Available online at www.sciencedirect.com
SCIENCE DIRECT®
Expert Systems with Applications 31 (2006) 684–696
Expert Systems with Applications
www.elsevier.com/locate/eswa

VoiceXML dialog system of the multimodal IP-Telephony—The application for voice ordering service ☆

Min-Jen Tsai

Institute of Information Management, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsin-Chu, 300, Taiwan, ROC

Abstract

The development of IP-Telephony in recent years has been substantial. The improvement in voice quality, the integration between voice and data, especially the interaction with multimedia has made the 3G communication more promising. The value added services of Telephony techniques alleviate the dependence on the phone and provide a universal platform for the multimodal telephony applications. For example, the web-based application with VoiceXML has been developed to simplify the human–machine interaction because it takes the advantage of the speech-enabled services and makes the telephone-web access a reality. However, it is not cost-efficient to build voice only stand-alone web application and is more reasonable that voice interfaces should be retrofitted to be compatible or collaborate with the existing HTML or XML-based web applications. Therefore, this paper considers that the functionality of the web service should enable multiple access modalities so that users can perceive and interact with the site in either visual or speech response simultaneously. Under this principle, our research develops a prototype system of multimodal VoIP with the integrated web-based Mandarin dialog system which adopts automatic speech recognition (ASR), text-to-speech (TTS), VoiceXML browser, and VoIP technologies to create user friendly graphic user interface (GUI) and voice user interface (VUI). The users can use traditional telephone, cellular phone, or even VoIP connection via personal computer to interact with the VoiceXML server. In the mean time, the users browse the web and access the same content with common HTML or XML-based browser. The proposed system shows excellent performance and can be easily incorporated into voice ordering service for a wider accessibility.

Keywords: IP-Telephony; Multimodal; VoiceXML

1. Introduction

1.1. Background

IP-telephony business was started on the account of cutting calling expense and now has become influential in the all-out IP based Internet Telephony industry. Various kinds of new applications have been developed rapidly and new services emerge continually in the whole industry. With the advantage of Internet, VoIP technology has developed many niche applications such as Internet/web call centers, e-commerce web messaging, chat rooms, e-learning, video streaming, etc. (Hassan, Nayandoro, & Atiquzzaman, 2000; Low, 1997). Nevertheless, among these applications, VoIP integration (Faulkner Information Services, 2001) with traditional PSTN in a visualized multimodal format is gradually used worldwide (e.g., Skype technologies S.A[1]).

While Internet and web browser are available everywhere as efficient information transferring tools, the issue of transmitting voice by packet is getting mature. Human beings start to apply the visualized multimodal interaction for the communication, a notable example of which is the Microsoft NetMeeting implementation. Through H.323 Protocol, it enables the users to send multimedia information to the PC counterpart by computer screen along with

☆ This work was supported by the National Science Council in Taiwan, Republic of China, under Grant NSC 91-2416-H009-012, NSC92-2416-H009-012 and NSC93-2416-H-009-009.
E-mail address: mjtsai@cc.nctu.edu.tw

[1] http://www.skype.com

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.01.010

the talking voice (Toga & Ott, 1999). In fact, PC platform and other IP device have been speedily taken as one of the main communicating tools (Georgescu, 2004; Rizzetto & Catania, 1999).

However, Internet and PC utilities are originally designed for data communication. Therefore, there are several fundamental problems on real time voice transmission. First, voice quality is not competitive to the one from traditional PSTN because of packet loss and latency (Perkins, Hodson, & Hardman, 1998). Secondly, long delay between PC speakers have left echo a problem, not thoroughly resolved yet (Ball, Bonnewell, Danielsen, Mataga, & Rehor, 2000). Though in recent years quite a few studies have suggested approaches for QoS issues to improve the real time voice quality, with the inbred shortage (Li, Hamdi, Jiang, Cao, & Hou, 2000), it is not yet competitive to the long standing PSTN in voice field relying only on uni-modal communication (Modarressi & Mohan, 2000; Wah & Lin, 1999). Therefore, apart from voice quality, IP-Telephony should emphasize the integration between data, voice and multimedia and implement the synchronizing voice as well as visualized multimodal interaction so as to enhance its competitiveness to the traditional telecommunication.

Among the multimodal applications (Kondratova, 2004), VoiceXML (Abbott, 2001; Sharma & Kunins, 2001) is an XML-based Internet markup language and can be applied to the speech interface that enables telephone access to the PSTN, Internet, or website contents (Maes, 2002; Privat, Vigouroux, Truillet, & Oriola, 2002). Users can access VoiceXML by dialing the associate phone number of the application. From the usage point of view, this phone number is similar to the uniform resource locator (URL) of the website. Therefore, VoiceXML applications provide another end-to-end telephone system solution for call centers.

The major encouragement of VoiceXML development has been from the telecommunication industry which wants to make existing telephone networks an essential component in the information age. However, using dial-up service to access the Internet is restricted by the slow bandwidth and unstable connectivity. Even the broadband service is growing, the whole architecture is not mature enough to replace the existing, well-dependent, and highly-tuned Plain Old Telephone Service (POTS) (Beasley, Farley, O'Reilly, & Squire, 2001).

To sum up, VoiceXML consolidates the strength of the Internet with the ubiquity of the telephone, and makes it possible for businesses to replace the expensive and proprietary interactive voice response (IVR) platforms with a unified architecture for delivering automated service from any telecommunication devices. From the cost-reduction point of view, VoiceXML applications have acquired great acceptance from many of the call centers (e.g., Tellme Network[2]) and many big corporations to improve their customer relationship management (e.g., AT&T Corporation).

1.2. The goal

This paper is proposing a new Internet Telephony platform and architecture which can provide VoiceXML dialogue as the multimodal application. When IP phone dials the non-PC-based PBX and traditional telephone device, it employs the built-in browser of IP-Phone and through other web server to obtain plenty of useful visualized information related to the device with extra multimodal telecommunication services. In this way, the integration between voice and data is easily achieved at the application layer. Without replacing or minimum upgrading the existing telecommunication equipment, the enriched web pages with markup language proposed by this article can provide the mentioned functionality to facilitate the IP-Phone service. In order to be compatible with pure voice channel of traditional telecommunication, this system tries to deploy an interpreter module like VoiceXML to bring the concept of transferring voice information to control instructions (Shan, Zhou, & Zhang, 2001). Compared with traditional data-link control system, our method is more practical in that all we needed is to pronounce the voice control instructions via traditional telecommunication channels. At the other end of the IP-Phone, PC will convert the voice into commands by means of automatic speech recognition (ASR) so as to be synchronizing with the enterprise's PBX, IVR device and provide multimedia and visualized interactive services (Georgescu, 2004). From the enterprise point view, as long as the IP-Phone user dials PSTN to the PBX in a company with this multimodal VoIP system, the enterprise can provide lots of useful information presented in multimedia mode via the web page on the web server. Coordinating with IVR procedure in PBX, the system realizes interaction between the voice and visualization.

In this paper, we implement a prototype of IP-Telephony system, a visualized multimodal VoIP with HTML/XML browser. While PC users dial traditional telephone line, computer will display relevant visualized information to assist the user's dialing and communication. Being compatible with the traditional telecommunication system, this structure will enable voice communication to integrate with multimodal in a simple and reasonable way. Thus, the enterprise is able to fully utilize the existing resource and realize integrated benefits. Besides, many Internet services are accommodated in "Web Service" architecture, this multimodal IP-Telephony will be able to provide many visualized data integrated application such as IP-Phone number inquiring, messages and even faxing which will be conveniently set up by enterprise itself.

There are quite a few of related works like Plum Voice Portal Technology[3] can present existing websites or intranet applications to a phone user. IBM's WebSphere provides HTML-to-VoiceXML transcoding that can be

[2] http://www.tellme.com

[3] http://www.plumvoiceportals.com

converted to speech by a VoiceXML browser. However, the commercial softwares are generally costly and unaffordable for researchers. In addition, none of these applications can provide the Mandarin dialog system which achieves the multimodal function as we intend to achieve.

To take the advantage of the proposed system, this study leverages the VoiceXML characteristics to integrate voice recognition and synthesis technologies with markup languages. In addition, we incorporate voice and graphical interfaces into current web architectures. The structure of this paper is as following: in Section 2 we explain how to construct the VoIP infrastructure and web-based Mandarin dialog system architecture in details. Consequently, we illustrate sample operations scenario in Section 3 and demonstrate the implementation results with comments and discussion in Section 4. Finally, the contribution and conclusion are stated in Section 5.

2. The system structure

Our major goal of this study is to design a web-based Mandarin VoiceXML dialog system. The ideal model for this dialog system has to integrate with traditional telephone system, Internet phone (VoIP), and existing VoiceXML dialog system for reducing the cost of introducing speech-enabled services. Thus, in the following sections, we will start with an analysis of the current telephone network architectures, including PSTN and VoIP, and current VoiceXML system architecture. Then we will move towards to describe the architecture and implementation of our web-based VoiceXML dialog system.

2.1. Pre-analysis

As mentioned above, PSTN (Faynberg, Gabuzda, & Lu, 2000), also known as POTS, is the most common telephone service in people's life. It supports our speech and dual-tone-multi-frequency (DTMF) interactions and responses from the callers. The traditional PSTN-based IVR platform has the following operational processes:

- End users make the phone call and IVR platform receives the incoming calls.
- IVR platform responds with greeting prompts.
- IVR platform waits for users' response and accepts the speech or DTMF input from users.
- The application takes action based on users' input into the IVR platform.

On the other hand, the new technique: VoIP, voice and data services can be used to provide the most cost-efficient, and flexible way of building networks. VoIP technology uses specific protocols, such as H.323, session initiation protocol (SIP) (Handley et al., 1999), and media gateway control protocol (MGCP) (Arango et al., 1999) for call setup and media gateway control. While people use PSTN connection, they must pay based on the time they stay on the lines. The longer they stay, the more they should pay. In contrast, leased Internet connection is generally a fixed rate. Thus VoIP application will not incur extra cost and the overall expense is much cheaper than PSTN service.

Unfortunately, there are some problems in the integration between VoIP architecture and Internet. A stable bandwidth provision among all time is generally not feasible or at high cost. As we can imagine, voice data communication requires a real time streaming service: this is in contrast to the heterogeneity of Internet architecture that can be made of many routers and have high round-trip time (RTT). Therefore, the quality of service (QoS) is still a crucial issue for VoIP applications.

2.2. System architecture

This section illustrates each module and operation of the system, which is shown as Fig. 1.

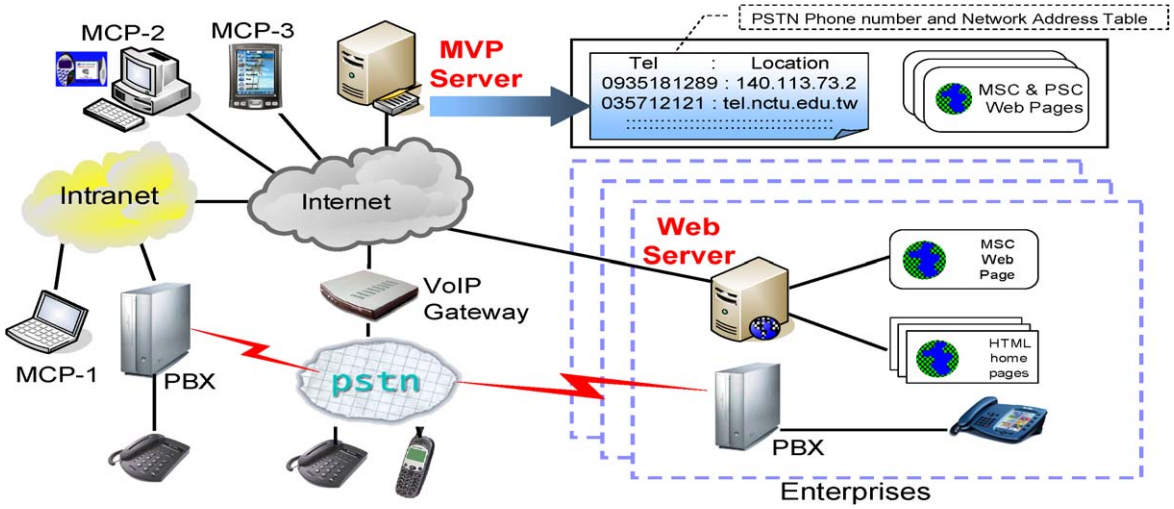

Fig. 1. Multimodal VoIP platform (MVP) service architecture.

2.2.1. Multimodal VoIP platform (MVP) architecture

MVP architecture discussed in this paper will enable the PC's VoIP software to provide visualized effect on screen. Under PC2Phone, dialing through Gateway to non-PC-Based PSTN circuit and device, in order to achieve multimedia integration; as the ability provided by PC-to-PC, and to simplify the service protocols of non-voice section, this architecture deploy the well-built web server of the enterprise to store interactive information which describes various kinds of PSTN telecommunication device or even the advertising web page. Such an arrangement makes the traditional telecommunication device and gateway equipment more independent with protocols. As the architecture is based on XML and text interface, it can easily reach interoperability under HTTP.

In addition, in pure voice service of VoIP, as gateways transmit its voice data only, this architecture can easily provide voice layer routing via ready-made gateways of different enterprises.

This architecture will enable PC-Phone connection to obtain the telecommunication service content from the enterprise's web page when dialing to PSTN. Therefore, the first task is that system will be able to provide corresponding website address when users dial the PSTN phone numbers. In fact, the greatest difficulty of PC2Phone calling mode lies in the inability to acquire the so called web numbering information. The reason is that IP2PSTN has nothing to do with Internet after signal past the gateway. Therefore, this architecture intends to build up an MVP server, as shown in Fig. 1. An important task of this MVP sever is to provide a lookup table from telephone numbers to the corresponding web site. Every PSTN phone number of enterprises should have a corresponding website which can be referred by the browser. Once the service website location is available, it will be able to provide lots of visualized multimedia information, and many other kinds of value-added service during the communication.

The second task of MVP server is to provide sample web page for multimodal service content (MSC) and public service content (PSC) at the initial stage. As this system is under the testing, system developers will provide many MSC and PSC sample pages stored in the MVP server for reference during the simulation. However, in the future, these web pages, like the situation of home pages or WAP applications in enterprises, will be set up and maintained by enterprises themselves. In this way, enterprises will be able to provide the most updated multimodal MSC information. In addition, third party and other manufacturers can develop many public services in assistance to PC-Phone users globally.

In Fig. 1, multimodal client phone (MCP) is the softphone for PC users or hardware with equivalent functionalities (MCP-1, MCP-2, MCP-3 in Fig. 1). When MCP starts dialing, phone number is sent to MVP server and acquires MSC web page address corresponding to the number. MCP browser will soon send HTTP to this address and receive response from MSC web page. MSC web page contains visualized multimedia interactive information and non-visualized telecommunication device information, both of which are labeled with HTML/XML markup language or Meta-Tag. Multimedia information will be displayed on the imbedded browser while controlling information, after being parsed from MCP, execute MSC instructions by the program to coordinate and control special communicating processes or even sent out DTMF voice signals. It enables the telecommunication devices on remote voice channel such as PBX and IVR, to switch the in-call to the specific extension number. In other word, the MSC web page will coordinate the related service and execute multimodal communication.

As shown in Fig. 2, enterprise sets up an MSC web page which will place the feature parameters of PBX and IVR, voice response instruction, extension number and other multimedia information into enterprise's web server using MSC format language as proposed in this article. The platform default different file appendix names, defined as "index.msc", so as not to effect the access of original home page. When MSC browser retrieves MSC web page

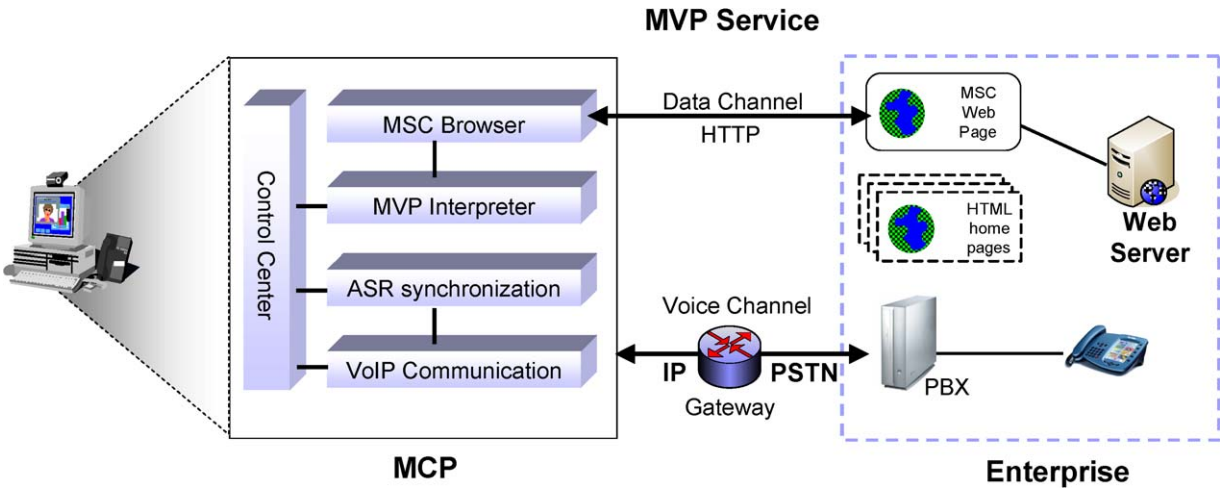

Fig. 2. MCP visual/voice multimodal flow.

through data channel by HTTP, with the assistance of MVP interpreter in MCP, it enables the browser of VoIP to provide much useful visualized information which is displayed on the screen in multimedia type. With visualized browser, blending with voice, mouse, keyboard, such as multimodal human–computer interaction, user obtains different kinds of information from enterprises and assists the process of dialing, talking, switching, and multimedia interactive transaction. All of these are communicating processes in multimodal IP-Telephony domain.

2.2.2. MCP platform

MCP platform has four basic modules with different tasks to achieve interactive VoIP function when connect to traditional telecommunication device with browser as shown by Fig. 3.

(1) *Telephone interface component*: telephone interface can be displayed on computer screen and simulates the user-friendly operation in actual phoning process. Its main functions includes dialing, sending DTMF signals, recording, receiving controlling instructions from browser and other expediently assisting operation. In view of the interaction with traditional telecommunication device, telephone interface is embedded with many functions that can resolve instructions into internal execution. These instructions partly come from MSC label script and partly from voice message sent by remote PBX device. So we may say, this system is to transplant the ability of protocol from the enterprise's telecommunication device entirely to PC VoIP software. Exploiting PC program and voice recognition technology, it achieves and coordinates the interactive and synchronizing communication between IP-telephony and traditional telecommunication device.

(2) *Browser*: this component mainly produce visualized browsing window and interpret MSC script language. When a user dials to the counterpart, the system will at the same time fetch the web location during the browser connection. Since MSC web page can be placed in the existing web server directly, this system will retrieve MSC web page according to URL or IP standard format, and display the visualized information on browser. Meanwhile, the system will resolve the implanted Meta-Tag controlling instructions and execute the procedures in cooperation with telephone module. For instance, when a user dials a certain enterprise, VoIP module create the voice connection between both sides, at the same time, the system will send back the MSC page location corresponding to the phone number. In general, the location will be the enterprise's homepage address. From the web page, there are extension numbers and names labeled as controlling instructions by special Meta-Tag. So PC users may see URL hyper-linked and then click to the extension number which the call will be switched to the intended person. This action can be conceived as phone user pressing the number key to switch the calling, but browser obtaining the extension table in this case automatically operates the switching.

(3) *VoIP communication component*: the VoIP communication module of this architecture seeks to establish voice-layer communication with gateway. The system is designed to switch voice via different gateways, without worrying the compatibility issues with supporting multimedia protocols in different gateway. Therefore, the user can select gateway servers provided by different ISP to connect PSTN lines to make the phone call. The purpose of this module is to be able to switch among different VoIP networks, make use of the existing mature standards and service systems without changing or minor upgrading of the existing enterprise infrastructure.

(4) *Network service interface*: the last component of this architecture is the network service interface. This interface can be connected with the MVP server of the remote site. Since this server stores a lot of lookup tables of telephone numbers and MSC address, MCP software platform can transfer telephone numbers to web addresses via this server interface during VoIP dialing. This interface also reserves additional MVP services that can be applied for the future usage. For example: controlling service for billing, IP addressing service while PC2PC connecting, or using LDAP for directory services.

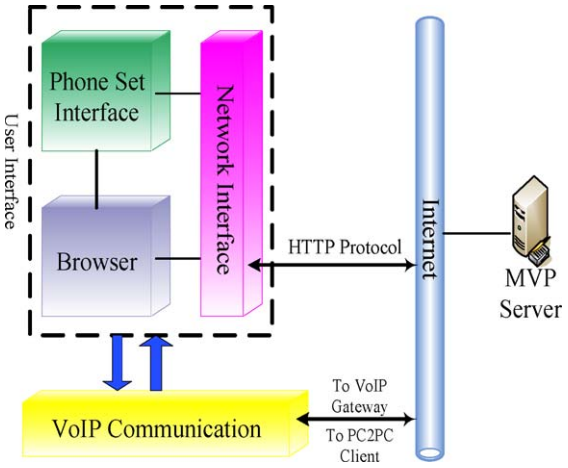

Fig. 3. Multimodal client phone (MCP) components.

2.2.3. VoiceXML dialog system architecture

To further utilize the proposed MVP multimodal architecture, Fig. 4 shows an end-to-end configuration for a VoiceXML application. The web server provides users the VoiceXML content through a voice server, VoIP gateway, and PSTN to the telephone (Lucas, 2000). The voice server has the TTS (Ball et al., 2000), ASR (Houlding, 2001; Leavitt, 2003), and H.323 telephony components along with a VoiceXML browser (Modarressi & Mohan, 2000). The VoIP gateway has the voice interface card (VIC) along with VoIP software and H.323 telephony components which interact with the PSTN. VoIP and H.323 components are available in one router or gateway (e.g., Cisco 2600 series, Dialogic D240) (Edgar, 2001).

The current end-to-end VoiceXML system architecture in Fig. 4 has some existing problems. At first, the existing

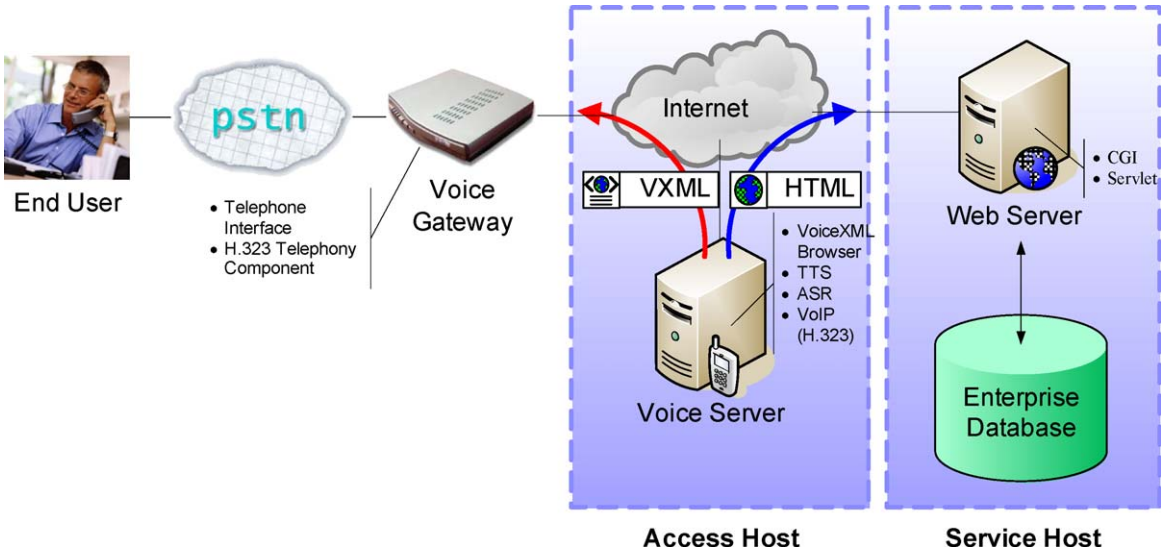

Fig. 4. The architecture of the end-to-end VoiceXML system.

VoiceXML system just has VUI representations; it is not suitable for users who are only familiar with GUI interaction. The second, it only supplies one-way access to the VoiceXML applications through the PSTN. It will be more convenient if end users (e.g., callers or PC-based users) can access the VoiceXML application in a bi-directional communication through the PSTN or Internet under the proposed MVP architecture. Hence, Fig. 5 shows our logical prototype of the web-based Mandarin dialog system. The system we proposed is platform independent. The key components involved are the ASR, TTS synthesis, VoiceXML Browser, VoiceXML Parser, and VoIP Internet phone (IBM, 2000; Tsai & Ho, 2000) which were mentioned in MCP of Fig. 2.

3. System operations

3.1. MSC description

This session briefly explains multimodal interactive language of MSC page and the operation of voice recognition protocol. It also explains how the system manages to set up linking on web address with actual phone number to facilitate web dialing.

3.1.1. MSC multimodal interactive script language

As MSC pages must be compatible with the IP-Phone platform, it is necessary to develop some easy and intuitive protocol standards and procedures so that different website

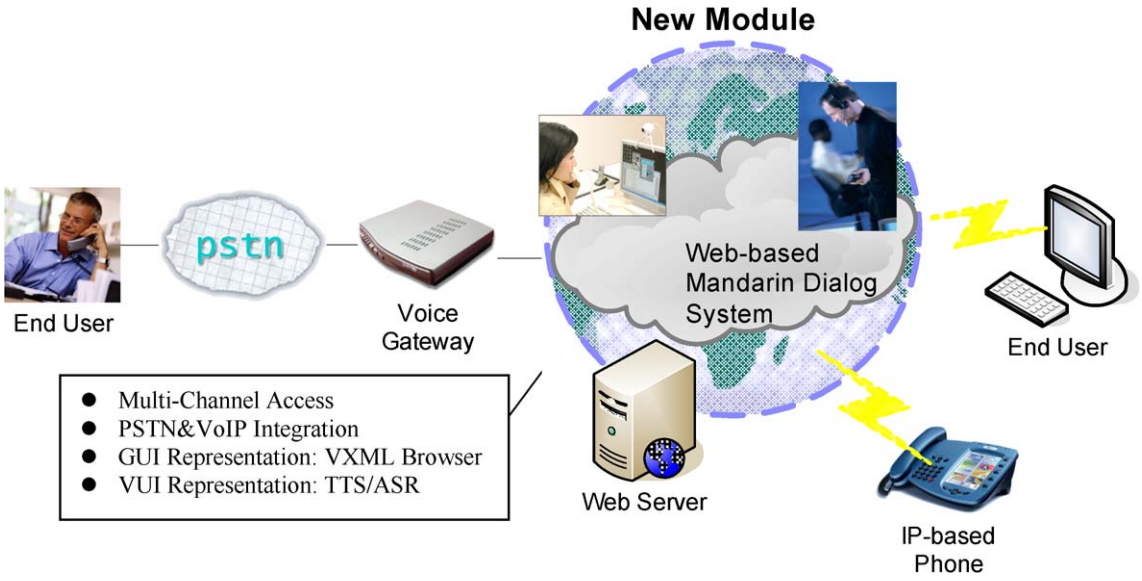

Fig. 5. The logical structure of the web-based Mandarin dialog system.

builders will have no troubles with MSC pages in controlling PBX coordinated with IP-Phone callers for multimedia information interaction. Therefore, the first thing to deal with is the location and name of MSC web page. For the sake of convenience, this paper denominates it as "index.msc," which takes the same position with index.html on homepage, such as www.nctu.edu.tw/index.msc.

In order to realize interaction between HTML and VoIP platform which has to be compatible with genuine browsers, this system deploy:

** content **

This hyper-link language format will intercept interactive commands to IP-Phone platform for resolving and responding. For example: write one sentence on the MSC page for a certain enterprise:

** Bill Wang **

Browser of VoIP platform will then emerge hyper-link letter Bill Wang. When mouse moves to the upside of the letter and click, browser will receive a web address and requirement for connecting. As long as the URL Type is **"mvp://"**, the system will intercept the information back to the platform and resolve it. As an example, the command is "pbx", data "102", so the platform will immediately send "102" DTMF signal routing the in-call to extension.

Some other script grammar may assist interaction of telephone communication as below:

"mvp://callto/tel": hangs up the phone, then dials to a new route described by "tel" to start phoning.

"mvp://voicemail/mailbox": start the recording device, send the recorded voice file to the address described in "mailbox" by means of email.

With interactive MSC page, telephone and computer integrated perfectly following the procedure described. When an IP-Phone user dials the enterprise, the user can obtain all kinds of real time interactive information related to the phone call. There will be more labels and grammars that can be defined for telephone-integrated service in order to make it more convenient for IP-Phone users in order to employ information integrated IP-Telephony.

3.1.2. Commands for MSC voice recognition language

MSC may realize simple protocol interaction with traditional voice device via voice or DTMF recognition instruction. There are two main utility of MSC voice recognition language in this multimodal IP-Telephony. First: send the voice of traditional telecommunication device to VoIP software as a command via ASR recognition or DTMF signal in order to determine the state of the remote telecommunication device. Secondly: extract voice command from both sides' conversation via the on-line ASR technology to provide auxiliary information automatically and synchronously.

For MCP platform, there will be two voice commands of output and input in different mode after recognizing the voice. One is the informing mode, a starting command in language **<asr state>**; the other is an sending event mode, a command in language **<asr event>**. For instance, on PBX of a certain enterprise, there is a simple IVR recording as: "Hello, This is NCTU Technology, please dial the extension number or wait for the operator", in cooperation with the traditional voice device, a label grammar like VoiceXML should be written on MSC Page then as the follows:

<asr state = "trans_state"> while <or> operator </asr>

Meanwhile, VoIP will execute ASR functions to check the keywords for "while" or "operator". If the key words are recognized, it will send "trans_state" command to inform MCP software for sending DTMF signal.

If users do not send any DTMF signals after traditional IVR waiting, or the signals are not fairly recognizable, The IVR will send the prompt voice once again, for example: "please try again". In this case, accordance with the actual situation, it should be written as following:

<asr event = "dtmf_again"> "Please dial again" </asr>

On receiving "dtmf_again" command, MCP will re-send DTMF signal clicked by the user automatically.

On receiving correct DTMF signal, IVR usually responds with "please wait". That is the time required for routing. IVR often plays welcome music during the waiting time and MCP may take the time to play some commercial videos of the enterprise. The following labels may be used in achieving the effects:

<asr state = "between_transfer"> "please wait" </asr>

On receiving "between_transfer", MCP is able to play the multimedia pages automatically, which has been set as the default page on the PC browser by the following script:

<choice next = "default.msc"> between_transfer </choice>

One can even pre-record the DTMF voice of assigned numbers and save it in a PBX to describe the status of the devices by the variable.

<asr state = "get_dtmf = 5s"> nn </asr>

"5 s" is the time for waiting seconds. If MCP receives a "102" DTMF signals from PBX within 5 s, the "*nn*" will be the value 102. One can easily define any number to certain status for a PBX or IVR.

When IVR gives the voice of "thank you, good-bye", it usually on-hook the phone while MSC sustains the state by the label below:

<asr state = "onhook_state"> thank you, good bye </asr>

After PBX successfully transfers the inbound call to the extension number, the bell rings with the phone conversation. If no one answers the phone, PBX/IVR will then take it over by giving the voice of "The line is busy, please dial later" or "the person you are calling is not available at this time, please try again later" or something else. This architecture allows writing ASR commands into MSC pages of the enterprise according to the actual voice recording. In this way, IP-Telephony offers hand-free interactive service during the phone conversation.

3.1.3. MSC web numbering mechanism

Since many researches have been emphasizing numbering system in communication domain, our method is fairly intuitive for MVP architecture to add a Meta-Tag for labeling the portal number of telephone of the enterprise on "index.msc" web page. For example:

<meta name = "pstn entry" content = "02-82269968">

Once people use MSC-supported IP-Phone, they may dial with web address, such as www.nctu.edu.tw, the platform will then automatically retrieve "index.msc" page from the location to extract phone number of the enterprise, with which, VoIP module will execute the dialing operation. If web address dialing information is not available from local enterprise's page, MVP server will locate the phone number from the database to provide the supplemented information. Even if the company is located overseas, the international call access is still achievable under this structure. This procedure manifests and indeed brings convenience of web address dialing to PSTN.

3.2. Illustration of multimodal communication

As shown in Fig. 6, this study has completed a prototype system of MVP and MCP system. The network facility is tabulated at Table 1. The specific function for each module is explained as following:

- MCP1, MCP2 and CP-7905G are the softphone and hardphone for this study. It can receive and transmit the message including SIP. The MCP softphone has a browser, which can display multimedia information for auxiliary use during the connection.
- Call server includes the proxy and registrar servers. It will handle the user's registration and message forwarding service.

Table 1
The network facility list

Facility	Specification
Server PC	CPU: Pentium4 3.20 GHz RAM: DDR-1GB NIC: Intel PRO/100 VE network connection OS: Windows XP SP1
Hardphone	Cisco IP Phone 7905G
Call server and BGCF	IPTel SIP Express Router
MVP server	NTP prepaid server with content service of MCP and PSC
PSTN Gateway	Cisco 3745 PSTN Gateway

- Breakout gateway control function (BGCF) is an IP multimedia subsystem (IMS) element that selects the network in which PSTN breakout is to occur. If the breakout is to occur in the same network as the BGCF then the BGCF selects a media gateway control function (MGCF), this will be responsible for interworking with the PSTN.
- MVP server include the NTP prepaid server with the content service for MSC and PSC. It will handle the calling payment issues.
- Voice server has the TTS, ASR and H.323 telephony components with VoiceXML browser.
- Media gateway transfers the message and in the mean time processes the analog and digital signal conversion if necessary.
- NCTU PBX deals with the PSTN network connection. By forwarding the signal from the media gateway, it provides the service to call PSTN and mobile phones.
- Taiwan Academic Network (TANet) provides the information infrastructure of the back bone and regional networks for the major research universities in Taiwan. NCHC is the National Center for High Performance Computing (NCHC) in Taiwan. Some of the simulations are performed across NCHC Network.

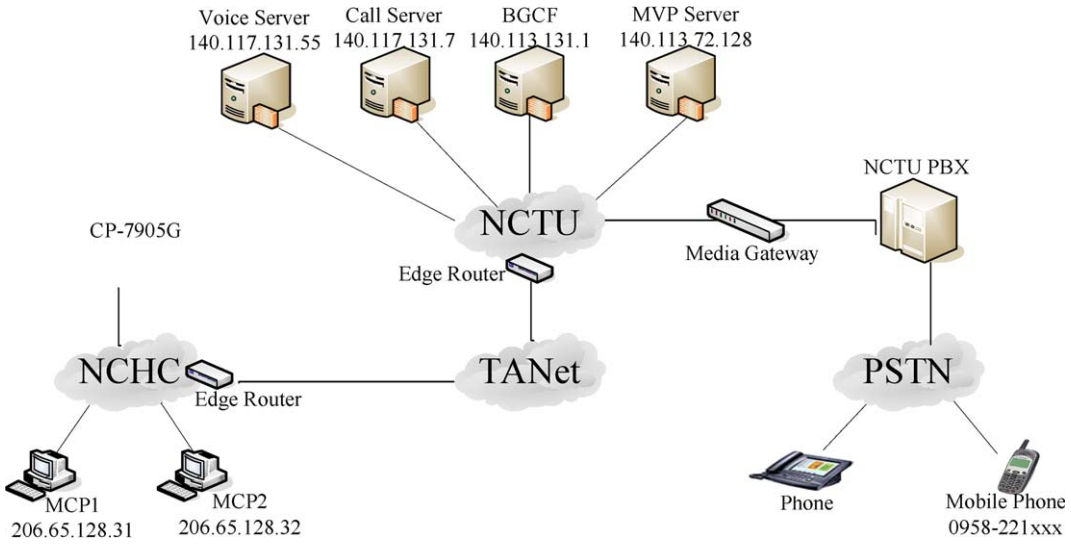

Fig. 6. The architecture of the proposed MVP system.

Fig. 7. Examples of an MCP calling to PSTN.

Fig. 7(a) shows that when MCP starts, the browser will display the default MSC page that can be a friends' phone list or be an extra function page provided by third parties. This software can easily modify the user's interface and make use of label grammar and Meta-Tags proposed by this paper to produce multimodal interactive effect.

As shown in Fig. 7(b), a user can apply the same scripts to set up personal MSC with the address lists of relatives and friends. When the PC phone software is executed, the user only needs to click the hyperlink for relatives and friends information on the browser, which produces the same effect as dialing on the IP-Phone, and then execute PC2PC or PC2PSTN connection. For example, when the user clicks this hyperlink of NCTU, the system will receive the descriptive instructions containing "Call-to" in Meta-Tag, therefore directly dial to the PSTN telephone portal of NCTU to establish a communication connection for voice conversation as common IP-Telephony software. On the other side, since this PSTN telephone portal of NCTU is 035712121, whether a user dials directly or as this example shows, clicks on the browser, this telephone number will be sent to the MVP server to retrieve the corresponding web location. An MSC address is then sent back, and the browser will connect to this MSC web page via the HTTP protocol. The content of MSC page in this NCTU example is to display specific extension numbers and their users' name when PBX plays IVR voice, as shown in Fig. 7(c). Therefore, VoIP dial-in users can see the extension number table with hyperlinks on the browser while listening to IVR. The user does not need to hear all IVR voice operations and only need to click the hyperlink on the browser, and then the IP-Phone will send the DTMF signal of this extension number and route to this extension position. Consequently, using the IP-Telephony proposed by this paper to dial to certain enterprise, you do not have to memorize people's extension numbers, or need the operator's help of a call center. All you need to do is to click the hyperlink on a browser of this specific soft-phone.

Apart from switching to extensions, IP-Telephony with browser can take advantage of the MSC page with built-in multimedia interactive information to offer many useful auxiliary functions during the dialing and communicating process. As shown in Fig. 7(d), when dialing in an enterprise's telephone system, this enterprise can display specific multimedia advertisement or other web pages with interactive capability with useful information.

3.3. Implementation of VoiceXML system

There are two major parts in the whole web-based dialog system: the VoiceXML Browser and the VoIP Internet phone. The VoiceXML Browser serves as the traditional web browser (e.g., Internet Explorer/Netscape) and supports VoiceXML-format files to be used in the browser. VoIP Internet phone is illustrated as the virtual phone appearance in Fig. 7 (The elliptical shape portion of the

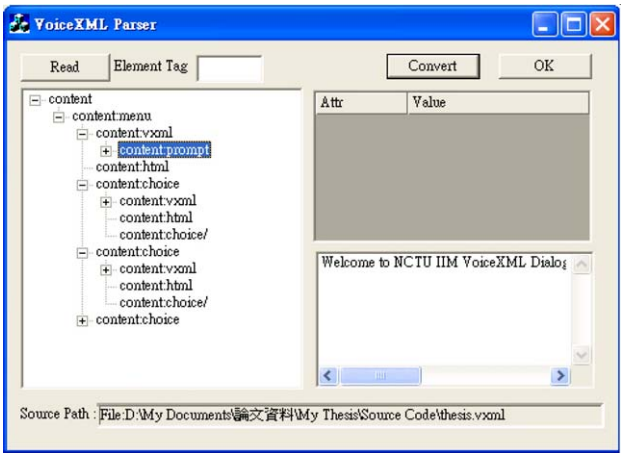

Fig. 8. VoiceXML parser.

NCTU RING in Fig. 7). VoIP Internet phone provides basic VoIP function so that the users can make the phone calls to mobile phones numbers or basic PSTN phone numbers. Besides, there are TTS, ASR, VoiceXML Parser, and XSL module in our VoiceXML system. We will explain each component in the following sections.

3.3.1. VoiceXML parser

The VoiceXML browser needs to parse VoiceXML pages. The VoiceXML pages are XML documents essentially. We utilize the Microsoft XML Core Services software—MSXML 4.0 SP1—to offer a number of new features and improvements over the MSXML 3.0, including support for the XML schema language and faster parser and XSLT engine.[4] Finally, we develop our own VoiceXML parser by MSXML 4.0 DOM technology. This mode gives a tree-like data structure of VoiceXML for the document in Fig. 8. The VoiceXML parser is used to parse a VoiceXML document. After parsing the document, it should retrieve the forms and fields content from the document. Our VoiceXML Parser supports reading or writing VoiceXML and validated XML files. Although it is very similar to the DOM centric class, each element it provides may be used independently. The specific function provides the programmer a great deal of flexibility.

3.3.2. XSL transformation processing

The prototype of the web-based Mandarin dialog system demonstrates an approach in which the static content is written in a content markup dialect. It will further be transformed into either HTML or VoiceXML format. The advantage of using this approach is that the static content is stored in one source file which improves maintainability. In addition, each of these static files contains various attributes and sub-elements that capture all the information needed to generate HTML or VoiceXML. So the root element also contains the elements of <html> and <vxml> tags.

Fig. 9 shows one XSL stylesheet that transforms any document into HTML while the web-based Mandarin dialog system is functioning. On the other hand, there is another stylesheet transformed to the VoiceXML file. In our prototype, the system will transform a file into an HTML file while the original file contains a sequence of anchors (<A HREF = ...>element) in the choice menu source code. In addition, the system will output a VoiceXML file while it contains the VoiceXML form with <choice> element. We use different presentation rules to express the HTML, and VoiceXML file. For example, the <content:html> element captures more verbose texts for GUIs and will be transformed into the HTML file, and the <content:vxml> element captures concise voice or speech content to be transformed into the VoiceXML file.

3.4. VoiceXML dialog system operations

When we start the VoiceXML Dialog System, it will display the fast food ordering menu as an example in Fig. 10, which we use as a voice ordering service demonstration. There are three fast food restaurants in our implementation, including the main menu (Fig. 10(a)), McDonald[5] (Fig. 10(b)), Pizza Hut[6] (Fig. 10(c)), and Yoshinoya[7] (Fig. 10 (d)). As shown in Fig. 10(a), whenever the user connects to the demo system, he can simply enter the McDonald website by pronouncing the correct keyword "McDonald". Then the system will open the corresponding corporation food ordering menu as shown in Fig. 10(b). After entering the McDonald fast food menu, Fig. 11(b) will display all the commodity items and price information through VoiceXML pages at the user's browser when the user plan to make an order. In our McDonald food ordering system, it offers four choices: Japanese hamburger, Korea hamburger, apple pie, and coke. Users can pronounce the specific keywords to indicate what they need and the system will reply with the verification information. Finally, Fig. 11(c) illustrates a successful transaction and then the user exits the system.

4. Discussion

4.1. MVP efficiency

During the study, the operation of clicking hyper-linked web page is thus a non-linear jumping process which provides more interactivity than the sequential mode of pure voice telephone communication. From the experiments, multimodal IP-Phone may bring benefits as the following:

[4] http://msdn.microsoft.com/xml

[5] http://www.mcdonalds.com/

[6] http://www.pizzahut.com.tw

[7] http://www.yoshinoya-dc.com/n_top.html

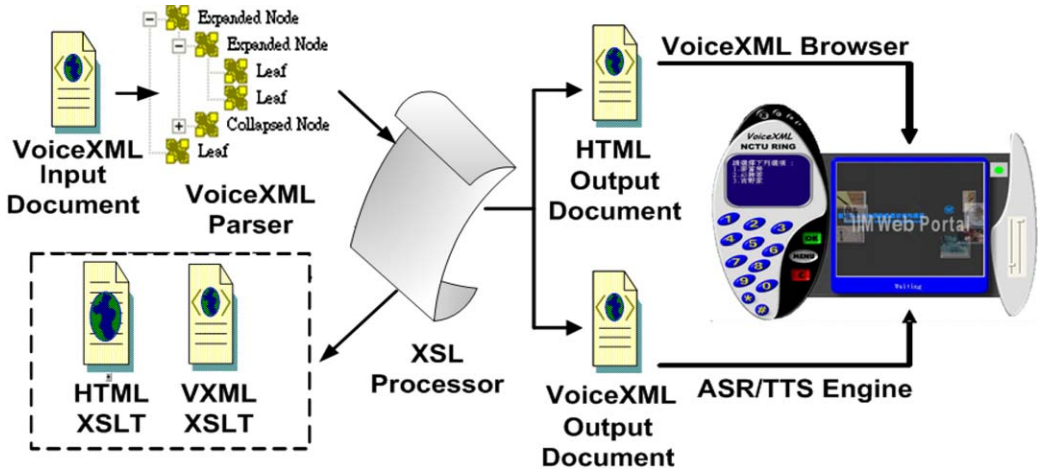

Fig. 9. XSL transformation processing.

Fig. 10. (a) VoiceXML Mandarin dialog system main menu, (b) McDonald, (c) Pizza hut, (d) Yoshinoya fast food menu.

(1) It relieves the users from the burden of memorizing phone numbers and saving the time for number searching. With the web numbering, users communicating with enterprise's individuals are in a more efficient way.
(2) Browser of Multimodal VoIP, which is maintained by the enterprise itself, can offer the visualized interactive information and automatically update web pages according to the status of PBX or conversation content. In this way, it offers the appropriate information constantly. This is really an efficient way of communication, like a smart secretary, laying out relevant files at any moment on the course of conversation.

We have found on experiments, there is 95% of accuracy of voice recognition in this architecture which will turn the keywords into proper instructions. The accuracy rate of ASR is compatible with the one from the speech related applications. Even when PBX takes wrong voice recognition or routing mistakes, the system will continue the oper-

Fig. 11. VoiceXML mandarin dialog system—example of McDonald's fast food ordering processes: (a) the greeting information, (b) the main food menu, (c) the completion of the transaction.

Table 2
Analysis table of system performance

Keywords no.	Keywords	How many times does the system recognize the order successful?
Task-1 McDonald		
1	McDonald	
2	Japanese hamburger	
3	Korea hamburger	
4	Apple pie	
5	Coke	
Task-2 Pizza hut		
6	Pizza hut	
7	Fresh seafood	
8	Super supreme	
9	Hawaiian	
10	Japanese	
Task-3 Yoshinoya		
11	Yoshinoya	
12	Beef meal	
13	Chicken meal	
14	Beef and chicken	
15	Main menu	

ations by repeated recognition or manual switching. The time delay is about 2 s which is tolerable for users. Therefore, the system exhibits its functionality and proves its potential in the interactive multi-media information operation. We are also aware that the delay, bandwidth congestion, packet loss, QOS issues will occur while the system is really implemented in real enterprise networks. However, the problems could be relieved while the broadband connection become universal and more powerful servers of Table 1 are implemented for the MVP platform.

4.2. VoiceXML efficiency

To verify the VoicdeXML multimodal capability, voice server was deployed in our performance evaluation environment as shown in Fig. 6. The result shows that the speech recognition and synthesis modules are fully tested with high accuracy. We also invited ten persons to attend the experiment for system evaluation. Half of the experiment samples are veterans who are familiar to operate or control the dialog system while the others are novices to our VoiceXML Mandarin dialog system.

We design our system performance analysis table and the testing items with pre-defined keywords. From Table 2, we divide three fast food restaurants ordering subsystems into three tasks in our system performance evaluation.

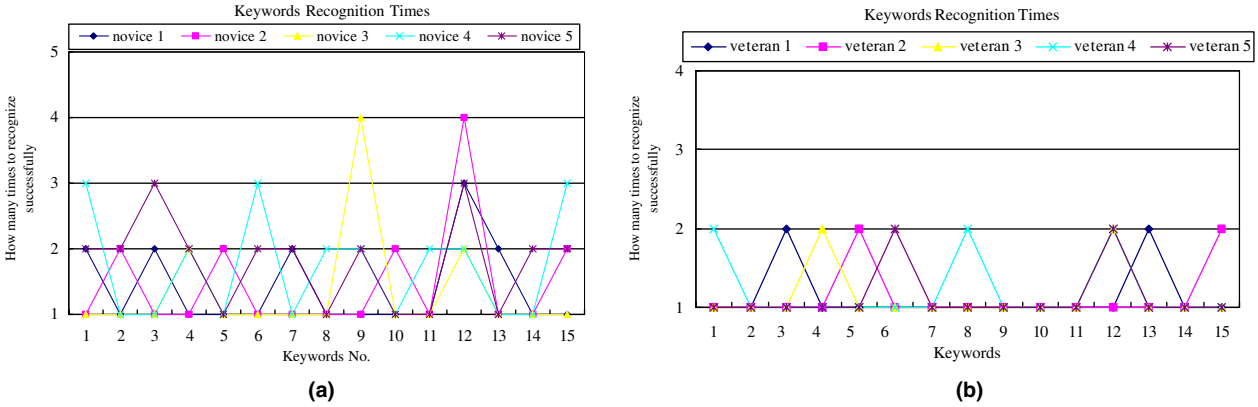

Fig. 12. Successful rate analysis of keywords recognition: (a) for Novices, (b) for veterans.

As Fig. 12(a) shows, the result indicates the novices sometimes have higher speech recognition errors which are caused by ambiguous input keywords or unfamiliarity with the dialog system. On the other hand, the result indicates the veterans generally have better speech recognition percentage in Fig. 12(b). Therefore, the consumers are better to be pre-trained or educated how to interact with the system, which is usually a prerequisite for speech-based applications.

5. Conclusion

This paper attempts to exploit more innovative applications for IP2PSTN to enhance the utility of Internet telecommunication from the view of multimedia interactive IP-Telephony. PC2Phone with screen display will create multimodal and other new applications that traditional telephone is not able to realize. It makes full use of the feature of Internet and multimedia so that IP-Telephony will achieve the goal that offers voice, data and multimedia integrated service as a whole package. The architecture proposed in this paper not only enables enterprises to write standard, visualized web service content but also facilitates many kinds of new IP communication services.

To extend the multimodal applications for wider usage like e-commerce, we deploy the web-based Mandarin dialogue system through the VoiceXML. The user can either use the telephone channels or personal computer with VoIP to access the voice server to simultaneously browse the information on the web server or enterprise database through the Internet. This approach provides multimodal access of directory management scheme for customer service or business management. The prototype system showed excellent performance from the experiments and can be easily constructed into a largely distributed telephone-based database and voice service provider for widely accessibility. The techniques and methodology we developed in this paper can take the advantages of the friendly speech interface and visual information to significantly improve the human machine interaction and communication.

Acknowledgements

The author likes to thank Mu-Yen Chen and Tien-Hwa Ho for data collection and simulation works.

References

Abbott, K. R. (2001). *Voice enabling web applications: VoiceXML and beyond*. Apress, pp. 33–40.

Arango, M. et al. (1999). Media gateway control protocol (MGCP) Version 1.0. RFC2705, October 1999.

Ball, T., Bonnewell, V., Danielsen, P., Mataga, P., & Rehor, K. (2000). Speech-enable services using TelePortal Software and VoiceXML. *Bell Labs Technical Journal*(July–September).

Beasley, R., Farley, M., O'Reilly, J., & Squire, L. (2001). *Voice application development with VoiceXML*. SAMS, pp. 10–30.

Edgar, B. (2001). *The VoiceXML handbook: Understanding and building the phone-enabled web*. CMP Books, pp. 79–110.

Faulkner Information Services (2001). VoIP Standards and Protocols, Faulkner Information Services.

Faynberg, I., Gabuzda, L., & Lu, H.-L. (2000). *Converged networks and services: Internetworking IP and the PSTN*. John Wiley & Sons.

Georgescu, S.-M. (2004). Multimodal access enabler based on adaptive keyword spotting. *Internet Technologies and Applications, 3262*, 349–357.

Handley, M. et al. (1999). SIP: session initiation protocol. IETF RFC 2543, March 1999.

Hassan, M., Nayandoro, A., & Atiquzzaman, M. (2000). Internet telephony: services, technical challenges, and products. *IEEE Communication Magazine*(April), 96–103.

Houlding, D. (2001). VoiceXML and the voice-driven internet. *Dr. Dobb Journal*.

IBM (2000). IBM WebSphere Voice Server with ViaVoice Technology Administrators Guide, Version 1.0. IBM Corporation.

Kondratova, I. (2004). Voice and multimodal technology for the mobile worker. ITcon, *Vol. 9*, Mobile Computing in Construction, pp. 345–353 (special issue).

Leavitt, N. (2003). Two technologies vie for recognition in speech market. IEEE Computer Society Press, Computer *36*(6), 13–16.

Li, B., Hamdi, M., Jiang, D., Cao, X. R., & Hou, Y. T. (2000). QoS enabled voice support in the next generation internet: issues, existing approaches and challenges. *IEEE Communications Magazine*(April), 54–61.

Low, C. (1997). Integrating communication services. *IEEE Communication Magazine, 35*(6), 164–169.

Lucas, B. (2000). VoiceXML for web-based distributed conversational applications. *Communications of the ACM, 43*(9).

Maes, S. H. (2002), IBM Thomas J. Watson Res. Center, Yorktown Heights, NY. A VoiceXML framework for reusable dialog components. In *2002 Symposium on Applications and the Internet (SAINT 2002)*. Nara, Japan.

Modarressi, A. R., & Mohan, S. (2000). Control and management in next-generation networks: challenges and opportunities. *IEEE Communications Magazine, 38*(10), 94–102.

Perkins, C., Hodson, O., & Hardman, V. (1998). A survey of packet loss recovery techniques for streaming audio. *IEEE Network, 12*(5), 40–48.

Privat, R., Vigouroux, N., Truillet, P., & Oriola, B. (2002). Accessibility and affordance for voice interactive systems with the VoiceXML technology. In *Computer helping people with special needs: 8th international conference (ICCHP), Vol. 2398*, pp. 61–63.

Rizzetto, D., & Catania, C. (1999). A voice over IP service architecture for integrated communications. *IEEE Network*(May/June), 34–40.

Shan, M., Zhou, Y., & Zhang, Y. (2001). CORBA based distributed computing model for multimodal speech recognition. In *Proceedings of 2001 international symposium on intelligent multimedia, video and speech processing*, 2–4 May 2001, Hong Kong.

Sharma, C., & Kunins, J. (2001). *VoiceXML: Professional developer's guide*. John Wiley & Sons, pp. 143–252.

Toga, J., & Ott, J. (1999). ITU-T standardization activities for interactive multimedia communications on packet-based networks-H.323 and related recommendations. *Computer Networks, 31*, 205–223.

Tsai, M.-J., & Ho, T.-H. (2000). WWW and telecommunication collaboration service for Mandarin automatic personal phonebook inquire dialogue system. In *ICME 2000 conference*. New York, USA.

Wah, B. W., & Lin, D. (1999). Transformation-based reconstruction for real-time voice transmissions over the internet. *IEEE Transactions on Multimedia, 1*(4), 342–351.