Media Conversion from Visual to Audio: Voice Browsers

Adel Youssef, Ben Shneiderman
{adel, ben}@cs.umd.edu
Department of Computer Science
University of Maryland
College Park, MD 20742
CS-TR-4426
April 2000

ABSTRACT
There is a large amount of information on the World Wide Web that is at the fingertips of anyone with access to the internet. However, so far this information has primarily been used by people who connect to the web via traditional computer. This is about to change. Recent advances in wireless communication, speech recognition, and speech synthesis technologies have made it possible to access this information from any place, and at any time. In this paper, we discuss voice browsers as compared to current web browsers. Some of the primary techniques of universal accessible design are listed with their relation to voice browsers and some ideas are offered to help authors implementing these considerations when designing web pages. The new voice markup language is briefly discussed.

Keywords
WWW, web browser, audio, voice, HTML, XML, aural style sheets, voice markup language

INTRODUCTION

In the world today, far more people have access to a telephone than have access to a computer with an Internet connection. In addition, sales of cell phones are booming, so that many of us have already or soon will have a phone within reach wherever we go. Voice browsers offer the potential to expand the reach of the Web beyond the desktop or laptop and offer information in ways that most content authors have never imagined: Telephone access to web pages; browsers for the visually impaired; hands-free web surfing while driving a car; reading and language instruction for children and adult learners; intelligent alarm clocks that parse the day's news and present summaries upon verbal request [10, 11].

Currently, it is common for companies to offer services over the phone via menus traversed using the phone's keypad [1, 5]. Voice Browsers [14, 16] offer a great fit for the next generation of call-centers, which will become portals to the company's services and related web sites, whether accessed via the telephone network or via the Internet. Users will able to choose whether to respond by a key press or a spoken command.

The Web will reach this state, soon. It is not a matter of if, but when. The web author's key to unlocking this potential is the concept of Universally Accessible Design: Creating pages which not only "look good" in today's browsers, but which are usable by both yesterday's simpler user agents and the diverse network access devices that will characterize the Web of the 21st century [4].

Voice browsers allow people to access the Web using speech synthesis, pre-recorded audio, and speech recognition. This can be supplemented by keypads and small displays. Voice may also be offered as an adjunct to conventional desktop browsers with high resolution graphical displays, providing an accessible alternative to using the keyboard or screen, for instance in automobiles where hands/eyes free operation is essential. Voice interaction can escape the physical limitations on keypads and displays, as devices become ever smaller.

WHAT IS A " VOICE BROWSER "?

> "*A device which interprets a (voice) markup language and is capable of generating voice output and/or interpreting voice input, and possibly other input/output modalities. [*16*]*"

The definition of a voice browser, above, is a broad one but what makes a software system that interacts with the user via speech a "browser"? From an end-user's perspective, the impetus is to provide a service similar to what graphical browsers of HTML and related technologies do today, but on devices that are not equipped with full-browsers or even the screens to support them. Much of the work done concentrates on using the telephone as the first voice-browsing device [6, 17, 18, 24]. This is not to say that it is the preferred embodiment for a voice browser, only that the number of access devices is huge, and because it is at the opposite end of the graphical-browser continuum, which highlights the requirements that make a speech interface viable.

WEB PAGE DESIGN STRATEGIES FOR VOICE BROWSERS

> "*A properly constructed web document is an accessible web document [*4*]*"

The major obstacle to wide-scale commercial deployment of voice browsers for the web is not the technology, but the ease (or difficulty!) with which web page designers can add speech support to their site [9]. Authoring a web page for any specific type of user agent or system configuration, should never be a completely separate subject with arcane new techniques developed for each special need, but rather an application of the common set of Universally Accessible Design principles that should be part of every web author's repertoire. With few exceptions, pages should never be designed for certain types (or brands) of browsers, but should instead be designed for all uses (and potential uses) of the information. All web documents should be equally accessible to voice browsers as to visual user agents [4].

The HTML Writer's Guild studies, and discussions with web authors, have shown that the primary obstacle to universal accessibility is ignorance. There are few cases where a conscious decision has been made to produce a

generally inaccessible web page; rather, the author is simply unaware of the need to create accessible pages and the techniques by which that is done. Once enlightened, most web authors eagerly embrace the concept of universal accessibility, since the benefits are many and obvious.

In this section, some of the primary techniques of Universally Accessible Design will be briefly listed as they relate to voice browsers, and offer ideas as to how authors can implement these considerations when designing web pages.

a. Aural Style Sheets

Aural Style Sheets [10, 21, 22, 23] are part of the Cascading Style Sheets, Level 2 specification, and provide for a level of control in spoken text roughly analogous to that for displayed/printed text. Aural rendering of a document is already commonly used by the blind and print-impaired communities. It combines speech synthesis and "auditory icons." Often such aural presentation occurs by converting the document to plain text and feeding this to a screen reader [20], software or hardware that simply reads all the characters on the screen. This results in less effective presentation than would be the case if the document structure were retained. Style sheet properties for aural presentation may be used together with visual properties (mixed media) or as an aural alternative to visual presentation [4].

The use of an aural style sheet (or aural style sheet properties included in a general style sheet document) allows the author to specify characteristics of the spoken text such as volume, pitch, speed, and stress; indicate pauses and insert audio "icons" (sound files); and show how certain phrases, acronyms, punctuation, and numbers should be voiced.

Combined with the @media selector for media types, a well-crafted aural style sheet can greatly increase the accessibility of a web document in a voice browser. Further investigation in this area is encouraged, especially in the area of example aural style sheets and suggestions for authoring techniques.

b. Rich Meta-Content

HTML 4.0 [15] gives the author the ability to embed a great deal of meta-content into a document, specifying information which expands on the semantic meaning of the content and allows for specialized rendering by the user agent. In other words, by using features found in HTML 4.0 (and to a limited extent, in other versions of HTML), an author can give better information to the browser, which can then make the document easier to use.

Judicious and ample use of meta-content within a document allows the author to not simply specify the content, but also suggest the meaning and relationship of that content in the context of the document. Voice browsers can then use that meta-information as appropriate for their presentation and structural needs.

c. Planned Abstraction

One use for meta-content information is the development of pages, which are designed to be *abstracted*. The typical web document found on the web can often be quite lengthy; finding information by listening to web page read out loud takes longer than visually scanning a page, especially when most web pages are designed for visual use.

Thus, most voice browsers will provide a method for *abstracting* a page; presenting one or more outlines of the page's content based on a semantic interpretation of the document.

Examples of potential or valid abstraction techniques include [20, 4]:

- Listing all the links and link text on a page.
- Forming a structure based on the H1, H2, ... H6 headers.
- Summarizing table data.
- Scanning for TITLE attributes in elements and presenting a list of options for expansion.
- Vocalizing any "bold" or emphasized text.
- Digesting the entire document into a summary based on keywords as some search engines provide.

There is any number of other options available for voice browser programmers to use to provide short, easily-digestible versions of web contents to the browser user. This suggests that the web author should provide as much meta-content as possible as well as careful use of HTML elements in their proper manner. Specific techniques include [20, 4]:

- Useful choices for link text (e.g., "the report is available" instead of "click here").
- Appropriate use of heading tags to define document structure, not simply for size/formatting.
- Use of the SUMMARY attribute for tables.

- Use of STRONG and EM where appropriate, providing benefits for both vocal and visual "scanability".
- Use of META elements with KEYWORDS and SUMMARY content.

d. Alternative Content for Unsupported Media Types

The "poster child" for web accessibility is the *ALT* attribute, which allows alternative text to be specified for images; if a user agent cannot display the visual image, the *ALT* text can be used instead. Widespread use of the *ALT* attribute by all sites on the Internet would likely double the accessibility of World Wide Web with such a simple change. Web authors who do not correctly use *ALT* text are seriously damaging the usability of the entire medium!

For voice browsers, *ALT* text is vitally important since images cannot be represented at all, aurally. Especially when used as part of a link, alternative content must be provided so that the voice browser can accurately render the page in a manner useful to the user.

In addition to *ALT* for *IMG* attributes, HTML 4.0 provides a number of other ways for specifying alternative content that can be used by a browser if an unsupported media type is provided. Some of those include [20, 4]:

- ALT attributes for image map AREAs, APPLETs, and image INPUT buttons.
- Text captions and transcripts for multimedia (video and audio).
- NOSCRIPT elements when including scripting languages, as voice browsers may be unable to process Java script instructions.
- NOFRAMES elements when using frame sets, as frames are a very visually oriented method of document display.
- Use of nested OBJECT elements to include a wide variety of alternative contents for many media types.

VOICE MARKUP LANGUAGE

It would be desirable for the voice browser to render interactive speech dialogs from standard HTML web pages. However, HTML has primarily been developed for visual rendering and it is conventionally used in this way. Hence, there will be elements which are not amenable to speech rendering (e.g. image maps) [15, 25], as well as web design practice which make speech rendering more awkward than visual rendering (e.g. the use of frames). Furthermore, there are many aspects of speech rendering which the designer will not have control over since there are no corresponding HTML tags. For example, using a specific type of recognizer or synthesizer, associating specific recognition grammar with input elements, control over the synthesizer volume, speed, pitch, etc and as well as interaction handlers for timeouts, errors, and so on.

One way of dealing with these problems is to deploy a specific markup language for speech rendering. This has been the strategy behind many initiatives emerging from telecommunication companies; for example, WML (Wireless Markup Language) [7] for small devices from the WAP consortium, and VoxML [13, 19] for voice browsers.

FUTURE DIRECTIONS AND ISSUES

There is much more to a voice browser being usable than devising a speech mark-up language or mimicking a telephone voice response system. Using audio only to browse the Web is much more than adding multimedia capabilities to a visual browser [8]. The following summarizes some future research directions to make voice browsers a reality:

- The problem of intelligently speaking a heavily formatted page that uses tables inside tables inside tables.
- Better WWW programming languages.
- Speaker verification for secure access.
- Requirements for interoperable mechanisms for controlling synthetic speech engines using markup language, style sheets or scripting.
- Support for multilingual documents.
- Ability to include non-speech audio samples (branding).

SUMMARY

Voice browsers offer the potential to expand the reach of the Web beyond the desktop or laptop and offer information in ways that most content authors have never imagined. Authoring pages for use by voice browsers is not difficult as long as care is taken to design for universal accessibility. Increased awareness of the need for Universally Accessible Design among the web authoring community is critical to the success of the web's expansion beyond desktop/laptop computer units. HTML as of today can be used with good results today to build intuitive voice dialogs. Adding voice capabilities to the web should be regarded as a natural evolvement of the web and not a

industry specific derivative

FURTHER READINGS

Dialog Requirements for Voice Markup Language

http://www.w3.org/TR/1999/WD-voice-dialog-reqs-19991223

This document establishes a prioritized list of requirements for spoken dialog interaction which any proposed markup language (or extension thereof) should address. Propose further work is on how the spoken dialogs can be integrated and synchronized with other input/output media to provide co-ordinated multi-modal interaction.

Grammar Representation Requirements and Preliminary Specifications for Voice Markup language

http://www.w3.org/TR/1999/WD-voice-grammar-reqs-19991223

The main goal of this document is to define a speech recognition grammar specification language that will be generally useful across a variety of speech platforms used in the context of a dialog and synthesis markup environment. The idea is to establish an appropriate set of requirements for grammar specifications, evaluate existing grammar languages for satisfaction of requirements, settle upon a language specification or modify as necessary, and finally deliver a specific language proposal to the full W3C working group.

Natural Language Processing Requirements for Voice Markup Language

http://www.w3.org/TR/1999/WD-voice-nlu-reqs-19991223

This document specifies requirements that define the capabilities of any component of a voice browser system which performs natural language interpretation, that is, the task of determining and representing the content of a natural language input from a user. Interpretation components include both stand-alone natural language understanding (NLU) components which receive text string results from a speech recognizer or keyboard as well as speech recognizers that incorporate natural language understanding functionality by returning interpretations rather than, or in addition to, text strings.

Speech Synthesis Markup Requirements for Voice Markup Language

http://www.w3.org/TR/1999/WD-voice-tts-reqs-19991223

The main goal of this subgroup is to establish a prioritized list of requirements for speech synthesis markup which any proposed markup language should address. This document addresses both procedure and requirements for the specification development.

The International Phonetic Alphabet

http://www.arts.gla.ac.uk/IPA/ipa.html

The International Phonetic Alphabet is defined by the International Phonetic Association (IPA). The IPA is the major as well as the oldest representative organization for phoneticians. The aim of the Association is to promote the scientific study of phonetics and the various practical applications of that science. In furtherance of this aim, the Association provides the academic community world-wide with a notational standard for the phonetic representation of all languages. The latest version of the IPA Alphabet was published in 1993 (updated in 1996).

VoiceXML

http://www.vxml.org/

The VoiceXML Forum formed by AT&, IBM, Lucent and Motorola to pool their experience. The Forum has published an early version of the VoiceXML specification. This builds on earlier work on PML, VoxML and SpeechML.

SpeechML

http://www.alphaworks.ibm.com/tech

SpeechML plays a similar role to VoxML [13], defining a markup language written in XML for IVR systems [1, 5]. SpeechML features close integration with Java.

TalkML

http://www.w3.org/Voice/TalkML

This is an experimental markup language from HP Labs, written in XML, and aimed at describing spoken dialogs in terms of prompts, speech grammars and production rules for acting on responses. It is being used to explore ideas for object-oriented dialog structures, and for next generation aural style sheets.

Java Speech Grammar Format

http://www.java.sun.com/products/java-media/speech/forDevelopers/JSGF/index.html

The Java™ Speech Grammar Format (JSGF) is used for defining context free grammars for speech recognition. JSGF adopts the style and conventions of the Java programming language in addition to use of traditional grammar

notations.

Java Speech Grammar Format Specification, Beta Version 0.6, April 1998 is located at
http://www.java.sun.com/products/java-media/speech/forDevelopers/JSGF/index.html

Java Speech Markup Language Specification, Beta Version 0.5, August 1998, is located at
http://java.sun.com/products/java-media/speech/forDevelopers/JSML/index.html

Java Speech Application Programming Interface, Beta Version 0.7, June 1998, is located at
http://www.java.sun.com/products/java-media/speech

SABLE
http://www.bell-labs.com/project/tts/sable.html

SABLE is a markup language for controlling text to speech engines. It has evolved out of work on combining three existing text to speech languages: SSML, STML and JSML.

MIT Spoken Languages Systems Group
http://www.w3.org/Voice/1998/Workshop/MIT-SLS.html

The Spoken Language Systems Group at the MIT Laboratory for Computer Science is devoted to research that will lead to the development of interactive conversational systems. The group formulates and tests computational models and develops algorithms that are suitable for human computer interaction using verbal dialogues. These research results are funneled into the development of experimental conversational systems with varying capabilities.

REFERENCES

1. Atkins, D., T. Ball, T. Baran, M. Benedikt, K. Cox, D. Ladd, P. Mataga, C. Puchol, J. Ramming, K. Rehor, and C. Tuckey. "Integrated Web and Telephone Service Creation." *Bell Labs Technical Journal*, pp. 19-35, Winter 1997.

2. Aural Style Sheets
   http://www.w3.org/TR/REC-CSS2/aural.html

3. Aural Cascading Style Sheets (ACSS). W3C Note
   http://www.w3.org/Style/css/Speech/NOTE-ACSS

4. Bartlett, Kynn. "Web Authoring Strategies for Voice Browsers." Position Paper for The W3C Workshop on Voice Browsers, Cambridge, Massachussets, October 1998.
   http://www.hwg.org/opcenter/w3c/voicebrowsers.html

5. Brown, Michael K. and B.M. Buntschuh. "A Speech Controlled World Wide Web Browser." *AT&T Bell Laboratories Internal Memorandum*, Nov. 28, 1995.

6. Brown, Michael K., Stephen C. Glinski, Bernard P. Goldman, and Brian C. Schmult. "PhoneBrowser: A Web-Content-Programmable Speech Processing Platform." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998.
   http://www.w3.org/Voice/1998/Workshop/Michael-Brown.html

7. Cover, Robin. "WAP Wireless Markup Language Specification (WML)." XML Cover Page, March 2000.
   http://www.oasis-open.org/cover/wap-wml.html

8. Danielsen, Peter, Nils Klarlund, David Ladd, Peter Mataga, J. Christopher Ramming and Kenneth Rehor. "Requirements for a markup language for HTTP-mediated interactive voice response services." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998.
   http://www.research.att.com/~klarlund/external/from_my_server/articles/W3C-requirements-markup-language-http-med-IVR.html

9. Glashan, Scott Mc. "Standards for voice browsing." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998.
   http://www.w3.org/Voice/1998/Workshop/ScottMcGlashan.html

10. Hemphill, C. and P. Thrift. "Surfing the Web by Voice." Proceedings of ACM Multimedia, San Francisco, CA, November 7-9, 1995, pp. 215-222.

11. Hemphill, C. and Y. K. Muthusamy. "Developing Web-based Speech Applications." Proceedings of Eurospeech '97, Rhodes, Greece, September 1997, Vol. 2, pp. 895-898.

12. James, F. "Lessons Developing Audio HTML Interfaces, Assets." Proceedings of the Conference on Assistive Technologies, Marina Del Ray, USA, April 1998.

13. Motorola, VoxML: The Mark-up Language for Voice.
   VoxML site: http://www.voxml.com/

VoxML references spec: http://www.w3.org/Voice/1999/VoxML.pdf

14. Raggett, Dave and Or Ben-Natan. "Voice Browsers." *W3C Working Draft*, Oct. 1998. http://www.w3.org/TR/NOTE-voice

15. Raggett, Dave, Arnaud Le Hors and Ian Jacobs. "Hypertext Markup Language (HTML) Version 4.0." December 1998. http://www.w3.org/TR/REC-html40/

16. Robin, Michael and Jim Larson. "Voice Browsers: An introduction and glossary for the requirement drafts." *W3C Working Draft*, December 1999. http://www.w3.org/TR/voice-intro/

17. Stallard, David. "BBN Position Paper on Conversational Web Access." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998. http://www.w3.org/Voice/1998/Workshop/DaveStallard.html

18. Thatcher, Jim, Phill Jenkins, and Cathy Laws. "IBM Special Needs Self Voicing Browser." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998. http://www.w3.org/Voice/1998/Workshop/PhilJenkins.html

19. The Extensible Markup Language (XML). http://www.w3.org/XML/

20. Vanderheiden, Gregg, Wendy Chisholm, and Neal Ewers. "Making Screen Readers Work More Effectively on the Web." Trace Research & Development Center Draft, March 1996. http://trace.wisc.edu/docs/screen_readers/screen.htm

21. Wium-Lie, Håkon and Bert Bos. "Cascading Style Sheets, level 1." December 1996. http://www.w3.org/TR/REC-CSS1-961217.html

22. Wium-Lie, Håkon, Bert Bos and Ian Jacobs. "Cascading Style Sheets, level 2." http://www.w3.org/TR/WD-CSS2/

23. Wynblatt, Michael, D.Benson, and A. Hsu. "Browsing the World Wide Web in a Non-Visual Environment." *Proceedings of the International Conference on Auditory Display (ICAD)*, Palo Alto, CA, USA, pages 135-138, November 1997. http://www.santafe.edu/~kramer/icad/websiteV2.0/Conferences/ICAD97/Wynblatt.pdf

24. Wynblatt, Michael and Stuart Goose. "Towards Improving Audio Web Browsing." *Position Paper for The W3C Workshop on Voice Browsers*, Cambridge, Massachussets, October 1998. http://www.w3.org/Voice/1998/Workshop/Siemens.html

25. Yankelovich, Nicole, Gina-Anne Levow, and Matt Marx. "Designing SpeechActs: Issues in Speech User Interfaces." *CHI'95 Proceedings*, Denver, CO May 7-11, 1995.