A pervasive Wiki application based on VoiceXML

Constantinos Kolias
University of the Aegean
Karlovasi, Samos, Greece
koliaskostas@gmail.com

Vassilis Kolias
National Technical University of Athens
9 Iroon Polytechniou St.
Zografou Campus, Athens, Greece
vkolias@medialab.ntua.gr

Ioannis Anagnostopoulos
University of the Aegean
Karlovasi, Samos, Greece
janag@aegean.gr

Georgios Kambourakis
University of the Aegean
Karlovasi, Samos, Greece
gkamb@aegean.gr

Eleftherios Kayafas
National Technical University of Athens
9 Iroon Polytechniou St.
Zografou Campus, Athens, Greece
kayafas@cs.ntua.gr

ABSTRACT
In this paper, we describe the design and implementation of an audio wiki application accessible via the Public Switched Telephone Network (PSTN) and the Internet for educational purposes. The application exploits mature World Wide Web Consortium standards such as VoiceXML, Speech Synthesis Markup Language (SSML) and Speech Recognition Grammar Specification (SRGS). The purpose of such an application is to assist visually impaired, technologically uneducated, and underprivileged people in accessing information originally intended to be accessed visually via a Personal Computer. Users may access wiki content via wired or mobile phones, or via a Personal Computer using a Web Browser or a Voice over IP service. This feature promotes pervasiveness to educational material to an extremely large population, i.e. those who simply own a telephone line.

Categories and Subject Descriptors
H.4.3 [**Communications Applications**]

K.3.1 [**Computer Uses in Education**]: Collaborative learning, Computer-assisted instruction, Computer-managed instruction and Distance learning.

General Terms
Design, Human Factors

Keywords
VoiceXML; Wiki; Information System; Audio Information System.

1. INTRODUCTION
Wikis are generally considered as collaboration platforms where users access or contribute knowledge on specific topics. Wikis are often used as web applications that allow registered (and sometimes even unregistered) users to create, edit, hyperlink and organize their content. Wiki applications are also used in many companies to provide effective Intranets and for Knowledge Management. Beyond doubt, the huge success of Wikipedia, a web encyclopedia and perhaps the most famous wiki application nowadays, proves in practice the suitability of wikis in education and knowledge exchange in general.

Wikis are generally designed with the philosophy of facilitation of access and contribution to knowledge. However the collaborative, and in many cases, open nature of wikis raises important issues such as accessibility, content validity and security. In most of the existing wiki implementations, access is as easy as browsing on a simple web page. Editing is also straightforward and it can be done by inserting information written in a specific (usually very simple) syntax, in the appropriate web forms that the wiki interface offers.

Nevertheless most of the existing wiki implementations are primarily dependent on well known World Wide Web (WWW) standards such as the Hypertext Markup Language (HTML), which present information in a visual manner, automatically knocking out visually impaired individuals from accessing their content. An undisputed truth is that the vast majority of modern web applications neglect the special needs of disabled people. Until now, visually impaired individuals who wish to interact with computers, usually rely on the features of the operating system they use, like the well known ShowSounds for Windows XP. While those features have been included in most of the modern operating systems, they do not offer support for dynamically generated content such as content from the WWW. Individuals are limited in using high cost systems that require special training as well.

Since wikis share a large amount of information among different people, the need to be accessible from the widest range of devices possible, is immanent. However, the integration of wikis to PDAs or smartphones is rather complicated due to the great number of different standards, technologies and operating systems that exist

in the market today. Web based wiki implementations are usually developed focusing on desktop scale HTML based browsers. This fact generates many problems when wiki content is accessed from mobile devices equipped with browsers supporting certain subsets of standard HTML, like WML or i-Mode HTML.

A wiki fundamental requirement, i.e. having some sort of access to the Internet, shuts out their content to a large population. These include various semi-literate and illiterate people in cities and rural areas of emerging economies or technologically uneducated people such as the elderly. The inability of underprivileged people to afford computers or web-enabled handheld devices and the disinterest of some people to acquire basic IT skills, lead to an undesirable blockade to a large amount of knowledge. However, access to a telephone line, wired or mobile, has certainly penetrated more among this population.

In this paper, a novel wiki implementation is presented which is based on VoiceXML and other W3C standards. The proposed system is accessible not only visually through an HTML web browser but also acoustically through wired and mobile phones. The only requirement, i.e. a telephone line, is generally considered as a low cost requirement and therefore suitable for a wide range of people. The acoustic representation of the wiki content contributes to the purpose of pervasive learning and offers great help to disabled individuals and people in developing countries.

The remaining of this paper is organized as follows: in section 2, previous work in the area is discussed. Section 3 provides an overview of the technologies and standards used in the proposed implementation. Section 4 describes the system architecture in detail, while section 5 presents some real-life scenarios. Section 6 identifies future work that can possibly improve the proposed application. Last section draws a conclusion.

2. PREVIOUS WORK

During the recent past, numerous efforts have been made to acoustically access information which was originally intended to be accessed visually. Motivated by the need to aid visually impaired individuals, applications were also developed to support individuals without having access to the Internet.

DAISY Consortium is an organization whose mission is to develop, integrate and promote standards, technologies and implementation strategies to enable global access by people with reading disabilities to information provided by mainstream publishers, governments and libraries [1]. A DAISY Digital Talking Book (DTB) is a file with a specific format which can be accessed either from an appropriate software in a PC or from a special device with DTB playback capability. Though feasible, to apply this approach in education would require a conversion of the material to this specific format. Unlike VoiceXML, DTB is not a widely adopted standard and therefore it is not appropriate for developers or suitable for dynamic systems like wikis. Additionally from the user's point of view DTB requires software which in turn requires a PC and the related skills or a special device. Both lead to extra costs for the end user.

In [2] a non-visual web browser is presented. It enables visually impaired people to navigate contents of web pages acoustically. A synthetic speech feature transforms the contents of web pages into sound and the open source Sphinx voice recognition engine [12] transforms the user's voice into signals which are recognized by the system. In this way one can navigate through pages with his voice only and can avoid the sequential narration of their content. This application provides significant advantages especially to partially blind users that are not willing to invest time and effort to learn new communication means. On the other hand it runs exclusively on PCs, making it inaccessible to users who do not own or do not know how to use a computer. In addition there is no (lightweight) version targeting to mobile devices making it inappropriate for roaming users, thus reducing pervasiveness.

In reference [3] a Wiki application similar to the one described in this paper is presented. It is based on the observation that mobile phones have penetrated more than the Internet into young people, becoming a fashion and in some countries, like the developing ones, has clearly dominated over it Under this assumption the proposed service waits for a Short Message Service (SMS) signal from the user's mobile phone with the title of the article he wishes to hear. After a while the service calls the user to his mobile phone and speaks the article content via a synthetic voice. During the call the user can navigate to the different sections of the article by pressing keys in his phone. The application is totally based on open source software components such as MediaWiki [4]. This application is addressed to students in emerging economies who usually are not familiarized with the use of PCs, or simply cannot afford one. On the other hand the fact that such a wiki is based on a mobile phone feature it makes it inaccessible to people, e.g. the elderly, who do not own one or cannot or do not know how to send an SMS. Although mobile phones are very popular nowadays, ordinary PSTN phones are still used by the majority of people.

3. RELATED STANDARDS

The proposed application harnesses a set of W3C standards such as VoiceXML, SRGS and SSML in order to satisfy the demands that such an application has. Note that besides VoiceXML the W3C has defined the following languages that can be combined with VoiceXML in order to enhance the user experience and the overall effectiveness of the application.

3.1 VoiceXML

VoiceXML is the most popular language for specifying audio user interfaces for web applications [5]. Actually, it is an official WWW Consortium (W3C) recommendation. VoiceXML is used to control the flow of the dialog between the user and the computer. Its intention is to be the audio equivalent of HTML. Applications based on VoiceXML require a voice browser which in turn has to include a speech synthesis and a speech recognition engine, in analogy to HTML applications which assume the existence of graphical browser, screen, keyboard and mouse. Today VoiceXML is widely used for the creation of voice interactive applications such as: phone banking, ticket reservation, news and weather information.

As an extensible language, VoiceXML inherits all the advantages of XML, such as portability among heterogeneous platforms. Since it is simple, it has a small learning curve and because it is an international standard, tested and matured over the years, it can be utilized by developers in order to extend existing web applications. In the context of the proposed application both static and dynamically produced VoiceXML files (with the help of ASP.NET scripts that reside on the web server) control the flow

of the application. These files contain instructions that among others manage the sequence of voice prompts, the expected timing of the user's voice input, redirection procedures to other pages, presentation of the results, error handling etc.

3.2 SSML

The Speech Synthesis Markup Language (SSML) [6] enables developers to specify instructions to the speech synthesis engine regarding the pronouncement of specific words or phrases. By providing tags which controls speech attributes such as pronunciation, volume, pitch and rate across different synthesis-capable platforms, it improves the overall aesthetic result of the speech synthesizer. It is stressed that SSML may be embedded in VoiceXML scripts or referenced as a standalone file. In the presented implementation SSML tags are relatively small in number, thus SSML code is embedded in the VoiceXML file.

3.3 SRGS

The Speech Recognition Grammar Specification (SRGS) [7] enables developers to specify words or phrases that should be expected or could be recognized by the speech recognition engine. At present, speech recognition engines are not capable of accepting any vocal input from the user, recognize it and possibly translate it to text. The developer must provide a predefined set of words as input that a user is expected to say at various stages of the application. This set of words is usually described in a corresponding grammar. A grammar can be embedded inside the VoiceXML file or in most cases it can be a single file and just be referenced in the appropriate VoiceXML file. The grammar file may also be static or it can be produced dynamically. A static file might contain a set of predefined words that are not expected to change during the application execution. In the proposed application such words are used as voice commands for browsing i.e. repeat, back, next to mention just a few. A dynamic grammar file is necessary when the expected terms may change, be added, amended or deleted very often. This is the case for the proposed wiki application in which article names that the user may search for, change frequently.

Table 1 provides a simple example of a VoiceXML document with some additional embedded SSML and SRGS tags. After a greeting message the application waits for half a second and prompts the user to select a category of topics. If the user pronounces one of the categories: history, science, environment, the application recognizes the user's choice and redirects him to the appropriate dialog for further choices.

4. SYSTEM ARCHITECTURE

Figure 1 depicts a high level view of the system architecture. The system is comprised of four major components.

Database: The database is a key system's architecture component in which application data are stored and from where they are retrieved. It holds information regarding the contents of the articles, their previous versions, the registered users (the users that have the privileges to create new articles) etc. The complexity of the database scheme is kept low. It is stressed that the content of the articles is stored in the database mixed with presentation information. These are nothing but special sequences of characters, usually referred to as WikiText that a user inserts during the editing of an article.

Web Server: On the web server resides the Wiki engine which is a web application responsible for the following tasks:

1. To constantly wait for the user's requests, either from a web browser, thus starting a web page session, or from a voice browser, thus starting a voice session.
2. To communicate with the database to retrieve or update article data.
3. In the case of voice session, the wiki engine has to generate a dynamic grammar of the available article titles, one of which will be the user's choice for presentation. In a wiki application the number of existing articles (and as a result the terms that the user is expected to ask for) might take large proportions. That would lead to the generation of a large grammar file which would affect the overall performance of the system and consequently the user's overall experience. For this reason, articles are categorized and the user is requested to first choose the category of the article and then ask for the articles title. In this way grammar files are kept smaller.

Table 1 - Example of a VoiceXML document (normal text) with embedded SSML tags (italics) and SRGS tags (bold)

```
<?xml version="1.0"?>
<vxml version="2.0" xmlns="http://www.w3.org/2001/ vxml">
 <form>
  <block>
   <prompt>
     Welcome to the <emphasis>Voice Wiki</emphasis>
     <break time='500ms'/>
     Please, provide a category.
      <grammar>
        <rule id="category">
          <one-of>
          <item>history</item>
          <item >science</item>
          <item >environment</item>
           .
           .
           .
          </one-of>
        </rule>
     </grammar>
   </prompt>
  </block>
 </form>
</vxml>
```

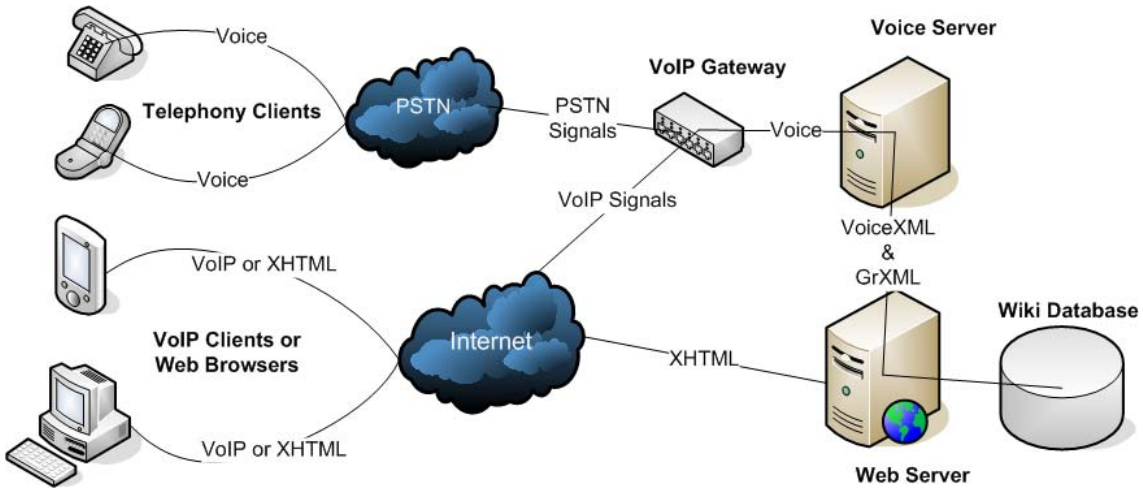

Figure 1 - System Architecture

Table 2 - The corresponding XHTML and VoiceXML forms of WikiText

WikiText	XHTML	VoiceXML
== Local Area Network == '"Contents"' * History * Technical aspects * See also * References === History === The first '" LAN "' put into service occurred in 1964 at the Livermore Laboratory to support atomic weapons research. LANs spread to the public sector in the late 1970s and were used to create high-speed links between several large central computers at one site. Of many competing systems created at this time…	Local Area Network <table> <tr> <td>Contents</td> </tr> <tr> <td> History Technical aspects See also <a href="# References"> References </td> </tr> </table> History The first LAN put into service occurred in 1964 at the Livermore Laboratory to support atomic weapons research. LANs spread to the public sector in the late 1970s and were used to create high-speed links between several large central computers at one site. Of many competing systems created at this time…	You are now listening to article entitled: <emphasis> Local Area Network </emphasis>. <break time='500ms'/> <menu> <choice next="#History" accept="approximate"> Say 1 to move to paragraph: History. </choice> <choice next="#Technical_aspects" accept="approximate"> Say 2 to move to paragraph: Technical aspects. </choice> <choice next="#See_also" accept="approximate"> Say 3 to move to paragraph: See also. </choice> <choice next="#References" accept="approximate"> Say 4 to move to paragraph: References. </choice> </menu> <break time='500ms'/> <form id="History"> The first <emphasis> LAN </emphasis> put into service occurred in 1964 at the Livermore Laboratory to support atomic weapons research. LANs spread to the public sector in the late 1970s and were used to create high-speed links between several large central computers at one site. Of many competing systems created at this time…

4. To transform wikitext sequences, retrieved from the database either in XHTML in the case of a web page session or in VoiceXML in the case of a voice session. An example of the transformation of wikitext into XHTML or VoiceXML is presented in Table 2.
5. To send the generated XHTML files to the user's web browser or the VoiceXML files to the system's Voice Server for further processing.

Voice Server: The voice server is the component responsible for the transformation of text documents to audio data. The voice server consists of a Voice Browser, a Text To Speech (TTS) engine and an Automatic Speech Recognition (ASR) engine. Additionally, a VoIP gateway is an additional component that plays an important role during the transformation, but is not an actual component of the voice server itself. The Voice Browser:

1. Receives a request from the user and executes multiple tasks which include:
2. Receives VoiceXML and grammar files from the Web Server.
3. Specifies the execution flow according to the instructions in the VoiceXML file. It isolates the text to be spoken, forwards it to the TTS engine and it forwards the grammar to the ASR engine.
4. Generates the request made by the user and forwards it to the Web Server.

The Text To Speech engine receives the text from the VoiceXML file meant to be spoken, transforms it into streaming sound and sends it to the VoIP Gateway to forward it to the end-user. On the other hand, the Automatic Speech Recognition engine receives a grammar, that is a set of terms that is able to recognize along with the client prompt and identifies if the prompt corresponds to any word in the grammar. If true it returns the term textually. The VoIP gateway receives calls from the Public Switched Telephone Network (PSTN), converts PSTN signals to VoIP signals and forwards them to the Voice Server. It is to be noted that the voice server will accept only VoIP signals. Signals originating from the Internet (from VoIP clients) might be Session Initiation Protocol (SIP) [9] signals, e.g. from Xlite softphone, or signals of some proprietary VoIP protocol, like Skype. The VoIP gateway is also responsible for the transformation of VoIP signals from the Internet to the protocol that the Voice Server recognizes. The Web Server and Voice Server components are depicted in Figure 2.

Clients: The system might accept different types of clients. A client may be a typical web browser installed on the user's PC or PDA for instance Firefox, Internet Explorer or IE mobile. It might also be a VoIP client program installed on a PC, like Skype. Finally, a client might be a wired or wireless connected phone that places its request through a PSTN network.

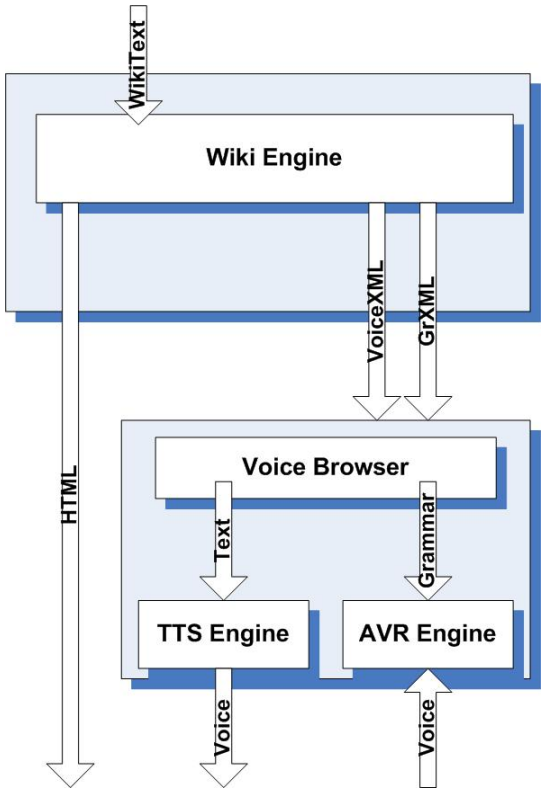

Figure 2 – Wiki's Web and Voice Server

In the proposed application the MS SQL Server 2005 [8] was used to store the wiki data. A wiki engine application was developed on ASP.NET scripts. MS Speech Server 2007 was selected for the voice server. MS Speech Server contains a powerful TTS engine as well as an advanced ASR engine. The application was tested with the Xlite softphone 3.0. Since this particular software uses the SIP protocol and Speech Server 2007 requires the messages to be in SIP as well the use of VoIP gateway was not necessary.

The proposed wiki implementation converts article text to audio on the voice server and sends it to the client through the Internet or the PSTN. This fact revokes the need for a voice browser with a TTS engine, or any other software that performs analogous tasks, to be installed on the client's PC. Also, enables clients to access the Wiki content via a phone. However, the client will receive voice streams instead of simple VoiceXML documents which will have a large size resulting to longer response delays. Also the architecture may have large implementation costs because it requires additional hardware and proprietary software like TTS and ASR engines.

5. REAL LIFE SCENARIOS

When a request is generated to the web server, the wiki engine analyzes it and produces the appropriate query to retrieve the corresponding data from the database. In the case when the request comes from a web browser then the web server produces the result in an XHTML page and sends it to the user through the Internet.

On the other hand, a call might originate either from the PSTN network or the Internet, so the VoIP gateway transforms the

signals to the appropriate VoIP protocol and forwards it to the Voice Server. The proposed application was tested by interacting with the Voice Server using the SIP protocol. Note that a dialplan that maps the phone number with the IP of the voice server must exist. If the server accepts the SIP invite message the application initiates and allocates resources for the speech recognizer, the speech synthesizer, and the Dual Tone Multi-frequency (DTMF) processing engine components. When a call is made to the voice server (by a wired or mobile phone) the voice server starts to interact with its voice browser component. Just like a normal web browser the voice browser places a request to the web server for a document containing the information requested. The web server responds with a dynamically produced document.

Unlike the first case where the web server produced an XHTML document in this case the server produces a VoiceXML document. The voice browser receives it, interprets the XML mark-up, and redirects the result to its TTS engine component. The TTS engine produces audio stream based on the results it received. The voice server will convert the audio signals to packets according to a VoIP protocol such as SIP or to analog voice signals with the help of the appropriate hardware component installed on it. If the client made the call from a SIP (soft)phone then the voice data will be carried through the Internet to reach the client. If the client made the call from a wired phone the voice signals will be carried through the PSTN network instead.

Table 3 presents an example of a user-system interaction scenario. The system will try to get a confirmation in each critical user's input such as the article category and title. Also, there exists a set of navigational commands such as NEXT, PREVIOUS and BEGIN. Finally, a table of contents is created and presented to the user under request if he wishes to hear only certain parts of the article.

6. FUTURE WORK

Although the presented implementation is functional there is space for improvement, in order to make it more ergonomic and thus enhance the user's overall experience. Many features can be added in order to better fulfill its purposes. Some proposals might be:

1. **Use of natural language:** Natural language lets user answer questions set by the application or speak application commands in an open way, that is without limiting the user's responses. Obviously this yields for sophisticated construction of grammar files but the improvement of the users experience will be significant.
2. **Multilanguage speech synthesis:** The implementation supports only recognition and speech synthesis for the English language. Expansion of the application to support other languages is considered a necessity since one of its main goals is to address and attract users from emerging economies most of who do not speak English.
3. **Provide the ability for voice annotations:** In its current form, the application supports pronunciation of the content of existing articles via synthetic speech. It does not support editing or creation of new articles due to the incapability of existing recognition engines to provide full speech recognition instead of recognition of words from a predefined set. Nevertheless, one of the basic characteristics of a wiki is the fact that it can be editable by almost every user. To overcome this, a possible solution is to provide the user with the ability to add voice annotations to the article by recording his voice. The large file size of the voice annotations in combination with the frequent article updates might cause problems to users accessing articles via standard web browsers.

Table 3 - A user - system interaction scenario

Wiki	Welcome to the Voice Wiki. Say HELP for navigation instructions or say SEARCH to search for article.
User	Help
Wiki	Say BEGIN to read an article from the beginning. While reading say CONTENTS to hear the article contents. Say NEXT to skip to the next paragraph. Say PREVIOUS to move back to the previous paragraph. Say HELP to hear navigation instructions.
Wiki	Say HELP for navigation instructions or say SEARCH to search for article.
User	Search
Wiki	Please say one of the following categories. ART, COMPUTERS, HISTORY, SCIENCE.
User	COMPUTERS
Wiki	Did you say COMPUTERS?
User	Yes
Wiki	Say your search term.
User	Local Area Networks
Wiki	Did you say LOCAL AREA NETWORKS?
User	Yes
Wiki	One article found. Article name is Local Area Network. Please say how to proceed.
User	Begin
Wiki	A local area network is a computer network covering small geographic area like home, office or group of buildings …
User	Next
Wiki	The first LAN put into service occurred in 1964 at the Livermore Laboratory to support atomic weapons research. LANs spread to the public sector…

4. **Migrate to open source components:** Our application is implemented in Microsoft's Speech Server 2007 which is a proprietary software, that demands from users (owners of the wiki application) to acquire a license in order to be used extensively. Open source alternatives do exist. For example, OpenVXI [10] is an open source VoiceXML interpreter toolkit. It is intended to be a component of Voice Browsers and provides APIs for speech recognition, speech synthesis and telephony services. Festival [11] is a general speech

synthesis system that offers several APIs for speech synthesis and a rich development environment. Also, Sphinx [12] is an open source speech recognition system. The main disadvantage is the absence of a single platform that integrates all these functions and collaboration problems may occur among these components.

7. CONCLUSIONS

In this paper, a novel wiki application that can be accessed by virtually any wired or wireless phone as well as by a common web browser was presented. A strong e-learning tool, such as wiki was made accessible from practically everywhere. Since common telephones are installed in almost every home, our application can bring wikis closer to a wider range of people who do not own or are not comfortable with the use of a PC. Finally, the audio nature of this application is expected to enhance the learning experience of visually impaired people. The advantages of the proposed implementation over similar existing implementations are: (a) it does not require installation of special software or a PC for someone to listen to wiki articles, (b) it is cost efficient for the end user, and (c) it accepts voice commands for navigation throughout articles and for controlling the application flow.

8. REFERENCES

[1] The DAISY Consortium. Retrieved on May 5, 2008 from http://www.daisy.org/

[2] Yevgen Borodin, Jalal Mahmud, I.V. Ramakrishman, Amanda Stent, 2007 The HearSay Non-Visual Web Browser, ACM International Conference Proceeding Series, Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A), Vol. 225, pp. 128-129, Banff, Canada, 2007.

[3] Teemu Leinonen, Eunice Ratna Sari, Francois Aucamp, Audio Wiki for Mobile Communities: Information System for the Rest of Us, Workshop on speech in mobile and pervasive environments, Mobile HCI 06 Conference, pp 3-5, Espoo, Finland, 2006.

[4] MediaWiki Retrieved on May 5, 2008. from http://www.mediawiki.org/wiki/MediaWiki

[5] Voice Extensible Markup Language (VoiceXML) Version 2.0 Retrieved on April 5, 2008 from http://www.w3.org/TR/voicexml20/

[6] Speech Synthesis Markup Language (SSML) Version 1.0 Retrieved on April 2, 2008 from http://www.w3.org/TR/speech-synthesis/

[7] Speech Recognition Grammar Specification Version 1.0 http://www.w3.org/TR/speech-grammar/

[8] Microsoft SQL Server 2005 Retrieved on April 13, 2008 from http://www.microsoft.com/sql/default.mspx

[9] Session Initiation Protocol Retrieved on May 5, 2008 from http://www.cs.columbia.edu/sip/

[10] Vocalocity's OpenVXI 3.0. Retrieved on May 5, 2008 from http://www.speech.cs.cmu.edu/openvxi/

[11] The Festival Speech Synthesis System. Retrieved on May 5, 2008 from http://www.cstr.ed.ac.uk/projects/festival/

[12] The CMU Sphinx Group Open Source Speech Recognition Engines Retrieved on May 5, 2008 from http://cmusphinx.sourceforge.net/html/cmusphinx.php