CU VOCAL Web Service: A Text-to-speech Synthesis Web Service for Voice-enabled Web-mediated Applications

Helen M. MENG*, Tin Hang LO*, Chi Kin KEUNG*, Man Cheuk HO*, Wai Kit LO* and P.C. CHING**

*Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
**Digital Processing Laboratory,
Department of Electronic Engineering,
The Chinese University of Hong Kong,
Hong Kong SAR, China

{hmmeng, thlo, ckkeung, mcho, wklo@se.cuhk.edu.hk, pcching@ee.cuhk.edu.hk}

ABSTRACT

This paper presents the implementation of the CU VOCAL Web service, one of the first Chinese text-to-speech synthesis Web services. The CU VOCAL Web service can be easily integrated with other Web services to develop innovative Web-mediated applications. We have developed a novel automatic voice alert system in the stocks domain by integrating CU VOCAL and several other Web services. This system can monitor a real-time financial information feed for alert conditions pre-specified in the user's personalized profile, and trigger synthesized spoken messages to alert the user via the (mobile) telephone.

Keywords

text-to-speech, Web services, interoperability

1. INTRODUCTION

A Web service is a modular software application that provides services over the Web with the use of the eXtensible Markup Language (XML). One can develop innovative Web-mediated applications (applications that have access to the Web) by integrating different Web services across platforms and systems. Text-to-speech (TTS) synthesis aims to generate synthetic speech based on input text. TTS is an essential component technology in spoken dialog systems (SDS) that supports human-computer interaction in terms of a conversation. We have developed a number of SDS including CU FOREX [1] – a trilingual (Cantonese, Putonghua and English)[1] system that allows the user to ask for real-time foreign exchange rates over the telephone; and ISIS [2] – a mixed-initiative spoken dialog system in the stocks domain. Both systems incorporate CU VOCAL [3,4,5], a TTS engine that can generate highly natural Cantonese speech based on Chinese textual input. This paper reports on our recent work in implementing CU VOCAL as a Web service. To the best of our knowledge, this is one of the first TTS Web services. It eases development of voice-enabled Web-mediated applications since it does not require local installation of resource-demanding TTS engines. Hence the TTS Web service is particularly suited for offering voice capability to thin clients.

2. BACKGROUND ON WEB SERVICES

A Web service is a software system identified by a URI, whose public interfaces and bindings are defined and described using XML [6]. XML is the core technology in Web services for data typing and structuring of interaction messages. A Web service enhances interoperability among software applications. A Web service is also self-describing and can be published, located, and invoked across the Web. Hence Web services facilitate the development of distributed applications by adopting a loosely coupled Web programming model. Systems developed in terms of Web services are language and platform independent. Additionally, they are easily scalable and extensible by establishing connections to new Web services when necessary.

Figure 1 depicts the Web service architecture that involves a Web service provider, Web service client and discovery agencies. The Web service provider executes and hosts access to the services. This is achieved via a WSDL[2] service description [7] that publishes the service with its API[3] to discovery agencies. The exposed API facilitates interoperability by allowing Web service clients to invoke other Web services at runtime. Through these discovery agencies, a Web service client can find the Web services that it needs. Web services can interact with one another by means of SOAP[4] [8], which acts as the messaging protocol to transport XML data.

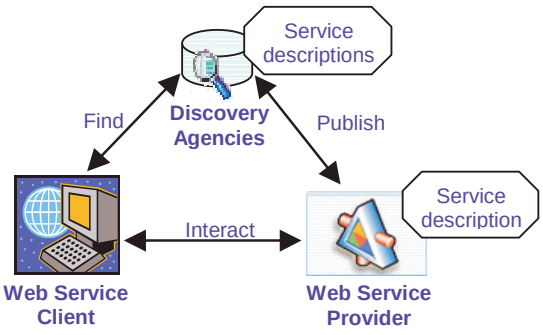

Figure 1. The Service-oriented architecture of Web services [6].

3. CU VOCAL: A CANTONESE TEXT-TO-SPEECH ENGINE

CU VOCAL is a Cantonese TTS engine that can generate highly natural and intelligible synthetic speech [3,4,5]. This is achieved by a syllable-based concatenative approach that considers both coarticulatory context and tonal context. This approach is

[1] These are the languages used in Hong Kong. Cantonese is a major dialect of Chinese predominant in Hong Kong, South China and many overseas Chinese communities. Putonghua is the official Chinese dialect.

[2] Web Services Description Language
[3] Application Programming Interface
[4] Simple Object Access Protocol

portable across Chinese dialects, e.g. from Cantonese to Putonghua. The approach can also be optimized for specific constrained domains (e.g. stocks, air travel, etc.) in order to improve the quality of the synthesized speech.

The Chinese language presents special challenges for TTS. Chinese text consists of a string of characters and each character is pronounced as a syllable. The character-to-syllable mapping is many-to-many. A Chinese word may contain one or more characters but there is no explicit word delimiter. Hence an input Chinese text string needs to be tokenized into a word sequence, and the lexical context helps determine the correct syllable pronunciation among the multiple possible candidates. It should also be noted that Chinese word tokenization/segmentation contains much ambiguity, and higher level linguistic knowledge is needed to disambiguate among the multiple possible segmentations. CU VOCAL has incorporated elaborate natural language technologies for processing the input Chinese text, in a procedure known as *text normalization*. Appropriate word tokenization helps automatic assignment of appropriate pause durations at specific word boundaries. This is important for prosodic control in the synthesized speech. In addition, CU VOCAL can appropriately verbalize special data objects (e.g. abbreviations, dates and times) and handle mixed-language textual input (between Chinese and English) that often appears in Hong Kong news text (e.g. in URLs, email addresses, English words and acronyms).

4. DEVELOPING THE CU VOCAL WEB SERVICE

The CU VOCAL Web service is developed by defining an API that encompasses the executable library of core engine. This is achieved by making use of Microsoft's Visual Studio .NET. The .NET framework handles all procedures common to Web services development, including the generation of WSDL for exposing the API. Synthesized speech is encoded as a base64 string in order to be transferred from the CU VOCAL TTS Web service to client applications as an XML response by means of the SOAP protocol.

The TTS Web service needs to transmit abundant audio data over the Web to client applications. However, the existing Web service architecture requires client applications to wait until all audio data is received before playing the audio. This often introduces significant time lags in the application. Since SOAP currently does not support data streaming, we have implemented a Web service for audio data streaming to integrate with the CU VOCAL Web service. Consequently, when an application invokes the CU VOCAL Web service, synthesized speech can be progressively generated and transferred to the client applications.

5. A NOVEL VOICE ALERT SYSTEM

In order to demonstrate the desirability of a TTS Web service, we have developed a novel voice alert system in the stocks domain. Voice alert systems should be applicable to domains with dynamic and time-critical information.

Figure 2 shows the architecture of the CU VOCAL voice alert system. It integrates the CU VOCAL Web service with four other Web services: 1) one that loads user-specified alert profiles; 2) one that retrieves real-time financial information; 3) one that triggers alert telephone calls and 4) one that streams encoded synthesized audio from the CU VOCAL Web service to the telephone alert Web service. For the sake of simplicity, Figure 2 shows only the CU VOCAL and telephone alert Web services.

To use the CU VOCAL voice alert system, users can input personalized alert profiles via a browser-based interface. The profile comprises of a financial index, its alert condition (i.e. above or below a pre-specified index level) and the alert telephone number(s). The voice alert system will then monitor the real-time financial index captured from a satellite downlink. When the alert condition is met, the system will generate a textual alert message that contains the index name and its real-time level, invoke the CU VOCAL Web service to generate a spoken form of this alert message and then pass the synthesized speech to the telephone alert Web service.

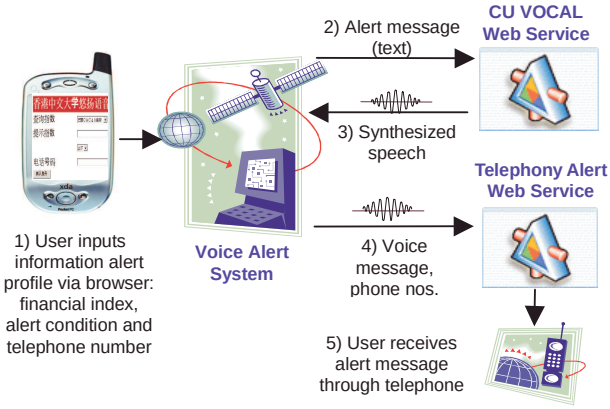

Figure 2. System architecture of the CU VOCAL voice alert system that integrates several Web services.

6. POSSIBLE FUTURE EXTENSIONS

The CU VOCAL voice alert system can serve as a reference implementation for integrating a TTS Web service in web-mediated applications. Voice alert systems can be used in other scenarios, e.g. dynamically changing flight arrival information, appointment reminders, reminder for book return to the libraries, etc. The CU VOCAL Web service can also be used in voice-enabled applications for the visually impaired, such as Web page reading, reading weather reports and news, etc.

7. ACKNOWLEDGMENTS

This work was partly funded by the Innovation and Technology Fund from the Hong Kong SAR Government (ITS/117/01) and a donation from Microsoft Research. We thank Kuansan Wang, Eric Chang, Michael Leung and Peter Ty from Microsoft for their input and contributions. We thank Reuters Hong Kong for providing the real-time data-feed via satellite.

8. REFERENCES

[1] Meng, H., Lee, S., and Wai, C. CU FOREX: A Bilingual Spoken Dialog System for the Foreign Exchange Domain. *Proc. of ICASSP 2000*.

[2] Meng, H. et. al. ISIS: A Multilingual Spoken Dialog System Developed with CORBA and KQML Agents. *Proc. of ICSLP 2000*.

[3] Fung, T. Y. and Meng, H. Concatenating Syllables for Response Generation in Domain-Specific Applications. In *Proc. of ICASSP 2000*.

[4] Meng, H. *et. al.* CU VOCAL: Corpus-based Syllable Concatenation for Chinese Speech Synthesis across Domains and Dialects. *Proc. of ICSLP 2002*.

[5] CU Vocal homepage. http:// www.se.cuhk.edu.hk/ cuvocal.

[6] W3C Web Services Architecture http:// www.w3.org/TR/ws-arch/.

[7] W3C Web Services Description Language (WSDL). http:// www.w3.org/ TR/ wsdl

[8] W3C SOAP Version 1.2 Part 1: Messaging Framework. http:// www.w3.org/ TR/ soap12-part1