Various modelling aspects of tutoring systems for people with auditory disabilities

Nelson Baloian and Wolfram Luther
Department of Computer Science, Universidad de Chile, Blanco Encalada 2120, Santiago, Chile, nbaloian@dcc.uchile.cl

Institut für Informatik und Interaktive Systeme, Gerhard-Mercator-Universität Duisburg, Lotharstrafle 65, D-47048 Duisburg, Germany, luther@informatik.uni-duisburg.de

Key words: Disability, Learning Environments, Learning Models, Recommendations, Tutor

Abstract: Following a survey of the existing tutoring systems for hearing impaired people the paper describes common aspects of modelling the real world when creating educational software. This involves testing and evaluating cognitive tasks with people who dispose of reduced auditory cues. To validate our concepts we used the multimedia learning system known as Whisper. We developed Whisper in 1997 and it has been evaluated extensively in special schools for the hearing impaired. Members of all the leading German rehabilitation centres have assisted in this evaluation process. Whisper implements an interactive educational system to help hearing-impaired people recognise speech errors.

1. INTRODUCTION

A number of systems have been developed with the intention of being used by people with disabilities. Systems created for hearing-impaired persons are oriented to train people by developing the necessary skills to overcome their disabilities.

We recognize in these systems similarities with the process of modelling the real world in a way with which the impaired can explore and interact.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35663-1_34

T. J. van Weert et al. (eds.), *Informatics and the Digital Society*

Mostly, this process consists of two steps: the first step involves modelling of the real world within the computer and the second projects this computer model as a multimedia type with which the impaired can interact.

In most systems the modelling process ends at this point. Systems supporting learning or skill acquisitions need a feedback with a meaningful response. That can only be provided if the system has an embedded student's model constructed upon the learner's answers to certain questions or even the requirements proposed by the system. In this article we address both issues, the modelling of the real world and of the learner.

We started developing a multimedia learning system in 1997 called Whisper (in both German and English). This system enhances the proficiency in speaking for postlingually deaf and in hearing for Cochlear implant users (Wans 1998; Wans 1999) Whisper offers a workplace for a hearing-impaired person to recognize speech errors. Words, sentences or small stories are taken from everyday life. The learner's best verbal productions spoken in the presence of a teacher or a hearing person are recorded and stored in a repository. The learner explores a typical situation presented by appropriate cartoons subtitled with names which are spelt using a phonetic alphabet. Then the learner has to repeat all names or short sentences. The utterances are recorded under similar conditions and an acoustic representation of the spoken words is compared with the patterns stored in the database. By means of intelligent pattern recognition algorithms the difference between the learner's best production and the current expression is evaluated against five important speech parameters. An appropriate graphic and textual output for the learner is generated. This visual feedback replaces the usual auditory outcome and enables the impaired person to improve their speech capabilities.

The use of Virtual Environments (VE) as an interactive technology to allow learners to explore, interpret and represent real and symbolic objects and actions is a promising tool for people with sensory disabilities. The virtual environment is a navigable world built by using appropriate modelling language like VRML 2.0, dynamic scene objects, and acting characters. Scenic objects are defined by graphic and acoustic attributes and the actions of each character are based on deterministic and non-deterministic plans described within an interactive hyperstory. According to Lumbreras and Sanchez (1999) a hyperstory occurs in a virtual environment as a hypermedia version of a literary story. A new version of Whisper will incorporate these ideas and allow the creation and acting out of short stories using special characters like cartoons.

This paper begins with a survey of existing systems for hearing-impaired people. Next we develop common modelling aspects leading to a model for Whisper and similar systems. This is followed by a discussion of all the

important steps of the modelling pipeline - the reduction of the original model to only graphic objects, the reconstruction of the internal computer model by the learner, the integration of an error diagnosis together with a user-adapted output, and the evaluation process.

2. RELATED SYSTEMS FOR HEARING-IMPAIRED PEOPLE

Our main interest group of hearing-impaired are the postlingually deaf who are familiar with the terms of the language because they have lost their sense of hearing after the acquisition of language and speech. This group of hearing-impaired has to live with severe communicative restrictions. The missing auditory comment often causes some deterioration in the control of speech production impacting both the intelligibility and social acceptability of the deafened speaker's voice. For deaf people the sign and gesture language was developed to allow nonverbal communication. The language provides all finger spelled letters used for proper names and rare words as well as thousands of word signs formed by hand and arm motions. To provide emphasis and to structure sentences, head motions and facial expressions are used. It is actually possible to create a dialogue based on hand gestures such as the dialogue between deaf human and deaf virtual human. A real time facial animation engine conceived to run on a common PC with a graphic accelerator card is described in Breton et al (2001).

Mehida (Alonso et al 1995) is an intelligent multimedia system for deaf or hearing-impaired children designed to assist them in acquiring and developing communication skills. It covers the following types of communication - finger spelling (representing the letters of the alphabet using the fingers), gestures or sign languages, lip reading (understanding spoken language through observing lip motion), and voice recognition.

The IBM SpeechViewer III (IBM 2000) is a powerful speech and language tool that transforms spoken words and sounds into imaginative graphics. SpeechViewer III is intended for people who have speech, language or hearing disabilities. It offers three main types of exercises. Awareness exercises are simple games concentrating on basic speech parameters and can be used by very young children. The skill building exercises ask the learner to work towards achieving a goal set by the expert and provide feedback on pitch, voicing, and vowel articulation. Patterning exercises offer a graphic representation of the learner's and the expert's voices for comparison concerning pitch and loudness.

ISAEUS (1997) is a speech training system for deaf and hearing-impaired people on a multilingual basis which has been developed through an EU-project involving groups in France, Spain and Germany.

The Visual Talker (Visual Talker 1999) and a recently developed product from the Liberated Learning Project provide real-time speech-to-text visual translator systems for classroom environments. The aim of this research is the development of a real-time speech-to-text system to be used by teachers, their students with hearing loss and/or speech impairments, and all other students in classroom settings to provide a 'full participation' communication system.

In summary, we can identify two types of enhanced systems for people with sensory disabilities: systems that try to complement common interfaces with other communication facilities and systems that support direct interactions with internal models or adequate representations of virtual models. Both try to transform the audible output into a textual or graphic format through which the user should navigate or communicate. We will now show how the modelling process should reflect the user's impairments.

3. MODELLING PIPELINE FOR DEVELOPING EDUCATIONAL SOFTWARE

The modelling process starts with the definition of the cognitive skills the learner has to acquire; then it considers the creation of a virtual environment composed by a navigable world and built by using an adequate modelling language, dynamic scene objects, and acting characters. Scenic objects are characterized by graphic and acoustic attributes and the actions of each character are based on deterministic and non-deterministic plans as in an interactive hyperstory. The learner explores the virtual world by interacting with appropriate interfaces and immediately obtains feedback.

The learner's actions, such as sound reproductions or utterances, are collected, evaluated, and classified, based on a student's modelling and diagnostic subsystems

In Whisper the modelling process is as follows. First, a virtual model **B** (a graphic, textual and phonetic model) is derived from the real or ficticious world scenario **A** (of objects, phonemes, words and stories) by means of abstraction and reduction - without taking into consideration the limitations of the potential users. The modelled objects are nouns or short animated stories which are named using phonemes and words. Then model **B** is projected to an adequate model **C** (a visual, frequency colour model) which can be explored by people with sensory disabilities. Obviously certain information channels cannot be used. Appropriate editors support the

modelling process and important model parameters must be identified at this stage. By interacting with the system, the learner makes an internal reconstruction of the model **C** called **D**. In Whisper, the learner explores the visual model of a short story and reconstructs the visual wave form representation by speaking words or sentences. Additionally, the learner may build an external representation of **D** which has to be evaluated. This can be done by using an appropriate multidimensional error measure $\mathbf{m_I(P,X,Y)}$ depending on the objects and their attributes, the parameters **I**, a learner **P**, the computer model **X**, the reconstructed model **Y**. Finally, the degree of similarity is derived from the error measure and the result is displayed in a learner-adapted output and used for updating the student model. The actual representation is checked by a human tutor or an intelligent computer module to look for any correspondence with the original model.

The error measurement (represented by the function $\mathbf{m_I(P,X,Y)}$) should reflect this difference. The index **I** stands for different properties of objects being a part of **X** and **Y**, **P** denoting the learner, **X** the internal model **C**, and **Y** the reconstructed model **C** consisting of the words or sentences. Then the 'distance' between the same object in models **C** and **D** will be calculated using an appropriate vector norm. Thus we can derive a candidate for an error measure $\mathbf{m_{Distance}}$.

3.1 Construction and evaluation of the internal model

Now to the issue of reconstructing the internal model and of finding out ways to measure how close to the computer model is the user's mental model. In learning oriented systems this task's goal is to give meaningful feedback to the user's performance. While in the presence of a human tutor, the reconstruction process is monitored and supervised. Practicing without assistance gives more independence and self-confidence, but requires some integration of student models together with diagnostic subsystems.

Studies on the speech production of the postlingually deaf have shown that the loss of auditory feedback affects the speech proficiency especially in the global parameters of articulation, accentuation (stress), intonation, pitch, rate, and volume. The most striking speech error involves the volume control. Because of the missing auditory feedback the postlingually deaf either speaks too softly or tends to shout.

The loss of the sense of hearing has also an important influence on articulation. In most cases, the sibilants are mis-produced. Another common type of error is slurring, which is elongation of segments with increased co-articulation giving the loss of the segment definition. Frequently, the prefix or suffix of a word is omitted, segments are added, omitted or substituted.

These errors tend to occur in clusters as if some aspect of the control of speech had temporarily become maladjusted. On the other hand, the articulation is sometimes exaggerated. The nature of speech determined by stress and intonation is often affected. Pitch is then monotonous, stresses are excessive or misplaced, or patterns of intonation are inappropriate. The intonation can be very individual - it may correspond neither to the sense nor the purpose of the statement and makes the statement hard to understand.

After each utterance the system has to find out the differences between the statement of the learner and the reference model in the database, analyse the reasons for the modifications, and provide information on how the learner can improve speaking. There are two important methods for analysing speech parameters and to detect errors - signal analysis or algorithms for speech recognition.

In the first case, the differences concerning the parameters of volume, articulation, stress, and intonation must be visualized. From these visual patterns important parameters are detected by means of picture processing analysis of, and in connection with, the stored reference patterns. An intelligent component of the system interprets these results, displays a visualization of the error function $\mathbf{m}_I(\mathbf{X},\mathbf{C},\mathbf{D})$, for each parameter **I**, and transforms them into instructions for the learner.

We sample the recorded speech signal at a rate of 22.05 kHz and 16 Bit encoding. Then we execute 40 fast Fourier transforms per second to obtain a frequency representation. For display we use an RGB colour map with three overlapping triangular characteristics for red, green, and blue in the intervals [0, 1000], [250, 4000], [1000, 10000] which can be freely modified. The resulting frequency-colour representation is smoothed and simplified by an appropriate cluster and threshold algorithm. Whereas the wave amplitude image provides in the xy-plane any feedback concerning volume and rate, omitted syllables are characterized by a lack of coloured areas, intonation and articulation by the parameters hue and saturation as well as the shape of the curves. Details are described in Hobohm (1993) and Gräfe (1998).

In recent years it has been shown as reported by Willett (2000) that the combination of neural networks with Hidden Markov Models (HMM) is a powerful approach to speech recognition. These so-called hybrid speech recognizers use multi-layer perceptron or recurrent neural nets to estimate the a posteriori probabilities for the phones in each feature. There are different new approaches to build three-state-phone HMMs in a hybrid recognizer, one state represents the beginning of the phone, one the static part of the phone, and the third state is used to model the transition into the following phone. When using context-dependent models, even small systems with a small vocabulary lead to an acceptable number of triphones in the output layer of the neural net. Additionally, this approach permits the

combination of discriminative training techniques with Maximum Likelihood methods for the HMM parameters using a phoneme-labelled speech database. However, in our case for the training procedure on the learner's best performances, phonetic transcriptions are necessary. The system provides a morpheme-to-morpheme transposition of an oral language into the visual channel. In this process all structures of the special oral language are preserved. Whereas the system adapts to the speaker (taking in consideration the collected training data), it does not have the intention of reducing the error recognition rate. On the contrary, the system is intended to detect errors concerning deformed or omitted phones or erroneous articulation. The system searches most probable phone sequences and the phones are displayed within a separate window.

The combination of a well-known and efficient speech recognition architecture together with new techniques of speech error detection seems to be a promising research field in the future and will be used in the context of speech error detection in a new version of Whisper.

3.2 User adapted output

The natural way to communicate and to comment on the outcome of a session would be to introduce a human expert. However, permanent assistance is expensive and not always available.

Currently our technical frequency-colour representation of speech is not quite adequate for young users. It seems to be more motivating to use special icons or comics. We have adapted forms and colours following observed error parameters. For example we might create visualisations of a learner's best production of the word aubergine a current production of the French word 'aubergine' with corresponding intonation error and add to the visual feedback a simplified representation of an aubergine as a cartoon which cleverly includes the most important features of the technical output. Then we give the following guidelines:

- Simplify the envelope of the speech pattern.
- Use the time and volume axes to indicate these errors by a simple transformation of the cartoon.
- Articulation errors can be represented by deformation of the contour.
- Omission errors are modelled by omission of control points in the B-spline representation of the graphic object.

Our task is to find pictorial elements which represent in a simple way the frequency-colour model. This seems to be feasible for nouns but is not easy for verbs and adjectives. An interesting method is the creation of a gallery of actors stored in a repository which are introduced into the scene by the learner to play small stories, for example the principal actors of mouse and

elephant in Figure 1. An articulation error is shown by altering the maximum value of the contour, height and length are related to volume and rate, and omission errors can be visualized by different proportion of the body.

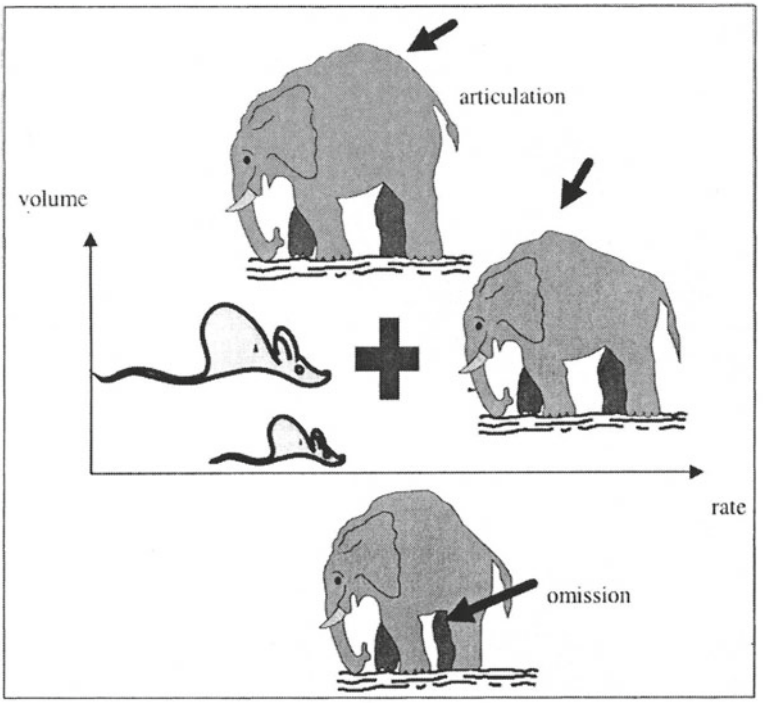

Figure 1. Error output by special representation of mouse and elephant

3.3 Evaluation and further work

To achieve a normative evaluation of Whisper, an interdisciplinary approach involving special schools, rehabilitation centers and self-help groups has been undertaken.

Whisper was presented at the REHA 99 exhibition in Düsseldorf and then evaluated during a six month period through special schools for hearing-impaired with the aid of all leading German rehabilitation centers. The usability of the English version was tested in association with the National Association for Deaf People in Dublin.

Most people admit to mistakes in volume, articulation, stress, and intonation and concede problems in understanding individual people, especially in group situations when voices are mixed. The test group plans to

work with Whisper at home, to improve the pronunciation, to learn with sentences and stories, and to modify the repositories. In the future, we plan to enhance the speech error recognition, to analyze stress and intonation, to include a calibration routine, and to provide appropriate visual feedback. To meet the expectations of hearing-impaired people, we will include other dictionaries and an automatic generation of phonetic descriptions, implement further foreign language versions, develop benchmarks, and evaluate user progress. This work will be done in accordance with our modelling guidelines and will allow for evaluation of the proposed modelling process.

Many systems maintain classic but inappropriate interfaces and try to fix them in a way the impaired can interact with. We think we should try to develop totally new interfaces that allow people to develop mental structures or models of the world being represented by using the channels available (Baloian and Luther 2002) We may use the same devices but with a different logic behind them. Another important issue is how to motivate the user when there are no sound resources to produce special effects? How to trigger surprising effects, feelings or certain moods that are normally transmitted by a combination of high quality image sequences and sounds when one of these channels is not available? We think some answers can be provided by studying classic silent movies or radio theatre pieces.

4. CONCLUSION

In this article we have described a state of the art example in the field of systems for hearing-impaired people, especially from the point of view of transferring the real world into computer representations for this interest group. We have also explained a unified methodology for modelling the real world and have illustrated important tasks of defining error measures and adapted output formats in these systems.

A critical mass of educational systems have already been developed for disabled people, which allows some generalizations and recommendations to be made. The development of systems for people with disabilities should no longer appear as isolated handcrafted efforts; instead, the construction of these types of systems must be systematized. Recent advances in hardware and software developments indicate that the technological foundation for such systems has already been laid.

Acknowledgements

This research and common multi-media software development for the disabled (AudioDoom for blind people and Whisper for postlingually deaf people) are being carried out by the authors in the course of a current

German-Chilean project funded by the German ministry BMBF and the Universities of Duisburg and (Santiago de) Chile. We thank Dr. M. Mühlenbrock and Dr. D. Willett for their valuable suggestions and our collaborators Dr. W. Otten and C. Wans for their valuable support, discussing these ideas and contributing their research during the development and evaluation of Whisper.

REFERENCES

Alonso, F., de Antonio, A., Fuertes, J. L. and Montes, C. (1995) Teaching Communication Skills to Hearing-Impaired Children. *IEEE Multimedia*, Vol. 2, No. 4, pp. 55-67.

Baloian, N. and Luther, W. (2002) Visualization for the Mind's eye. To appear in *Dagstuhl SV Lecture Notes in Computer Science*. Springer, Berlin.

Breton, G., Bouville, C. and Pelé, D. (2001)FaceEngine. A 3D Facial Animation Engine for Real Time Applications. In *Proceedings of 2001 Web3D Symposium*, Paderborn, Germany, pp. 15-22.

Gräfe, J. (1998) *Visualisierung von Sprache und Erkennung sprechtypischer Parameter und ihre Veränderung bei Spätertaubten*. Diploma dissertation, GMU Duisburg.

Hobohm, K. (1993) *Verfahren zur Spektralanalyse und Mustergenerierung für die Realzeit-visualisierung gesprochener Sprache*. PhD dissertation, TU Berlin.

IBM (2000) *Speech viewer III*
[http://www-3.ibm.com/able/snsspv3.html]

ISAEUS (1997) *Speech Training for Deaf and Hearing-Impaired People*
[http://www.ercim.org/publication/Ercim_News/enw28/haton.html]

Lumbreras, M. and Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. *CHI '99*, Pittsburg PA, USA, pp. 318-325.

Visual Talker (1999)
[http://www.ed.gov/offices/OERI/SBIR/FY99/phase1/ph199t02.html]

Wans, C. (1998) An Interactive Multimedia Learning System for the Postlingually Deaf. *Proceedings of ITiCSE'98, 6th Annual Conference on the Teaching of Computing, 3rd Annual Conference on Integrating Technology into Computer Science Education*, ACM, New York, Oxford.

Wans, C. (1999) Computer-supported hearing exercises and speech training for hearing impaired and postlingually deaf. *Assistive Technology Research Series*, Vol 6 (1), IOS Press, pp. 564-568.

Willett, D. (2000) *Beiträge zur statistischen Modellierung und effizienten Dekodierung in der automatischen Spracherkennung*. PhD dissertation, GMU Duisburg, 2000.