

WAV: Voice Access to Web Information for Masses

Himanshu Chauhan Pankaj Dhoolia
Ullas Nambiar Ashish Verma
IBM Research - India, New Delhi

{himchauh, pdhoolia, ubnambiar, vashish}@in.ibm.com

ABSTRACT

One of the main reasons for a large section of the world population to be left out of the internet revolution is, limited or no access to a computer due to economic, educational, cultural and age factors. Enabling masses to extract information from the web via voice will bring the Internet revolution to additional billions of people. In this paper, we describe a system called WAV (Web Access via Voice), that is a step in this direction. Departing from the traditional approaches of manually building a VoiceXML based site, the WAV system uses information from existing web sites to serve the user. Challenges to overcome include extracting contextually relevant information from the user and also from the pages returned by websites, reducing amount of information relayed to user over phone and maintaining the context of the conversation for easy refinement based on feedback from the user. Our prototype system not only shows successful integration of many different technologies such as automatic speech recognition, scripts for web navigation, text to speech conversion, but also introduces a novel way of extracting information from web via voice in a programmatic manner. We describe initial solutions developed to tackle above challenges and demonstrate the feasibility of the system by describing prototype implementations on two popular web sites in India.

1. INTRODUCTION

The Internet and the world-wide-web, arguably the most important inventions of the last century, have changed the lives of billions of people in the world. Unfortunately, there are still billions of people who are unable to access the information on the world-wide-web. Any invention that allows the billions that are left out in the current Internet revolution to benefit from the Internet will have a huge positive impact.

One of the main reasons for a large section of the world population to be left out of the internet revolution is, limited or no access to a computer/internet due to economic, educational, cultural and age factors. Although the penetration of Internet is low in the poor segment of the society, the use of telephones; particularly basic mobiles phones, is growing at an astounding rate. By November 2007, the total number of mobile phone subscriptions in the world had reached 3.3 billion, or half of the human population. Most of these subscriptions are in the developing regions of the

Copyright is held by the author/owner(s).
W3C Workshop 2010, May 6-7, 2010, New Delhi.

world and for such people, the only easy-to-use medium for information retrieval is the basic mobile phone¹.

Web Access by Voice (WAV) is a system that focuses on decoupling the *Information Content* available on the World Wide Web from the *Web Browsing* methodology used to access it. It combines the *ubiquity* of the phone (both analog and mobile) with the information content available on the Web. WAV uses a novel methodology where the system performs web browsing instead of the user so that the user doesn't need to be familiar with the browsing and its nuances. The system works by taking the user's query, identifying the websites corresponding to user's query, gathering the required inputs from the user to extract the information, extracting the information related to user's query from websites, transforming it in a form consumable by the user over phone and supplying it to the user. Another important aspect of WAV is that the user doesn't need to know which websites have the information that the user wants. This really helps a large number of users who are not familiar with the Internet or the browsing concept and hence do not have the knowledge of various websites.

Challenges: The ultimate goal of providing access would be to build an intelligent natural language understanding system that parses a natural language query, understands the query, retrieves the information from the Internet and then extracts the answer for the query from the extracted information. With the current natural language parsers, such a system looks infeasible in the short term. However, speech recognition in a domain specific manner with a bounded number of words has been quite successful in spoken dialog systems [11]. WAV uses a carefully designed dialog manager to restrict the number of possible inputs by the user in a domain-constrained manner. There are still many other technical challenges:

1. *Extracting Information for the User* : How can the system navigate the website to determine the information required by the user? Since many websites provide information through dynamic content generation, this step may require filling of forms on the user's behalf to extract the information. For example, suppose that the user is interested in finding out the schedule of trains from point A to point B. The user may not even be aware of the form that needs to be filled to get the schedule.

¹In this paper, all references to mobile phones are to basic mobile phones that do not give advanced capabilities like browsing Web pages or reading and responding to emails etc.

2. *Reducing Information for Voice-based Interaction* In many examples, the information returned from the website may be too much to read out to the user. Whereas a user can quickly choose the required piece of information in a visual medium, the voice medium necessitates that the information either be summarized or be broken down into chunks so that the user is not given excessive information in a sequential manner.
3. *Refining the Query based on Previous Results* Just as in visual web browsers, the users typically refine or change their query based on the results obtained in the previous query. If no state is retained from the previous query, the user would be forced to provide all the details of the query again. By retaining the context from the previous query and providing the missing arguments automatically the user is spared of repeating the information after each query.

In this paper, we introduce a website independent method of extracting information from the website. There are two aspects of this method. First, we use a language which decouples the placement of form elements from the form element itself. The second aspect of our method is to keep a domain-specific translation table to allow the same method to work for multiple websites. For example, one travel website may use "Origination Point," and the other website may use "From" as the listbox for the user to choose the destination. By keeping the web page profile, the user is insulated from such differences between the websites.

2. SYSTEM ARCHITECTURE

As Figure 1 illustrates, WAV system has three main architectural components - The WAV configuration studio, the WAV database, and the WAV engine. Rest of this section describes these architectural components, the usage scenarios they participate in, and the details of the comprising subsystems and components.

2.1 WAV Configuration Studio

While our primary focus in the WAV system is to address the needs of the masses, it is also important to make sure that the task of extending and configuring the reach of the WAV system doesn't become too daunting for a WAV service-provider that it starts to impact its profitability. WAV Configuration studio is a web-based application that is intended to address this. Powered by form-detector, result-detector, click-thru interaction recorder, and schema matchers it assists a WAV system configurator by providing substantial automation in configuring the WAV system. A set of generators further assist in configuring new domains, by transforming detected set of forms, results, and interaction patterns to domain ontology, and domain conversation interface, based on configurator's training interactions with representative content-provider sites. The Studio populates the WAV database with the configurations, which are then picked by the WAV Engine to support end-users with corresponding conversations.

2.2 WAV Database

WAV database is the configuration repository where the artifacts of the WAV model described earlier, are stored. These artifacts are produced by a WAV service provider using the Configuration studio, and used by the WAV engine

in its interactions with both the end-users and the specific WWW sites. Apart from the domain, and WWW-site configurations, WAV database also contains the actual user-interaction data. User Profiler observes the patterns of interaction in the recorded data and derives personalized information and interaction reduction rules from it.

2.3 WAV Engine

WAV engine is the executional backbone of the WAV system. On one end it enables optimal voice conversations for the masses addressing the form-factor, while on the other, it integrates with the configured WWW-sites seamlessly to support those voice conversations.

2.3.1 Voice Interaction Engine

The Voice Interaction Engine component carries out the conversations with end-users. Domain focused voice configurations are leveraged by the engine to increase the performance of speech recognition. Further the learnt user and usage profiles are leveraged to enhance the user-experience. As a voice driven user interface this engine has to provide three key functionalities : *Recognition of User Speech, Relay of textual information in audio format, Management of conversation.*

Automated Speech Recognition (ASR) is responsible for recognition of user speech. For WAV we use finite state speech grammars to limit what users can speak [5], thus improving recognition accuracy. Audio relay of textual information is achieved by use of Text-to-speech (TTS) component. VoiceXML [12] format provides both ASR and TTS functionalities with the help of voice-browsers. WAV uses VoiceXML (VXML) dialogs to interact with the users.

Dialog Manager manages conversations between the system and users to provide better user experience. It stores caller specific details as User Profile and uses it for personalized menu choices and quick retrieval of user data without prompting for them.

2.3.2 Information Retrieval Kernel

The Information Retrieval Kernel governs the conversations that the WAV system has with the configured WWW-sites. Receiving the user-query and inputs from the voice interaction engine, the information retrieval kernel transforms the query into corresponding site-specific queries, using the provider-profile configurations, and federates the query to the relevant sites. It then receives the site-specific responses, transforms them to the site independent domain conversation format and aggregates them, removing duplications. The Kernel then leverages the information reduction rules to optimize and sort the resulting response, and passes it to the voice interaction engine to communicate to the end-user.

3. WAV PROTOTYPE IMPLEMENTATION

We have implemented a prototype of WAV system based on the architecture presented above. The current implementation provides useful insights for building a scaled up system. At present the system does not allow free speech conversations and expects the user to provide precise answers to prompts. Most of the extracted information text is relayed as is on the web-page. The prototype incorporates information retrieval from web-sites of three different domains. For each domain, the conversation interface was created by observing label data (visible on the rendered page) for HTML

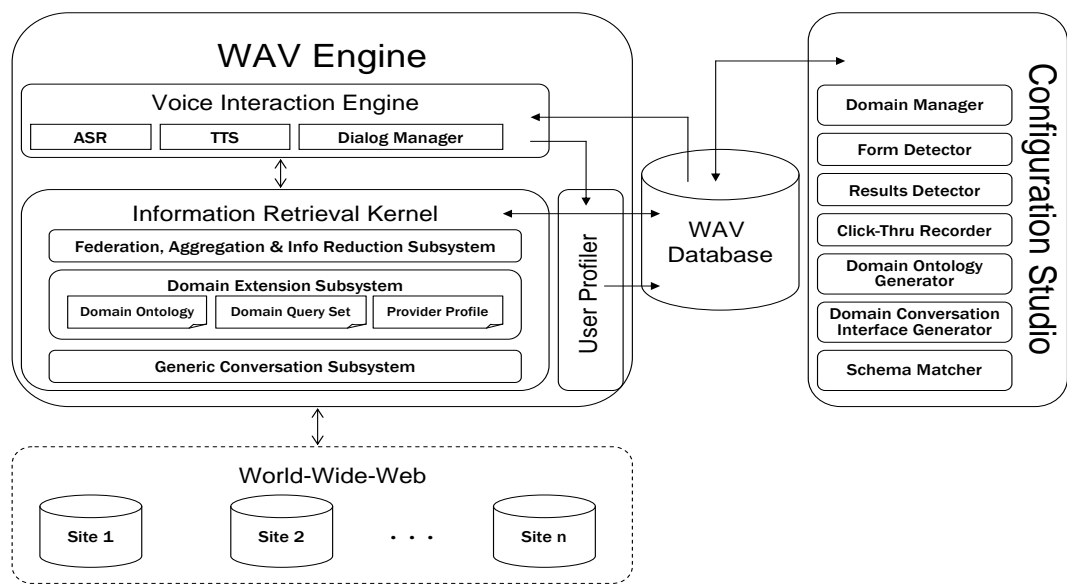


Figure 1: WAV System Architecture

inputs as captured in the CoScripter script. Static rules are used for reduction of information being relayed. At present these rules are manually configured based on domain knowledge, for example *in travel domain flights are read out in increasing order of fares*. Using techniques like context based information extraction [13] and domain ontology an extraction schema for the domains retrieves relevant and concise information to be relayed back to the user.

Prototype for cheapest flight between a source-destination pair: Configuration process for extracting flight information is started by creating a representative CoScripter script by manually performing a flight search. Figure 2 shows actual web-site form being filled and corresponding script, generated using *Configuration Studio*.

On generation of the script, the domain ontology is augmented with label data from script such as ‘From’, ‘To’, ‘One Way’, ‘Depart’, ‘Class’ etc. Creation of *Provider Profile* follows the script creation step by using script to identify provider and inherit the URL. A common domain input schema is created using script statements and user provided input as parameters. These inputs are then mapped to common domain terms using inference models of domain ontology.

After completion of configuration step the domain is incorporated in the WAV system. Upon receiving a generic query for flights, it is translated to provider specific query format for each provider and then query Federation System federates specific queries for all the provider profiles.

As seen in the figure, each row contains one flight detail. Extraction schema for the domain is configured to extract tuples of airline, departure time, fare and duration from each of these rows. These details are then collected and sorted based on the fare of each flight, putting cheapest flight on top, and relayed back to the user in audio format.

Prototype for train seat availability: Indian Railways provides information about all the trains schedules and availability of seats for reservation in trains through their website www.indianrail.gov.in. This prototype imple-

ments a train-domain conversation “Seat Availability Query”, in a similar manner to that of flight details implementation. In this two step conversation, the conversation is initiated by providing source and destination stations along with the date of the travel to retrieve the list of trains available between the given source-destination pair on the particular date. In the second step of the conversation based on the train selection by the user, seat availability details are retrieved and communicated back.

4. RELATED WORK

Speech interface to mobile web browsers has been provided provided to perform simple web searches [9]. However this mode of internet access still requires familiarity with the web browsing concept and also it is difficult to use due to the tiny display available with the mobile phones. There also have been some efforts to perform web search through spoken query [1]. Speech interface have also been developed for standard web browsers [10]. W3C Voice Browser Activity Group [5] has been working in this area and has come out with various standards, such as, VXML, SRGS, CCXML etc., mainly towards speech recognition, text-to-speech synthesis and natural language understanding aspects. However, there is not much work regarding how to perform the information extraction from dynamic web pages which are getting richer and more complex with usage of technologies, such as, Javascript and Ajax. There have also been some focussed work to improve the browsing efficiency of visually impaired users starting from screenreaders [2] to more sophisticated approaches using content filtering and semi-automation [4, 13]. In [4] an accessible interface is developed for CoScripter, a programming by demonstration solution, which helps the blind user perform a pre-define internet task more efficiently. In [13] the browsing time for a blind user is significantly reduced as compared to using a normal screenreader by using smart content filtering mechanisms and content-aware web browsing concepts.

In a related effort at our lab, called WorldWide Telecom

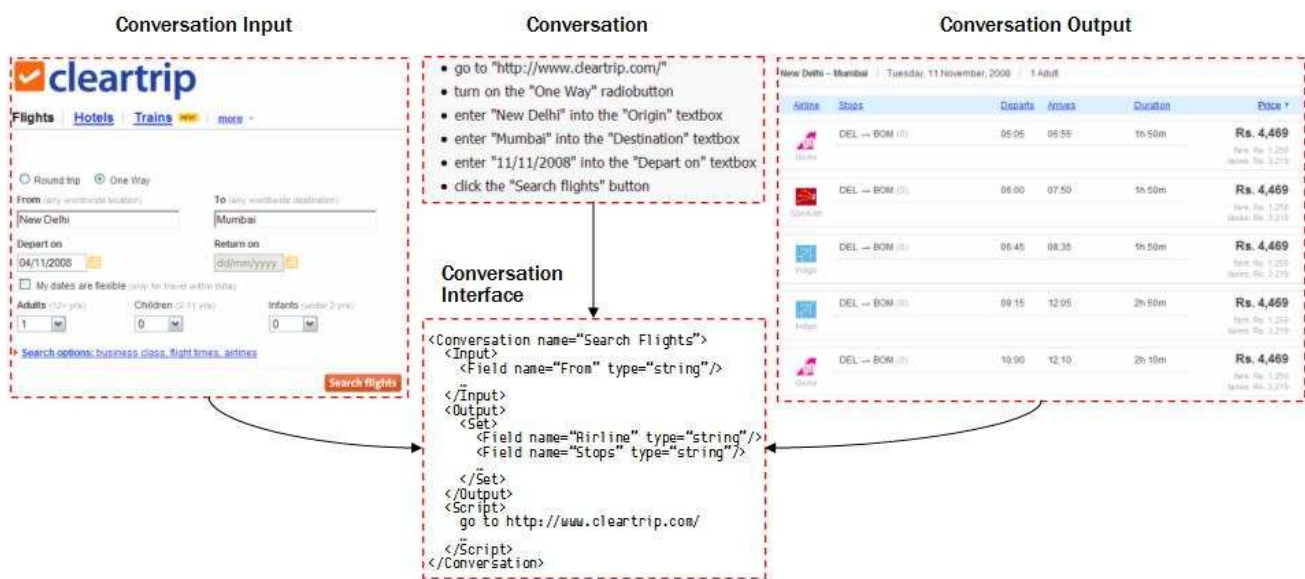


Figure 2: Representative Script for 'Flights' domain

Web [7], a parallel worldwide web in the telecom domain is worked upon where the information would be stored completely in audio format, called voice sites, and would be browsed through a telecom browser [8]. The main difference between SpokenWeb and WAV is that while in SpokenWeb the information resides in newly conceptualized voice sites, in WAV the information is extracted from the traditional worldwide websites.

5. CONCLUSION

In this paper, we present a system to access the information content on the Worldwide Web through voice for consumption by people who either don't have access to computer or/and are not familiar to web browsing for various reasons. We described how the system extracts information related to a user's spoken query in a domain-dependent but website independent manner. This is performed by developing a domain ontology and a mapping between website independent terms and website dependent terms. We have described how designing such a system in a domain dependent manner helps in addressing the issues in information extraction from different websites. We have also shown how pseudo-natural language script produced by CoScripter, can be used for such a system. We described initial solutions developed and demonstrated the feasibility of the system by describing prototype implementations.

Currently, we are working on developing hierarchical domain structuring and ontology inheritance to develop a scalable framework of WAV covering a large number of domains and hence a larger number of websites for information access.

6. REFERENCES

- [1] Huixiang Gu, Jianming Li, Ben Walter and Eric Chang, "Spoken Query for Web Search and Navigation", *Proc. of International WWW Conference*, HongKong 2001
- [2] <http://www.freedomsscientific.com>
- [3] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper Induction for Information Extraction. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 729-737. San Francisco, CA: Morgan Kaufmann, 1997.
- [4] J. P. Bigham, T. A. Lau and J. W. Nichols, "TrailBlazer: Enabling Blind Users to Blaze Trails Through the Web", submitted to International Conference on Intelligent User Interfaces, Florida, 2009.
- [5] <http://www.w3.org/voice>
- [6] G. Leshed, E. Haber, T. Matthews, T. Lau, "CoScripter: Automating and Sharing How-To Knowledge in the Enterprise", *CHI 2008*, Florence, Italy, April 2008.
- [7] Arun Kumar, Nitendra Rajput, Dipanjan Chakraborty, Sheetal K. Agarwal and Amit Anil Nanavati, "WWTW: The WorldWide Telecom Web", *NSDR 2007 (SIGCOMM workshop)*, Kyoto, Japan 2007
- [8] Sheetal Agarwal, Arun Kumar, Amit Anil Nanavati, Nitendra Rajput, "The WorldWide Telecom Web Browser", *Proc. of International WWW Conference*, Beijing, China, 2008
- [9] <http://mobile.yahoo.com/onesearch>
- [10] Dong Lin, Lin Bigin, Yuan Bao-Zong, "Using Chinese Spoken-Language Access to the WWW", *Proc. of International Conference on WCCC-ICSP*, Volume 2, pages:1321-1324, 2000
- [11] Y. Gao, H. Erdogan, Y. Li, V. Goel and M. Picheny, "Recent Advances in Speech Recognition System for IBM DARPA Communicator", *Proc. of EUROSPEECH*, Denmark, 2001
- [12] <http://www.w3.org/TR/voicexml20/>
- [13] Jalal Mahmud, Yevgen Borodin and I.V. Ramakrishnan, "CSurf: A Context-Driven Non-Visual Web-Browser", *International World Wide Web Conference WWW*, (<http://www.www2007.org>)