TC-GXML

A Transcoder for HTML to XML Grammar

Raghuraj Singh, Prabhat Verma, Avinash Kumar Singh
Computer Science and Engineering Department
Harcourt Butler Technological Institute,
Kanpur-208002, India
rscse@rediffmail.com,pvluk@yahoo.com,avinashsingh_87@hotmail.com

Abstract--**This paper discusses a transcoder, TC-GXML, designed and developed by us as a part of an ongoing research project related to design and development of a speech based web-browsing system for visually challenged.**

It is always easier to develop a visual application than a non-visual application based on speech interface. There are issues like Non-Standard format of HTML, due to which it is difficult for a developer to design a common non-visual interface. In the design of TC-GXML, we have considered all these issue in transforming HTML to GXML and then to VoiceXML or any other XML type. TC-GXML is an improvement over existing Transcoding techniques in terms of its capability of parsing a wider range of websites of varied structure.

Keywords: Transcoder, HTML to VoiceXML, HTML to Grammar XML, GXML, VoiceXML

I. Introduction

The impact of Internet on the lives of visually-challenged has not received the attention of the mainstream society especially in Indian context. Internet can become a very effective tool for visually challenged empowering them with the knowledge and information of choices available as regards employment, independent living etc [1]. Using Internet, they can have access to the same wealth of information as sighted people and on the same terms. So we can say that the web has become an indispensable source of information and we use it more and more in our daily life. The primary mode of interaction with the web is via graphical browsers, which are designed for visual interaction. As we browse the Web, we have to filter through a lot of irrelevant data. Sighted individuals can process visual data in no time at all. They can quickly locate the information that is most relevant to them. But, this task can be time consuming and extremely difficult for people with visual disabilities. Speech based browsers are essentially sequential in processing. Therefore, clever techniques must be applied for presenting the items available on the website as per the need of the user. Important technique to do so is Transcoding. It is the direct digital-to-digital conversion of one encoding to another. This is usually done to incompatible or obsolete data in order to convert it into a more suitable format. The Voice eXtensible Markup Language (VoiceXML) is an XML-based markup language for creating two way voice applications in which the input and/or output are through a spoken, rather than a graphical user interface. VoiceXML is an emerging industry standard that has been defined by the VoiceXML Forum (http://www.voicexml.org) and accepted for submission by the World Wide Web Consortium (W3C) as a standard for voice markup on the Web. Users can access deployed VoiceXML applications.

II. Existing TrancoderS for HTML to Voice XML

A large amount of work has been done in the development of Transcoding techniques and tools. Some of them are discussed here:

The IBM WebSphere Transcoding Publisher (WTP) is a commercial product that supports an HTML to VoiceXML transcoder [2], [3]. Client-side browsers can use this proxy to obtain transcoded content. The proxy intercepts the HTTP from the client-side browser, fetches the requested HTML document, transcodes it, and forwards on the transcoded document to the browser. In the case of the HTML-to-VoiceXML transcoder, a voice browser could be configured to use the proxy to receive VoiceXML versions of HTML web pages. IBM has an HTML to VoiceXML transcoder for use with WTP that splits the HTML into two main sections in the VoiceXML code it produces: a main content section and a listing of all the links on the page. In addition, a menu is added to the VoiceXML file to allow users to navigate between the two other sections, or to exit.

Aurora Transcoding System [11] is another transcoding technique which adapts a webpage and builds it's equivalent XML document. Aurora Transcoding system supports Internet Explore and IBM Home Page Reader. Thus, it is not wrong to say that Aurora Transcoding system is specific for some particular tasks like interaction and information gathering.

Speech Application Language Tags (SALT) [5] is different from Transcoding Techniques but in some aspects it resembles. SALT directly adds Tags to HTML document and presents it to us using a special browser which interacts using graphics and voice at the same time. SALT uses programming approach to do so.

XHTML+Voice [7] is another approach recommended or provided by W3C and similar to SALT. In this approach, existing VoiceXML tags are integrated into XHTML as

978-0-7695-3958-4/10 $26.00 © 2010 IEEE
DOI 10.1109/DSDE.2010.69

compared to SALT which tries to integrate new tags to HTML.

Several Interface, Single Logic (Sisl) [8] has architecture to support single logic for multiple user interfaces. It uses Domain Specific Language (DSL) for interfacing. The idea is to convert DSL into several interfaces like VoiceXML and HTML. Sisl supports to build a system from the scratch and does not directly convert existing HTML to VoiceXML.

UIML is a language designed to build multi platform interface. It works on transformation of High Level Interaction to platform specific Transcoding. It resembles to Sisl strategy and it carries the same problem of creating multiplatform interface from scratch and not from the existing content.

Emackspeak uses a different approach than Transcoding. It analyzes the webpage and gathers the information to give auditory interface. It fails to provide voice recognition interface and requires interface through keyboard.

Tell-me [6] and BeVocal [10] are two centralized approaches that gather information on the web via phone user interface. They take advantage of VoiceXML to provide a more flexible voice interface with Interactive Voice Response (IVR).

III. WEBSITE STRUCTURE

We analyzed about 50 different websites and observed that different websites have different structure in their data representation, content structure and have non standardized format. Due to these reasons, development of TC-GXML becomes difficult. Based on the analysis, web pages can be categorized as below:

A. *Text Part*

This is part of the webpage that contains pure text i.e. text found between the tags like <h1> </h1>, <p> </p>, <b> </b>, <i> </i> etc. This part generally contains information about the website or webpages or about links on the webpages.

B. *Menu/Links Part*

This part is another interesting and difficult part to understand. Websites generally have different structures to represent their menu items like table view, hierarchical view, tree view etc. But, the good thing is that a link is always represent using < a href=” ” > Tag which helps us to develop TC-GXML.

C. *Webpage Structure*

This is last and very important part of the webpage that tells about the organization of various webpage elements such as Text, Images, Menu etc. Sometimes a webpage contains repeated links or text e.g. links represented on header as well footer of the webpage. So, the design of perfect and universal regular expression for webpage becomes even more difficult.

IV. METHOD USED FOR TRANSCODING

TC-GXML is a programming approach based on regular expression, which is a complete transcoding package. The schematic diagram of TC-GXML is shown in Fig. 1 and the whole process is described below.

A. *Convert Web Page into HTML*

Webpage (refer from Fig 2) conversion into HTML (refer from Fig 3) is a basic technique that can be done using programming languages like C# and Java. To achieve this conversion we follow the following steps:

- Request from HTTP for given URL.
- Open the connection of specified URL using methods like GET, POST, PUT etc and get content of the webpage.
- Create a HTML Document and save it at a place specified by the user of TC-GXML
- Check the encoding of HTML Document and pass it to next step.

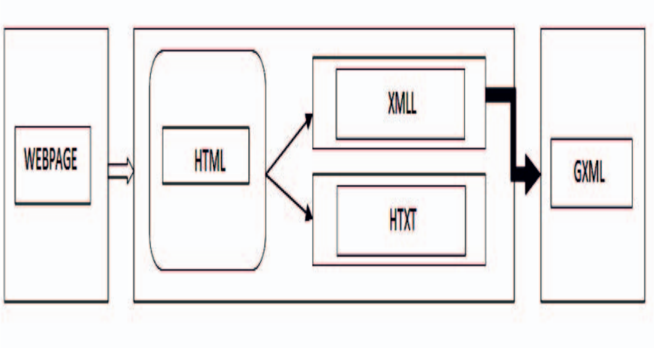
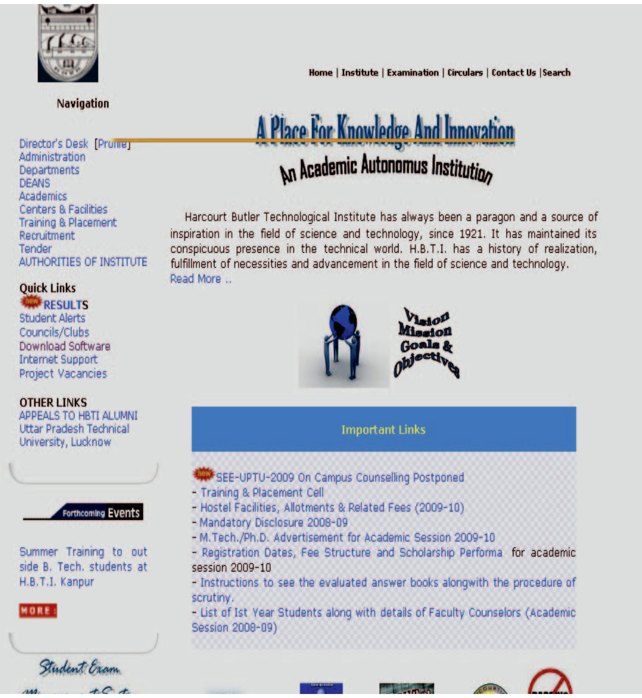


Figure 1. TC-GXML Structure



Home | Institute | Examination | Circulars | Contact Us |Search

Navigation

A Place For Knowledge And Innovation

An Academic Autonomus Institution

Director's Desk [Prun
Administration
Departments
DEANS
Academics
Centers & Facilities
Training & Placement
Recruitment
Tender
AUTHORITIES OF INSTITUTE

Harcourt Butler Technological Institute has always been a paragon and a source of inspiration in the field of science and technology, since 1921. It has maintained its conspicuous presence in the technical world. H.B.T.I. has a history of realization, fulfillment of necessities and advancement in the field of science and technology.
Read More ..

Quick Links
NEW RESULTS
Student Alerts
Councils/Clubs
Download Software
Internet Support
Project Vacancies

Vision Mission Goals & Objectives

OTHER LINKS
APPEALS TO HBTI ALUMNI
Uttar Pradesh Technical University, Lucknow

Important Links

NEW SEE-UPTU-2009 On Campus Counselling Postponed
- Training & Placement Cell
- Hostel Facilities, Allotments & Related Fees (2009-10)
- Mandatory Disclosure 2008-09
- M.Tech./Ph.D. Advertisement for Academic Session 2009-10
- Registration Dates, Fee Structure and Scholarship Performa for academic session 2009-10
- Instructions to see the evaluated answer books alongwith the procedure of scrutiny.
- List of Ist Year Students along with details of Faculty Counselors (Academic Session 2008-09)

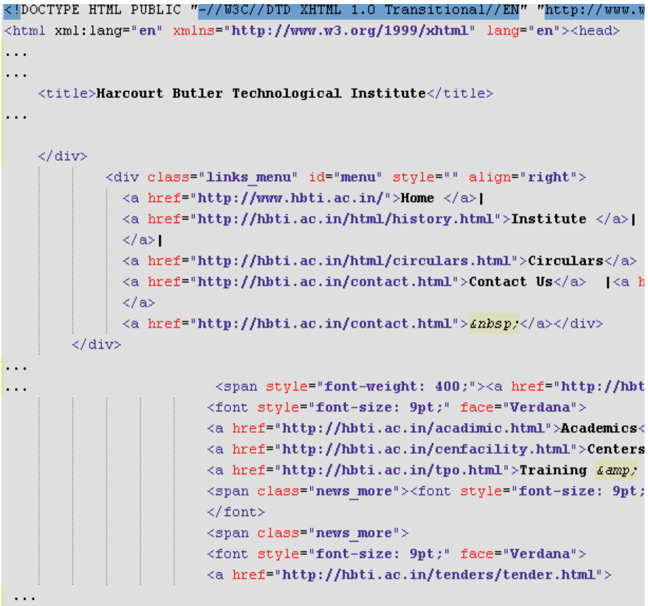
Forthcoming Events

Summer Training to out side B. Tech. students at H.B.T.I. Kanpur

MORE

Student Exam

Figure 2. Website use for get HTML



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w
<html xml:lang="en" xmlns="http://www.w3.org/1999/xhtml" lang="en"><head>
...
...
    <title>Harcourt Butler Technological Institute</title>
...

    </div>
            <div class="links_menu" id="menu" style="" align="right">
              <a href="http://www.hbti.ac.in/">Home </a>|
              <a href="http://hbti.ac.in/html/history.html">Institute </a>|
              </a>|
              <a href="http://hbti.ac.in/html/circulars.html">Circulars</a>
              <a href="http://hbti.ac.in/contact.html">Contact Us</a>  |<a h
              </a>
              <a href="http://hbti.ac.in/contact.html">&nbsp;</a></div>
        </div>
...
...                          <span style="font-weight: 400;"><a href="http://hbt
                            <font style="font-size: 9pt;" face="Verdana">
                            <a href="http://hbti.ac.in/acadimic.html">Academics<
                            <a href="http://hbti.ac.in/cenfacility.html">Centers
                            <a href="http://hbti.ac.in/tpo.html">Training &amp;
                            <span class="news_more"><font style="font-size: 9pt;
                            </font>
                            <span class="news_more">
                            <font style="font-size: 9pt;" face="Verdana">
                            <a href="http://hbti.ac.in/tenders/tender.html">
 ...
```

Figure 3. HTML Document

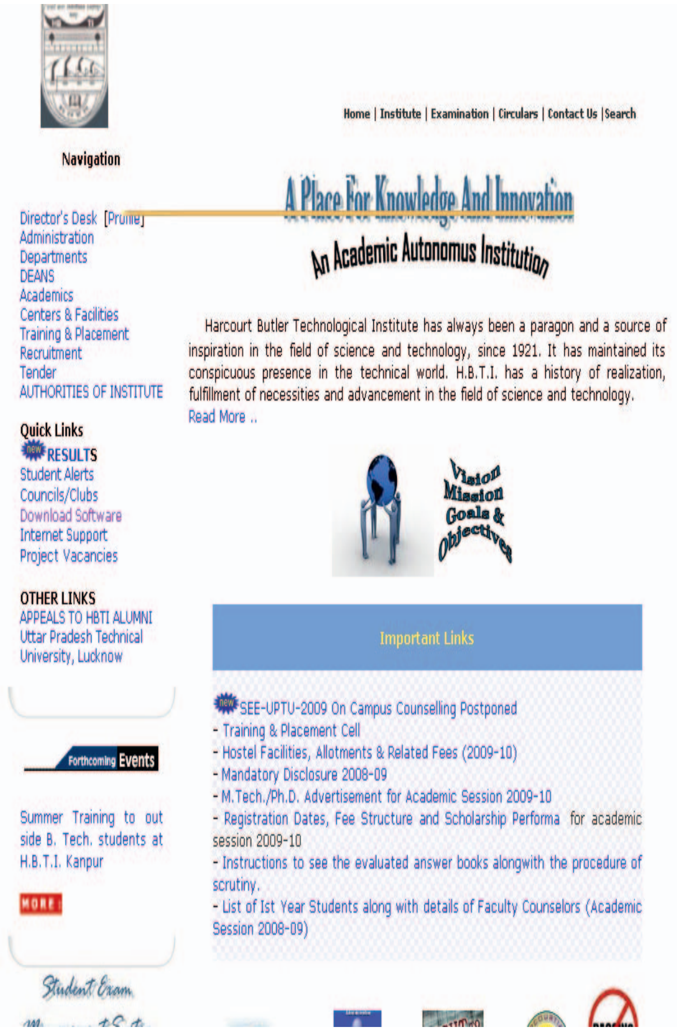
B. *Convert HTML into HX (HTXT and XMLL)*

This is the second phase of TC-GXML which coverts a pure HTML document into the XML format specified by w3c. This can be done using XSLT Schema but it is not always useful and fast. If there are 'N' HTML documents and we want 'M' XML type object documents then it requires (M x N) XSLT stylesheet to do this work.

To resolve this problem, we have used the power of regular expressions in the development of TC-GXML. Although it is not faster than XSLT Schema, but using this method we do not require creating a schema for conversion from HTML to XML every time. This conversion has two parts:

i) HTXT (HTML to Pure Text)

This part of TC-GXML works directly with webpage and converts the whole webpage into text document. This part uses the programming approach with regular expression method to parse HTML document containing Tags like <title>, <h1> , <P> etc to converts it into the text document, which may be used in the applications like Text to Speech Application for Web. The Webpage to HTML shown in Fig 5 and Fig 6.



Home | Institute | Examination | Circulars | Contact Us |Search

Navigation

A Place For Knowledge And Innovation

An Academic Autonomus Institution

Director's Desk [Prun
Administration
Departments
DEANS
Academics
Centers & Facilities
Training & Placement
Recruitment
Tender
AUTHORITIES OF INSTITUTE

Harcourt Butler Technological Institute has always been a paragon and a source of inspiration in the field of science and technology, since 1921. It has maintained its conspicuous presence in the technical world. H.B.T.I. has a history of realization, fulfillment of necessities and advancement in the field of science and technology.
Read More ..

Quick Links
NEW RESULTS
Student Alerts
Councils/Clubs
Download Software
Internet Support
Project Vacancies

Vision Mission Goals & Objectives

OTHER LINKS
APPEALS TO HBTI ALUMNI
Uttar Pradesh Technical University, Lucknow

Important Links

NEW SEE-UPTU-2009 On Campus Counselling Postponed
- Training & Placement Cell
- Hostel Facilities, Allotments & Related Fees (2009-10)
- Mandatory Disclosure 2008-09
- M.Tech./Ph.D. Advertisement for Academic Session 2009-10
- Registration Dates, Fee Structure and Scholarship Performa for academic session 2009-10
- Instructions to see the evaluated answer books alongwith the procedure of scrutiny.
- List of Ist Year Students along with details of Faculty Counselors (Academic Session 2008-09)

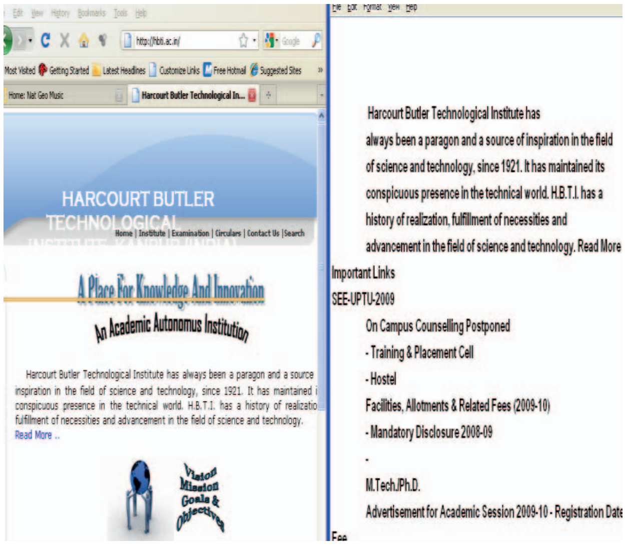
Forthcoming Events

Summer Training to out side B. Tech. students at H.B.T.I. Kanpur

MORE

Student Exam

Figure 5. Institute Home Page


HARCOURT BUTLER
TECHNOLOGICAL
Home | Institute | Examination | Circulars | Contact Us | Search
A Place For Knowledge And Innovation
An Academic Autonomus Institution
Harcourt Butler Technological Institute has always been a paragon and a source inspiration in the field of science and technology, since 1921. It has maintained i conspicuous presence in the technical world. H.B.T.I. has a history of realizatio fulfilment of necessities and advancement in the field of science and technology.
Read More ..
Vision Mission Goals & Objectives
File Edit Format View Help

Harcourt Butler Technological Institute has
always been a paragon and a source of inspiration in the field
of science and technology, since 1921. It has maintained its
conspicuous presence in the technical world. H.B.T.I. has a
history of realization, fulfillment of necessities and
advancement in the field of science and technology. Read More
Important Links
SEE-UPTU-2009
On Campus Counselling Postponed
- Training & Placement Cell
- Hostel
Facilities, Allotments & Related Fees (2009-10)
- Mandatory Disclosure 2008-09
-
M.Tech./Ph.D.
Advertisement for Academic Session 2009-10 - Registration Date
Fee

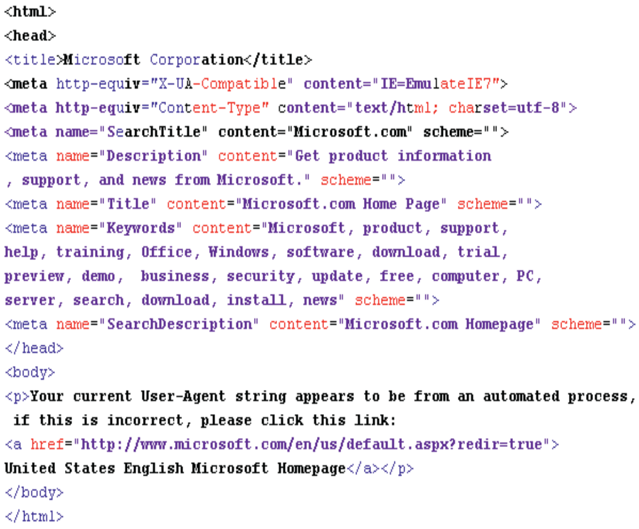
Figure 6. HTML to Pure Text

ii) XMLL (XML Document for Links)

XMLL Tags have been developed by us for XML Language. This part of TC-GXML creates a XMLL document for Links obtained from the HTML document. Here again we use regular expressions to do so.

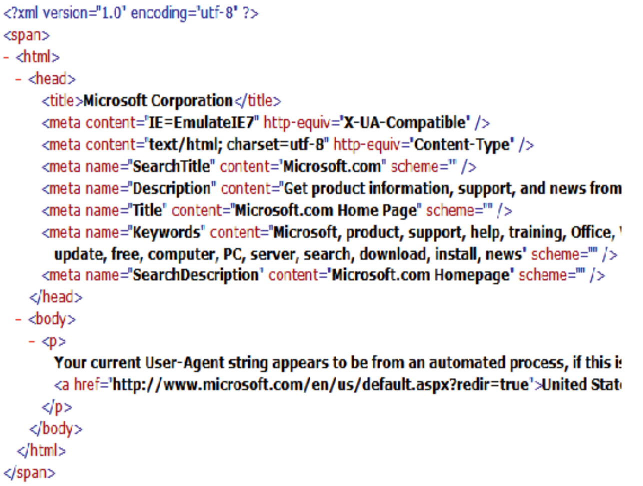
Both the above methods work simultaneously and give two documents HTXT and XMLL for Text and Links of the webpage respectively.

The HTML to XMLL conversion shown in Fig 7 and Fig 8.



```
<html>
<head>
<title>Microsoft Corporation</title>
<meta http-equiv="X-UA-Compatible" content="IE=EmulateIE7">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<meta name="SearchTitle" content="Microsoft.com" scheme="">
<meta name="Description" content="Get product information
, support, and news from Microsoft." scheme="">
<meta name="Title" content="Microsoft.com Home Page" scheme="">
<meta name="Keywords" content="Microsoft, product, support,
help, training, Office, Windows, software, download, trial,
preview, demo,  business, security, update, free, computer, PC,
server, search, download, install, news" scheme="">
<meta name="SearchDescription" content="Microsoft.com Homepage" scheme="">
</head>
<body>
<p>Your current User-Agent string appears to be from an automated process,
 if this is incorrect, please click this link:
<a href="http://www.microsoft.com/en/us/default.aspx?redir=true">
United States English Microsoft Homepage</a></p>
</body>
</html>
```

Figure 7. HTML Document



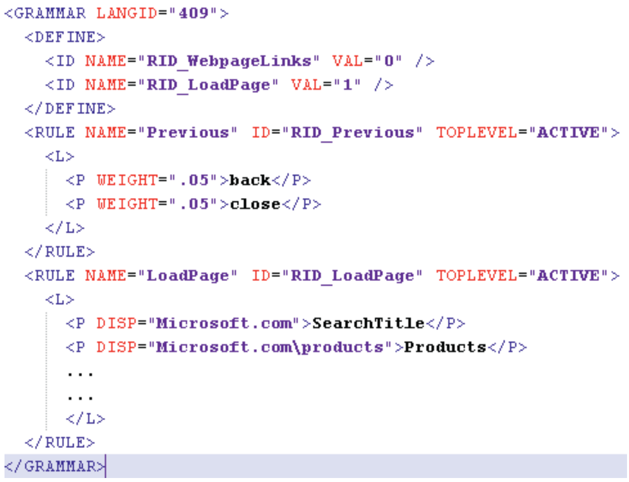
```
<?xml version="1.0' encoding='utf-8' ?>
<span>
- <html>
  - <head>
      <title>Microsoft Corporation</title>
      <meta content="IE=EmulateIE7" http-equiv='X-UA-Compatible' />
      <meta content="text/html; charset=utf-8" http-equiv='Content-Type' />
      <meta name="SearchTitle" content='Microsoft.com" scheme='' />
      <meta name="Description" content="Get product information, support, and news from
      <meta name="Title' content="Microsoft.com Home Page" scheme="" />
      <meta name="Keywords' content="Microsoft, product, support, help, training, Office, '
        update, free, computer, PC, server, search, download, install, news' scheme="" />
      <meta name="SearchDescription' content='Microsoft.com Homepage' scheme="" />
    </head>
  - <body>
    - <p>
        Your current User-Agent string appears to be from an automated process, if this is
        <a href='http://www.microsoft.com/en/us/default.aspx?redir=true'>United Stat
      </p>
    </body>
  </html>
</span>
```

Figure 8. XMLL Document

C. *XMLL to GXML (XML Grammer)*

GXML is another XML language that is developed by Microsoft Corporation Inc and recommended for two way voice application. This is the final phase of TC-GXML where we covert XMLL (which we get from HTML) into GXML. GXML (Grammar eXtensible Markup Language) is specific format for developing two way Voice Application. GXML does not completely follow the W3C grammar format. It is based on the Microsoft SAPI 5.0 grammar format and provides flexibility in developing voice applications using SAPI or equivalent speech engines. The task becomes easier after the conversion of HTML to XMLL. TC-GXML searches and analyzes the XMLL tags and converts the XMLL document to the specified GXML document. The GXML may be converted to any type of XML or HTML document as specified in the schema or stylesheet.

The XMLL to GXML conversion shown in Fig 8 correspond to Fig 9.



```
<GRAMMAR LANGID="409">
  <DEFINE>
    <ID NAME="RID_WebpageLinks" VAL="0" />
    <ID NAME="RID_LoadPage" VAL="1" />
  </DEFINE>
  <RULE NAME="Previous" ID="RID_Previous" TOPLEVEL="ACTIVE">
    <L>
      <P WEIGHT=".05">back</P>
      <P WEIGHT=".05">close</P>
    </L>
  </RULE>
  <RULE NAME="LoadPage" ID="RID_LoadPage" TOPLEVEL="ACTIVE">
    <L>
      <P DISP="Microsoft.com">SearchTitle</P>
      <P DISP="Microsoft.com\products">Products</P>
      ...
      ...
      </L>
  </RULE>
</GRAMMAR>
```

Figure 9. GXML Document

V. RESULTS AND DISCUSSIONS

We analyzed about 40 websites with our TC-GXML and found that it gives about 88% accuracy to transcode the webpage into GXML and about 98% accuracy in transcoding the webpage into HTXT. Thus, TC-GXML gives 93% accuracy overall efficiency to transcode a webpage.

VI. CONCLUSION

TC-GXML is a perfect tool for the developers and programmers who develop two way web application or windows application. Existing Transcoding tools and techniques are not able to cope with the variations in webpages due to their dependency on overall structure, content structure, and links/Tabs View etc. It has been observed that existing tools and techniques many times fail to do the required work. TC-GXML has been designed to do these works together because of its independence to the aforesaid factors. It works uniformly over structure of webpage, content and links. This tool will be useful in the handling of transcoding of a broad range of websites for the proposed speech based web browser.

ACKNOWLEDGMENT

This research is a part of the Major Research Project entitled "Design and development of Web Browser for Visually Impaired" funded by University Grants Commission, New Delhi running in Computer Science & Engineering Department at Harcourt Butler Technological Institute, Kanpur.

REFERENCES

[1] Usage of Computers and Internet by the Visually Challenged: Issues and Challenges in the Indian context, Findings of a study conducted by Enabling Dimensions, January, 2002.

[2] Hopson, Nichelle. WebSphere Transcoding Publisher:HTML-to-VoiceXML Transcoder. January 2002.Accessed on October 4, 2002

[3] from:http://www7b.boulder.ibm.com/wsdd/library/techarticles/0201_hopson/0201_hopson.html.

[4] IBM Corporation. WebSphere Transcoding Publisher Version 4.0 Developer's Guide. Available at: http://www-1.ibm.com/support/docview.wss?uid = swg27000533&aid=1

[5] Raman, T.V. Emacspeak- Toward The Speech-enabled Semantic WWW. Available at http://www.cs.cornell.edu/Info/People/raman/publications/semantic-www.html. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[6] SALT. Available at http://www.saltforum.org/..

[7] Tellme. Available at http://www.tellme.com.

[8] XHTML+voice. Available at http://www.w3.org/TR/xhtml+voice/.

[9] Ball, T., Colby, C., Danielsen, P., Jagadeesan, L.J., Jagadeesan, R., Laufer, K., Mataga, P. and Rehor, H. Sisl: Several Interfaces, Single Logic, International Journal of Speech Technology 3, 93-108, 2000.

[10] M. Yankelovich, N. How do Users Know What To Say? ACM Interactions, Volume 3, Number 6, November/December 1996.

[11] Bevocal. Available at http://www.bevocal.com

[12] Huang, A.W. and Sundaresan, N. Aurora: A Conceptual Model for Web-Content Adaptation to Support the Universal Usability of Web-based Services, CUU '00 Arlington VA.

[13] Plomp, C.J. and Mayora-Ibarra, O. A. Generic Widget Vocabulary for the Generation of Graphical and Speech-Driven User Interfaces, International Journal of Speech Technology 5, 39-47, 2002.

[14] Zhiyan Shao, robert Capra, Manuel A. Perez-Quinines, Annotations for HTML to VoiceXML Transcoding:Producing Voice WebPages with Usability in Mind.

[15] Narayan Annamlai, B.Prabhakaran and Gopal Gupta, Extensible Transcoder for HTML to VoiceXML Conversion.

[16] Mduduzi E. Nxumalo and Daniel Mashao, Adapting Web Content for Telephone Users by transcoding XSLT.