Integration of Multimodal Interaction as Assistance in Virtual Environments

[+]Kiran Pala
[+]Sachin Joshi
Ramnaresh Pothukuchi
[+]Suryakanth V Ganagashetty
[+]International Institute of Information Technology Hyderabad, Hyderabad India
{kiranap, rushtosachin, prnaresh.prnaresh} @ gmail.com, svg@iiit.ac.in

Abstract

This paper discusses the significance of the multimodal interaction in virtual environments (VE) and the criticalities involved in integration and coordination between modes during interaction. Also, we present an architecture and design of the integration mechanism with respect to information access in second language learning. In this connection, we have conducted an experiential study on speech inputs to understand how far users' experience of information can be considered to be supportive to this architecture.

1 Introduction

In the era of globalization education has taken a different path from the traditional space of teaching and learning. A nation's commerce and its market with respect to global changes, the implications of global needs are all demanding to policy makers for them to change educational policies accordingly.

In the above scenario, technology also has a significant role to play. Rapid development and use of new technologies have helped the human learning trajectory to take a complete shift from the classrooms to communities, personalization etc. There the e-learning and learning through technologies can be television and internet technologies, gadgets, tablets etc. E-learning has, with certainty, become a major entity in personal and community based learning. In addition, these days most of the classrooms have adapted itself to the concept of personalization with the help of technology assistive mechanisms in education, that is, the education sector shapes their face as e-education. Learning is a differently nuanced concept from teaching and instruction. Also, learning is a continuous interactive process; it cannot be a discretely developing process as we see that the definition of learning has shifted to a kind of entertainment activity. As shown in Pala (2012a) the interaction can be active or passive. We know that environments play a more significant role in facilitating the interaction with the learner as an interface between learners and communities. A learner receives information from environment through their senses such as visual, tactile and auditory with different activities which can directly affect their memory both declarative and procedural (Ullman, 2001). The activities blend with an interaction continuous with the environment. The tremendous development of information and communication technologies (ICTs) and its applications have made it possible to replicate the real environments on virtual platforms. The virtual environments facilitate the interaction for communication and information processing more or less like real environments.

Generally, whatever information is received through senses from the environment will be redirected to memory in the form of experience and then it is modulated with respect to the form of both production and perception states of a learner (Miller, and Johnson, 1976). But, whether the virtual environments provide an experience to the learners similar in these respects to the real environments is an answerable question to the community. Such experience is only possible when the multimodal interaction and assistance take place at the learner level from the environment. This communication, interaction and assistance can be peer-to-peer or person-to-person or peer-to-person etc. In any interaction or communication, assistance will be harnessed to rethink and rehearse

Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments, pages 8–12,
Jeju, Republic of Korea, 8-14 July 2012.

the information which has been received. Since the rehearsal process is directly related to memory, it helps learner to be fluent and expert in the related domain.

2 Assistance in Accessing of Information

The assistive technologies played an important role in the olden days and even today with emerging information technology it does play a significant role. The assistive technologies are used not just for those who have physical or cognitive difficulties, but even in areas of information access and representation. Some of the assistive technological devices include speech recognition, screen reader, touch screen, on-screen keyboard, word prediction, transliteration etc. In the virtual environment, the resources considered are image database, text database, and video or action data (Bartle, 2004). VE will support the learner in many aspects and would boost learners' abilities. VE would be helpful in many ways such as providing immediate feedback, experimentation, grabbing focus, furthering exploration, and would also suit the learner requirements.

Accessing information and assistance with an eye on the representation of the accessed information is highly interrelated in "*understanding the meaning*". For example consider a sound-meaning relationship, if a naïve learner wants to learn the sounds of a new language and listens to a sound like */a/*. Users may not be able to immediately utter the same sound. For that we will use "*/a/* for */apple/*". Sometimes we need to show the picture of /*apple*/ also to make the learner better "*understand the meaning*" i.e. pragmatic information of the condition or statement like shown in figure-1. This instance easily and naturally occurs in real environments. But it is possible in VE by integrating multimodal interaction (tactile, visual, auditory) as assistance for the purpose of representing the accessed content from the crawled database extracted from the web according to the level of the learner and requirements like games or only content or meaning etc.

However, in the personalization of learning and facilitation according to content representations, the expected naturalness is still far away from what occurs in real situations. In this paper, we propose a naïve architecture with the reference to Indian languages and the target group is second language learners (L2).

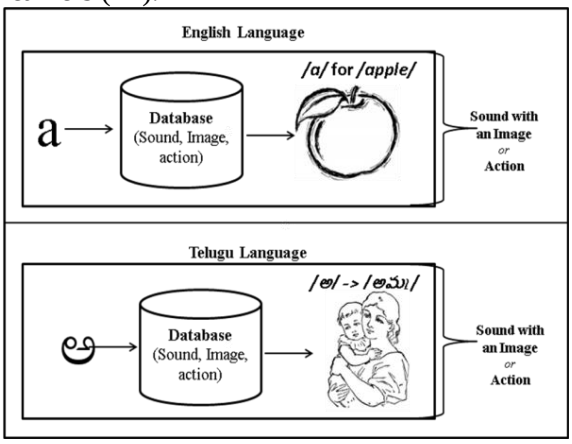

Figure 1: An example, environment required for understanding of the meaning with assistance.

3 Architecture

Here we discuss the details of the proposed architecture with the reference to each module's functions. This architecture mainly focuses on the integration of multimodal interaction as assistance to individuals who are adult learners. We have considered in the designing of this architecture learners' behavioral profiles, cognitive abilities and technological traits to pave the way for a more personalized interaction with the environment. Pala (2012a) has shown that these learners can be from any age group after the stage of puberty including even those who do not have much experience in use of virtual environments.

Input Devices: All these input devices like Automatic Speech Recognition (ASR) touch screen, mouse, keyboard etc. are interconnected to each other to ensure avoidance of information loss during non-linear interaction as well. Generally, adult individual learners move towards multitasking and non-linear interaction at a time and it has been expected that it should be a continuous activity. For example, the learner can give a speech input which is recognized by the ASR, at the same time the learner can utilize touch screen, keyboard and mouse to give another input. The input of the learner can be an alphabet or a word. Here we are dealing with sound-meaning relationships and conceptual structures and their types in languages at the lexical level. The multiple input facilities will assist the learner to provide versatile inputs of their own choice. It also has a

significant role in furthering or initializing learning in learners who have physical disorders. This combined interaction of the visual and tactile senses is directly connected to the procedural memory (Christiansen and Chater, 2008; Tomasello, 2008).

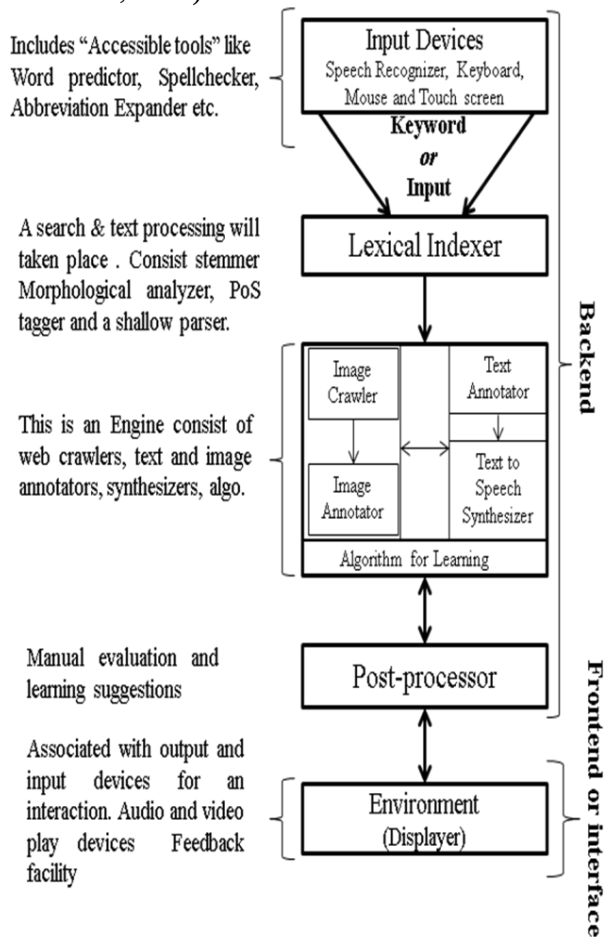

Figure 2: A Block diagram of Virtual Environment with Multimodal interaction as assistive.

Lexical Indexer: It is a kind of database with the linguistic categories and relations of each lexeme as has been discussed in Pala (2012b). It consists of a morphological analyzer and a stemmer. At the functional level it extracts the root word from the given input and verifies it in the indexer for its category and relationships in order to search for the same category-oriented examples and images from the web through crawlers. Additionally, the same keywords will be indexed again for ranking purposes of a specific learner. If a keyword is not available with the indexer, it sends the keyword directly to the web with a new index and later learns the relations and categories with the help of parts-of-speech taggers (POS) and shallow parses (Parser/Hindi, 2012; Akshar, Chaitanya, and Sangal, 1995).

An Engine: This **e**ngine consists of web crawlers for content resources, annotators, synthesizers (Text-to-Speech) and a predictive learning algorithm which has been built on self-organizing maps. Speech synthesizers receive information from the text annotator. The examples are provided in the form of phones, lexical items and sentences, it converts them into a signal form to speak it aloud when the learner requests.

Here annotators have a significant responsibility in handling information. In the process of building image annotators, we have used regular expressions for replacing the names. In addition, we have used wavelet transforms to verify the quality i.e. pixel depth, colors hue etc. of the image. Some other parameters like size and weight of the image have also been taken into account. Similarly, according to Pala (2011a) the text annotators have been constructed with an eye on parameters like removal of punctuations and special symbols etc. through an inclusion of the heuristic mechanism for anaphora references. The projection of video for action-related lexical items has been dealt with in the post-processing section.

Post-Processor: In this module we will have a verification process at initial stages, i.e., in the developmental state of the application a manual check up will be carried out along with auto verification process by the content developers who will look into the pragmatic and semantic aspects of example sentences, action videos and images very carefully. In the case of videos, the post-processing stage is more important in that when the input keyword contains a verb, making the action through image or text understandable is highly difficult. Thus, we have chosen the video form for lexical items related to action and motion. This categorical information will be received from the lexical indexer. The video codecs, definition of the video or animations quality, the length of the video and the mixture of audio clarity are very important parameters in selection and building of such action oriented contents.

In this paper we are dealing with the content representation modes but there is a similar significant role that mediates having a "*kind of content and presentation model for presenting to*

understand the meaning" in learning process. This will streamline the process of the constant review process by the domain experts as shown in Pala (2012b).

Displayer: It is a space to interact with the user or the leaner, i.e., it is an interface between the learner and the application. It is embedded with all interaction modes (input and output tools) which we have discussed above for the assistance purpose. It projects the output in all types of modes which affect different senses (visual and auditory) of the user on screen according to user input requests. The displayer is crucial as the learners get distracted and lose interest in learning if the size of the screen, projection and the level of pixel value are to be defined according to user requirements. This requires a meticulous design so that the users' attention and their rehearsal activity gravitate towards the learning content.

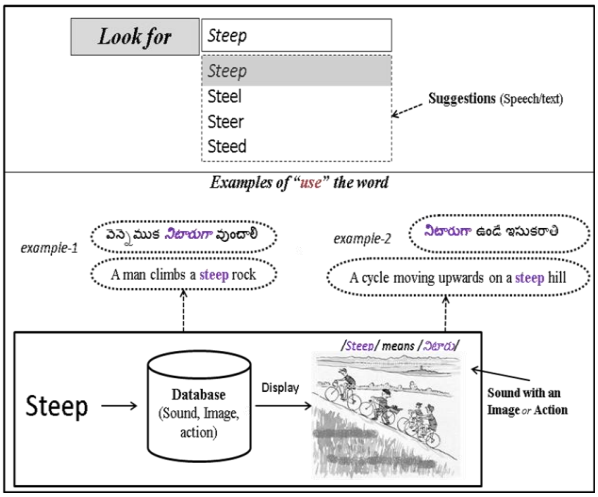

Figure 3: Example for Bilingual environment (English to Telugu)

Since this application is multilingual, the learner can make a request in any language. At this moment we have built an application for two major Indian languages and English. If, for example, a user asks for a meaning and use of the lexical item in English and their target language is Telugu, the "meaning" and "use" of the lexical items will be shown in what we see in figure-3 below. Native speakers generally look for the synonym for a "regular use" of a lexical item. We consider this factor to be of much importance and build a database which consists of the synonyms with their "regular use" as shown in the figure-4 below.

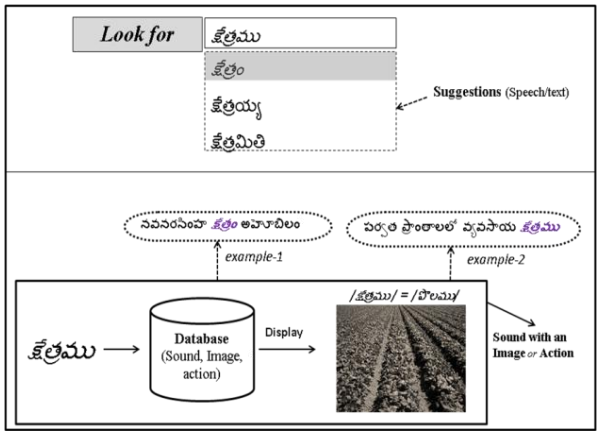

Figure 4: An example process of monolingual environment (Telugu to Telugu (Robert and Wyatt, 1956))

4 User Experience Study on Multimodal Interaction

To demonstrate the impact of multiple input modes on the quality of users' experience we have performed an experiential study to elicit users' perceptual inference- through speech and keyboard. We have built an English ASR using CMU Sphinx. For this we have used 1000 isolated words for the testing of the ASR which is used for training. The study was executed by providing the isolated words recorded by speakers. In this study we have passes since we would like to test user experience after the integration of the multimodal input mechanisms (here we have integrated a keyboard with ASR) to an individual computer. In the first pass the spoken word was decoded using the entire vocabulary of 1000 words given to the recognizer. Then the user was asked to type the first character of the spoken word. The words starting with that character were segregated. In second pass, the spoken word was decoded with only segregated words given as input vocabulary to ASR. As expected, the second pass decoding showed a major accuracy improvement because of reduction in search vocabulary size. The relative improvement in accuracy was 36.61% percent. The entire procedure has been designed in such a manner that each lexical item will be selected from a bag of lexical items. As the entire procedure is executed, significant parameters for evaluation of the responses from the participants are drawn up for further analysis. All users reported that they

were much more satisfied with multimodal items than with using speech recognition alone, since the system performs better with a minimal additional effort of pressing a single key. Not only accuracy but speed of the system was better.

5 Implications

Results accrued from such a study are believed to have ramifications for the interface between decision making behavior at the level of the individual and the organization in a more specific sense. Thus this observation shows that multi-modal interfaces can lead to better user experience. Human experience is labile and malleable in that it can be harnessed in different modes and through different media with the added advantage that the same content can be harnessed, molded and manipulated for differentially oriented purposes and tasks at hand. This character of experience is fine-tuned for multimodal learning of linguistic structures the underlying cognitive structures of which can be observed to shape and be reshaped by such experiences in VEs as the study has revealed. This is extremely valuable for any study that aims at figuring out how cognitive structures during learning can be seen to behave in vivo.

6 Future Work

There are several limitations and problems with the current study. Language learning especially lexical learning is a very complicated and multi-dimensional process requiring representationality at several levels of architectural specification. This has been attenuated by orders of magnitude for the sake of modeling and initialization of the processes within the architecture of the current VE. This needs a further elaboration within the current architecture that will lead to multi-layered sub-architectures for lexical learning cutting across syntactic, morphological, semantics/pragmatic and other cognitive levels of representation.

References

Akshar, B, Chaitanya, V and Sangal, R., 1995, Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi, 65-106.

Bartle, R.A., 2004, Designing virtual worlds, New Riders Pub.

Christiansen, M.H. and Chater, N., 2008, Language as shaped by the brain. Behav. Brain Sci. 31, 489–509

Miller, G, Johnson, L. P., 1976., Language and Perception. Cambridge: Cambridge University Press.

Pala. K., and Gangashetty S.V., 2012a (In Press), Virtual Environments can Mediate Continuous Learning, Technology Inclusive Learning. IGI, USA.

Pala K., Gangashetty S.V., 2012b (In press), Challenges and Opportunities in Automatically Building Bilingual Lexicon from Web Corpus, in Interdisciplinary Journal on Linguistics, University Press.

Pala, K. and Begum, R., 2011a An Experiment on Resolving Pronominal Anaphora in Hindi: Using Heuristics, Journal on Information Systems for Indian Languages, 267-270, Springer.

Pala, K. and Singh, A.K. and Gangashetty, S.V., 2011b, Games for Academic Vocabulary Learning through a Virtual Environment, Asian Language Processing (IALP), 2011 International Conference on, 295-298, IEEE

Parser/Hindi, 2012, Hindi Shallow Parser source, Retrieved 1 March 2012 from, Hindi Shallow Parser-source, http://ltrc.iiit.ac.in/analyzer/

Robert, C. and Wyatt, JL, 1956, A Comparative Grammar of the Dravidian or South Indian Family of Languages, Robert, Revised and edited by Rev, JL Wyatt and T. Ramakrishna Pillai, Reprint ed., (Madras:. University of Madras, 1961)

Tomasello, M., 2008. The Origins of Human Communication, MIT Press

Ullman, M.T., 2001. The Declarative/Procedural Model of Lexicon and Grammar, Journal of Psycholinguistic Research, 30(1).