

## **Introduction**

### **Definition**

In its most generic sense a voice portal can be defined as “speech enabled access to Web based information”. In other words, a voice portal provides telephone users with a natural language interface to access and retrieve Web content. An Internet browser can provide Web access from a computer but not from a telephone. A voice portal is a way to do that.

### **Overview**

The voice portal market is exploding with enormous opportunities for service providers to grow business and revenues. Voice based internet access uses rapidly advancing speech recognition technology to give users any time, anywhere communication and access-the Human Voice- over an office, wireless, or home phone. Here we would describe the various technology factors that are making voice portal the next big opportunity on the web, as well as the various approaches service providers and developers of voice portal solutions can follow to maximize this exciting new market opportunity.

## Why Voice?

Natural speech is modality used when communicating with other people. This makes it easier for a user to learn the operation of voice-activate services. As an output modality, speech has several advantages. First, auditory input does not interfere with visual tasks, such as driving a car. Second, it allows for easy incorporation of sound-based media, such as radio broadcasts, music, and voice-mail messages. Third, advances in TTS (Text To Speech) technology mean text information can be transferred easily to the user. Natural speech also has an advantage as an input modality, allowing for hands-free and eyes-free use. With proper design, voice commands can be created that are easy for a user to remember. These commands do not have to compete for screen space. In addition unlike keyboard-based macros (e.g., ctrl-F7), voice commands can be inherently mnemonic (“call United Airlines”), obviating the necessity for hint cards. Speech can be used to create an interface that is easy to use and requires a minimum of user attention.

## VUI (Voice User Interface)

For a voice portal to function, one of the most important technology we have to include is a good VUI (Voice User Interface). There has been a great deal of development in the field of interaction between human voice and the system. And there are many other fields they have started to get implemented. Like insurance has turned to interactive voice response (IVR) systems to provide telephonic customer self-service, reduce the load on call-center staff, and cut overall service costs. The promise is certainly there, but how well these systems perform-and, ultimately, whether customers leave the system satisfied or frustrated-depends in large part on the user interface.

Many IVR applications use Touch-Tone interfaces-known as DTMF (dual-tone multi-frequency)-in which customers are limited to making selections from a menu. As transactions become more complex, the effectiveness of DTMF systems decreases.

In fact, IVR and speech recognition consultancy Enterprise Integration Group (EIG) reports that customer utilization rates of available DTMF systems in financial services, where transactions are primarily numeric, are as high as 90 percent; in contrast, customers' use of insurers' DTMF systems is less than 40 percent. Enter some more acronyms. Automated speech recognition (ASR) is the engine that drives today's voice user interface (VUI) systems. These let customers break the 'menu barrier' and perform more complex transactions over the phone. "In many cases the increase in self-service when moving from DTMF to speech can be dramatic," said EIG president Rex Stringham.

The best VUI systems are "speaker independent"-they understand naturally spoken dialog regardless of the speaker. And that means not only local accents, but regional dialects, local phrases such as "pop" versus "soda," people who talk fast (you know who you are), and all the various nuances of speech. Those nuances are good for human beings; they allow us to recognize each other by voice. For computers, however, they make the process much more difficult. That's why a handheld or pocket computer still needs a stylus, and why the 'voice dialing' offered by some cell-phone companies still seems high-tech.

Voice recognition is tough. And sophisticated packages not only can recognize a wide variety of speakers, they also allow experienced users to interrupt menu prompts ("barge-in") and can capture compound instructions such as "I'd like to transfer a thousand dollars from checking to savings" in one command rather than several.

These features are designed to not only overcome limitations of DTMF but to increase customer use and acceptance of IVR systems. The hope is that customers will eventually be comfortable telling a machine "I want to add a driver to my Camry's policy." Besides taking some of the load off customer service representatives, VUI vendors promise an attractive ROI to help get these systems into insurers' IT budgets.

ASR systems can be enabled with voice authentication, eliminating the need for PINs and passwords. Call centers themselves will likely transform into units designed to support customers regardless of whether contact comes from a telephone, the Web, e-mail, or a wireless device. At the same time, the 'voice Web' is evolving, where browsers or Wireless Application Protocol (WAP)-enabled devices display information based on what the user vocally asks for. "We're definitely headed toward multi-modal applications," Ehrlich predicts. ASR vendors are working to make sure that VUI evolves to free staff from dealing with voice-related channels; it's better to have them supporting the various modes of service that are just now beginning to emerge.

Good VUI systems have multiple fallback strategies for speech recognition failure, such as asking callers to spell names or breaking a question into parts. The interface could revert to DTMF, especially if ASR was added to an existing DTMF-enabled system. Still, transferring an unintelligible customer to a human agent is frequently an early fallback option in order to minimize customer frustration.

Facilitating this evolution is XML's telephonic flavor, VoiceXML. As its name implies, VoiceXML is a markup language specification that defines the key elements of voice-enabled transactions, which also allows repurposing of this data across any number of platforms and systems

## **Next Generation Network Services**

Today's business environment is more competitive – and complex than ever before. Customer service is the key to success. And demand is growing for powerful new communications services as a strategic way to enhance customer service and build a competitive edge. At the center of these new services is the next-generation network (NGN).

What is Next generation network? It is the next step in the world communications traditionally enabled by three separate networks: the public switched telephone network (PSTN) voice network, the wireless network and the data network (the Internet). NGNs converge all three of these networks-voice, wireless and internet- into a common a common packet infrastructure. This intelligent, highly efficient applications and service opportunities.

Three types of services drive NGNs: real-time and non real-time communication services, content services, and transaction services. The services-driven NGN gives service providers greater control, security, and reliability while reducing their operating costs. Service providers can quickly and costs effectively build new revenue.

Built on open modular elements, standard protocols, and open interfaces, the NGN caters to the specific needs of all users - enterprises, remote offices, telecommuters, and small office / home office (SOHO) customers. It unites traditional wireline and wireless voice, video, and data using a packet-based transport. The new class of services it enables is more flexible, scalable, and costs efficient than services that have been offered in the past.

One key NGN service is voice portal, which provide callers with anywhere, anytime access to information like news, weather, stock quotes, and account balances using simple voice commands and any telephone. Voice portals are poised to become the next big thing in communications. The organizations using them will have a very real edge in the market place-differentiating themselves from their competitors, attracting loyal customers and growing their revenues.

Voice based service for the next-generation network; Intel Corporation has the hardware, software and services for the next

generation network solutions, including a complete voice portal platform.

### **Next-generation Network Benefits**

- More choices-Open systems promote multiple-vendor participation in the marketplace.
- Lower cost-Solutions built on open standards cut development time for solution providers.
- Innovative Services- Working to standards frees developers to concentrate on adding value.
- Lower Risk- Compatibility with products and technologies from many suppliers decreases the risk of ownership.

### **Deploying NGN Services**

Next-generation communications solutions are built with open, flexible, standards-based building blocks. The reasons are clear- with open solutions, there's no need to start from scratch to add next-generation voice enabled E-business services. Using modular building blocks makes it easy and affordable to add new features, services, and value to existing systems. It all adds up to powerful, affordable solutions that protect your- and your customers – investments.

The NGN begins with media servers, which provide advanced media-processing capabilities. The value of a media server is its flexibility for supporting advanced media-processing services like basic voice announcements, interactive voice response (IVR), conferencing, messaging, text-to-speech (TTS), and speech recognition. Built with open, standard computing and voice-processing boards, media servers can be deployed many ways. For example, a voice portal platform is a media server that provides a speech driven user interface with simple speech recognition and TTS capability, providing voice-client access to Internet content, messages, or both.

## Voice Recognition and Authorization

Speech recognition and speaker authentication are often thought of as the same thing, but that this is a myth. Speech recognition's focus is on what is being said, and can not determine who is speaking. Speaker authentication's focus is on the speaker –“and does not know what is being said.”

Both, however, face similar hurdles. Markowitz said a handy rule is that "if it is a problem for speaker authentication, it is difficult for speech-recognition.”

Markowitz dispelled a second myth: that voice biometrics "can not resist tape recorders." But in fact, a challenge-response program for released criminals actually automates "which one of the questions it will ask you to repeat, so you never know what it is going to ask you say!" The challenge-response system phones home-arrestees, who are sentenced to house-arrest, and asks them to say something over the phone that they would not and could not know in advance - thus not being able to be duped by a pre-recorded taped response.

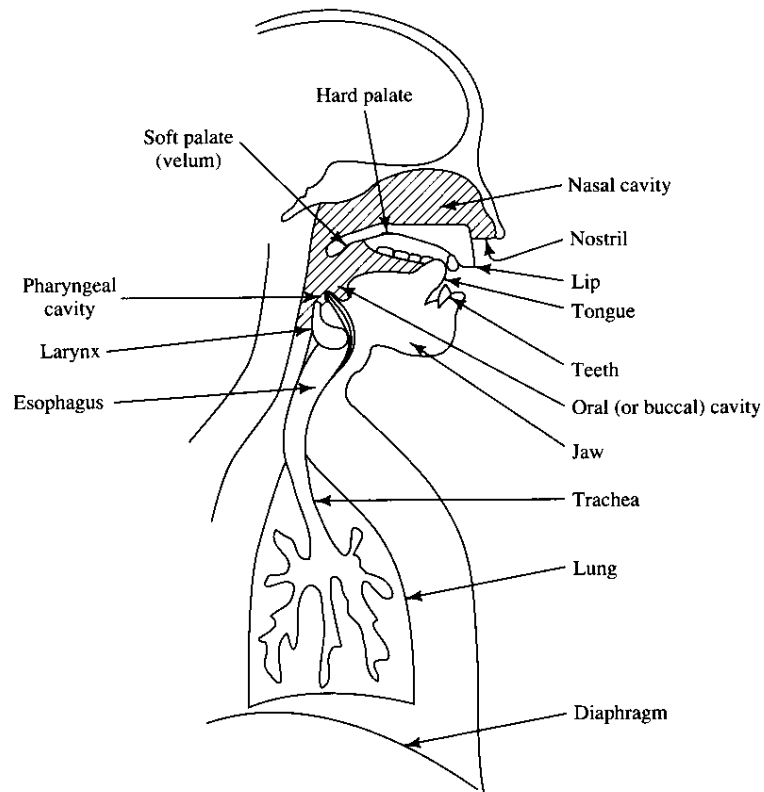
A third myth Markowitz dispels is that speaker authentication is not as accurate as other biometrics. "Voice is doing pretty darn well compared to other biometrics." She reminds us that "no form of security is 100 percent secure - but there are ways to combine belts and suspenders, and blend in partnership different kinds of techniques to create multi-level security." Here we would be talking only about the speaker verification that is speaker authentication and not speech recognition.

## Speaker Verification

The speaker-specific characteristics of speech are due to differences in physiological and behavioral aspects of the speech production system in humans. The main physiological aspect of the human speech production system is the vocal tract shape. The vocal tract is generally considered as the speech production organ above the vocal folds, which consists of the following: (i) laryngeal pharynx (beneath the epiglottis), (ii) oral pharynx (behind the tongue, between the epiglottis and velum), (iii) oral cavity (forward of the velum and bounded by the lips, tongue, and palate), (iv) nasal pharynx (above the velum, rear end

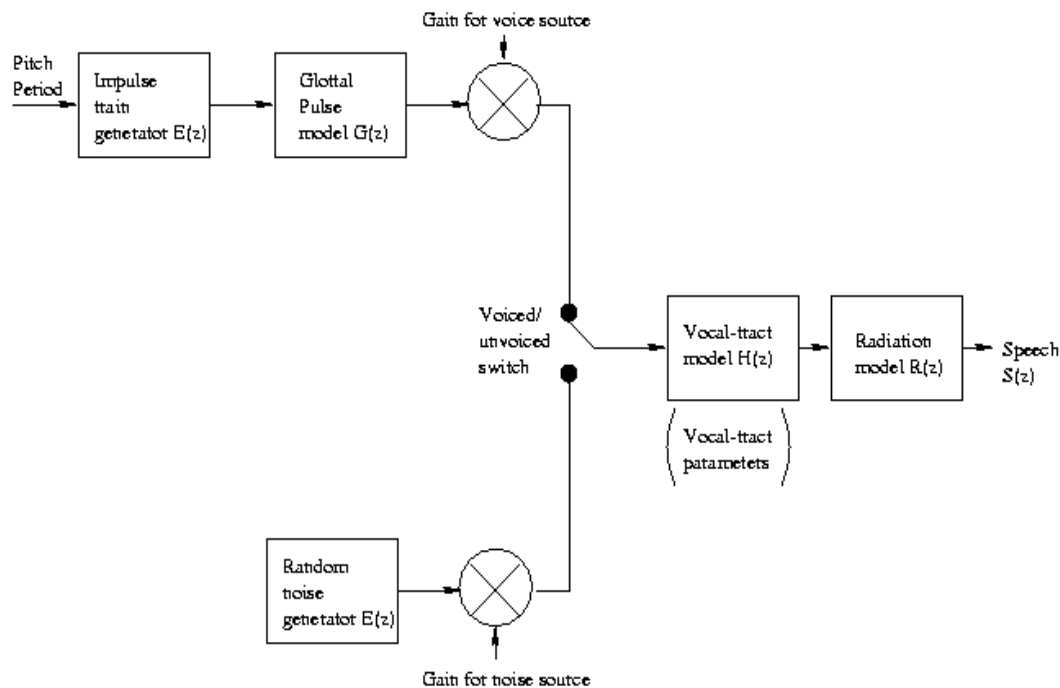
of nasal cavity), and (v) nasal cavity (above the palate and extending from the pharynx to the nostrils). The shaded area in figure 1 depicts the vocal tract.

Figure 1:



The vocal tract modifies the spectral content of an acoustic wave as it passes through it, thereby producing speech. Hence, it is common in speaker verification systems to make use of features derived only from the vocal tract. In order to characterize the features of the vocal tract, the human speech production mechanism is represented as a discrete-time system of the form depicted in figure 2.

Figure 2:



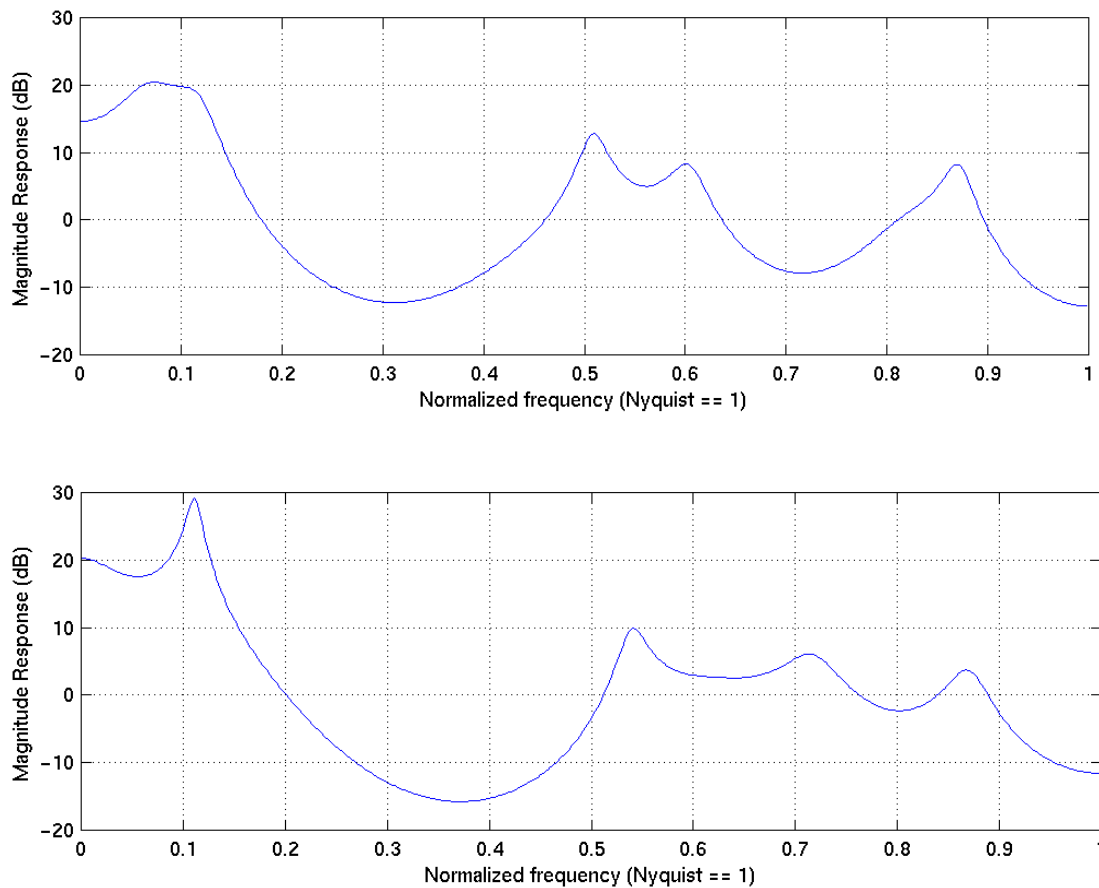
The acoustic wave is produced when the airflow from the lungs is carried by the trachea through the vocal folds. This source of excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these. Phonated excitation occurs when the airflow is modulated by the vocal folds. Whispered excitation is produced by airflow rushing through a small triangular opening between the arytenoid cartilage at the rear of the nearly closed vocal folds. Frication excitation is produced by constrictions in the vocal tract. Compression excitation results from releasing a completely closed and pressurized vocal tract. Vibration excitation is caused by air being forced through a closure other than the vocal folds, especially at the tongue. Speech produced by phonated excitation is called voiced, that produced by phonated excitation plus frication is called mixed voiced, and that produced by other types of excitation is called unvoiced.

It is possible to represent the vocal-tract in a parametric form as the transfer function  $H(z)$ . In order to estimate the parameters of  $H(z)$  from the observed speech waveform, it is necessary to assume some form for  $H(z)$ . Ideally, the transfer function should contain poles as well as zeros. However, if only the voiced regions of speech are used then an all-pole

model for  $H(z)$  is sufficient. Furthermore, linear prediction analysis can be used to efficiently estimate the parameters of an all-pole model. Finally, it can also be noted that the all-pole model is the minimum-phase part of the true model and has an identical magnitude spectra, which contains the bulk of the speaker-dependent information.

The above discussion also underlines the text-dependent nature of the vocal-tract models. Since the model is derived from the observed speech, it is dependent on the speech. Figure 3 illustrates the differences in the models for two speakers saying the same vowel.

Figure 3:



## **Choice of features:**

The LPC features were very popular in the early speech-recognition and speaker-verification systems. However, comparison of two LPC feature vectors requires the use of computationally expensive similarity measures such as the Itakura-Saito distance and hence LPC features are unsuitable for use in real-time systems. Furui suggested the use of the cepstrum, defined as the inverse Fourier transform of the logarithm of the magnitude spectrum, in speech-recognition applications. The use of the cepstrum allows for the similarity between two cepstral feature vectors to be computed as a simple Euclidean distance. Furthermore, Atal has demonstrated that the cepstrum derived from the LPC features results in the best performance in terms of FAR and FRR for a speaker verification system. Consequently, we have decided to use the LPC derived cepstrum for our speaker verification system.

## **Speaker Modeling :**

Using cepstral analysis as described in the previous section, an utterance may be represented as a sequence of feature vectors. Utterances spoken by the same person but at different times result in similar yet a different sequence of feature vectors. The purpose of voice modeling is to build a model that captures these variations in the extracted set of features. There are two types of models that have been used extensively in speaker verification and speech recognition systems: stochastic models and template models. The stochastic model treats the speech production process as a parametric random process and assumes that the parameters of the underlying stochastic process can be estimated in a precise, well defined manner. The template model attempts to model the speech production process in a non-parametric manner by retaining a number of sequences of feature vectors derived from multiple utterances of the same word by the same person. Template models dominated early work in speaker verification and speech recognition because the template model is intuitively more reasonable. However, recent work in stochastic models has demonstrated that these models are more flexible and hence allow for better modeling of the speech production process. A very popular stochastic model for modeling the speech production process is the Hidden Markov Model (HMM). HMMs are extensions to the conventional Markov models, wherein the observations are a probabilistic function of the state, i.e., the model is a doubly embedded stochastic process where the underlying stochastic process is not directly observable (it is hidden). The HMM can only be viewed through another set of stochastic processes that produce the sequence of observations.

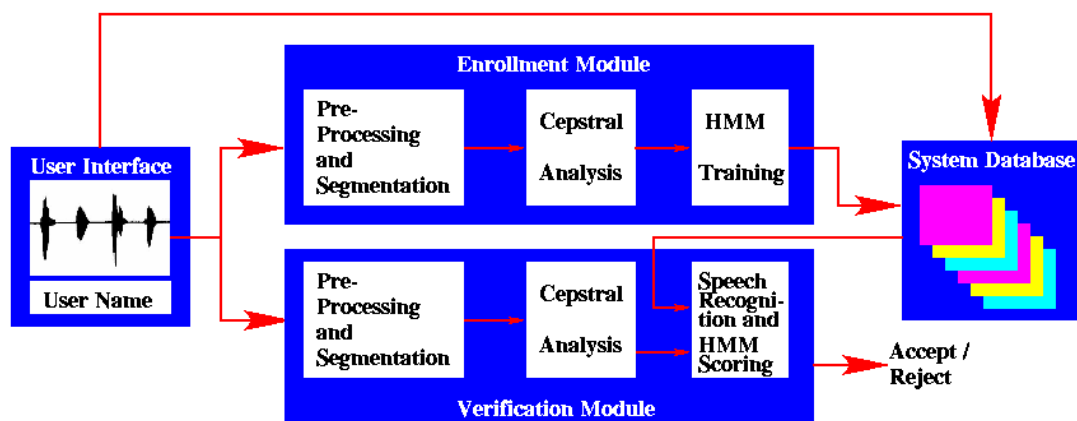
Thus, the HMM is a finite-state machine, where a probability density function  $p(x | s_i)$  is associated with each state  $s_i$ . The states are connected by a transition network, where the state transition probabilities are  $a_{ij} = p(s_i | s_j)$ . A fully connected three-state HMM is depicted in figure 4.

For speech signals, another type of HMM, called a left-right model or a Bakis model, is found to be more useful. A left-right model has the property that as time increases, the state index increases (or stays the same)-- that is the system states proceed from left to right. Since the properties of a speech signal change over time in a successive manner, this model is very well suited for modeling the speech production process.

### Pattern Matching:

The pattern matching process involves the comparison of a given set of input feature vectors against the speaker model for the claimed identity and computing a matching score. For the Hidden Markov models discussed above, the matching score is the probability that a given set of feature vectors was generated by the model.

### A Speaker Verification System:



## Speech Recognition Challenges

Bell labs designed the first speech-recognition system in the early 1950s. the technology was amazingly accurate at determining single digit spoken numbers. By early 1970s natural language speech understanding was demonstrated by Terry Winograd's SHRDLU system, a robot that understood commands such as "move the red block on top of the small green one." By the 1990s speech recognition was being applied to internet. In 1996 California State University at Northridge was able to demonstrate an experimental voice controlled web interface. By 2000 many voice driven portals were being developed.

Until recently, a number of issues made it almost impossible to develop a speech-recognition engine that would recognize fluent and natural speech. The basic challenges faced by voice application developers included the following:

1. Variability of speech patterns: Different people speak the same language differently and even speak the same word in many different ways. Interpreting speech variability has led to the development of complex pattern analysis. Understanding natural pauses, speaking rates and changes in volumes has been complex and difficult.
2. Processing power: In the mid-1980s, a new technique known as Hidden Markov's Models improved the ability to recognize word relationships. This computation intensive technology eventually led to powerful speech recognition applications. Achieving real time voice recognition found in voice portals requires processing power that is not commercially viable until recently.
3. Extracting meaning: Very few speech Recognition applications are able to accurately determine the meaning of words. The quality of speech interpretation depends on the ability of the speech recognition engine to properly choose the best match for spoken sounds from its list of expected text phrases. A more advanced process was required to extract meaning from those words. Because of the many possible ways that people speak and the many words that are used to communicate the same concept, a full understanding of human factors was critical to properly interpret the meaning.

4. Background Noise: Before mobile phone users often access voice portals, background has been difficult for voice recognition developers to filter out. The development of better microphones has helped, but issues such as wind, murmurs and music have made it a challenge to properly isolate voice from noise.
5. Continuous speech recognition: Designing systems that are powerful enough to understand and respond to continuous speech requires a large amount of processing power that was not available at reasonable cost in the past. When a person speaks at a natural rate, it has been difficult to distinguish which sounds were associated with specific words. For example the phrase “to recognize speech” could have been misunderstood to be “to wreck a nice beech.” Users do not naturally speak with pauses between words. As a result processing phrases in real time as they are naturally spoken has been a major challenge.

Many of the solutions to problem associated to speech recognition are still being fine tuned. Powerful computers have provided the processing power to overcome many of the limitations of the past. However the most common speech-recognition systems of today are still very different from the way in which people naturally interact.

## **Conclusion.**

Speech-enabled Internet portals or voice portals are quickly becoming the hottest trend in E-commerce – broadening the access to internet content to everyone with the most universal communications device of all, a telephone. Voice portals put all kinds of information at a customer's fingertips anytime, anywhere. Customers just dial into the voice portal's prescribed number and they use simple voice commands to access whatever information they need. It's quick, easy and effective, even from a car or the airport.

The potential for voice portals is as wide as the reach of telephones, which today number 1.3 billion around the world. Compare that to the 250 million computers with Internet access and it is easy to understand while analysts believe voice-enabled web access will take off. Feasibility studies exhibit that by 2005, 45 million wireless users around the world will regularly use voice portals to handle their everyday cyber chores.

Today's voice portals are just the tip of the iceberg-the first step in changing the way people access Internet content and, ultimately, how businesses and consumers will conduct business over the Internet. Over the next few years, voice portals-and core technologies behind them are poised to businesses view and Internet with their customers.

Voice portals are changing the telephone interaction from a vendor-centric to a customer-centric experience-increasing satisfaction for customers while improving efficiency and cutting costs for businesses. A voice portal provides telephone users with a natural language interface to access and retrieve web content. An Internet browser can provide web access from a computer but not from a telephone. A voice portal is a way to do that. Of course simple access and retrieval of information is just the beginning. A voice portal can also provide users access to virtual personal assistance and web-based unified messaging applications

Voice portals can also cut operating expenses by freeing up agent time and replacing human operators with an easy-to-use automated solution. They also provide new revenue opportunities by opening up the

possibility of new subscription services or building revenue through advertising.

Voice portals are the next frontier in convergence, the intersection of the Internet And Telecommunications, blurring the distinctions among voice and data,

Computers and telephones and like any frontier, the rewards are great for Staking your claim early.

The voice portal reference system is a packaged, integrated hardware and

Software reference system for building hardened E-Business and speech-enabled voice portal solutions. Combining the power of server technology with telephony interface boards in an integrated server reference system, it is embraced by leading speech-technology providers