Dialogic®

Application Note

Development Tools

Introduction to Dialogic® Continuous Speech Processing

Executive Summary

Dialogic® Continuous Speech Processing (CSP) is a feature available with select Dialogic® boards supporting high-performance, speech-enabled applications. Speech technologies can benefit from CSP because it provides board-level firmware that processes real-time voice signals to identify human speech input and present it to the host platform for speech recognition. This integrated speech processing support approach offloads host platform resources for more complex speech recognition tasks, such as analyzing and recognizing the speech input in support of the application.

CSP is also available with Dialogic® Host Media Processing (HMP) Software, enabling customers with board-based speech solutions to easily add VoIP to their product portfolio.

Table of Contents

Introduction 2

Feature and Benefits........ 2

Applications........ 2

Continuous Speech Processing 2

CSP on Dialogic® Products........ 3

Speech Technologies........ 3

Functional Description 4

Traditional Speech Processing 4

Competitive Speech Processing Feature........ 4

How Continuous Speech Processing Works........ 5

Echo Canceller 6

Voice Activity Detector 7

Technical Details 8

Hardware System Requirements........ 8

Introduction

This application note describes the Dialogic® Continuous Speech Processing (CSP) software and firmware feature. The CSP features and benefits, applications, and supported Dialogic® products are discussed. A functional description contrasts traditional speech processing with how CSP works in Dialogic products.

CSP, when packaged with open computing platforms, supports high-performance, speech-enabled applications. This technology enhances existing speech technologies by providing board-level firmware that processes real-time voice signals to identify human speech input and present it to the host platform for speech recognition. The real-time functions include both echo cancellation and Voice Activity Detection (VAD). This approach offloads host platform resources for more complex speech recognition tasks, such as analyzing and recognizing the speech input in support of the application.

Features and Benefits

CSP has the following features and offers these benefits:

- **Provides low implementation cost and enhanced system performance** — Integrated software features can be delivered on open, industry-standard computing platforms
- **Provides software for select Dialogic® boards** — Robust speech processing applications can be deployed without dedicated speech hardware or extensive host system resources
- **Reduces system latency, increases recognition accuracy, and improves overall system response time** — Firmware features can offload critical real-time signal processing in speech-enabled applications to onboard Digital Signal Processors (DSPs)
- **Is Scalable** — Unified Application Programming Interface (API) can enable applications on small- to large-scale systems to fit business requirements
- **Is Flexible** — Range of port densities on each board can provide high-performance, cost-effective solutions

Applications

CSP can be employed in applications such as:

- Voice portals: weather, traffic, movies, restaurant guides, etc.
- Speech-enabled Interactive Voice Response (IVR)
- Speech-activated dialing

CSP supports various application-friendly features, such as the ability to interrupt speech prompts by speaking over them — known as "barge-in". Barge-in lets callers control the pace of the conversation and complete the interaction more quickly, resulting in a more pleasant experience and more efficient use of the platform. Barge-in saves host-system resources, improves system usage, and can reduce phone charges by shortening calls.

VAD technology in CSP includes a pre-speech buffer that produces better voice recognition using less host processing and enhances the accuracy of speech detection. Designed to be flexible, CSP lets VAD be disabled or used in conjunction with the speech detection algorithms developed by speech technology developers. Developers can develop and deploy enhanced speech technology platforms that are enabled for voice commands with first-rate accuracy and performance.

CSP now includes a silence compressed streaming feature, which removes the silence between the caller's utterances before streaming to the host. This saves host processing cycles and allows for increased port density. The amount of silence compression and energy thresholds are configurable. Enhanced Echo Cancellation (EEC) is also available on select media loads and products. EEC lets developers select longer echo cancellation tail lengths of 32 ms and 64 ms (beyond the standard 16 ms) to further improve and refine audio quality. In addition, a feature available on select media loads and products is streaming to the CT Bus, which lets echo cancelled data be streamed into the TDM bus.

Continuous Speech Processing

CSP can be the basis for high-quality, high-performance speech-enabled applications. The software consists of a library of functions, device drivers, firmware, sample demonstration programs, and technical documentation to help create Automated Speech Recognition (ASR) applications. CSP is an enhancement to existing Echo

Cancellation Resource (ECR) and barge-in technology from Dialogic.

CSP is available on Dialogic HMP Software and select boards, and supports existing speech technologies. CSP supports Windows® and Linux operating systems.

CSP on Dialogic® Products

The following is a list of Dialogic products and their datasheets that provide a high-density, feature-rich environment for voice portal and other speech-enabled applications:

- Dialogic® Host Media Processing Software Release 3.1LIN — http://www.dialogic.com/products/ip_enabled/docs/9323_HMP3-1_ds.pdf
- Dialogic® Host Media Processing Software Release 3.0 for Windows® — http://www.dialogic.com/products/ip_enabled/docs/8762_HMP_3_0_Windows_ds.pdf
- Dialogic® DM/V600BTEP Media Board, Dialogic® DM/V600BTEC Media Board, Dialogic® DM/V1200BTEP Media Board, Dialogic® DM/V1200BTEC Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/8848_DMVB_PCI_cPCI_ds.pdf
- Dialogic® DM/V600BTEPEQ Media Board, Dialogic® DM/V1200BTEPEQ Media Board, Dialogic® DM/V300BTEPEQ Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/10339_DMVxxxBTEPEQ_ds.pdf
- Dialogic® DM/V3600BP Media Board, Dialogic® DM/V4800BC Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/8842_DMV3600BP_4800BC_ds.pdf
- Dialogic® D/120JCT-LS Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/6046_D120JCTLS_ds.pdf
- Dialogic® D/480JCT-1T1 Media Board, Dialogic® D/600JCT-1E1 Media Board, Dialogic® D/480JCT-2T1 Media Board, Dialogic® D/600JCT-2E1 Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/7131_D480_600_ds.pdf
- Dialogic® D/240JCT-T1 Media Board, Dialogic® D/320JCT Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/6038_D240_300_ds.pdf
- Dialogic® D/160JCT Media Board, Dialogic® D/320JCT Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/7647_D160_320_ds.pdf
- Dialogic® D/41JCT-LS Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/7870_VFX41JCTLS_ds.pdf
- Dialogic® VFX/41JCT-LS Media Board — http://www.dialogic.com/products/tdm_boards/media_processing/docs/7870_VFX41JCTLS_ds.pdf
- Dialogic® DM/IP241-1T1-PCI-100BT IP Board, Dialogic® DM/IP301-1E1-PCI-100BT IP Board, Dialogic® DM/IP481-2T1-PCI-100BT IP Board, Dialogic® DM/IP601-2E1-PCI-100BT IP Board, Dialogic® DM/IP601-CPCI-100BT IP Board — http://www.dialogic.com/products/ip_enabled/docs/3940_DMIP_ds.pdf

The boards are well-suited for developers choosing to provide cost-effective, highly scalable communications applications requiring multimedia resources such as voice, software-based speech recognition, and network interfaces.

Speech Technologies

Speech recognition provides host-based recognition engines and a variety of tools for developing and implementing robust applications. Dialogic, in addition to leading speech software suppliers, offers developers a wide range of technologies with proven and successful track records for helping to bring the power of ASR and Text-To-Speech (TTS) applications to market in a wide variety of industries.

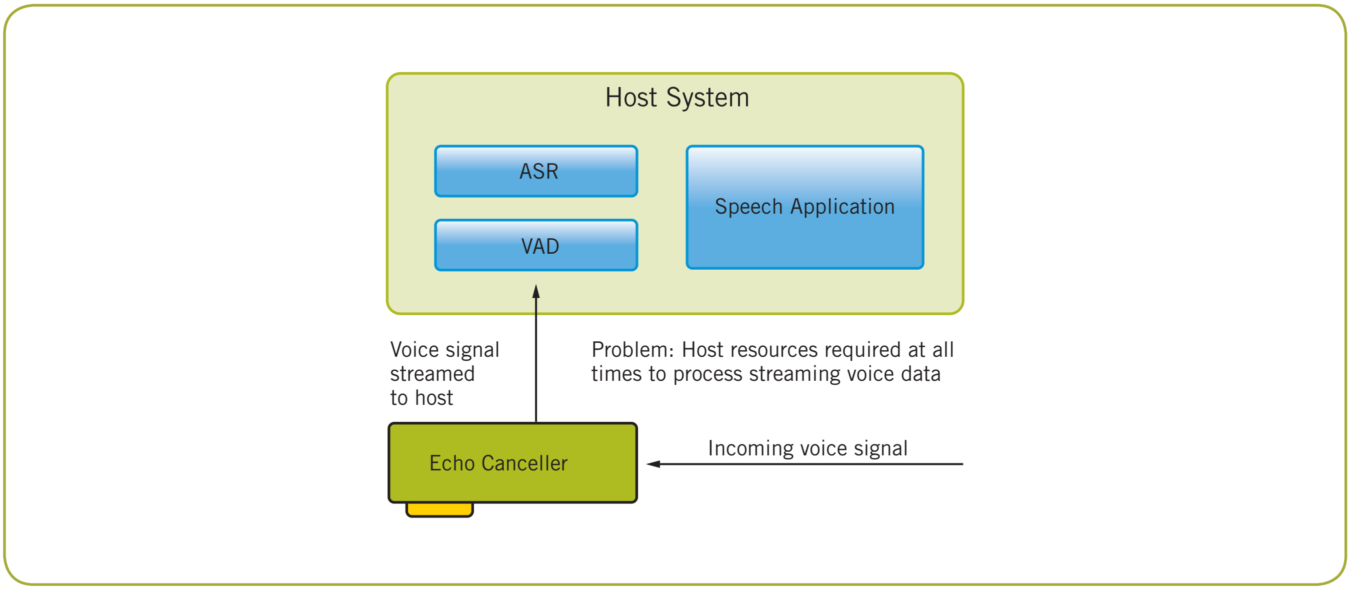
Host System
ASR
VAD
Speech Application
Voice signal streamed to host
Problem: Host resources required at all times to process streaming voice data
Echo Canceller
Incoming voice signal

Figure 1. Traditional Speech Processing

Functional Description

Traditional Speech Processing

To better understand the benefits offered by CSP, see the traditional approach to speech processing shown in Figure 1. Most speech-enabled systems are built so the processing required for voice recognition depends heavily on the host processor of the system in which they are installed. A DSP-resident echo canceller on a system voice board processes incoming voice data. The incoming voice signal is then streamed directly to the host system where the VAD and ASR engines reside. The host system must process voice events as they are received. This approach requires dedicated host resources and consumes much processing power, reducing the performance and scalability of the host system.

Although this approach has been used by a multitude of vendors, it requires the host system to be configured with high-performance, high-cost resources that are dedicated to a single purpose. This means that speech-enabled platforms are costly, hard to install, and require expensive, dedicated hardware in order to perform high-density applications.

Competitive Speech Processing Feature

In contrast, CSP offers a different paradigm for designing ASR applications and other speech processing solutions. Dialogic products provide base-platform functionality with integrated speech processing support. Specifically, with CSP, the traditional approach of using dedicated hardware to support voice recognition is unnecessary. Instead, board-resident resources are provided, letting the echo cancellation and VAD functions front-end the host system, so that the host system is only engaged when voice energy has been detected (see Figure 2). CSP differs from traditional approaches to ASR because it uses board-level speech processing features and barge-in technology, and does so without heavily loading valuable host resources.

With CSP, the incoming voice data from the caller is processed using the DSP-based echo cancellation and VAD integrated on the voice card. The incoming voice signal is then streamed to the host system and the ASR engines only when voice energy is detected. Higher accuracy and higher efficiency is attained using features such as the pre-speech buffer and the onboard VAD. This approach can reduce the processing load on the host, improve accuracy, and provide for higher density solutions.

An advantage is that the host system and its resources are reserved for the complex tasks associated with analyzing speech commands and language semantics. The communications boards are enabled to off-load the host and allow for greater accuracy and higher density solutions.

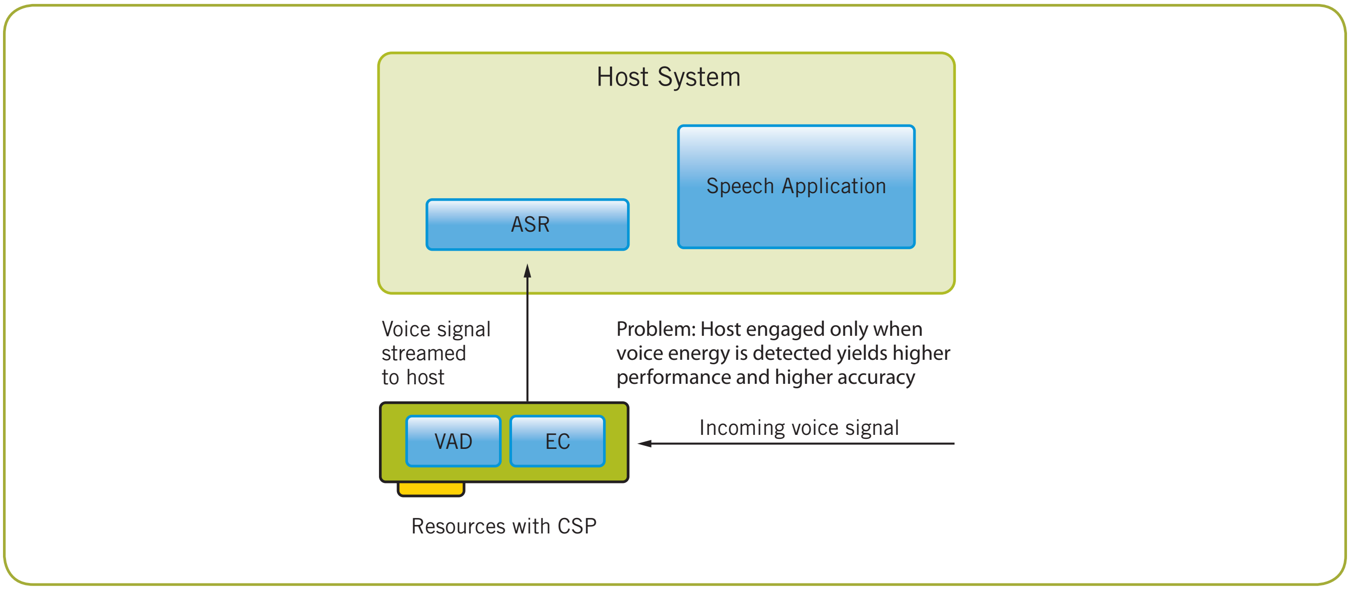
Host System
Speech Application
ASR
Voice signal streamed to host
Problem: Host engaged only when voice energy is detected yields higher performance and higher accuracy
VAD
EC
Incoming voice signal
Resources with CSP

Figure 2. Continuous Speech Processing (CSP)

Speech processing and the technologies that support speech, such as ASR and TTS, are not new. In fact, speech-enabled solutions built on these technologies have existed for years. Dialogic has been delivering products that include speech processing since the early 1990s, and has thrived in this market segment through technology leadership and by cultivating a broad set of speech technology relationships. Dialogic® technology has enabled Computer Telephony (CT) applications to run in hundreds of service providers' networks with these open, industry-standard components.

Several technological advances have ushered in an era of innovation and performance in CSP:

- **Cost reductions in computing platforms** — Lower cost hardware enables high-performance host systems for speech platforms
- **Full-duplex operation** — Provides the capability to simultaneously play/record voice data on a single channel
- **Enhanced echo cancellation** — Eliminates up to 64 ms of echo in the incoming signal
- **Voice Activity Detector (VAD)** — Detects audio energy and triggers data transmission only when speech is present
- **Pre-speech buffer** — Significantly reduces the problem of clipped speech and increases recognition accuracy
- **Barge-in capability** — Lets a party speak or enter digits without waiting for the prompt to finish
- **Voice event signaling** — Works in conjunction with the VAD to let the CSP firmware send speech-detected messages to the host speech application
- **Voice-activated recording or streaming** — Streams voice data to the host system only when voice energy is detected
- **Silenced-compressed streaming** — Removes silence between a caller's utterances, before streaming voice energy to host

These advances have served as a catalyst for a breed of speech applications offering cost savings, improved performance, greater accuracy, scalability, and higher density.

How Continuous Speech Processing Works

To better understand how CSP works, consider the call flow sample from an ASR automated attendant application detailed in the following example:

Event/Action	Flow	Description
1. Caller dials into “Any Enterprise.com” for information.	➡	Activates the ASR auto-attendant functions and the Continuous Speech Processing server software.
2. Caller listens to the welcome greeting: “Hello, and thank you for calling Any Enterprise.com. If you know the name of the person you wish to reach, say that name now. For a directory of contacts, say…”	⬅	The outgoing welcome greeting and initial menu for the auto attendant are played to the caller. Continuous Speech Processing begins removing prompt echo from the incoming voice signal while monitoring for speech input.
3. Caller interrupts the prompt by speaking the name “Stephanie Smith.”	➡	Interrupts the outgoing prompt. Continuous Speech Processing provides the following functions: • Terminates the prompt using barge-in • Forwards the “Stephanie Smith” signal, with pre-speech buffer data, to the ASR application
4. Caller is connected to “Stephanie Smith.”	⬅	The ASR application recognizes the name, correctly responds to the request, and connects the caller to Stephanie Smith.

Echo Canceller

The echo canceller is the component in CSP that eliminates echoes in the incoming signal caused by the prompt. In the sample call flow just described, the incoming signal is the utterance “Stephanie Smith.” Because of the echo canceller, the “Stephanie Smith” signal has insignificant echo and can be processed more accurately by the speech recognition engine.

Figure 3 shows how an echo canceller works. After the echo canceller processes the incoming signal, the resulting signal no longer has significant echo and can be sent to the host application.

If echo cancellation is not used, the incoming signal usually contains an echo of the outgoing prompt. Without echo cancellation, an application must ignore incoming speech until the prompt and its echo terminate. With these types of applications, there is typically an announcement that says, “At the tone, please say the name of the person you wish to reach.”

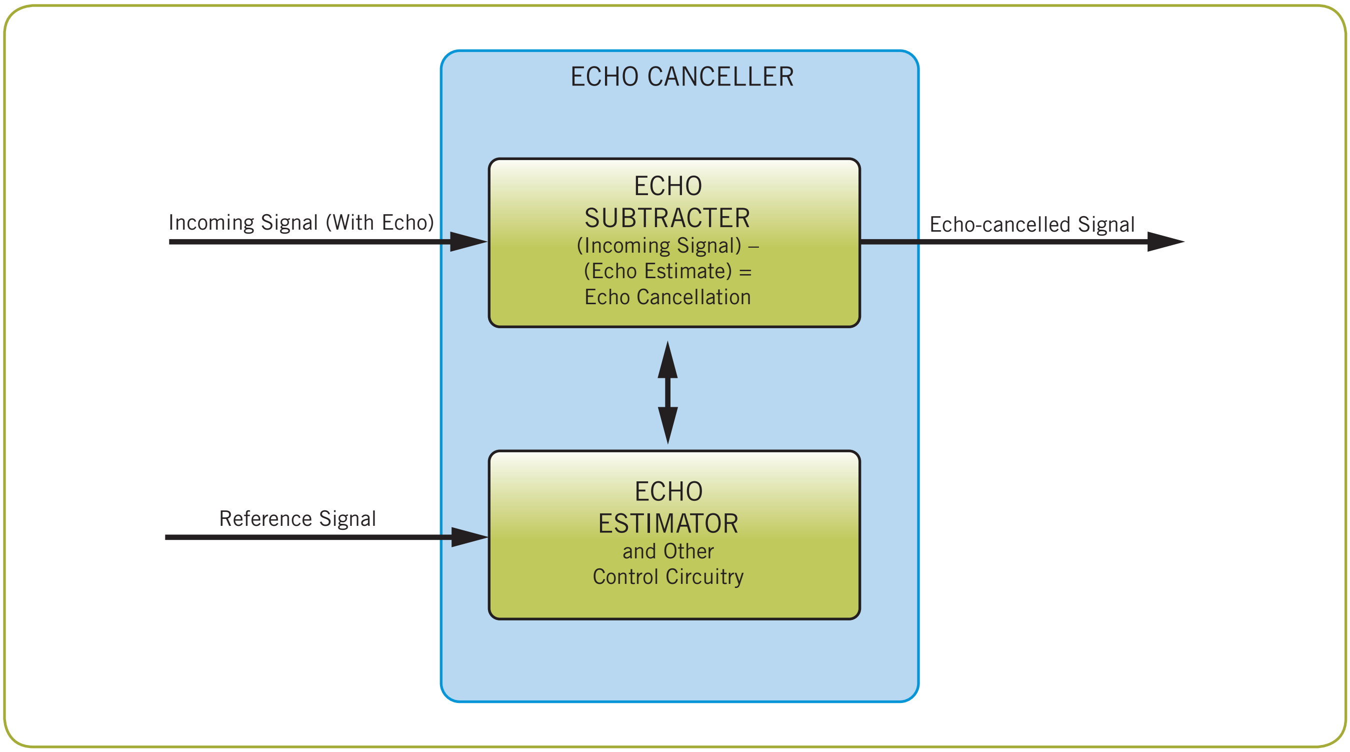

Figure 3. Echo Cancellation

With echo cancellation, the caller can interrupt the prompt, and the incoming speech signal can be passed to the ASR application.

Tap Length

The duration of an echo is measured in tens of milliseconds (ms). The number of milliseconds an echo canceller removes is known as the length of the echo canceller. The length of an echo canceller is measured in "taps" (each tap is 125 microseconds).

Adaptation Modes

The echo canceller has two adaptation modes:

- **Fast mode for rapid convergence** — Used immediately after the echo canceller is reset and energy is detected on the reference signal.
- **Slow mode for slower convergence** — Used the rest of the time. This mode is entered automatically after a few hundred milliseconds of fast mode.

Developers can design their applications to enter fast mode each time a new utterance is collected, or to perform rapid convergence just once, as the call begins, and remain in slow mode during the remainder of the call.

Voice Activity Detector

When a caller begins to speak over a prompt (also known as barge-in), the application stops playing the prompt so that it does not distract the caller.

A VAD is the CSP component that examines the caller's incoming signal and determines if the signal contains significant energy to be identified as speech. The VAD has several configurable parameters (such as the threshold of energy) that are considered significant during prompt play and after the prompt has completed.

Pre-Speech Buffer

The VAD does not usually detect an utterance immediately. Instead, the energy of the utterance builds until the utterance triggers the VAD. For example, the name "Stephanie", when pronounced, begins with a low-energy hiss.

Because the VAD sends the signal to the application only after the beginning of an utterance is detected, the low-energy start of the utterance is likely to be missing. This can create a problem because the ASR engine requires the complete speech utterance to correctly process the signal and fulfill the caller's request. To avoid this, CSP stores a pre-speech buffer; that is, a recording of the echo-cancelled incoming speech signal prior to the VAD trigger (see Figure 4). The data in the pre-speech buffer is sent to the application along with subsequent speech signals. The pre-speech buffer is an integral part of VAD. There is one pre-speech buffer per voice channel. CSP was designed with flexibility in mind. It allows the VAD to be disabled or used in concert with a speech technology developer's proprietary speech detection algorithm.

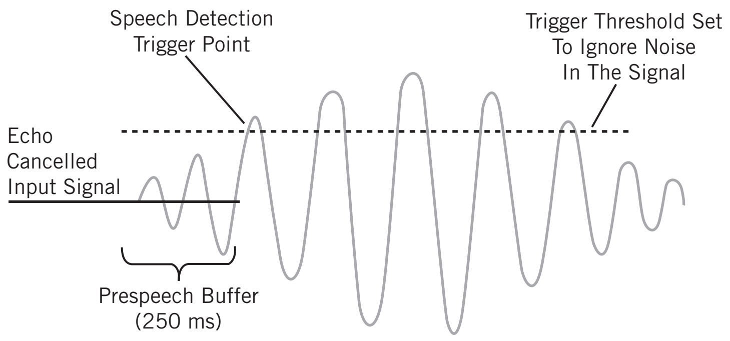

Figure 4. Pre-Speech Buffer

Barge-In and Voice Event Signaling

The combination of echo cancellation and VAD can be used to effect barge-in. Echo cancellation significantly reduces the echo of the prompt from the incoming speech signal. The VAD detects the beginning of an utterance and can send a VAD event to the host application.

The barge-in feature stops the playing of the prompt upon detection of audio energy exceeding the threshold. For some applications, the prompt can be halted based on other criteria. For example, after the caller utters a valid vocabulary word. In this case, CSP can be set to inform the application that voice energy has been detected without terminating the prompt.

Technical Details

Table 1 provides the CSP technical details.

Echo Cancellation	
Input dynamic range	+3 dB to –35 dB in compliance with G.165 and G.168
End-path delay	Eliminates up to 64 ms
Echo Return Loss Enhancement (ERLE)	Greater than 34 dB, and infinite with NLP enabled
Convergence rate	Greater than 20 dB of ERLE in 200 ms
Audio Signal	
Usable receive range	(Analog) –40 dBm0 to +2.5 dBm0 nominal, configurable by parameter** (T1) –40 dBm0 to +2.5 dBm0 nominal, configurable by parameter** (E1) –43 dBm0 to +2.5 dBm0 nominal, configurable by parameter**
Silence detection	–38 dBm0 nominal, software adjustable**
Transmit level (weighted average)	(Analog) –9.5 dBm0 nominal, configurable by parameter** (T1) –9 dBm0 nominal, configurable by parameter** (E1) –12.5 dBm0 nominal, configurable by parameter**
Transmit volume control	40 dB adjustment range, with application definable increments and legal limit cap
Audio Digitizing	
48 kbps	G.711 PCM (μ-law for T1 and A-law for E1) @ 6 kHz sampling rate
64 kbps	G.711 PCM (μ-law for T1 and A-law for E1) @ 8 kHz sampling rate
Digitization selection	Selectable by application on function call-by-call basis
Playback speed control	Pitch controlled Available for G.711 PCM 8 kHz coder Adjustment range: ±50% Adjustable through application or programmable DTMF control

**Configurable to meet country specific PTT requirements. Actual specification may vary from country to country for approved products.

Table 1. CSP Technical Details

Hardware System Requirements

For Dialogic® boards with Continuous Speech Processing:

- Single or dual processor PCI or PCI Express bus, or Compact PCI bus computer
- Operating system hardware requirements vary according to the number of channels being used
- System must comply with PCISIG Bus Specification Rev. 2.1 or later or PCIe 1.1 electromechanical specification

For more information about Dialogic® products, visit **www.dialogic.com.**

Dialogic Corporation
9800 Cavendish Blvd., 5th floor
Montreal, Quebec
CANADA H4M 2V9

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH PRODUCTS OF DIALOGIC CORPORATION OR ITS SUBSIDIARIES ("DIALOGIC"). NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN A SIGNED AGREEMENT BETWEEN YOU AND DIALOGIC, DIALOGIC ASSUMES NO LIABILITY WHATSOEVER, AND DIALOGIC DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF DIALOGIC® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY INTELLECTUAL PROPERTY RIGHT OF A THIRD PARTY.

Dialogic products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications.

Dialogic may make changes to specifications, product descriptions, and plans at any time, without notice.

05/08 10106-01