Voice Portal Solution

Table of Contents

Technology Brief ..................................................................................................................... 2
What is Automated Speech Recognition? ................................................................................. 3
Key Service Delivery Components (hardware, software and network elements) ................. 6
Architecture Diagram .............................................................................................................. 7
Call flow example ................................................................................................................... 7
Speech Technology in IVRS .................................................................................................... 8
Glossary ................................................................................................................................. 9
Technology Brief

As Internet access becomes a basic necessity, alternative ways to go on-line without a computer will be required by mobile professionals, the visually impaired, and people without access to computers. Also, more companies will look for new ways to expand access to their Web sites, using voice technology.

The last decade brought an incredible convergence of communications and computers, with the World Wide Web arguably being the single most important development of our times. With the advent of easy access to the Internet, vast resources of information, both public and private, have already become readily available. As the pool of accessible information continues to grow, and better methods of selecting and filtering the desired portions are developed, the Internet is becoming an irresistible magnet for all variety of users. In fact, being connected is fast becoming a basic necessity.

The primary method of access today continues to be the computer, which has certain advantages as well as some limitations. Computers offer a visual Internet experience that is usually rich in content. Some basic computer skills and knowledge are needed to access the Internet. But, computer-based access is proving insufficient for the professional on the move. When in the car or away from the office or computer, accessing the Web is difficult, if not impossible. And, an increasing number of people prefer an interface that allows them to hear and speak rather than see and click or type.

The computer-based Internet experience also does not meet the needs of another segment of the population – the visually impaired. Neither visual displays of information, nor keyboard-based interactions naturally meet their needs and this segment is often unable to benefit from all that the Information Age has to offer.

Some existing Internet users have also identified problems with the visual Internet experience. Pages are increasingly full of graphics, advertisement banners, etc., which move, flash, and blink as they vie for attention. Some find this “information overload” annoying, and lament the delays it creates by severely taxing the available bandwidth.

While computers and their use are on the rise, they’re not ubiquitous yet. A large segment of the population still doesn’t have access to the Internet. In some cases, the barrier is cost, although the price of a computer has come down significantly in recent years. Other consumers have a basic distaste for complex technology, which prevents them from accessing Web-based information via a computer. A more natural, less cumbersome way to interface with the net would provide them an opportunity to experience the Internet as well.

In addition, much of the world population prefers information in a language other than English, which currently dominates the World Wide Web. A service that translates the accessed information into the desired language would clearly add value to these users.

As one industry analyst summed it up:

"Having easy access to information is what we’re looking for."

As the need for alternative access to the Internet becomes more evident, several technology companies are pursuing solutions. Their products include “smart” cell phones with visual displays, intelligence built into the handset, and voice-activated Web sites. These products address different aspects of the problems outlined above.
While these alternative technologies are in the pipeline, few are ready for market. But the very existence of a race to market by many companies is evidence of a large potential market.

The challenge therefore is to integrate existing technologies, or develop new technologies, to make simple, affordable, alternative Internet access possible. As the need for an alternative access method to the Internet has become evident, progress continues to be made by technologists to provide such solutions. One key area of focus has been voice-based technology, which would allow a very natural interface for most people, and address the limitations described earlier. The following article describes the Voice Portal system, its components and usage scenario.

**Voice Portal** is a system that enables customers to access information on the Internet through a telephone interface. It uses technologies like Speech Recognition and Text to Speech (TTS) conversion to create a user interface and let users navigate through the “voice web page” using a phone and voice commands. The idea behind a voice portal is to enable customers to use Internet information anytime and anywhere using the most common access device and the most natural way of communicating.

Judging from the various Internet articles, there seems to be a lot of hype around voice technology and voice portals, particularly in the USA. In Europe, voice portals are less common, possibly partly because of popular SMS (and WAP) services.

The market is expected to grow. Often cited estimate by researcher International Data Corp. is that by 2004, some 600 million people worldwide will hook up to the Net via PCs, but 1.4 billion will connect through cell phones and another 1.4 billion will get on through wired phones. Kelsey Group estimates that revenues from voice commerce are over $30 billion by 2005 and from speech portals about $4.6 billion by year 2003.

**What is Automated Speech Recognition?**

Automated speech recognition (ASR) is a technology that allows users of information systems to speak entries rather than punch numbers on a keypad. ASR is used primarily to provide information and to forward telephone calls.

**How does ASR work?**

Speech recognition works in a manner similar to human speech. Specifically, it works like the ear and human brain; the ear takes in the sounds in the form of vibrations, and the brain decodes the signals and determines meaning. In a basic sense, this is how speech recognition works.

At the core of all modern high-performance speech recognition systems are complex statistical models that are able to characterize the properties of the sounds of the language to be recognized. This is an especially challenging problem because of the large degree of variability introduced along the way. Things such as speaker differences (male Vs female, tone of voice etc), accents, speaking rate, background noise, telephone handset variability, transmission channel differences, etc. all affect the properties of the signals that reach the speech recognition system. Because of this variability, there is no simple way to model what a “b” sound or a “d” sound should be. It is the job of the statistical models, which are trained on large amounts of real speech, to take into account all of the factors described above. These models judge the probability that a given segment of speech is a particular phoneme (the basic unit of the sounds of a language).
The overall process of speech recognition is shown in figure below. The stages of recognition are:

1. **Capture and Digitization**: This stage interacts with the telephony hardware to listen for and detect the incoming speech. In order to support barge-in (to allow the user to interrupt the prompt), this stage must perform echo cancellation to remove the echo of the outgoing prompt from the signal, and must support speech detection in the presence of noise and any residual echo from the outgoing prompt. In order for barge-in to be effective, these algorithms must detect speech and cut off the outgoing prompt very quickly, ideally within 100ms from the beginning of the user’s speech. Because of these latency requirements and because this stage must be listening for the user’s speech for the duration of the call, this processing is best done on programmable DSP (Digital Signal Processing) cards, or on line-interface cards (which already have special purpose DSP capabilities on the card).

2. **Spectral Representation**: The spectral representation stage converts the input signal into a representation which better represents the aspects of the acoustic signal which are important for distinguishing among speech sounds. This stage mimics many of the known properties of the human ear. It first converts the signal into the spectral domain (energies over time in 128 different frequency bands) and then maps that onto a nonlinear spectral scale, which matches the sensitivity of the human ear. The signal then goes through one more transformation to allow the subsequent modeling stages to operate more efficiently. A number of techniques are used in the spectral representation stage to reduce the variability cause by noise and channel conditions.

3. **Segmentation**: The recognition system is based on work on segmental systems performed for many years at MIT and other research groups. The primary difference between this segmental approach and frame-based Hidden Markov Models (which is the original way that statistical models were applied to speech recognition) is that the statistical model is able to take into account the entire phonetic segment (the basic units of speech). Frame-based approaches apply their statistical model on uniform time segments (usually 10 millisecond frames – the typical phonetic speech segment is 10-100 milliseconds long). The advantage of segmental modeling is that by considering the entire phonetic segment, the statistical models are better able to account for the static and dynamic properties of the individual speech sounds. So, the primary motivation is increased accuracy, but another benefit is significantly reducing the amount of computation needed in the modeling and search phases (by modeling longer, and therefore fewer units in a given amount of input speech).

4. **Phonetic Modeling**: This is the core of modern high performance speech recognition systems. This phonetic modeling is performed by measuring various properties of the speech signal (in a segmental system - a feature vector that captures the important information about an entire phonetic segment) and then modeling the probability distributions of each of the phonetic units in this multi-dimensional feature space. This modeling is performed by using very complex statistic models that are trained on large amounts of previously collected speech data. This training sets the millions of parameters in the statistical model to allow the model to best match the characteristics of each of the speech sounds. This training allows the model to account for all the remaining variability in the speech signal including noise and channel variability which was not normalized out by the signal processing phase, speaker differences, speaking rate, accents, etc. Much of the work on improving speech recognition accuracy is put into improving these models to better match the actual properties of the speech signal.

5. **Search and Match**: The prior stages are able to take a speech signal and produce a
network of possible segmentations, each with associated probabilities of every possible speech sound. The job of the search stage is to compare that with everything the user might have said and find the best match. In general, this would require a huge amount of computation (We now perform tasks with tens of thousands of words. If you allow sentences composed of these, the system must search through all possible combinations of these tens of thousands of words). Fortunately, over the last 20 years, we and others have developed very efficient techniques for performing this search. We have continued to refine these techniques to further reduce the computational requirements and to allow things such as dynamic word additions, dynamic language and semantic models, N-Best output, etc.

Source: Speechworks

PS Note: The Speech Recognition process described above is as implemented by Speechworks Inc. Other speech recognition engines would not necessarily follow the same algorithms/logic but would have a similar analogy built into their recognition engines.

Key Solution Delivery Components (hardware, software and network elements) in a Voice Portal System

A Voice Portal system consists of the following components:

1. Telephony Network

This can be a PSTN (Public Switched Telephony Network), a regular analog line or lines coming through a PBX (Private Board Exchange) system, ISDN (Integrated Services Digital Network) lines or VoIP (Voice over IP) network. The telephony network is connected to the VoiceXML gateway. The telephones can be regular phones or IP (Internet Protocol) phones if connected to the VoIP network.
VoiceXML Gateway (Voice Server)

The VoiceXML browser is the key component that requests VoiceXML documents, interprets them and controls the dialog flow. It also controls speech and telephony resources. These resources include ASR (Automated Speech Recognition), TTS (Text-to-Speech), play/record audio, and telephony network interface.

The telephony component is responsible for all the telephony features like DTMF (Dual Tone Multi-Frequency) extraction and detection, call placing, call transfer and call termination. Typically, a single instance of VoiceXML browser has an instance of the ASR and TTS engine. If there is a need to grab user input, it passes control to the ASR engine. For generating speech output, it passes the request to the TTS engine.

The ASR engine is responsible for recognizing the user utterance and converting it into text, which is forwarded to the VoiceXML browser.

The TTS engine is responsible for generating the speech output from text that is sent to the engine by the VoiceXML browser. If the audio output is prerecorded audio, the VoiceXML browser forwards the raw data to the telephony component.

Web/Application Server

This is typically a Web server that runs the application logic, and may contain a database or interface to an external database or transaction server.

Internet-style Network

This is a TCP/IP (Transport Control protocol/Internet protocol) based packet network that connects the application server and voice server via HTTP.

VoiceXML Documents

These define the voice user interaction and dialog flow control.

Grammar Files

These files define the valid commands that are allowed during the voice interaction. Grammar can be defined at the development stage or generated dynamically at the run time.

Audio Files

These are prerecorded audio files that are played back, or the recordings of the user’s input.

VoiceXML

VoiceXML is a standard technology to make Internet content and information widely accessible via voice and phone. VoiceXML uses speech recognition and touch-tone keypad (DTMF) for input, and pre-recorded audio and text-to-speech synthesis (TTS) for output. It is based on the XML, and leverages the Web paradigm for application development and deployment. VoiceXML is mainly defined for developing voice applications over telephone but can also be extended to other platforms. VoiceXML language provides features for four major components of Voice Web:

- voice dialogs
- platform control
- telephony
- performance

Each VoiceXML document consists of one or more dialogs. The dialog features cover the collection of input, generation of audio output, handling of asynchronous events, performance of client-side scripting and dialog continuation.
Telephony features include simple connection control (call transfer, add 3rd party, call disconnect) and telephony information like Automatic Number Identification (ANI) and Dialed Number Information Service (DNIS).

VoiceXML provides authors close control on performance by providing features for caching and prefetching (fetchhint) the documents and grammar files.

Architecture Diagram

Call flow example

1. A caller uses a phone to access the voice-enabled application.
2. The platform answers the call and assigns it to a speech server running that application.
3. The speech resource sends a request for the associated URL via the Internet or a local area network (LAN) to the customer’s Web server.

Example Call Flow

4. Speech resource then executes the VoiceXML script, passing data and recorded prompts between systems and the caller. Other scripts may be accessed or the call may be transferred to a live agent via IP or PSTN.
Speech Technology in Interactive Voice Response Systems

The most promising application of speech recognition technology is in Interactive Voice response Systems (IVRS). Many companies have already heavily invested in human-powered call centers or DTMF (touch-tone) IVR systems.

Why switch to Speech Technology?

Because it saves money...

- In most cases, the operating costs of a speech recognition system are about 10% of a comparable human-powered call center and half to 1/3 the cost of touch-tone or web systems.
- No operator training is required, and turnover is never an issue.
- Speech recognition powered call centers are available 24 by 7 and automatically expand and contract to handle peak loads and off-hours.
- Shorter call times mean lower long distance charges.
- With some hosting services, one would pay only for actual customer connect time, not for surplus port capacity, hence no minimum charges or guarantees.

Because it improves the customer experience...

- Never busy, never on "hold", even during peak hours.
- Studies show speech recognition applications are far more popular with callers than DTMF ("touch tone hell") menu systems. In some parts of the world 80% of callers encountering a touch-tone menu system will quit or elect to transfer to the operator. With speech recognition that can drop to 20% or below.
- Easier site navigation. No numerical menus, the customer just asks for the product or service they need.
- Shorter call times. The easy navigation and elimination of hold time makes the average call shorter than with either a human operator or a DTMF system.
- Ultrafast service for experienced users. Well-designed "natural language" systems allow experienced users to move very quickly through to their desired service or product. They need never again listen to the phrase "our menu options have changed".
- Help is always available. Properly designed applications allow the user to ask for help - or even ask specific questions - and get a fast, useful answer. The user is never rushed and they are welcome to ask for questions or information to be repeated as often as needed

For Click & Mortar businesses, speech recognition would...

- Provide self-service alternatives for improved customer support
- Increase the opportunity for additional revenue by freeing up CSR's to take more complex ordering calls
- Improve the options and quality of customer support thereby increasing customer retention and durability
- Reduce costs by completely automating simple, repetitive calls
- Cut seconds off complex calls needing a live operator
Glossary:

This glossary would aid in familiarizing with key speech technology terms.

- **Barge-In:**
  Feature allowing caller to interrupt a prompt and still be recognized. For example, if a banking application normally plays a prompt such as "Welcome to the 24-hour banking system. Please listen to the following choices and speak your preference," the caller can interrupt at any time.

  The result might look like the following:

  System: "Welcome to the..."
  Caller: "Checking account, please"
  System: "What is your checking account number?"
  "Barge-in" technology might sound simple, but it's not. The system must distinguish the caller's speech from the line noise, echoes, and even the sound of its own prompt being played. The implementation of this feature is a characteristic of the ASR.

- **Caller**
  The human who is talking to the speech recognition system.

- **Confidence score**
  Measure of how sure the recognizer is about what the caller said. Usually a number between 0 and 999, a very high number indicates high confidence in the answer. If the recognizer is highly confident, then the application can often accept the results and continue. If it is moderately confident, the application might want to confirm. If not confident, the application will reject the utterance and reprompt the user. The confidence score is compared to the confidence threshold to determine the application behavior.

  **Confidence threshold**
  The value of the confidence score below which the utterance is rejected. Usually expressed as a number between 0 and 999. In speaker verification, the value below which the caller is rejected. Adjusting the confidence threshold trades off between false acceptances and false rejections. (In recognition, there might actually be two thresholds, one below which to confirm and one below which to reject).

- **Confirmation**
  Asking the caller if the system understood him or her. This is useful and necessary in several situations:
  - When confidence levels indicate that the system probably, but not definitely, got the right answer.
  - When the caller indicates an irreversible operation. For example, "hang up" or "sell 500 shares."
  - When the system wants to assure the caller that it has understood. For example, the system might say, "I think you wanted to transfer $200. Is that correct?"

- **Continuous speech**
  Refers to naturally spoken words, phrases, and sentences. Continuous speech means speaking fluently, without pausing between each word. Technology that recognizes continuous speech is more advanced than a "discrete" speech recognition system (which does require pauses), but it also requires more tuning to improve accuracy.
Directed dialog
A technique of guiding the caller step-by-step for information. For example:

System: "Welcome to Widgets International. Do you want Orders or Support?"
Caller: "Orders."
System: "Is that a New Order or are you checking an Existing Order?"
Caller: "New Order."
System: "OK, please speak your Customer Number."

A Directed Dialog is helpful to novice users of a system, but can impede experienced callers. For this reason, the ASR allows callers to interrupt prompts and speak complete sentences. An experienced caller might skip parts of the Directed Dialog:

System: "Welcome to Widgets..."
Caller: "New Orders, please."
System: "OK, please speak your Customer Number."

The technique of allowing a caller to use complete sentences to skip parts of the directed dialog is called a natural language shortcut.

Disambiguation
Deciding which of several possible items was intended by the caller. For example:

Caller: "I want to speak to John."
System: "We have two Johns at this office. Do you want John Smith or John Wesley Harding?"

DSP
Abbreviation for "digital signal processor". A special-purpose chip designed for handling real-time tasks like detection for a speech recognizer. ASR’s recognizer can be configured with DSP boards or host-only, depending on the system’s requirements.

DTMF
Dual Tone Multi Frequency, i.e., touch-tone.

Echo cancellation
When a prompt is played, the circuitry within the telephone network echoes it back to the recognizer. An echo cancellation algorithm is required to soften the echo so that the recognizer does not treat the prompt as speech from the caller when the caller barges-in during the prompt. A good algorithm is thus crucial for barge-in to work.

End-of-speech detection
The process of detecting when the caller stops talking. Detecting the end of speech immediately is important so the caller does not experience a delay in response time.

Endpoint
Used to describe both speech detection and end-of-speech detection.

Grammar
The sequences of words allowed by a recognition context. For instance, the grammar for an ASR could allow you to say things like "thirty four dollars and twelve cents" but not "cents dollars forty-four".
Host-only
A recognizer configured to do almost all its computation on the host computer (e.g., a PC running Windows NT or UNIX). Even host-only configurations will often use special-purpose hardware to handle the telephony interface and the real time processing required for speech detection.

Interactive Voice Response (IVR)
A term for any telephone-based application that prompts the inbound caller for information using a recorded or synthesized human voice. Most IVR systems do not allow the caller to respond by voice, instead most IVR input is by DTMF.

Language model
A statistical model that is used in the recognizer to bias it appropriately towards more frequently spoken phrases. This increases overall recognition accuracy.

Mixture Gaussian model
The statistical method used in phonetic classification to represent how each phoneme sounds. Most state-of-the-art recognizers use this method.

Natural Language
Speech-enabled applications in which the call can ask questions or provide information using ordinary sentences.

Neural network
A statistical method used in phonetic classification to represent how each phoneme sounds. It can be used in conjunction with a mixture Gaussian or on its own.

N-best List
A list of best guesses for what a caller said, in the most likely order of correctness.

Phoneme
Basic unit of speech, such as the "b" in "bike" or the "th" in "father". ASR represents words as sequences of phonemes. Unlike older technologies, this allows you to recognize any English word without having a special model for it.

Phonetic classification
The process of determining each segment's phoneme.

Port
The resources required to handle a single call. A system with 24 ports, for example, can handle 24 simultaneous calls. The more ports that a given computer can handle, the more economical the system is to buy and maintain.

Prompt
Speech played to a caller either to ask a question or to provide feedback. For example, "What is your account number?" or "I think you said 'Boston.' or "Please wait while I check for flight information."

A prompt can be played from a pre-recorded file, generated from text, or a combination of the two. Often, a single prompt is created by connecting phrases of speech; for example, "Welcome"."Caroline"."to your personal mailbox." The preceding example joins three phrases; the
advantage is that the first and last phrases can be used for any caller and do not need to be recorded for each person who has a personal mailbox.

- **Recognition context**
  A specific set of vocabularies, speech models, and grammars used when recognizing a specific utterance.

- **Recognition resource**
  A host CPU or DSP card used to handle speech recognition computation. Recognition resources are often shared to save cost and physical space. For example, one Digital Signal Processor (DSP) might handle 8 simultaneous telephone calls (i.e., the calls on 8 ports). Sharing works well because typically only a few of the callers on different calls are speaking simultaneously. The remainder are listening to prompts or waiting for database queries to finish, and therefore not using the recognition resources.

- **Response time**
  The time between when the caller stops speaking and when the recognizer returns a response. The response time depends on several factors, including (a) the recognizer's efficiency, (b) the difficulty of the recognition task, (c) how busy the recognizer is handling speech on other ports, (d) the end-of-speech detector, (e) other computation required to return the response, e.g., querying a database.

- **Retry**
  After a recognition timeout or rejection, the application will often prompt the user again to speak. This is called a retry. A good user interface design will give the caller information about why the retry is needed and sometimes alert the caller to fallback methods. For instance:
  - System: "Please enter your personal ID now."
  - Caller: "I don't know what you mean."
  - System: "Sorry. I didn't understand. Please enter your 12-digit ID now or enter it on your keypad."

- **Segmentation**
  The process of breaking speech into pieces which are then used to perform recognition. There are several methods for segmentation. Some ASR's use an approach where speech is broken into phonemes. Some other vendors break speech into pieces of fixed-duration that are independent of where the phonemic boundaries are.

- **Service**
  A synonym for "application."

- **Speaker Dependent Recognition**
  Speech systems which recognize the speech of a single individual. Such systems may have large vocabularies, but require a lengthy training period.

- **Speaker Independent Recognition**
  Speech systems capable of recognizing the speech of most any caller. Such systems generally have fairly small recognition vocabularies.

- **Speaker verification**
  A system for verifying a caller's identity based on the characteristics of his or her speech.
I **Speech density (also called recognizer duty cycle)**
The percentage of time during a call occupied by the caller’s speech (as opposed to prompts or database queries, etc.). Since recognition resources are only being used when the caller is speaking, applications with high speech density tend to require more resources.

I **Speech detection**
The process of detecting that the caller has stopped listening to the prompt and started speaking. The algorithm must discriminate against background noise.

I **Speech models**
Statistical models of how people speak different phonemes. The system uses speech models to determine the correct phrase from the utterance. Creating models is a challenging aspect of speech technology that requires the sophistication and expertise of a speech scientist.

I **Synonyms**
Two items in a vocabulary, which have the same meaning in a system. For example, nicknames such as Bill, Billy, William, and Will; commands such as purchase and buy or hang up, quit, exit and logoff.

I **Utterance**
A distinct piece of caller speech usually collected in response to a prompt, which is recognized using a specific Recognition context.

I **Vocabulary**
The set of words that can be understood as a part of an application. The grammar determines the sequence of these words that are allowed. For example, both cents and dollars can be in the vocabulary of the Currency Dialog Module, but they can only be said in particular locations in the phrase.
About Wipro Infotech

Wipro Infotech, a division of Wipro Limited (NYSE:WIT), provides enterprise customers with high value Information Technology Products, Services and Solutions in India, and provides Technology & Software Services in the Asia Pacific and the Middle East. Wipro Infotech’s position of leadership in this business is built on a strong foundation of Quality processes, including Six Sigma, ISO, Knowledge Management & People processes (PCMM- People Capability Maturity Matrix). Headquartered at Bangalore, India, Wipro Infotech has business offices in Dubai, Saudi Arabia, Singapore, Sydney, Taipei & Hong Kong in the Asia Pacific & the Middle East.

© Copyright 2002.Wipro Infotech. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted in any form, by any means, electronic, mechanical, photocopying, recording, or otherwise, without express permission from Wipro Infotech. Specifications subject to change without notice. All other trademarks mentioned herein are the properties of their respective owners.

India- Corporate Headquarters

Wipro Infotech
Doddakannelli, Sarjapur Road,
Bangalore-560 035
Tel: 91-80-8440011
Fax: 91-80-8440216

Middle East

Wipro Infotech
Office No.124, Building 1,
First Floor, Dubai Internet City,
PO Box 500119, Dubai,
United Arab Emirates
Phone: 9714-3913480
Fax: 9714-3913482

Asia Pacific

Wipro Limited (Singapore Branch)
22, Malaca Street #06-01
Royal Brothers Building
Singapore -048980
Tel: 65- 5320604
Fax: 65- 5327610

Visit us at: www.wipro.co.in
E-mail: info@wipro.co.in