

**Recognize the Difference:
Why Telephony Boards Matter
in Speech Applications**

Keith Byerly

Senior Market Development Manager



table of contents

Abstract	3
Introduction	3
Improving the Performance of Speech Systems	4
Performance Roadblocks in Speech Systems	5
Improving Client Performance: How Telephony Boards Make a Difference	5
Voice Activity Detection (VAD)	6
Putting Telephony Boards to the Test: Measuring the Results	7
– Client System	7
– Client CPU Utilization	7
– PCI Bus Loading	8
Designing a Speech-Enabled VAD	9
– Multimodal Operation	9
– Initialization	9
– Windowing	9
– Adjustable Detection Threshold	9
Echo Cancellation	9
Conclusion: Telephony Boards Do Matter	11
Appendix 1: Emerging Speech Standards: VoiceXML and SALT	11
What are VoiceXML and SALT?	12
Why Web-Based Standards?	13
Issues and Alternatives	13
Cantata and Speech Standards	13

abstract

The successful deployment of speech-driven applications requires careful attention to the design of the entire system, not just the speech user interface and application software. Network capabilities and configuration can determine call transfer capabilities and the choice of application integration approaches will define the system architecture and development environment for years to come. But one critical decision that developers need to consider—and that often gets overlooked—is choosing the right telephony board.

The right telephony board can make the difference between slow, inaccurate speech applications with frustrated users versus ones that deliver a satisfying user experience. It can also mean the difference between an oversized, costly system for client recognition versus a smaller, optimally sized system that works with extreme efficiency.

This white paper describes how the latest generation of advanced telephony boards have been specifically designed to improve accuracy, scalability, system responsiveness and performance to improve the development of optimized speech systems that meet the high expectations of today's demanding users.

introduction

Choosing the right telephony board when developing a speech system can mean the difference between problematic installations with low accuracy and frustrated users versus ones that deliver a satisfying user experience with high accuracy. It can also mean the difference between an oversized, costly system for client recognition versus a smaller, optimally sized system that works with extreme efficiency. The latest generation of advanced telephony boards can improve accuracy, scalability, system responsiveness and performance to improve the development of optimized speech systems that meet the high expectations of today's demanding users.

A common problem in speech processing is voice activity detection (VAD)—detecting the presence or absence of a voice, and identifying when a user begins or ends speaking. With typical telephony boards, a simple energy-based threshold is used for this purpose due to CPU and memory limitations. However, these tend to be overly sensitive to background noise and non-speech sounds from the user, and so must be disabled for use with speech applications in order to avoid false barge-in. This is where the speech-enabled VAD of advanced telephony boards can really make a difference in system performance, by freeing up the host system resources and increasing the system's overall call capacity.

Systems without speech-enabled VAD send a continuous audio stream to the client CPU, which then must process the entire signal, including unnecessary sounds such as silence and noise. And because the VAD can be integrated with on-board echo cancellation, residual echo can be eliminated as well. By using a telephony board with speech-enabled VAD, the vast majority of non-speech audio (echo, silence and noise) is filtered out by the board, allowing the CPU to process only the user's voice.

During a typical IVR session, the user speaks for a small percentage of the total time; the majority of the session typically consists of prompts being played, silent pauses while the user considers their options, and other non-speech events. Telephony boards with speech-enabled VAD, which filter out most of the unnecessary audio, allow more simultaneous calls to be processed by the host, resulting in increased system capacity. In a recent test (detailed in this white paper), the speech-enabled VAD reduced the number of audio samples transferred to the speech server by 84%, which resulted in a 30% reduction in host CPU load. For developers, this means that those additional CPU resources can be used to improve recognition accuracy or run additional applications (such as VoiceXML or SALT gateways) on the host system.

Echo cancellation is another key factor in the performance of any speech application that supports barge-in. For example, when a speech system plays a prompt (“What city please?”) and the user responds (“Austin”), the prompt echo will mix with the user’s voice, affecting the speech recognition server’s ability to recognize the response. This can trigger an inaccurate recognition (“Boston” instead of “Austin”), which ultimately results in a frustrated user. By using advanced telephony boards with on-board long-tail echo cancellation, developers can help eliminate this problem because the telephony board can preprocess the audio signal and eliminate echo before it is sent to the speech system, resulting in more accurate recognition of user responses. This can also save money by eliminating the need for an external echo cancellation system.

By selectively eliminating noise, silence and echo from the audio stream, the speech-enabled VAD and long-tail echo cancellation capabilities of advanced telephony boards can clearly improve both recognition accuracy and overall system performance. Cantata uses the term “audio scrubbing” to describe this combination of functions. Developers need to recognize these differences when selecting telephony boards for their speech applications.

improving the performance of speech systems

As service providers and enterprises consider speech-enabling their mission-critical applications, they must be convinced not only that they can achieve a rapid Return on Investment (ROI), but also that they can deploy reliable, cost-effective and scalable systems.

The fact that speech systems are typically based on industry-standard servers addresses many of these concerns. Robust, state-of-the-art PCI and CompactPCI servers are offered by a variety of manufacturers at very competitive prices. These systems benefit from the huge industry investment in constantly advancing the underlying technology; faster processors, multiprocessor configurations, and highly evolved operating systems all contribute to the constantly improving performance of speech systems.

System architectures have also evolved to address the demanding requirements of deployable speech solutions. Distributed, IP-based client/server architectures have become the norm, allowing different components of the speech solution—network gateway, speech application servers, and back-end data repositories—to scale independently as appropriate for a particular application. The emergence of web-based application standards such as VoiceXML and SALT (*see Appendix 1*) leverages the advantages of this architecture within the application domain, simplifying application design and maintenance, and facilitating the use of best-of-breed technology.

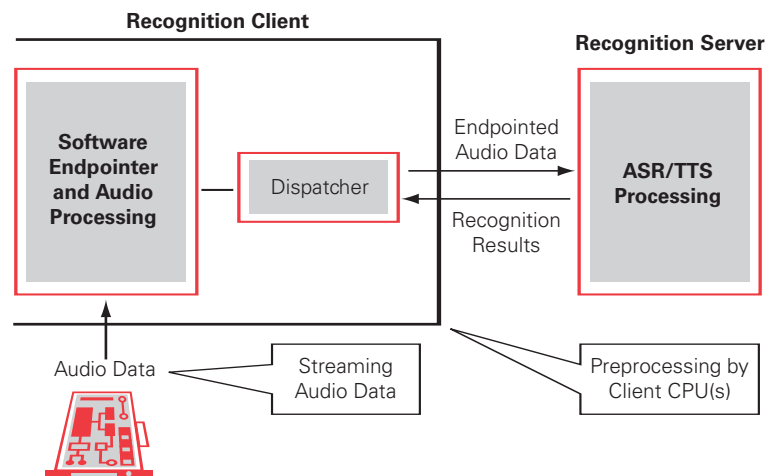
Although many of the underlying system components have become standardized, those related directly to speech processing continue to evolve rapidly as innovation by individual vendors exceeds the limited scope of existing standards. Several vendors now offer proprietary versions of the latest generation of speech recognition software, delivering “natural language” capabilities that were only a dream five years ago. Text-to-Speech (TTS) vendors offer increasingly realistic-sounding synthesized voices, now extending towards a broad range of regional languages and dialects. Advanced voice verification and audio mining techniques are gaining traction as security, monitoring and surveillance needs have become more acute. Even telephony hardware, considered merely a “necessary evil” by many speech developers, now incorporates innovative technology specifically designed to improve the performance and scalability of speech systems and applications.

performance roadblocks in speech systems

The potential performance of a speech system is ultimately defined by its system configuration, which effectively determines which system component of the solution will become the performance bottleneck. An “all-in-one” system configuration, in which speech recognition and call termination are performed on the same host system, is relatively simple, efficient and responsive. However, the processing and memory available on a single host system limits the maximum number of channels that it can support. This can be an issue for systems handling high traffic volumes, applications using very large vocabularies, or deployments based on low-cost server hardware.

In a client-server configuration (see Figure 1), line termination and speech processing are effectively decoupled, allowing each to scale independently. The client typically runs on the system where the calls terminate, sending preprocessed speech data over a network to one or more servers where the actual speech recognition is performed. This architecture requires careful system engineering to avoid potential latency issues that would impact responsiveness during barge-in, and to efficiently manage the distributed system resources. This networked approach simplifies the addition of recognition servers, and the client system ultimately becomes the limiting factor in determining speech system scalability and performance.

Figure 1: Client-Server Configuration



improving client performance: how telephony boards make a difference

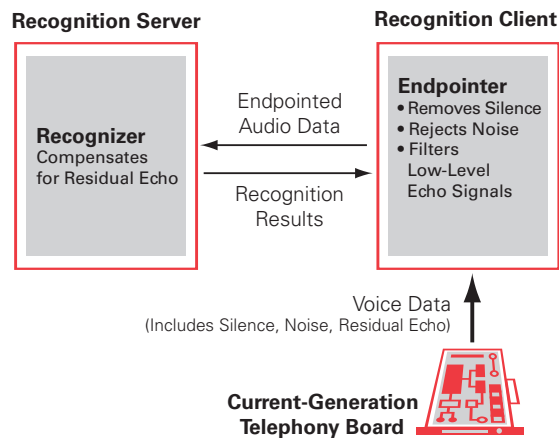
In a typical client-server speech system, the client provides connectivity to the telephone network and performs preprocessing of the audio signal before it is sent to the recognition server(s). This preprocessing reduces the echo caused by outgoing prompts, and performs “endpointing” so that only voice signals—not signals containing just residual echo, background noise, or silence—are transmitted over the network for detailed analysis by the expensive server resources. The DSP resources on the telephony board typically perform the echo cancellation. Although it’s critical that the vendor-specific client software perform endpointing to ensure recognition accuracy, a properly implemented first-pass endpointer on the board can significantly reduce both host and bus loading, improving client performance and scalability.

voice activity detection (VAD)

Although board-based VAD functionality has been available for some time, it has until recently been incompatible with speech systems. In comparison with the specialized endpointer of the speech client, these detectors typically use simpler, energy-based threshold detection algorithms, are constrained by CPU and memory limitations, and tend to be overly sensitive to background noise and non-speech sounds from the caller. Because they will generally trigger on voice as well as any similar sounds, board-based VADs cannot be relied upon to detect barge-in. However, when properly designed and configured, a board-based VAD can be used as a first-pass endpointer for the client system, so that the host CPU will not need to process audio containing only echo, silence or noise.

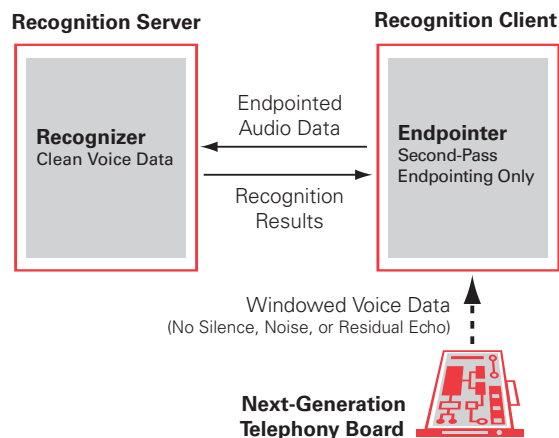
As shown in Figure 2 below, telephony boards without a speech-enabled VAD send a continuous audio stream to the recognition client throughout the call. In this case, the audio streams from the active channels represent a continuous, static load that must be transmitted by the host bus and processed by the host CPU.

Figure 2: Without VAD



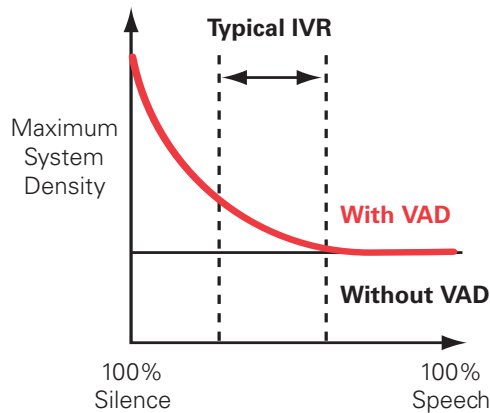
However, as shown in Figure 3, a speech-compatible VAD essentially filters out most non-speech audio from each channel. Although the board-based VAD will generally pass along some amount of extraneous audio samples, it will eliminate the vast majority of those consisting only of noise or silence—and because the VAD can be integrated with the onboard echo cancellation, residual echo can be eliminated as well.

Figure 3: With VAD for First-Pass Endpointing



Because the amount of audio data processed by the client is no longer a function of the number of active channels, maximum system channel density can be increased. During a typical IVR session, the user speaks (and speech recognition is performed) for a small percentage of the total session time; the majority of the session typically consists of prompts being played, silent pauses while the user considers their options, and other non-speech events. As you can see in Figure 4, this allows more calls to be handled on the client simultaneously.

Figure 4: The Effects of Dynamic Loading



putting telephony boards to the test: measuring the results

In order to validate the impact of a speech-enabled VAD on client system performance, a test program was created to measure changes in CPU utilization and bus traffic on a representative system (see client system configuration below). The test program answers the line, plays one of several prompts, and performs speech recognition on one of several recorded files using a static grammar.

Although there are many variables that can affect system performance, this test was designed to isolate the specific effects of an onboard VAD. Confidence scores were slightly improved with the VAD enabled, indicating that this improved performance was achieved without harming recognition accuracy.

Client System:

Processor: Pentium III 866 MHz (TYAN S1854 Trinity 400)

O/S: Microsoft Windows 2000 Professional (5.0.2195)

RAM: 512MB (total)

Software: SpeechWorks® OpenSpeech Recognizer™ 1.1.3; Brooktrout Bfv API Rel. 3.0.3

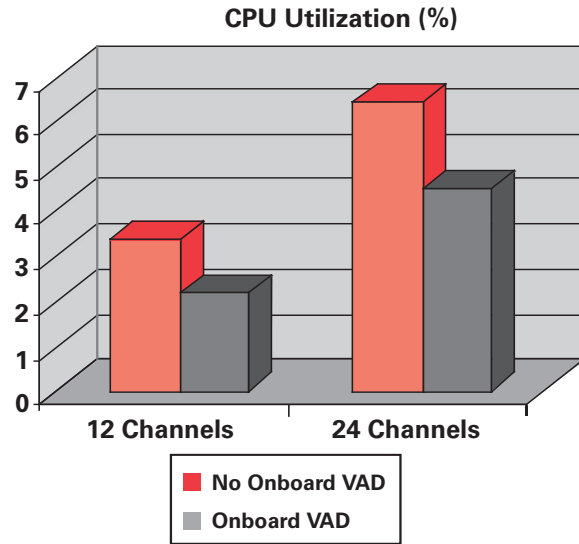
Telephony: Brooktrout TR1000+P24vH-T1

Client CPU Utilization

These results demonstrate a 30%+ reduction in host CPU loading; the client performs endpointing only on audio passed from the onboard VAD, reducing wasted processing time. This can improve system performance because additional CPU resources can be applied to improving recognition accuracy, or running additional applications (*such as VoiceXML or SALT gateway*) on the host system.

Percentage CPU Utilization

	12 channels	24 channels
No Onboard VAD	3.37	6.45
Onboard VAD	2.21	4.50
Improvement	34%	30%



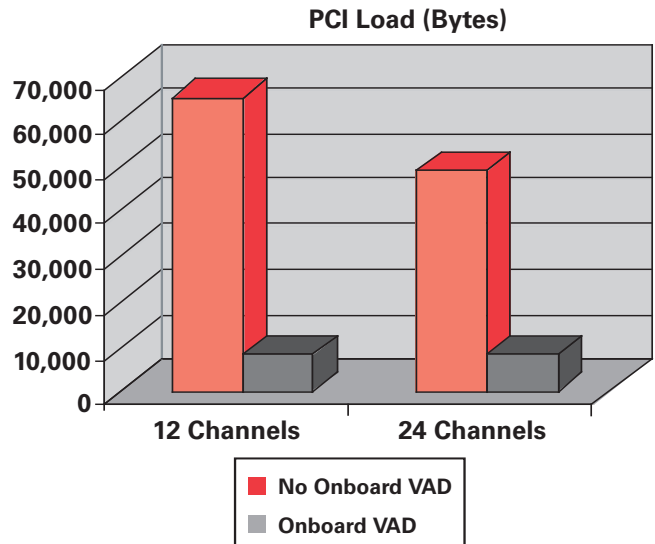
PCI Bus Loading

As expected, the speech-enabled VAD reduced the number of audio samples transferred to the client by 84%+. Although not an issue for smaller systems, streaming hundreds of full-duplex 64 kbps audio channels represents a significant load on PCI-based systems.

Note that the total number of bytes transferred in the 24 channel test was less than in the 12 channel test due to differences in prompts and scripting. By using telephony boards with speech-enabled VAD, most of the unnecessary audio is filtered out and more simultaneous calls can be processed by the host, resulting in increased system capacity.

Bytes Transferred

	12 channels	24 channels
No Onboard VAD	65,254	49,012
Onboard VAD	8,193	8,052
Improvement	87%	84%



designing a speech-enabled VAD

In order to provide compatibility with speech systems, several specific features must be incorporated into the design. These are needed to provide flexible operating modes, ensure unaffected operation of the client endpointer, compensate for the limitations of the onboard VAD algorithm, and allow for varying levels of background noise.

Multimodal Operation

To provide the functionality needed to address a variety of applications, the operating modes of the VAD should be configurable in real time. For example:

- Mode 0: VAD is disabled; a continuous audio signal stream is sent to the host
- Mode 1: VAD is enabled; voice start/finish events and continuous audio signal stream is sent to the host
- Mode 2: VAD is enabled; voice start/finish events and windowed audio signal stream is sent to the host
- Mode 3: VAD is enabled; voice start/finish events and windowed audio signal stream is sent to the host; board stops prompt immediately when barge-in detected by onboard VAD (for faster prompt cutoff)
- Mode 4: VAD is enabled; voice start/finish events and windowed audio signal stream is sent to the host; VAD threshold adjustable by application

Initialization

At the start of each call, the client endpointer will typically need to set an initial noise threshold based on an estimate of the background noise. The onboard VAD must pass audio continuously for approximately one second at the start of the call to allow the software to adapt its noise threshold properly.

Windowing

An energy-based VAD will generally be effective at detecting spoken vowels, but may often miss the beginning or end of utterances that terminate with relatively short, soft high-frequency consonants, resulting in misrecognition (“Austin” vs. “Boston”). A back-up buffer, typically containing audio samples for the 100–300 msec. prior to the detection of speech by the VAD, makes sure that the weak onset of speech is preserved. A similar hold buffer, approximately 1 second of audio samples following the end of detected speech, allows the client endpointer to adapt to changing levels of background noise.

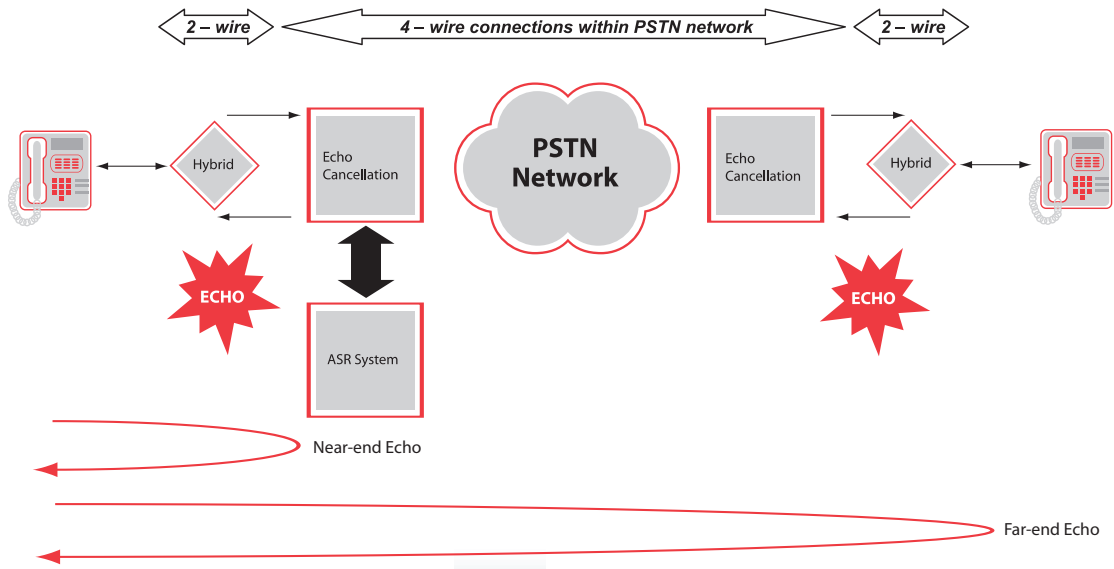
Adjustable Detection Threshold

Because of the wide variability in the levels of background noise, it’s important to allow application control over the VAD detection threshold. In general, the speech client’s endpointer will be more effective at discriminating between speech and noise, especially as background noise levels increase.

echo cancellation

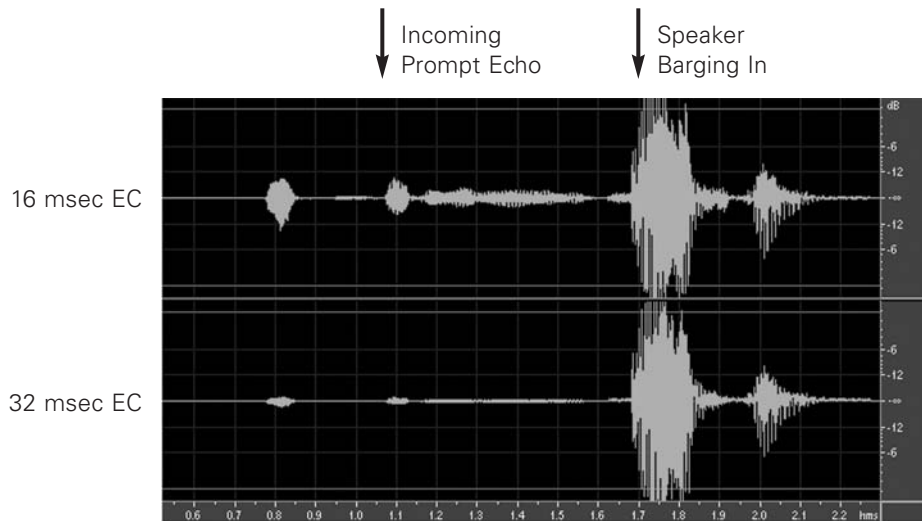
Echo cancellation is a key function of any speech application. Hybrid, or line echo is the primary source of echo over the PSTN. Just as you can see your reflection in a pool of water, the impedance mismatch between the two-wire local loop and the four-wire network causes a reflection of the outgoing signal. Components called “hybrids” are used to join the local loops and network connections; as you can see from Figure 5 below, near-end echo results from reflections from the closest of these hybrids, while far-end echo results from the hybrid at the other end of the network connection. Because round-trip propagation delay can be significant, far-end echo is the primary echo problem to be considered.

Figure 5: Hybrid Echo in the PSTN Network



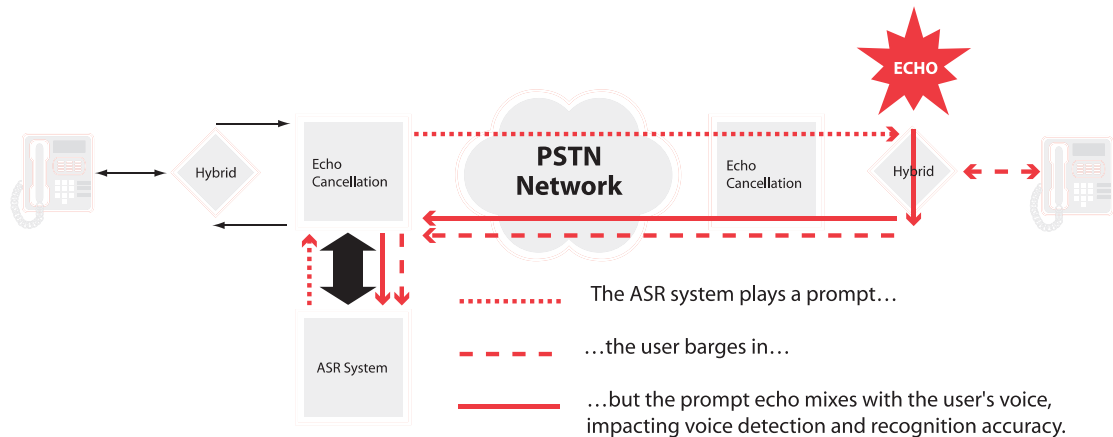
Although some echo cancellation is performed within the network, the remaining echo delay must be handled by echo cancellation on the telephony board. In most cases, cancellation of 12–16 msec. delays is adequate, although some connections may introduce 30 msec. or more of delay. If the tail length of the echo canceller is insufficient for the combined delay and impulse response of the echo path, echo cancellation will be compromised. Figure 6 shows an example where a 16 msec. echo canceller cannot effectively filter an echo signal that has a 28 msec. delay.

Figure 6: Long-Tail Echo Cancellation



Speech systems need effective echo cancellation to perform accurate recognition during barge-in. The unique nature of typical interactive speech applications provides special challenges and considerations.

Figure 7: Echo Cancellation in an ASR System



First, speech systems are particularly susceptible to echo. Unlike human users, who find some echo actually improves the sound, speech systems work best with echo eliminated. Echo cancellation in systems designed without speech in mind may not provide sufficient cancellation performance for accurate recognition under all circumstances. In some cases, expensive external echo cancellation systems must be added when echo-related recognition problems are discovered during system tuning.

Second, because prompts are typically human voices (or synthesized versions of them), their echo sounds like a person speaking softly into the speech system. This may trigger a false recognition, or may combine with the user's voice to cause misrecognition—e.g., "Austin" could be recognized as "Boston" in an IVR application.

Echo cancellation problems in speech systems are not mere annoyances; they can destroy the accuracy, usability and value of a speech system. Advanced telephony boards offering the option of long-tail 32 msec. echo cancellation ensure effective barge-in performance under nearly all real-world conditions.

conclusion: telephony boards do matter

Today's end users demand speech systems that are accurate, fast and easy to use. Choosing the right telephony board can make a difference in the performance and scalability of speech applications.

By providing audio-scrubbing—innovative voice activity detection and long-tail echo cancellation—today's next-generation telephony boards can free up limited client resources that can be used to run value-added applications, support additional speech channels on a given hardware platform, or reduce the cost to deploy a desired number of channels.

appendix 1: emerging speech standards: voiceXML and SALT

During the recent dot-com boom, Web-based architectures and standards expanded rapidly into every area of information technology. From its very beginning, the Web's logical structure and protocols were open and grew rapidly into extensible standards, supporting widespread innovation that enhanced its power, scope and

usability. At the same time, the ongoing evolution in speech recognition algorithms and processing power finally made speech technology accurate and cost-effective enough for real-world applications. At the convergence of these two trends are two web-based speech standards: the Voice Extensible Markup Language (VoiceXML) and Speech Application Language Tags (SALT).

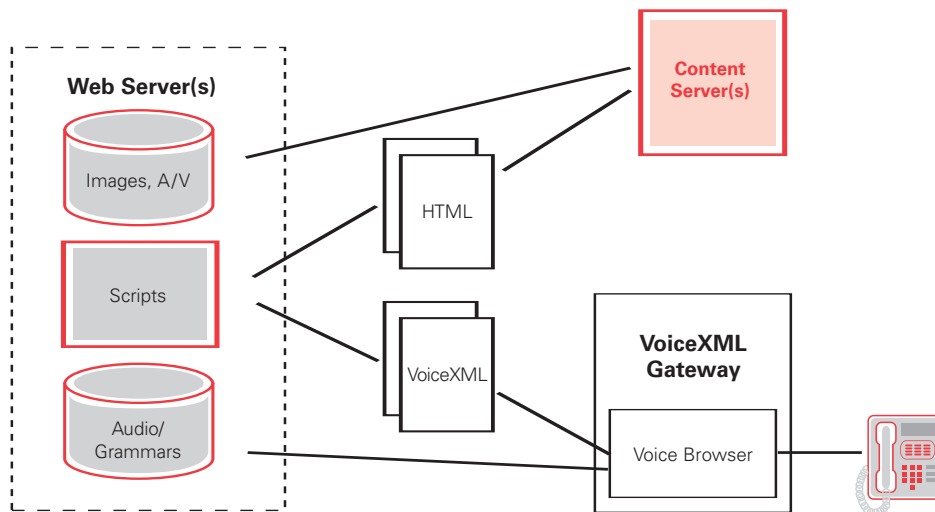
What are VoiceXML and SALT?

VoiceXML, developed under the auspices of the Voice XML Forum and the World Wide Web Consortium (W3C) allows speech application developers to use familiar, Web-based tools and techniques to build interactive speech user interfaces. VoiceXML specifies language elements in XML syntax that are interpreted by “voice browsers” to execute interactive voice dialogs. VoiceXML is designed around a form-filling methodology, in which audio prompts are played to elicit input from users via speech recognition or DTMF input, while complementary Web protocols (such as HTTP) handle non-voice-specific functions.

VoiceXML Version 2.0 was released as a true Web standard in March 2004 by the World Wide Web Consortium (W3C), the preeminent Web standards body, and has been rapidly adopted by a broad range of VoiceXML application and platform vendors.

However, although speech is effective for relatively simple input and selection tasks, complex output is often presented more effectively in text and/or graphical format. The SALT standard is specifically designed to support multimodal applications, and has been contributed to the W3C by the SALT Forum. SALT, like VoiceXML, leverages the Web’s architecture, standards, installed base and developer expertise to simplify the addition of speech to Web-based applications. However, unlike VoiceXML, SALT is a pure markup language—because it has no scripting capability of its own, it relies on the use of complementary XML-based standards to create a complete solution. This provides added flexibility, especially in supporting multimodal applications, but at the cost of increased complexity.

Figure 8: VoiceXML Architecture



Why Web-Based Standards?

Open standards are intended to allow portability and interoperability, so that software components from different vendors should run on a variety of hardware platforms. The proven three-tier Web architecture—presentation through a “browser,” middleware consisting of the Web server and the application’s business logic, and the back-end data storage tier—greatly simplifies system design and maintenance, and facilitates a multi-vendor implementation. At least in theory, any VoiceXML or SALT browser should be able to handle content from any Web server, so users can select best-of-breed components from competing vendors to deliver the best product at the lowest cost.

Issues and Alternatives

Complementary Web-based standards that address specific limitations of VoiceXML and SALT are still in early stages of development.

For example, XHTML+Voice (X+V), extends VoiceXML to provide multimodal support by integrating speech synthesis, recognition grammar and XHTML components.

And to address a major limitation of both VoiceXML and SALT—that they don’t support complex call control functions (answer, transfer, conference, etc.)—the W3C is in the process of creating a Call Control XML (CCXML) standard.

Sophisticated visual development tools are available from a variety of third-party vendors that can produce raw VoiceXML and/or SALT output. These can simplify and speed web-based speech development, especially for those who are less familiar with speech technology and more familiar with the application itself.

Cantata and Speech Standards

Cantata actively supports both the VoiceXML and SALT standards. Cantata speech boards integrate with the voice “browser,” transparently handling media and telephony signaling tasks within these environments. Although this abstraction means that applications based on VoiceXML or SALT do not need to integrate directly with telephony APIs, those who deploy speech applications must consider how the capabilities of the telephony board can impact the accuracy, performance and cost of the overall speech system.



cantata
TECHNOLOGY

Corporate Headquarters

410 First Avenue
Needham, MA 02494
USA

Tel: +1 (781) 449-4100

Fax: +1 (781) 449-9009

Email: info@cantata.com

Cantata Technology maintains multiple locations worldwide in North America, Asia and Europe.

www.cantata.com