Speech Enabled E-Learning for Adult Literacy Tutoring

Paul Walsh and Jason Meade
Cork Institute of Technology, Ireland
pwalsh@cit.ie; jmeade@cit.ie

Abstract

It is estimated in a recent OECD International Adult Literacy Survey that up to 500,000 Irish adults are functionally illiterate, that is many people have difficulty in reading and understanding everyday documents. We address this problem by allowing users to interact with speech enabled e-Learning literacy content using multimodal interfaces. We present two experimental prototypes that explore technical solutions and identify an application architecture suitable for literacy e-Learning. The implementation of an evolutionary prototype that uses client side technology is described and feedback from this phase of the project is reported.

1. Introduction

The National Adult Literacy Agency (NALA) is the national training co-ordinating agency for all adults involved in literacy tuition in the Republic of Ireland. It was established in 1980 and receives its core funding from the Irish Department of Education and Science. NALA policy is guided by a recent OECD report, which found that up to 500,000 adults in the Irish Republic are functionally illiterate [1]. Currently only 20,000 Irish adults are in basic literacy tuition [2]. This figure represents just 4% of adults with weak literacy skills, which highlights the need for more literacy options. Therefore the vast majority of people with literacy difficulties are not in literacy provision but are *potential* independent learners; people who do not necessarily want to go to a literacy scheme for tuition, but may wish to study in the privacy of their own home. Moreover, many literacy schemes now provide IT equipment and Internet access to literacy students who do not have access to IT resources. Hence, one potential way in which to facilitate both independent learners and literacy students in tuition is to use e-Learning technology.

E-Learning has been defined as the use of Internet technologies to deliver a broad array of solutions that enhance knowledge and performance and can provide benefits such as reducing travel, infrastructure and training expenses, while allowing wide access and scalability [3]. These are just some of the potential benefits of e-Learning that can be applied to literacy education sector. For example, e-Learning in literacy can save independent learners in remote communities from travelling, can be tailored to different interest groups, can be available anytime of the day through web browsers in community centres, libraries or schools and can be shared as a resource with organisations similar to NALA in any English speaking country, with minor modification. To this end an e-Learning resource and Learning Management System (LMS) www.literacytools.ie has been developed in conjunction with NALA, to provide content, assessment and support for literacy students and tutors. This resource is complemented by literacy tutor IT training courses, literacy television programs broadcast on national television, free phone telephone support and free literacy text books.

However one obvious limitation in providing IT access to learners with low literacy levels is the difficulty that these learners have in reading text. This difficulty is compounded when the learner is faced with computer technologies and concepts, often for the first time. Speech technology is a possible solution to this problem as it can enable students to learn literacy skills using a more natural human-computer interface. This has motivated us to use speech technology to build a multimodal interface to literacy content. Multimodal interfaces mix elements of visual and voice interaction into a single interface and the development of these interfaces is considered architecturally and technically complex [4]. Hence in this paper we outline our experiences in developing multimodal interfaces for literacy e-Learning and present our findings thus far.

2. Speech Technology

Speech technologies include Text-To-Speech (TTS), and Speech Recognition (SR) software [5]. Text-to-Speech systems (also known as Speech Synthesis) convert digitised text to spoken words, whereas Speech Recognition is the identification of spoken words by a machine, whereby speech is digitised and matched against coded dictionaries in order to identify the spoken words.

The initial phase of our research will focus on TTS technology for enabling e-Learning literacy content, as it represents the most practical and feasible element of the technology; current SR conversational accuracy is only between 50% and 80% accurate [6].

There are 2 main types of TTS application: static and dynamic. A static application uses pre-recorded audio files as output, whereby a dynamic application creates audio files at runtime. A dynamic system is more robust,

Proceedings of the The 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)
0-7695-1967-9/03 $17.00 © 2003 **IEEE**

and adjusts to runtime changes, while a static system is simpler to implement [7]. The research presented here is based on dynamic TTS, as we are concerned with the automatic production of new sentences, based on user input.

3. Pedagogical Considerations

It has been stated that one of the most important factors for the successfully implementation of e-Learning is the need for careful consideration of the underlying pedagogy [8]. The methodology employed in the development of the literacy e-Learning environment is based on the Instruction Design methodology described in [8] and Knowle's Theories of Adult Learning [9].

The goal of our research is to make the e-Learning content developed with these methodologies accessible to literacy students using speech technology. The basic challenge is to make web content available to literacy students by converting any text that the student may have difficult reading into speech. Students are also assessed using speech enabled Multiple Choice Questions (MCQ), True/False Questions (TFQ) and Matching Questions (MQ). The student's answer, result and the correct answer are given both in text and speech format.

4. Software Prototypes

Many software projects fail to due to poor communication between the software developer and the end user. Software prototyping addresses this problem by implementing crucial aspects of the system such as the user interface and core functionality and serves as a system model for co-ordination between the developer and end user [10]. To this end there are three categories of prototype that are commonly used in software engineering [11]:

- Explorative prototypes that promote an understanding of the requirements and propose different solutions.
- Experimental prototypes that investigate possible technical solutions.
- Evolutionary prototypes that are milestones in an iterative process of development.

Experimental prototyping was used in this research to investigate both client side and server side technologies. Such prototyping has proved useful in evaluating potential speech and Internet technologies. Evolutionary prototyping is also used on an ongoing basis to investigate user satisfaction with the system and guide development, by elucidating feedback from literacy students. Providing prototypes *online* also facilitates feedback from users, making prototyping a valuable tool in the development process.

4.1 Experimental Prototypes

Client side TTS technologies are those that implement TTS functionality on the client, whereas server side TTS implement that functionality mainly on the server. One of the main benefits of client side technology is that the system is more scalable as much of the processing is off-loaded to client machines. While client side technology is more scalable than server side technology, it is not platform neutral and depends on non-standard plug-ins. Server based components, while not as scalable, can be used to generate platform and browser independent HTML and can uses de-facto standard audio files such as the WAV format. Such server side rending of speech is common in many web based speech application in servers that act as gateways to web clients [4].

Literacy Exercise	File Size in Bytes	
	Server	Client
Exercise 1	634,476	25,710
Exercise 2	391,331	20,952
Exercise 3	370,616	46,662
Exercise 4	207,318	10,599

Table 1 Comparison of download file size between server side TTS components and client side TTS components for sample literacy content.

Both client side and server side prototypes were implemented and demonstrated to users. However, as expected the main drawbacks with the server side approach was the scalability and efficiency of the system. Table 1 shows a comparison of download file sizes for identical literacy e-Learning content implemented using client and server side components. It was found that on average, the prototype that employed client side TTS technology was 17 times more efficient, in terms of bandwidth use, than the prototype that used server side TTS technology. Moreover, the CPU intensive nature of TTS operations make the server side implementation of the system less practical.

Both prototypes were demonstrated to end users, including both literacy students and their tutors. Informal discussion and questionnaires were used to elucidate feedback from these users. Respondents who expressed a preference found the client side TTS prototype more usable than the server side system.

Hence it was decided to focus development efforts on an evolutionary prototype that uses client side TTS technology, as this is the most practical alternative in terms of usability and bandwidth, although a prototype based on server side TTS technology is also available to users who do not use the Microsoft platform: www.literacytools.ie/speech/.

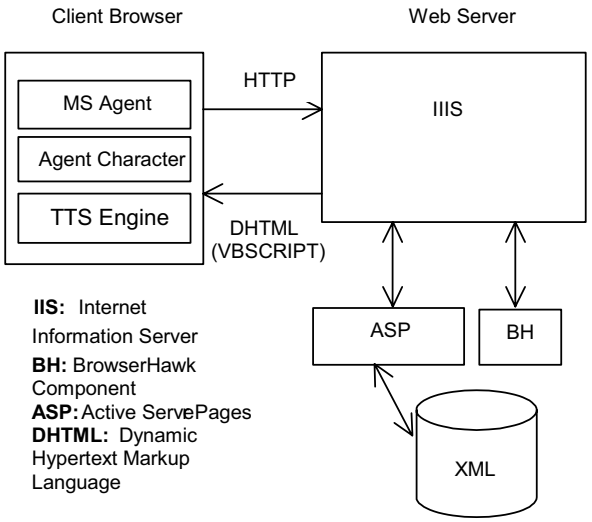

Figure 1 Speech Application Architecture.

4.2 Application Architecture

The application architecture of the client side TTS system is described in Figure 1. The technologies used includes XML, Internet Information Server (IIS), Active Server Pages (ASP), VBScript, Microsoft Agent 2.0, Cyscape's BrowserHawk 6.03 component and Scansoft's TTS Engine.

The type of platform and bandwidth speed available to learners is detected by a server side ActiveX component: Cyscape's browser compatibility component BrowserHawk 6.03. This allows the server to detect the details of connecting users configuration and connection speed. Users that do not have the required browser (IE5 or later) are prompted to either download this browser, or are redirected to the content that is supported by server side TTS components.

Client side VBScript is then used to detect whether the required client side plug-ins are installed. If not then the user is guided through the installation process, which consists of downloading the client side TTS components.

4.3 Data Formats

Speech content is stored in extensible Markup Language (XML) format [12]. XML is a meta-markup language; a set of rules for creating semantic tags used to describe data. One of the key advantages of this technology is the extensibility of the format. This feature is exploited in this application by using elements of VoiceXML and the Instructional Management System (IMS) XML schema. XML also allows the system to be developed using an evolutionary software development process; the data format and system functionality can be readily extended using this approach.

VoiceXML is utilised in this prototype to tag speech-based elements of the literacy content. VoiceXML is designed for creating audio dialogs that feature synthesized speech, digitised audio, recognition of speech, and mixed-initiative conversations. Although VoiceXML 1.0 was designed primarily for speech-based telephony applications we use this standard to add meaningful mark-up tags to the relevant content used in this application. This approach can also be extended to incorporate tags from Speech Synthesis Markup Language Version 1.0.

Instructional Management System XML tags are used to define the structure of the content in the prototype system [13]. The IMS is a consortium of vendors and implementers who focus on the development of XML-based specifications. These XML specifications provide a structure for representing e-Learning meta-data (data that describes data). The IMS consortium has also specified a Question and Test Interoperability (QTI) specification, which is structured to support a wide range of question types including MQCs.

4.4 Client Side Components

A number of TTS systems were reviewed when implementing client side speech accessibility. One client side option is to utilize screen reader accessibility software. A screen reader is software that works together with a speech synthesizer to read aloud everything displayed on a computer screen, including icons, menus, text, punctuation, and control buttons. Such software is used by the visually challenged and people with literacy difficulties. There are currently a number of commercial products on the market that provide this functionality:

- JAWS by Henter-Joyce (http://www.hj.com/)
- Home Page Reader by IBM
- (http://www-3.ibm.com/able/hpr.htm)
- Hal by Dolphin (http://www.dolphinusa.com)
- ReadPlease (http://www.readplease.com/).

While these screen reader provide much of the functionality required for literacy tutoring, they do not provide the flexibility of a bespoke TTS implementation. Hence it was decided to develop a TTS based multimodal interface to the literacy content.

After reviewing a number of TTS components, it was decided to use Microsoft's Agent API. MS Agent is a speech enabled user interface element that presents an animated character to the user. The benefits of using this technology are:

- Provides industry standard speech technology interface; Speech Application Programming Interface (SAPI).
- Agent character appears in own window; does not mask e-Learning interface.
- Provides an ActiveX control that can be embedded into DHTML.
- Supports TTS and Speech Recognition.
- Allows rapid development of prototype.
- Agent is royalty free.

The Agent API provides all of the functionality required for the prototype phase of the project and includes support for prosody markup elements, which is a key feature of natural sounding speech. Prosody is the set of features of speech output that includes pitch (also called intonation or melody), timing (or rhythm), pauses, speaking rate, emphasis on words and many other features. Producing human-like prosody is important for both making speech sound natural and correctly conveying the meaning of spoken language.

Prosody and other speech characteristics such as sound and volume are specified in the agent by the use of speech output tags. The output tags are used within the text parameter of the agent Speak method. These tags begin and end with a backslash character "\"and are case-insensitive. The escape sequence for a backslash character in the output text is a double backslash "\\". These tags are embedded as VBScript by server side ASP pages into the literacy e-Learning content.

The following code segment illustrates the `\spd`, `\emp`, `\pit`, `\chr` and `\vol` tags, which are used to alter the speed, emphasis, pitch, character and volume of the associated text, as shown in Figure 2.

```
Genie.Speak ("New Baby \emp\Girl. A gentle
\emp\cry lets you know that she's there.")
Genie.Speak ("\chr="+chr(34)+"
whispervoice"+chr(34) + "\\spd=160\\pit=65\A
little hug lets her know you care, So fragile and
helpless, like a little porcelain doll,"
Genie.Speak ("\pit=90\She looks so peaceful
\pit=1000\all curled up asleep in a little
ball!")
```

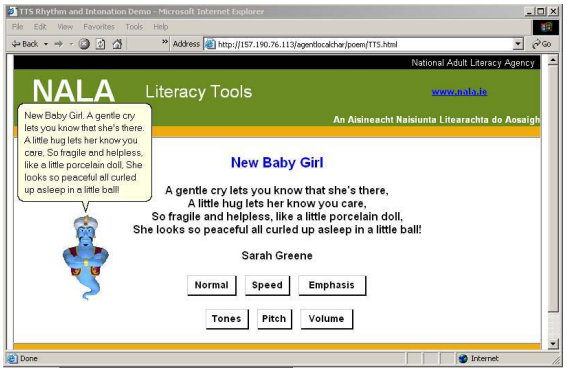

Figure 2 Rendering **of literacy content with MS Agent API (http://www.literacytools.ie/speech/TTS.html).**

Note that the `\chr` tag is used to alter the character of the voice, which is specified by the succeeding string parameter. In the above example the "`whisperVoice`" parameter is used to specify a quiet whispering voice.

VBScript is used to facilitate user interaction with the agent. This is important, as cognitive load considerations must be considered for the efficient structuring of instructional presentation involving more then one modality. It has been suggested that there are at least 2 cases where extra information can make a course more difficult to learn:

1. When equivalent auditory and visual explanations are presented concurrently, and
2. When an instructional format is not matched to learner experience.

These cases "increased the risk of overloading some of the sensory channels and might have a negative learning effect" [14]. Hence it is important that users are able to choose when speech output is presented. This is done by using VBScript to detect when the speaker symbol is clicked, as shown in Figure 3. This event invokes the MS Agent to speak the associated text in the appropriate format. Students can also select individual words to be spoken by the agent by double clicking in the word, or select blocks of text to be spoken by highlighting the text with the mouse. VBScript is also used to assess student progress by means of MCQ, TFQ, MQ and Cloze test questions.

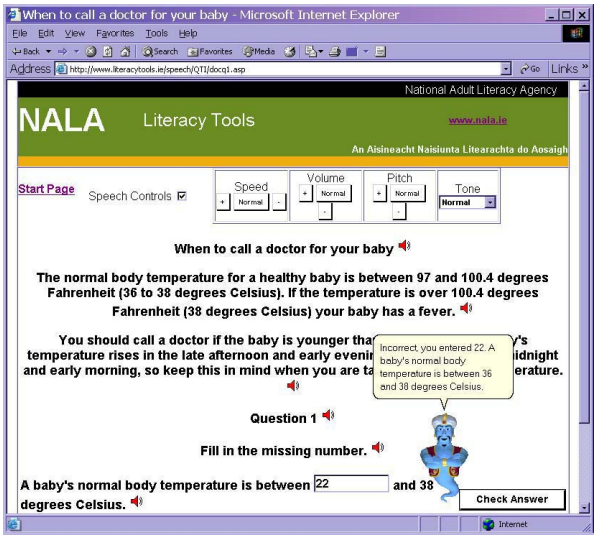

Figure 3 Dynamic generation of feedback with agent interaction (**www.literacytools.ie/speech/index.htm**).

The current working prototype was developed using ASP. Functionality is currently being re-implemented using XML and XSL Transformations (XSLT) as this technology offers a number of advantages [15]:

- No need to embed presentation code (html) directly into application logic.
- XSLT is a declarative high-level language that can be highly optimised.
- XLTS is standardised by the W3C and has implemented on a number of platforms.
- XSL transformations and rendering can be offloaded to XSLT enabled clients.

The client side processing of XSL is a significant point as the study performed in [15] shows that up to 1,000 pages per second can be delivered by Internet Information Server 5.0, whereas a server side implementation could only process 120 pages per second.

5 Related Work

Literacy is not just an issue in Ireland as levels of literacy in many countries are far from satisfactory, as reported in the OECD's recently completed International Adult Literacy Survey (IALS) covering 20 mostly developed countries or regions [16]. To this end there have been a number of e-Learning sites implemented in OECD regions, as reported in [17]:

- www.thestudyplace.org
- www.cyberstep.org
- www.tv411.org
- www.bbc.co.uk/skillswise/

However the level of dynamic speech synthesis integrated directly into these sites is limited and these sites do not directly address the needs of Irish learners.

6 Evaluation

The system is currently in the early stages of evaluation. Results from informal interviews and questionnaires show that users found the speech e-Learning prototype useful and most of those surveyed found that speech technology helped them to complete literacy lessons. Feedback from the prototypes has had significant impact in the design of the site and the development of e-Learning content. For example one group of users found that text was been spoken back too quickly. We responded to this by including a control panel interface that allows users to adjust the speed, pitch and characteristic of the speech engine, see Figure 3. Users also indicated that they would like to select individual words for speech synthesis. This functionality was also added to the site.

An online feedback form is also provided on the site and more formal analysis based on attitudinal questionnaires [18] and the Adult Literacy Resource Institute software review criteria [19] is currently underway. It is also possible to track the progress of on-line literacy students using LMS tools developed in the site and by analysis the web server log files.

7 Conclusion

In this paper the results of our experience with experimental prototypes for speech based literacy e-Learning have been presented and an application architecture suitable for literacy based e-Learning has been identified. Our findings so far findings indicate that:

- Client side TTS technology is a scalable solution for implementing a speech enabled literacy e-Learning system.
- XML is suitable for semantically tagging literacy content and supports evolutionary development.
- Software prototyping facilitates communication between the developers and users.

Future work will focus on extending the evolutionary prototype into a more comprehensive e-Learning site for literacy students. A related TV series is already in place and complementary programmes are planned to accompany the site. The largest audience for any one programme in the series so far was 273,676 viewers [20] and future broadcasts will publicise the completed system. Emerging standards such as Speech Application Language Tags (SALT) [21] will be evaluated as potential technology for future development. A more detailed usability analysis guided by [22] will also be carried out.

Bibliography

[1] International Adult Literacy Survey, Irish Results, OECD 1997, http://www.nala.ie.

[2] Department of Education and Science, Further Education Unit, Literacy Results 2001.

[3] M. J. Rosenberg E-Learning: Strategies for Delivering Knowledge in the Digital Age, McGraw-Hill, 2001.

[4] K. Abbott, Voice Enabling Web Applications: VoiceXML and Beyond, APress, 2002.

[5] R. Peacocke, and D. Graf, An Introduction to Speech and Speaker Recognition, *IEEE Computers 23(8), 26-3, 1990.*

[6] R. M Baecker et al, Human-Computer Interaction: Towards the year 2000, 1995.

[7] M. Broughton, Measuring the Accuracy of Commercial ASR Systems, Human Factors Conference, Melbourne, 2002.

[8] T. Govindasamy, Successful Implementation of E-Learning Pedagogical Considerations, *Internet and Higher Education*, Elsevier Science, 2002.

[9] M. Knowles, *The modern practice of adult education.* New York , Association Press, 1974.

[10] M. Heinrichs, Extensible Technology for Electronic Patient Records, MSc Thesis, Cork Institute of Technology, 2002.

[11] C. Floyd, A Systematic Look at Prototyping, Springer-Verlag, Germany, 1984.

[12] I. Graham, L. Quin, *XML Specification Guide*. New York, NY: John Wiley & Sons, 1999.

[13] C.Smythe, L.Brewer and S.Lay *IMS Question & Test Interoperability,* Final Specification, Version 1.2, 2002, http://www.imsproject.org/.

[14] S. Kalyuga, When using sound with a text or picture is not beneficial for learning, *Australian Journal of Educational Technology*, 16(2), 161-172, 2000.

[15] C. Lovett, A Practical Comparison of XSLT and ASP.NET, http://msdn.microsoft.com, 2003.

[16] OECD, Literacy in the Age of Information Age. Final Report of the International Adult Literacy Survey, OECD, ISBN 92-64-17654-3 (81 00 05 1), Paris 2000.

[17] C. Holland, Designing an Online Literacy Interface, BERA Conference, University of Exeter, 2002.

[18] James R. Lewis, IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation, International Journal of Human-Computer Interaction, Volume 7, 1995.

[19] The Adult Literacy Resource Institute, Boston Massachusetts, Adult Education *Software Review,* http://www.alri.org/softreview/softreview.html, 2003.

[20] Read Write Now TV Literacy Series, Evaluation Report, www.nala.ie, 2002.

[21] A. Kooiman, K. Wang et al, SALT: The Light In Speech Mark-Up, *Internet Telephony Magazine,* 2002.

[22] Kirakowski, J, Estimating the Cost of Quantitative Evaluation, http://www.ucc.ie/hfrg/resources/powerp.pdf, 2003.