pen, handwritten gesture), olfaction, taste, haptic and input from conventional devices are the components of input module. The most popular output modalities are in audio and video formats. Interaction manager plays an important role in synchronization of various signals from independent sources. Interaction manager has two major components, which analyzes the semantics of the communicated message and manages the input output functionality. Semantic analyzer will analyze the meaning of the communicated message and I/O monitor will manage the input output functionality for interacting with the application program.

4.1 Speech

Speech is the most prominent mode of communication in human–human interaction. As we have discussed earlier, human prefers speech based interaction with the machine too because of ease of use. Automatic speech recognition is being a challenging research problem for past few decades. However the domain based speech recognition systems are available now with a reasonably good performance. High accuracy speaker independent speech recognition with emotion identification is still a research problem. There are many matured speech recognition systems/ frameworks available where you can plug the desired language and acoustic models. CMU's Sphnix-4 is one of the widely used open source system for speech recognition. User can create his own acoustic and language models and plug in to the sphinx speech engine (http://cmusphinx.sourceforge.net/sphinx4/). There are many commercially available systems like Dragon naturally speaking, IBM's ViaVoice etc.

Considering the usability point, speech recognition system can be classified into two types, namely, small vocabulary/large users and large vocabulary/limited users. The small vocabulary program is perfect for automated answering on the telephone. It can identify different accents and variations in speech patterns. Sensibly, it is restricted to basic menu and generic responses. In larger vocabulary program, the system can identify more words with greater accuracy but it can identify fewer users (http://www.accuconference.com/resources/speech-recognition.aspx).

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. The analog signal received through microphone will be digitized and sent to the pre-processing module. The system filters out unwanted noises, categorizes the frequency levels and normalizes the sound. Since people speak at different speeds, this should be aligned to the speed stored in the computer. It is really difficult to identify exact phoneme boundary in the spoken unit (word). Hence comparing the phonemes in the context to other phonemes is really a computationally intensive task which involved complex mathematical and statistical methods.

In the initial stages of speech recognition studies, speech recognition was a mere signal processing problem where language attributes were not considered. Later statistical and mathematical methods were incorporated which considerably improved the recognition accuracy. Statistical models like Hidden Markov models and neural networks are extensively used to solve various non-linear pattern recognition problems. In this chapter we have considered a small vocabulary speech recognition system with semantic analyzer for explaining the multimodal concept.

4.2 Gesture

Gesture is one of the ancient modes of communication before the evolution of spoken languages. The idea of interacting with machines via gestures is also a four decade old problem. Gesture recognition enables humans to communicate with the machine (HMI) and interact naturally without any mechanical devices. Gestures have long been considered an interaction technique that can potentially deliver more natural, creative and intuitive methods for communicating with computers (Sreekanth, Gopinath, Supriya, & Narayanan, 2011). The gestures are commonly used in human-human interaction and plays a major role in communication when the participants are unable to speak, or the situation does not allow the participant to speak etc. The gestures also play as an offset-input to the other mode of communication, for example gesture and speech are co-expressive and they form a part of rich human conversational features (Quek, 2003; Quek et al., 2002). If we look at the evolution of gesture based interaction initially there were glove based devices, but they lacked the naturalness factor as they had introduced an additional hardware constraints on the user. The models employed for gesture processing are either 3D models or image based processing. The former lacks the computational efficiency and the simplicity compared to other. In the image based processing method there are several techniques based on color, contour and correlation for identifying gestures (Chen, 2008).

Gestures are mainly classified into two based on the origin. Gestures are generated with the help of external objects like pen, coloured objects or with parts of the human body. Generally gesture recognition process means ability to track and identify the movements of user's body parts, which plays a major role in gesture based communication. As far as human user is concerned the hands, face, lips and heads are the major organs that take part in gesture based interaction. As we discussed earlier the device or object based gesture lacks the naturalness, focus is shifted more towards the computer vision based gesture processing. In gesture recognition technology, a camera reads the movements of the human body and communicates the data to a computer that uses the gestures as input to control devices or applications. Complex image processing techniques are used for extracting the meaning of the communicated message.

Gestures can be classified into three based on their functionality as

1. Symbolic gestures which convey a single meaning like American Sign Language gestures
2. Deictic gestures are pointing gestures which are mainly used in HCI
3. Iconic gestures which convey information about the size, shape or orientation of the object. Iconic gestures cannot be understood without accompanying speech.

Gesture recognition finds application in the emerging gaming scenarios, as it highly enhances the entertainment experience as well as simplifies the human–computer interaction. SixthSense is a wearable gestural interface device by Pranav Mistry, a PhD candidate in the Fluid Interfaces Group at the MIT Media Lab. SixthSense augments the physical world around us with digital information and lets us use natural hand gestures to interact with that information (http://www.pranavmistry.com/projects/sixthsense/).

Pen based gestures are another form of input mechanism for computers. Computer interaction through the drawing of symbols with a pointing device like a pen is taken as an input pen gesture. It provides an alternative to the direct manipulation or point and click method of interacting with a computer, allowing gestures or strokes of the pen to be translated into direct commands. The current technology is advanced in such a way that the user can even write the commands directly to the console.

4.3 Haptic

The word "haptics" refers to the capability to sense a natural or synthetic mechanical environment through touch. Haptic technology or haptics refers to the technology that connects the user to a computerized system by the application of sense of touch such as force, vibration or motion. Haptic information is a combination of tactile information as well as kinesthetic information. Kinesthesia is the ability to perceive one's body position, movement and weight. Haptic interfaces generate mechanical signals that stimulate human kinesthetic and touch channels and thus enable the human–machine communication through touch in response to user movements. The applications of haptic devices are mainly in mobiles, games, medicine, robotics etc. Haptics provides improved usability, enhanced realism and restoration of mechanical feel (http://www.cim.mcgill.ca/~haptic/pub/VH-ET-AL-SR-04.pdf).

4.4 Artificial Nose and Tongue (e-Nose, and e-Tongue)

Electronics nose and electronics tongue functionally imitate human nose and tongue for detecting the smell (odour) and taste respectively. The sensing system in the both the devices can be an array of several different sensing elements (e.g., chemical sensors), where each element measures a different property of the sensed chemical, or it can be a single sensing device (e.g., spectrometer) that produces an array of measurements for each chemical, or it can be a combination. Each chemical vapour/substance presented to the sensor array produces a signature or pattern characteristic of the vapour or substance. By presenting many different chemicals to the sensor array, a database of signatures is built up. Like any pattern classification system database of labeled signatures for various vapours or gases or substance should be maintained for training (Keller, Kangas, Liden, Hashem, & Kouzes, 1995).

4.5 Semantic Analyzer

As we have discussed, multimodal signal will be a combination of signals from independent sources which should be logically synchronized to achieve the goal of communication. Human cognition is capable of analyzing the signals from various sensors (eye, ear, skin, tongue, and nose) as well as synchronizing them logically. Functionally, the semantic analyzer module will mimic the human cognition system. Signals from individual sensors are recognized and it will to be sent to the semantic analyzer for understanding the message. Multimodal grammar has to be defined to understand and parse the multimodal signal. Before analyzing the meaning of the message, the signals from various sources has to be combined and it has to be represented in the system understandable format. Generation of a multimodal signal is discussed in following section. The recognized signals from various sensors are converted and represented in XML format. The recognized words from the speech are embedded in the tag < s>,</s>. The tag < w1>,<w2 > contains the recognized words in the communication. Similarly for gesture, the recognized gesture will be converted and embedded in between the tags < G > <w1 > . . ...</w1>, <w2 > . . ..</w2 > . . .. </G>, where w1, w2 are the recognized words in the gesture vocabulary. The temporal information will be encoded with each event for proper synchronization. Semantic analyzer will convert the communicated messages from user vocabulary space to system vocabulary space (Sreekanth, Supriya, Girish, Arunjith, & Narayanan, 2008).

4.6 Input Output Monitor (I/O Monitor)

The multimodal sentence generated by semantic analyzer will be given to I/O monitor for issuing necessary signals to perform desired operations on the application program. Error or ambiguity in the input signal will be communicated to user via the feedback module. For example, if the user issued a command via speech *"copy this file",* but the gesture corresponds to the deictic *"this"* is not given, i.e., gesture is missing, and then it should be notified to the user. The semantic analyzer will seek signals from the input modules to substitute the word "this" in speech. If it is not found in the stipulated time, the same will be notified to the user. The user can give the corresponding gesture if possible during this notification period. If nothing is received during a given interval, the system will go back to idle or safe mode so that it can listen to new input. If an invalid input pattern is recognized, the system will not respond to it.

4.7 Output Modality

While interacting with the machine, the interaction cycle will get complete only if the user gets a valid output, in the desirable format. The preferable output formats are audio visual signals. The current display technology along with speech synthesizer can provide the output in user desirable format. Advancements in the 3D display technology actually added to the effective ways of information display. 3-D displays are really effective for product design, complex scientific simulations, DNA/ chemical structure analysis, aircraft design and gaming. Coupling of speech synthesizer with 3D display will virtually create a real-time world for a better user experience. Discussion of output modalities is presently not considered in the scope of this chapter.

5 Implementation of Multimodal Interface

There are plenty of interface mechanisms available for HCI and the effort to couple the various modalities for enriching the interaction mechanism with computers lead to the development of Multimodal Frame work. It has been proven that adding more modalities always improves the quality of interaction and also helps to resolve the ambiguity in communications. By mimicking the way of human–human interaction via speech, researchers are more interested in incorporating Automatic Speech Recognition system, for natural and easy way of man–machine interaction. For example adding gesture recognition improves the quality of speech recognition also. The ambiguity in decoding an input speech signal can be resolved by accepting the gestures so that the more accurate word or sentence can be picked up from the vocabulary list. The lips reading or lips modeling is one among the several other gesture offset methods to improve the recognition accuracy of a speech recognition system (Cetingul, Erzin, Yemez, & Tekalp, 2006). In addition to that, hand and head gestures also improve the performance of interaction with computers by coupling with the Automatic Speech Recognition system. Simply adding or overloading the modalities one up on the other will not improve the quality of interaction. More systematic and intelligent models are required to couple various modalities and identify the semantics of the communicated message.

The typical implementation scenario for interacting with desktop systems for normal operations like simple file operations (open, close, copy, delete, move etc.) and other operations like search for a key word, zooming, copy a selected image location in the image, selection etc are discussed here. The present case study is based on a system which accepts input through speech, hand gesture and pen gesture.

5.1 Multimodal Signal Representation

As we have discussed earlier a typical multimodal signal have candidate elements from various participating modalities e.g. from speech, gesture, olfaction etc. Physically all these signals are independent in nature but semantically they are coupled. To process this signal from independent sources, this has to be synchronized properly. Moreover in real time scenario, these individual signals have a high temporal relation. The combined multimodal signal can be represented as an XML file or a markup language (W3C, 2009). Consider a scenario for deleting a file from the system, user can choose various ways since various modalities are incorporated. Suppose a user says "*Delete this file*" followed by a gesture pointing to the required icon using hand or finger. The string generated for processing will have the candidate members from speech vocabulary as well as gesture vocabulary. In this example the words "*delete*", "*this*" and "*file*" are the members of the set of speech vocabulary set *S* which are recognized by a speech recognition system. Similarly the gesture recognition system will return a string with location reference and temporal information to a semantic analyzer. An example of a gesture vocabulary data base is given in Fig. 3. Depending on the recognized gesture, corresponding strings will be generated as per the database.

The generated multimodal message for "*delete this file*" is given below (Fig. 4).

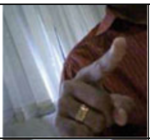
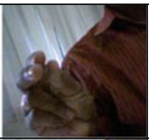
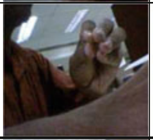
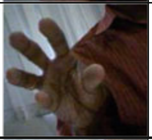
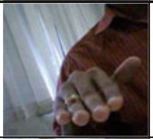
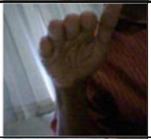
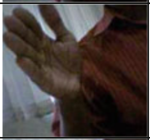
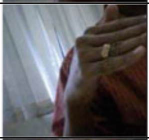
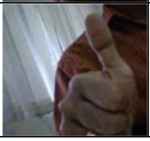
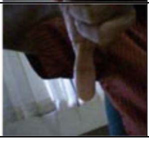
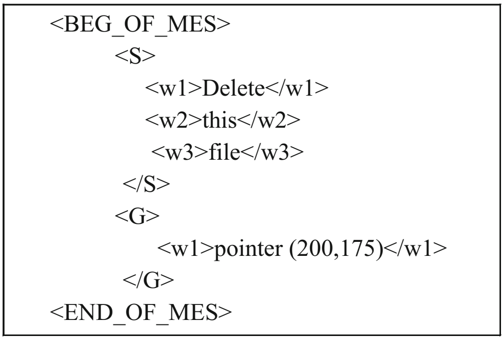
	Pointer (x,y)		Select Icon at the Present location of the courser
	Zoom Out The Selected area		Zoom the selected area
	Page Down (Hand Expected to move from top to bottom		Page up(Hand Expected to move from bottom to top
	Previous / Move right Hand moves left to right		Next / Move left Hand moves right to left
	OK		Cancel

Fig. 3 Issuing commands through gesture



```
<BEG_OF_MES>
        <S>
            <w1>Delete</w1>
            <w2>this</w2>
             <w3>file</w3>
         </S>
        <G>
              <w1>pointer (200,175)</w1>
         </G>
<END_OF_MES>
```

Fig. 4 Multimodal Message

5.2 Message Parsing and Understanding

Once multimodal message is generated, it will be sent to the semantic analyzer for understanding the meaning of the communicated message. Consider the message "Delete this file" followed by a gesture to point the required icon through hand or finger, the operational keyword "delete" is the significant part of this multimodal message. This can be represented as DELETE arg1, [arg2, arg3,. . .] (here at least one argument is must and others are optional) where DELETE is the operational key word.

In the above message the word after "*delete*" is "*this*", which is an deictic whose resolution is done using a multimodal reference resolution parse tree. The three level multi modal reference resolution parse tree for the above string can be represented as follows. This parse tree has a root node with Multimodal message and has three or more immediate child nodes depending on the number of input modalities involved in a typical multimodal message as shown in Fig. 5.

Here the circled word "this" is an deictic and after applying the reference resolution the antecedent equivalent to the deictic "this" is "pointer (200,175)". In the above example the pen gesture input is not present so it is marked as ε (null).

The above discussed is a simple multimodal reference resolution parse tree. Consider a multimodal parse tree which is highly time dependent which means the structure of the tree depends on the time of occurrence of input event. The user issues a command to the system as speech "*Copy this file to that folder*" and gestures for the source and destination corresponds to the deictic "*this*" and "*that*" in the utterance. Here the pointing gesture is used twice where the first points to source and second to destination. The corresponding multimodal string is generated by the system considering the time of occurrence of the event and the corresponding time stamp is added. The sequencing diagram for the above discussed scenario is shown in Fig. 2. The chronological ordering of these two gestures is mandatory as they specify the source and destination folders sequentially. So generation of the multimodal string should be with respect to the temporal aspects of the event. The multimodal message with temporal information for the

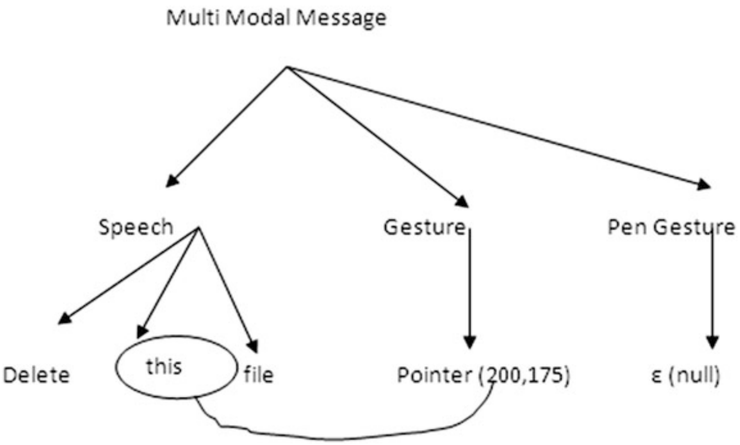
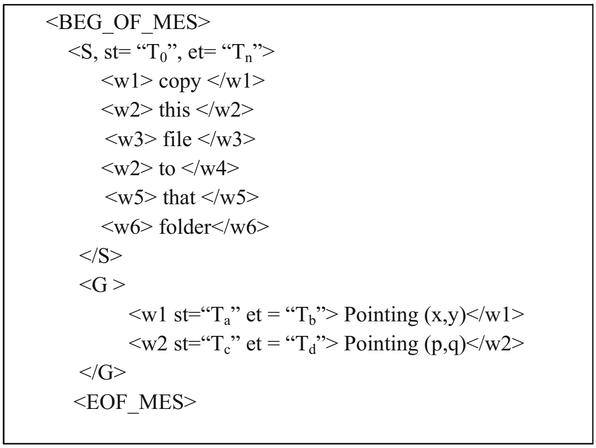


Fig. 5 3-level multimodal reference resolution parse tree



```
<BEG_OF_MES>
   <S, st= “T0”, et= “Tn”>
       <w1> copy </w1>
       <w2> this </w2>
        <w3> file </w3>
       <w2> to </w4>
        <w5> that </w5>
       <w6> folder</w6>
    </S>
    <G >
           <w1 st=“Ta” et = “Tb”> Pointing (x,y)</w1>
           <w2 st=“Tc” et = “Td”> Pointing (p,q)</w2>
    </G>
   <EOF_MES>
```

Fig. 6 Multimodal message with Temporal Information

above tasks can be represented as follows. In this example “st” and “et” are start time and end time respectively (Fig. 6).

From the above multimodal message the operational keyword, the time stamp and the attributes associated with various tags can be found by performing left to right parsing. The multimodal dependency parse tree for the above example is given in Fig. 7.

Consider a case in which pen-gesture is also involved.

“Open this file and search for” rose

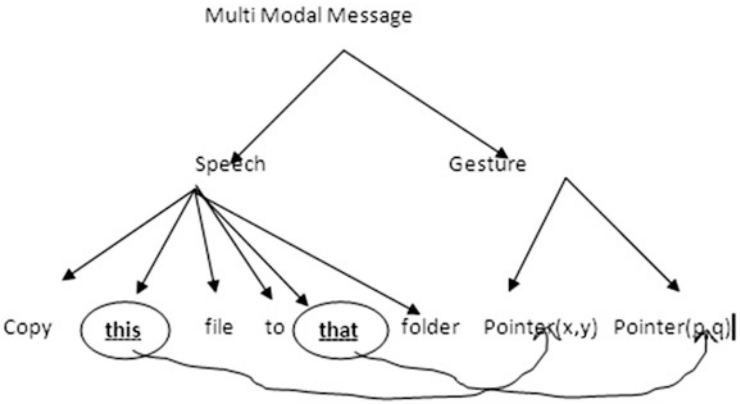


Fig. 7 Multi modal dependency parse tree

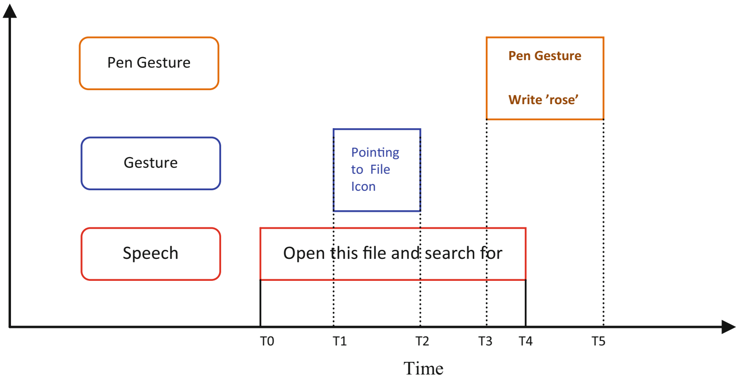
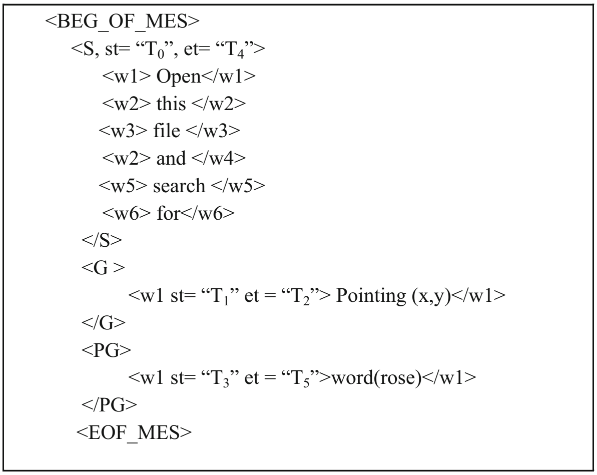


Fig. 8 Sequence diagram with the occurrence of speech gesture and pen gesture

In this case three modalities are involved, they are speech, finger gesture and pen gesture. The speech recognition system will return the string "*Open this file and search for*", and finger gesture will give the value of the location of "*this*" in the string. The pen gesture will recognize the stroke and will return the word "*rose*". The sequence diagram is shown in Fig. 8.

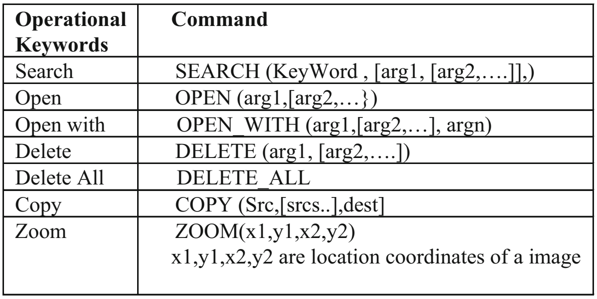
The inclusion of pen gesture in multimodal message is indicated by the tag < PG > </PG>. The multimodal message is represented in Fig. 9.

In this message the operational keyword is "*search*" which is defined as SEARCH (arg1, pattern1,[pattern2, pattern3,. . ..]) where arg1 is the file to be searched and pattern1 is the pattern to be searched in the file specified in arg1. The operational keyword vocabulary for multimodal desktop interaction is listed in Fig. 10



```
<BEG_OF_MES>
   <S, st= "T0", et= "T4">
      <w1> Open</w1>
      <w2> this </w2>
      <w3> file </w3>
      <w2> and </w4>
      <w5> search </w5>
      <w6> for</w6>
   </S>
   <G >
        <w1 st= "T1" et = "T2"> Pointing (x,y)</w1>
   </G>
   <PG>
        <w1 st= "T3" et = "T5">word(rose)</w1>
   </PG>
   <EOF_MES>
```

Fig. 9 Multimodal message with many modalities



Operational Keywords	Command
Search	SEARCH (KeyWord , [arg1, [arg2,….]],)
Open	OPEN (arg1,[arg2,…})
Open with	OPEN_WITH (arg1,[arg2,…], argn)
Delete	DELETE (arg1, [arg2,….])
Delete All	DELETE_ALL
Copy	COPY (Src,[srcs..],dest]
Zoom	ZOOM(x1,y1,x2,y2) x1,y1,x2,y2 are location coordinates of a image

Fig. 10 Commands through speech

6 Multimodal Interaction in Consumer Electronics

6.1 Mobile and Hand-Held Devices

As a result of increasingly capable networks, devices, and speech recognition technology, the number of existing multimodal applications, especially mobile applications, is rapidly accelerating. Especially while using mobiles, user has very limited access to input space, and it is a cumbersome effort to use mobile keypad for text compilation. The developments in the haptic and speech based technologies have given a different dimension for interaction with the mobile systems. Speech offers one-handed and hands-free operation (W3C n. d.).

A related effort has recently been completed in the W3C by the HTML Speech Incubator Group (HTML Speech XG). The focus of the XG was developing proposals for accessing speech recognition and speech synthesis from HTML5 browsers, and Voice Search and Speech Command Interfaces are possible use cases for these technologies in the browser. However, the XG did not attempt to address modalities other than speech, such as handwriting, emotion, or the wide variety of present and future input modalities. Similarly, it didn't attempt to address non-browser contexts. In contrast, the Multimodal Architecture provides a generic framework for modality integration and control. Speech in the browser can be seen as a special case of modality integration covered by the MMI Architecture (W3C, 2014).

Dynamic gestures like waving and fist hitting gesture recognition are integrated by Microsoft Kinet, Sony PSP, etc. in their consoles. Dynamic hand-shape recognition is addressed in American Sign Language recognition in game development for deaf children (Brashear et al., 2006).

6.2 Home Appliances, e.g., TV, and Home Networks

There has been a tremendous effort from the players of consumer electronics industries for incorporating the multimodal interface to interact with the electronic gadget. Multimodal interfaces are expected to function as a remote control for home appliance and entertainment systems. The smart TV introduced by Samsung is a good example for this, where speech and gesture based interaction provides a hands free interaction. The fusions of various modalities for interacting with the system are the remarkable changes that we can observe in the forthcoming versions of electronic gadgets and home appliances. The gesture pendant (Starner, Auxier, Ashbrook, & Gandy, 2000) is a wearable device for the control of home automation systems via hand gestures. This solution has many advantages over traditional home automation interfaces as it can be used by those with loss of vision, motor skills, and mobility.

6.3 Enterprise Office Applications and Devices

Multimodal has benefits for desktops, wall mounted interactive displays, multifunction copiers and other office equipments which offer a richer user experience and the chance to use additional modalities like speech and pens to existing modalities like keyboards and mice. W3C's standardization work in this area should be of interest to companies developing client software and application authoring technologies, and who wish to ensure that the resulting standards live up to their needs.

Dialogue-Assisted Visual Environment for Geoinformation (DAVE_G) (Rauschert et al., 2002) that uses different interaction modalities, domain knowledge and task context for a dialog management that supports collaborative group work with GIS in emergency management situations. DAVE_G, a multimodal, multiuser geographical information system (GIS), has an interface that supports decision making based on geospatial data to be shown on a large-screen display. Interactions with Robot assistants (Rogalla, Ehrenmann, Zöllner, Becher, &

Dillmann, 2002) will be effective if they resemble natural human dialogue with gestures and speech.

6.4 Intelligent IT Ready Cars

With the emergence of dashboard integrated high resolution colour displays for navigation, communication and entertainment services, W3C's work on open standards for multimodal interaction should be of interest to companies working on developing the next generation intelligent car systems. Ford Model U Concept Vehicle (Pieraccini, Dayanidhi, Bloom, Dahan, & Phillips, 2003) was first shown at the 2003 North American International Auto Show in Detroit. The system, including a touch screen and a speech recognizer, is used for controlling several non critical automobile operations, such as climate, entertainment, navigation, and telephone. The prototype implements a natural language spoken dialog interface integrated with an intuitive graphical user interface, as opposed to the traditional, speech only, command-and-control interfaces deployed in some of the vehicles currently on the market. Hyundai has also come up with their concept car HCD-14 with integrated eye-tracking and 3-D hand-gesture recognition to satisfy driver commands.

6.5 Medical Applications

Mobile healthcare professionals and practitioners of telemedicine will benefit from multimodal standards for interactions with remote patients as well as for collaboration with distant colleagues. Wheelchairs, as mobility aids, have been enhanced through robotic/intelligent vehicles (Kuno, Murashima, Shimada, & Shirai, 2000) able to recognize hand-gesture commands. "Gestix",(Wachs et al., 2008) a vision-based hand gesture capture and recognition system that interprets in real-time the user's gestures for navigation and manipulation of images in an electronic medical record (EMR) database. Navigation and other gestures are translated into commands, based on their temporal trajectories, through video capture. A novel human–machine interface, called "FAce MOUSe"(Nishikawa et al., 2003), for controlling the position of a laparoscope was designed which allows nonintrusive, nonverbal, hands off and feet off laparoscope operations, which is more convenient for the surgeon.

7 Conclusion

The advances in information and communications technologies, computing and proliferation in use of internet have been one of the biggest contributors to the media convergence phenomenon. It helped to bring various modes of communication like audio, video, text based communication etc. under a single platform.

It introduced a different perspective of information sharing over conventional media like newspaper, radio and television. All those media were unidirectional, monotonous and had very limited scope for user interaction. The emergence of online version of newspaper, converged with social media like facebook, twitter, etc. provided a platform for people to interact and communicate their thoughts on a topic effectively. It is found that the convergence of various communication technologies have wider acceptance in the present society. The standard input output mechanisms for man machine interaction created a bottleneck in effective utilization of the full potentials of convergence of communication technologies and media. This can be overcome by the convergence of various modalities. The multimodal interaction provides varieties of input modalities like speech, gesture, haptic, etc. along with standard input output mechanisms.

Multimodal interaction framework provides a natural way to interact with the computers and electronic systems. The notion of converging different electronic and mobile devices accelerate the necessity of investing more time of researchers, to bring naturalness in human–machine interaction. By mimicking the human information perception and dissemination model we can design systems that are intelligent and effectively user friendly. From a human computer interaction point of view it is interesting to look at the various multimodal ways people interact with the environment and each other and to design systems that are sensitive to what the user wants without having been given explicit commands. The advancement of multimodal interaction paves the way to progress towards interfaces that are capable of human like perception.

References

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics, 14*(3), 262–270.

Brashear, H., Henderson, V., Park, K., Hamilton, H., Lee, S., & Starner, T. (2006). American Sign Language recognition in game development for deaf children. In *Proceedings of ACM SIGACCESS Conference on Assistive Technologies (Portland, OR, Oct. 23–25).* ACM Press, New York, 2006, pp. 79–86

Cetingul, H. E., Erzin, E., Yemez, Y., & Tekalp, A. M. (2006). Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal Processing, 86*, 3549–3558. Science direct, Elsevier.

Chen, Q. (2008). PhD thesis, Canada.

Dumas, B., Lalanne, D., & Oviat, S. (2009). *Human machine interaction* (pp. 3–26). Berlin: Springer.

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In D. Lalanne & J. Kohlas (Eds.), *Human machine interaction: Research results of the MMI Program* (LNCS, Vol. 5440, pp. 3–26). Berlin: Springer.

Flippo, F., Kerbs, A., & Marsic, I. (2003). A framework for rapid development of multimodal interface. In *Proceedings of ICMI'03*, *Vancouver, British Columbia, Canada. November 5–7, 2003*. http://www.caip.rutgers.edu/disciple/Publications/icmi2003.pdf

Ito, E. (2001). Multi-modal interface with voice and head tracking for multiple home appliances. In P*roceedings of INTERACT2001 8th IFIP TC.13 Conference on Human-Computer Interaction*, pp. 727–728.

Keller, P. E., Kangas, L. J., Liden, L. H., Hashem, S., & Kouzes, R. T. (1995). Electronic noses and their applications. In *IEEE Northcon/Technical Applications Conference (TAC'95) in Portland, OR, USA on 12 October 1995*.

Krahnstoever, N., Kettebekov, S., Yeasin, M., & Sharma, R. (2002). A real-time framework for natural multimodal interaction with large screen displays. In *International Conference on Multimodal Interfaces Proceedings of the 4th IEEE International Conference on Multimodal Interfaces-2002*, pp. 349–354.

Kuno, Y., Murashima, T., Shimada, N., & Shirai, Y. (2000). Intelligent wheelchair remotely controlled by interactive gestures. In *Proceedings of 15th International Conference on Pattern Recognition (Barcelona, Sept. 3–7, 2000)*, pp. 672–675.

Maat, L., & Pantic, M. (2007). Gaze-X: Adaptive affective multimodal interface for single-user office scenarios. In T. S. Huang, A. Nijholt, M. Pantic, & A. Pentland (Eds.), *Artificial intelligence for human computing* (Lecture Notes in Computer Science, Vol. 4451, pp. 251–271). Berlin: Springer.

Nishikawa, A., Hosoi, T., Koara, K., Negoro, D., Hikita, A., Asano, S., et al. (2003). FAce MOUSe: A novel human-machine interface for controlling the position of a laparoscope. *IEEE Transactions on Robotics and Automation, 19*(5), 825–841.

Pieraccini, R., Dayanidhi, K., Bloom, J., Dahan, J.-G., & Phillips, M. (2003). A Multimodal Conversational Interface for a Concept Vehicle. In *Eurospeech 2003*.

Quek, F. (2003). The catchment feature model for multimodal language analysis. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set*.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X. F., Kirbas, C., et al. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on ComputerHuman Interaction (TOCHI), 9*(3), 171–193.

Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., & MacEachren, A. M. (2002). Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems (McLean, VA, Nov. 8–9)*. ACM Press, New York, 2002, pp. 119–124.

Reitter, D., Panttaja, E. M., & Cummins, F. (2004). UI on the Fly: Generating a multimodal user interface. In *Human Language Technology Conference Proceedings of HLT-NAACL 2004*. pp. 45–48.

Rogalla, O., Ehrenmann, M., Zöllner, R., Becher, R., & Dillmann, R. (2002). Using gesture and speech control for commanding a robot assistant. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (Berlin, Sept. 25–27, 2002)*, pp. 454–459.

Sreekanth, N. S., Gopinath, P., Supriya, N. P., & Narayanan, N. K. (2011). Gesture based desktop interaction. *International Journal of Machine Intelligence, 3*(4), 268–271. ISSN: 0975–2927, E-ISSN: 0975–9166.

Sreekanth, N. S., Supriya, N. P., Girish, K. G, Arunjith, A, & Narayanan, N. K. (2008). Performing operations on graph through multimodal interface: An agent based architecture. In *Proceedings of ICADIWT 2008. First International Conference on the Applications of Digital Information and Web Technologies*, pp. 74–77.

Sreekanth, N. S., Supriya, N. P., Thomas, M., Haassan, A., & Narayanan, N. K. (2009). Multimodal interface: Fusion of various modalities, multimodal interface fusion of various modalities. *International Journal of Information Studies, 1*(2), 131–137.

Starner, T., Auxier, J., Ashbrook, D., & Gandy, M. (2000). The gesture pendant: A self-illuminating, wearable, infrared computer-vision system for home-automation control and medical monitoring. In *Proceedings of the Fourth International Symposium on Wearable Computers (Atlanta, Oct. 2000)*, pp. 87–94.

Thomas, M., Hassan, A., Sreekanth N. S., Supriya, N. P. (2008). Multimodal interface to desktop. In *Proceedings of International Conference on Opensource Computing-2008, NMAMIT-Nitte Mangalor, Indiae,* 2008, pp. 26–30.

Turunen, M., Hakulinen, J., Kainulainen, A., Melto, A., & Hurtig, T. (2007) Design of a rich multimodal interface for mobile spoken route guidance. In *Proceedings of Interspeech 2007–Eurospeech*, pp. 2193–2196.

W3C. (2003). Multimodal interaction frame work. http://www.w3.org/TR/mmi-framework/

W3C. (2009). EMMA: Extensible multimodal annotation markup language http://www.w3.org/TR/emma/

W3C. (2014). Multimodal interaction working group charter. http://www.w3.org/2011/03/mmi-charter

W3C. (n. d.). Multimodal Access http://www.w3.org/standards/webofdevices/multimodal

Wachs, J., Stern, H., Edan, Y., Gillam, M., Feied, C., Smith, M., et al. (2008). A hand-gesture sterile tool for browsing MRI images in the OR. *Journal of the American Medical Informatics Association, 15*(3), 321–323.