SALSA Version 1.0:
A Speech-based Web Browser for Hong Kong English

Pascale FUNG, CHEUNG Chi Shun, LAM Kwok Leung, LIU Wai Kat, and LO Yuen Yee
{pascale,eepercy,cpegeric,eekat,eeyy}@ee.ust.hk

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology (HKUST)
Clear Water Bay, Kowloon
Hong Kong

ABSTRACT

In this paper, we present a prototype speech-based Web browser, SALSA1.0, and describe some of the research issues we need to address while building this system for Hong Kong users. SALSA1.0 allows the user to speak English command words as well as partial or complete link names on any page. The research issues involved in building SALSA1.0 are mainly (1) how to handle large accent variations and mixed-language and (2) how to handle unknown words, especially proper names, in Web links. The recognition engine for SALSA1.0 is trained on WSJ data, and then retrained on a small amount of Hong Kong accent WSJ data to handle accent variations. An edit-distance algorithm is used to replace all unknown words by the closest known word in the word network for recognition. With these methods, link name recognition rate is at 91.20% for links without unknown words, and 82.40% for links with unknown words. SALSA is currently being developed into a multilingual, natural language-based Intranet service provider for HKUST campus information access.

1. INTRODUCTION

The World Wide Web is generally regarded as a giant repository of information. Information transmission on the Web can be categorized into (1) user search for specific information via search engines and category services; (2) user look for related information by browsing and following the Web links; (3) provider actively disseminate information to specific target users. Various search engines are developed to serve the first purpose. Browsers with keyboard and mouse input help users achieve the second goal. Various push technologies are implemented to realize the third goal. A lot of research is being done in the field of information retrieval to help improve search engines, and to facilitate user access of Web information. Meanwhile, there is a new trend in recent years to augment user-friendliness of the browsers through natural language-based interfaces. In addition to text-based natural language understanding, speech recognition interface to the Web browser is also being trumpeted as an important step in improving user-friendliness, especially for naïve users.

We want to point out that there is another, stronger, reason for building a speech-based Web browser---it is a powerful tool for Intranet service and information access. Many universities, government agencies, libraries and companies are already using the Web to provide information to restricted internal audiences. We believe that this rather under-exploited advantage of the Web platform can be greatly enhanced by the development of a spoken language understanding interface. Intranet usage tends to be goal-specific, and requires fast turn-around and large throughput of users. The audience is very often a busy user who, while accessing Intranet information, is also working on document editing or updating another database. In this situation, a hands-free mode of information access is even more important than mere Web surfing.

Since 1996, the HKUST campus has been populated with Express Stations---PC clients connected to a central Web server. These Express Stations provide Internet services as well as HKUST Intranet access. These stations are installed in public areas, and are being used by mostly students and visitors who need quick access to some specific information. In building SALSA, our goal is to add a spoken language understanding interface to this Intranet and Internet service.

The requirements for such an interface are:

1. **speaker-independence,** since the turn-around time for each user is typically 15 minutes;
2. **accent-tolerance,** since the students and visitors might have different cultural background;
3. **dynamic vocabulary-based,** since the information on the Web can be in any domain;
4. **recognition of unknown words,** such as proper names since Web pages contain an unpredictable amount of such words;
5. **multilinguality,** since the majority of Express Station users are either English speakers, Cantonese speakers, or Mandarin speakers.
6. Understanding of **spontaneous speech.**
7. **Hands-free** access.

8. Easily extendable.

To facilitate extendibility, all versions of SALSA are implemented using a client-server architecture. In addition, SALSA1.0 fulfills requirements 1 to 4. SALSA2.0 is multilingual and does not require push-to-talk mouse clicks. SALSA3.0, under development, will provide spontaneous speech understanding beyond link name recognition. In this paper, we focus our discussion on SALSA1.0 and the issues raised in points 1 to 4.

2. SALSA1.0 ARCHITECTURE

The Express Stations, like other Intranet terminals, are thin clients with access to a more powerful central server. To leverage off of such an Intranet architecture, we use a client-server architecture for SALSA. SALSA is divided into three modules: client, server, and recognition engine. SALSA1.0 clients are PCs with SoundBlaster sound card, a microphone, running on Win95. The browser supported by Express Station server and hence supported by SALSA is Netscape4.0 or above. SALSA1.0 server and recognition engine are implemented on a Unix machine.

Figure 1 shows the overall system architecture for SALSA1.0.

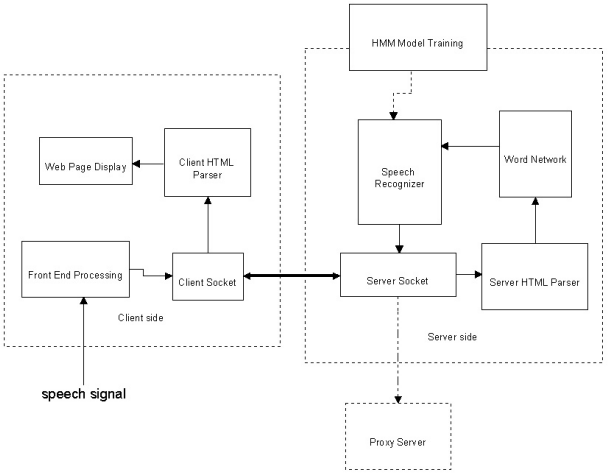

Figure 1: SALSA1.0 system architecture**.**

2.1 Client

SALSA1.0 client is a browser application. It is built on top of the browser. It simulates most of the functions of a Netscape browser, in addition to commands. SALSA client is consisted of the following sub-modules:

1. User Interface
2. Client Cache
3. HTML Parser (Client Side)
4. Audio Recording
5. Client/Server Communication

2.2 Server

SALSA1.0 server is implemented using C. Its sub-modules are:

1. Word Lattice Generator
2. Server Cache
3. Result Matching
4. Server/Client Communication

2.3 Recognizer

The recognition engine for SALSA1.0 is built from HTK modules. We use 39 features: 12 mfcc, 12 delta mfcc, 12 delta delta mfcc, energy, delta energy, delta delta energy. The feature vectors of the speech signal together with its transcription are used to train a HMM-based speech recognizer. There are 46 HMMs, including models for 44 English phones, one short pause and one silence. For each HMM, there are three states and only a single Gaussian mixture is used for each state.

Figure 2 illustrates the important modules related to the recognizer.

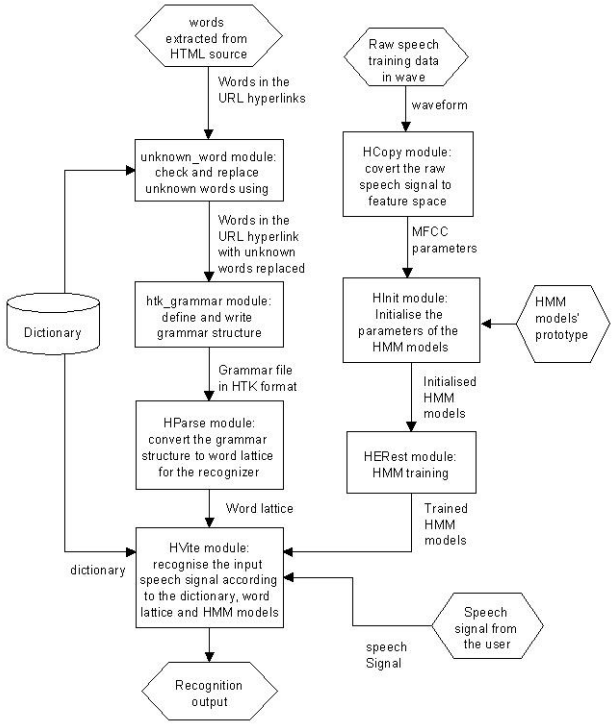

Figure 2: Architecture of the SALSA1.0 recognition module.

We use 9800 utterances from 96 speakers in the Wall Street Journal Continuous Speech Recognition Corpus Cambridge

version (WSJCAMO) to train the initial SALSA1.0 recognizer. 13125 English words which frequently appear in Web pages are included in the dictionary of SALSA1.0. The grammar allows for partial as well as full link names.

Initial tests show a phoneme accuracy of 63.08% and link name recognition rate of 60%. These low recognition rates are due to accent variations among test speakers. In addition, the recognition rates are even lower when link names contain words not found in the SALSA1.0 dictionary. In the following sections, we will focus our discussions on two main issues we have to address while building SALSA1.0 for English speakers in Hong Kong.

3. ACCENT VARIATIONS

In Hong Kong, most PC users can read English. We carried out a pilot test with HKUST students in which the subjects are instructed to "talk to the Express Station" in the most natural way via a simulated speech recognition interface. The result shows that when using this simulated speech recognition interface to the Express Station, users speak almost entirely in English. Since 98% of Web material is in English and since Hong Kong users prefer using English as browser input, we choose to implement an English speech recognition engine for SALSA1.0. However, more than 98% of Hong Kong people are native Cantonese speakers. The majority of English speakers in Hong Kong have some accent in English. We propose to initialize the baseline models by training with WSJCAMO, and then re-train the models with data from a WSJHK corpus, locally collected by our group.

Like all training data, various accented speech data tend to bias the recognizer towards the particular accent group configuration present in the training set. SALSA1.0 at first performed better on Hong Kong students than on visitors with other accents. To obtain an overall increase in accuracy in all accent groups, we normalize all accent effects to an acoustically neutral set. This neutral set does not correspond to any existing natural accent but is an acoustic average of all accents.

Using the recognizer trained on WSJCAMO, with Hong Kong English speech, the phoneme recognition accuracy is 63.08% and the link name recognition rate is 60%. To improve the performance of our recognizer, we initialize our HMMs with WSJCAMO training data, then retrain it using a second, smaller set of WSJHK. After retraining, the phoneme accuracy is increased to 90.95% and the link name recognition rate is increased to 91.2%.

4. UNKNOWN WORD HANDLING

A speech-based Web browser has to recognize link names from a vocabulary of unlimited size. However, at each instance, the user is constrained to utter only a small set of command words and follow links on a Web page. So the vocabulary size, while being unlimited and large over all, is constrained at any instance to a small, dynamic set of key words. In our system, when an HTML page is loaded, the server extracts all the link URL and link names, appended by a garbage model, to form a word network.

Web links for campus information contain many abbreviated, Hong Kong-specific words and lot of proper nouns, especially English transliteration of Chinese names, which are not found in the WSJ dictionary. Tests show that even with accent retraining, the phoneme accuracy drops to 73.55% and the link name recognition rate drops to 75.2% when the links contain unknown words.

The small vocabulary size of a single Web page suggests an easy solution: for each WORD1 not found in the pronunciation dictionary, the most morphologically similar dictionary word, WORD2 is found by an edit-distance matching method. Phonetic transcription of WORD2 is used for WORD1. This quick-and-dirty method works quite well for our system, enabling us to circumvent the construction of phonetic pronunciation rules. With unknown word handling, the phoneme accuracy is improved to 81.95% and the link name recognition to 82.4%, even in the presence of superfluous speech.

One side benefit of this method is Web pages in languages similar to English (e.g. French) can be accessed without additional implementation. Since the link names in French are orthographically and phonetically similar to English, and the number of links per page is limited, our recognizer is able to process user queries for French pages without the implementation of a French speech recognizer.

5.CONCLUSION

In this paper, we briefly show the architecture of a speech-based Web browser, SALSA1.0. We propose that speech recognition interface to the Web can be used to enhance Intranet services based on the Web platform by allowing hands-free access, in addition to making a browser more user-friendly. We also discuss the effect of accent variation between Hong Kong English speakers and American speakers. We propose a quick-and-dirty method for handling unknown words without constructing phonetic pronunciation rules. The baseline and improved recognition results of SALSA1.0 are shown in Table 1. Currently, we have finished developing SALSA2.0 and are developing SALSA3.0. Both SALSA2.0 and SALSA3.0 handle multilingual input in English, Cantonese and Mandarin. These later versions also use utterance detection and verification methodology. They do not require a push-to-talk mechanism, thereby allowing total hands-free usage. Details of SALSA2.0 and SALSA3.0 will be reported in our future papers.

SALSA configuration	Unknown word	Phoneme Accuracy	Link Name Accuracy
WSJCAMO	No	63.08%	60.00%
WSJCAMO+WSJHK	No	90.95%	91.20%
WSJCAMO+WSJHK	Yes	73.55%	75.20%
WSJCAMO+WSJHK+ Unknown word handling	Yes	81.95%	82.40%

Table 1: Evaluation results of SALSA 1.0 with different training sets, and with and without unknown word handling.

7.REFERENCES

1. The HTK BOOK
2. Huang, Raymond "Common Errors in English Pronunciation for Cantonese Students" , Dept of Extramural Studies, the Chinese University of Hong Kong
3. Fung, Pascale, Bertram Shi, Dekai Wu, Lam Wai Bun, Wong Shuen Kong "Dealing with Multilinguality in a Spoken Language Query Translator", ACL 97 Workshop on Spoken Language Translation, Madrid,1997, Jul, pp. 40--43.
4. Mehling, Herman "Intranet helps fine-tune product development", Computer reseller News 1997, Feb 17, p. S39
5. Avishai, Bernard "Get what you want from the web", Fortune 1997, Oct 27, p. 283-284; European 123-124
6. DiDio, Laura "OS/2 lets users talk back to 'net", Computerworld 1997, Dec 23, 1996-Jan 2, p. 12
7. Zelinka, Doug "Netscape Navigator 2.02 for OS/2 Warp", InfoWorld 1997, Jan 20, p.88
8. Semilof, Margie "NetPhonic gives Web servers a voice", CommunicationsWeek 1996, Mar 25, p.36-38
9. Bosher, Peter "See hear", Wired 1997, Jan p.167