Voice as the user interface – a new era in speech processing

Debbie Greenstreet
Product Marketing Manager
Digital Signal Processors
Texas Instruments

John Smrstik
Marketing Manager
Digital Signal Processors
Texas Instruments

As the evolution of computing continues from the desktop era and into the mobile arena, the technology advancements that have led to these changes are impressive to witness.

It is hard to imagine that in nearly every person's pocket is several times the computing power once needed to get a crew of American astronauts to the moon and safely back home. This computing power isn't just embedded in our phones – it's in our tablets, music players, watches and appliances, and is becoming more pervasive all the time.

If technology is going to be part of everyday life, making electronic gadgets easy to use with a simple, intuitive, natural user interface is necessary. Historically, designers addressed user interfaces through hardware devices (a keyboard and mouse), graphical interfaces (icons, buttons), and touch screens that combine both.

Several next-generation alternatives are emerging to make interaction with computing devices more accessible, easier and safer. Some of these are familiar, like gesture-based interfaces (Nintendo® Wii or Kinect gaming systems). Others serve specific market niches, like a brain-computer interface that maps electric signals of specific commands to help paralyzed patients interact with the world.

The user interface technology that is arguably showing the most promise today has been around for many years – voice and speech recognition. While intelligent voice recognition (IVR) systems have been around since at least the 1980s, many frustrations with the technology have kept voice as the user interface limited to applications like call-center management. These frustrations include inaccuracy and misinterpretation in speech detection, and the limited vocabulary used in menu-driven systems. As designers continue to address these frustrations, they have become more comfortable to use, and consumers are relying more on the capabilities that voice recognition has to offer.

Voice as the user interface

The focus for creating a voice-based user interface must be to make the user experience simple, comfortable and reliable. To deliver a good experience, designers need to consider questions such as: where the interaction will take place, how much information is being communicated, whether the user is near-by or far away from the device and how the interaction should happen. Each type of voice-enabled system may have different answers to these questions, and any system designer should attempt to account for the most likely use cases and then design to meet the expected challenges.

Voice recognition systems

Voice or speech recognition systems fall into two main categories.

The first and simplest is called "grammar-based" recognition and is based on programming the recognition engine with a few pre-defined rules to drive the activity. Sometimes called discrete word systems, these systems have a finite vocabulary and may not efficiently separate commands from other speech that it does not understand.

Consider a voice-controlled portable media player without a Wi-Fi® or cellular connection. This player might respond to commands like “play,” “pause” or “skip.” The vocabulary of commands for this type of system could be only a few words, or potentially dozens of words and phrases (“track nine,” “raise volume,” “new playlist”). The recognition engine and its vocabulary would need to be embedded in the device and run from local memory. These recognition engines use a pattern-matching algorithm that performs a “best fit” of the digital signature of the captured phrase with a reference command signature.

The second category is “natural language” recognition. In natural language systems, the recognition engine can process continuous speech, identify key words and in some systems interpret simple context. Natural language systems are more resource-intensive to implement than discrete word systems, but have the potential to provide a richer user experience. If the system’s local processing, memory and storage resources are large (for example, in a PC), then a natural language system can be embedded without the need for additional resources. However, as electronics scale smaller, connected resources such as cloud-computing resources may be necessary in order to accommodate natural language.

A good example of a natural language system is OK Google, which requires a connection to Google Cloud resources to provide its voice user interface. Locally, the voice is sampled, digitized and compressed, and then sent over a wireless network to the cloud infrastructure where the voice processing engine resides. The voice processing engine decodes, processes and analyzes the voice for content, matching it to an index for a response. The response is re-encoded and sent back over the network, and the local device plays back the response. Delivering a rich user experience in this case requires the resources of a cloud infrastructure, while a smartphone is acting as a voice capture and playback device.

Some more sophisticated systems today employ both techniques. These systems use grammar-based engines as a trigger for certain actions, and use natural language processing and cloud computing resources to deliver more advanced responses. A simple command library processes actions that are handled locally and that have only a local context. An example could be to ask a voice-enabled thermostat to “raise the temperature to 74 degrees.” Decoding and taking action on this phrase could be handled locally with no other intervention. And in this example, the thermostat does not need to provide an audio acknowledgment of the temperature change.

Employing cloud speech recognition services could add richness to this interaction. As an example, consider this string of speech: “Thermostat, raise the temperature to 74 degrees. It seems cold … what is the forecast today in Dallas?” These sentences could represent a simple command (“raise the temperature”) while also acting as a trigger (“thermostat”) to the system to listen to the full speech and send it to the cloud to retrieve current weather forecast information.

This experience requires a few additional things. First, the thermostat needs to be connected to the Internet to access the cloud. Second, there would need to be a return path for the response, most likely via audio playback. Although the electronics are a bit more complicated, if it improves the user interface of the thermostat and provides additional value (the weather forecast), the additional costs can be worth the effort.

Amazon Echo with Alexa Voice Service and Google Home are two types of systems that employ trigger phrases as well as connectivity to cloud-enabled

voice services that comprehend natural language. These devices are marketed as home automation hubs that coordinate with other cooperating devices built within the same software ecosystem.

Capturing the voice signal

A system's ability to recognize speech, interpret commands, process natural language and provide a relevant response in proper context will make interactions with computing systems much simpler and much more tangible for many types of users of all abilities. The computing power stored in the cloud will enable these systems to evolve further to produce even more enriched experiences. As evidenced by products available today, voice recognition provides a comfortable, valuable user interface for all kinds of systems.

But interpreting the voice signal is only part of the technical challenge. For voice recognition systems to work well, the input to the recognition engine – the core algorithm that maps the incoming signal to a known signature – must be clean and give the recognition engine the best opportunity to do the job it was created to do. Background noise and the location of the voice relative to the microphones can contribute to attenuation of the target signal and produce errors. The presence of an audio signal in the playback path, the reflection of the original voice, and even acoustic coupling from the vibrations of a loudspeaker can bring additional impairments. Speakerphones and other conferencing systems that handle audio have dealt with many of these considerations.

Ideally, to provide a smooth user experience, a voice recognition system should work effectively in the presence of ambient noise (and occasionally loud background noise) as well as when the speaker's voice is coming from relatively long distances, such as from across the room. The system also needs to account for echo effects that may be present. Fortunately, digital signal processing (DSP) techniques commonly found in traditional communication systems can improve the performance of voice recognition systems.

The far-field effect

Smartphone users take for granted the performance of accurate speech processing. However, accurately filtering a voice from other audio sources among adjacent noise – when that voice is meters away from the microphones – is another level of processing complexity that requires a technology known as far-field microphone arrays. This technology also employs signal-processing techniques such as beamforming and spatial filtering, which are commonly used to reliably receive a directional signal.

A signal received across an array of sensor elements – in this case, microphones – that are spatially separated, has geometric characteristics that can be used to improve reception of the audio signal. Each element receives the signal at slightly different times based on the relative positions of the sensors (microphones) and the source (voice) signal. Piecing these time-shifted signals together determines the location of the source relative to the sensors.

Applying adaptive filtering techniques can effectively tune the system to listen more closely to certain microphone sensors, while ignoring sensors farther away from the voice. This strengthens the input signal, especially when compared to a single element that only listens to audio coming from all directions (the omni-directional case). While there are certainly diminishing returns to the number of microphone elements added to a system, more microphones can theoretically provide better input that is more immune to ambient noise.

Selecting how many microphones to use and their configuration requires consideration of the expected

use case, balanced against the practical realities of system cost. For example, a system mounted to a wall could assume that the voice signal of interest is most likely coming from a 180-degree field in front of the system. In this case, you might use a linear array of two to four microphones with outward pointing beams (at 90 degrees relative to the microphone array). A system designed to be used in the center of a room in which the target voice could be coming from a 360-degree field might employ a circular array of six or more microphones, as shown in **Figure 1**. Well-thought-out choices for the number, configuration and placement of microphone elements should consider how the system is likely to be used, and can make a significant impact on the user experience.

***Figure 1**. Example of a working eight-microphone system configured in a circular array delivering digital audio output.*

Implementing spectral noise reduction behind the microphone array and beamforming contributes to further improvement. Spectral noise reduction uses the frequency domain to effectively remove ambient and unwanted noise components from an audio signal, making it easier to hear the desired signal. Using adaptive filtering techniques creates an estimate of the signal's noise component. This noise component, when subtracted from the composite signal, reveals the best possible representation of the actual speech signal.

Combining spectral noise reduction with beamforming can improve the system's signal-to-noise ratio and produce a high-quality input signal to a speech recognition engine. End users will experience better performance in noisy environments and/or while speaking farther away from the system – the far-field effect that many systems are striving for. Dealing with environmental factors addresses one of the key attributes of a good user experience.

Echo and reverberation

Many systems implementing voice recognition also incorporate an audio-playback path. This represents another challenge to voice recognition systems, as it adds other noise components. Acoustic echo occurs when the audio being played back is picked up by the microphone elements and sent back into the system. It can be exacerbated when the audio playback volume is high and when the speaker and microphone elements are in close proximity.

Reverberation occurs when audio-signal reflections are fed into the microphone.

A system can't work well without eliminating the effects created by audio playback. See **Figure 2**. Acoustic echo-cancellation algorithms and de-reverberation filters are signal-processing routines traditionally employed in conferencing, mobile and speakerphone applications. They are finding a new application in speech recognition systems.

The embedded voice-trigger and recognition system

An embedded voice system can be broken up into three different functional areas. On the front end

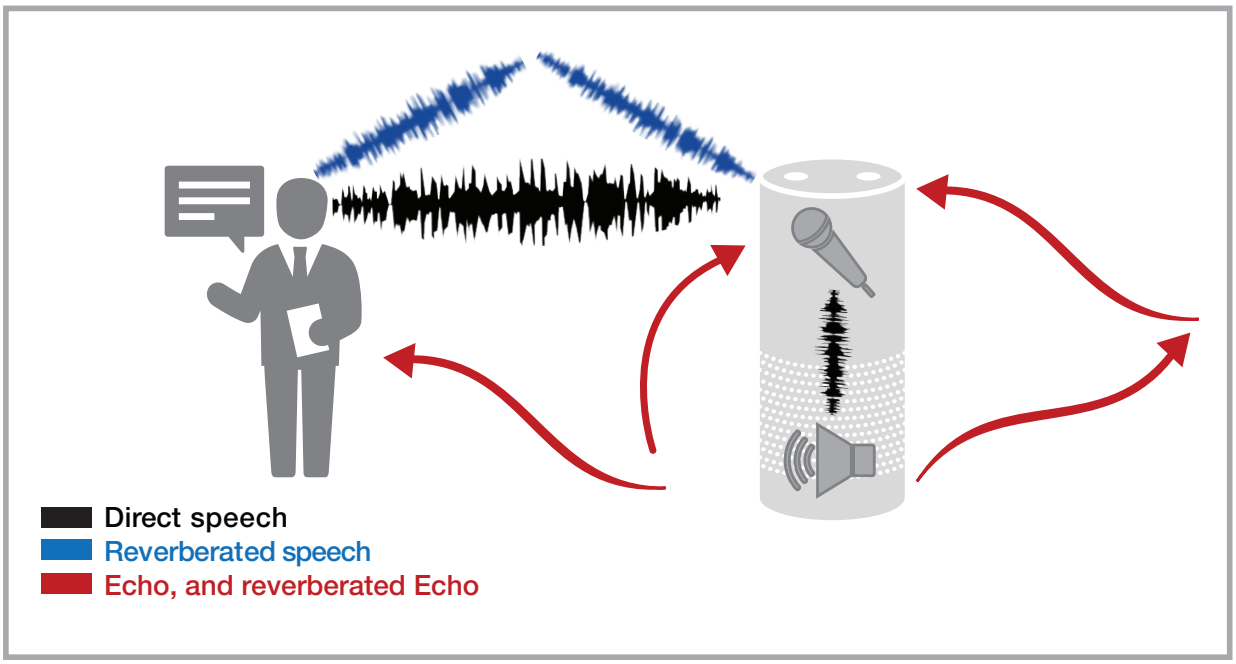

***Figure 2**. Echo and reverberation complicate the capture of a clean voice signal at the microphone.*

of the system are signal-processing components responsible for acquiring and improving the quality of the input speech signal. As discussed, these can include beamforming, spectral noise reduction, acoustic echo cancellation and de-reverberation. DSPs can efficiently implement these common algorithms. An example of such system is shown in **Figure 3** below.

The speech signal, once processed, is passed to an embedded speech recognition engine. This speech recognition engine implements a small vocabulary of commands or trigger words that kick off specific actions. Actions may include a set of locally executed commands like "turn on ..." and usually include a trigger alerting the rest of the system to begin listening. Siri, OK Google and Alexa are three common triggers in use today.

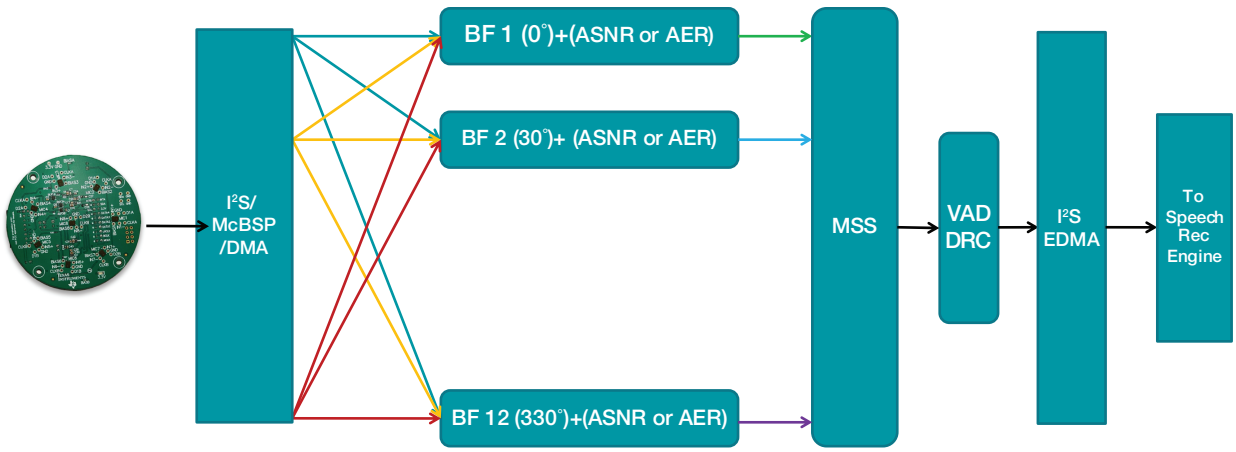

***Figure 3**. Functional diagram of a far-field voice pre-processing application running on a DSP.*

Upon acceptance of the system trigger, an Internet connection (or network connection) needs to be established in order to send the incoming speech to a network-based natural language processor. Cloud-based systems have the amount of processing capability necessary to apply to recognition and context interpretation, an ability to learn and update algorithms for accuracy and robustness, and the storage available for the massive amounts of speech data that pass through the system to accelerate learning. Examples of cloud-based systems include Google Voice, Alexa Voice Service with Amazon Web Services (AWS) and IBM Watson. The cloud-recognition engine processes natural language and translates it to actions that can be fulfilled through web integration or applications built to work with that service.

The cloud returns a response based on the action requested; in many cases this is an audio response. An audio playback path is often part of the system, including a speaker and amplifier solution, as shown in **Figure 4**. Depending on the desired quality of audio playback (for example, playing back music), you may want to employ audio post-processing and equalization technology to improve the listening experience. There are many alternatives for audio processing; again, DSPs or standard microprocessors can implement most of them easily and efficiently.

Putting the system all together

The applications for this technology are diverse, rapidly growing and expandable beyond consumer products like in-home virtual assistants. Home security, building automation and industrial-based uses of this technology abound. However, the range of solutions needed to address voice as the user interface applications varies; no one size fits all. From the number of microphones, to the sophistication of beamforming, to the need for acoustic echo cancellation, the elements are highly dependent on the product requirements.

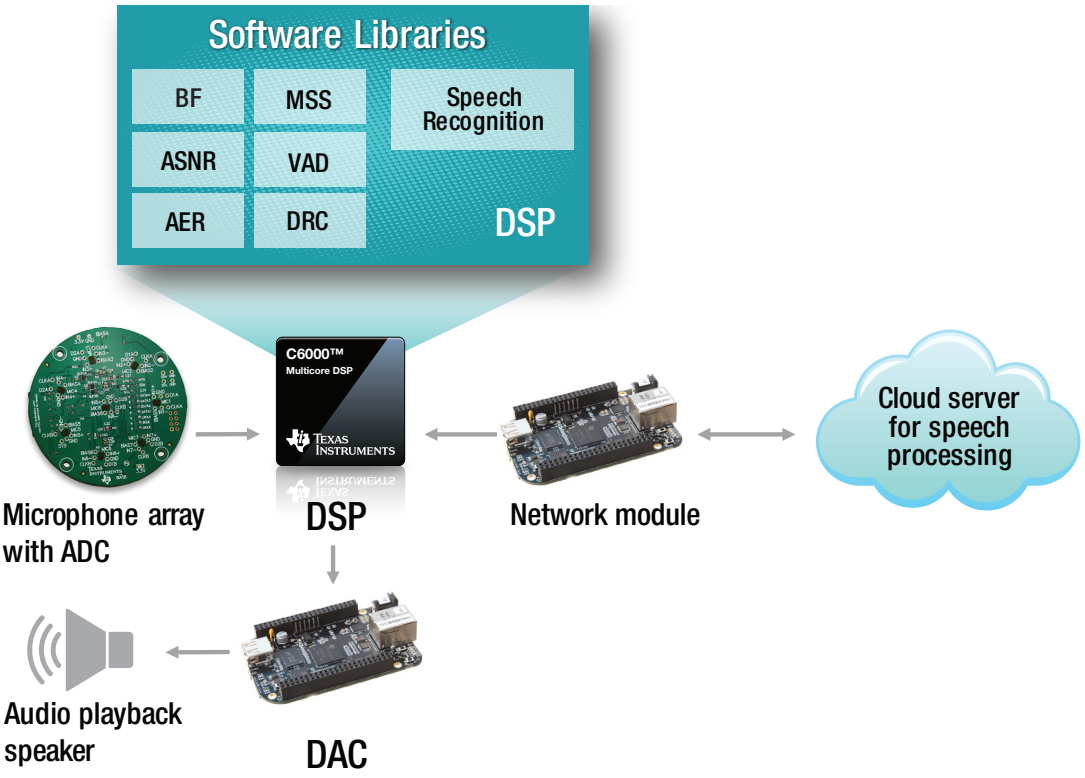

Figure 4. A complete voice user interface with cloud-recognition services and an audio playback path.

Fortunately, TI has a suite of silicon and software components, from the analog front end to processed speech output, that you can integrate into a complete yet efficient solution for a given application.

A variety of audio-specific analog-to-digital converters (ADCs) are applicable to different system needs. TI's portfolio offers high-dynamic-range, low-noise products that can handle single, dual or four microphone designs in a single package. These software-controlled ADCs feature advanced features such as energy signal detection and infinite impulse response (IIR) linear-phase first impulse response (FIR) filters for interfacing with analog or digital microphones.

TI DSPs enable speech audio pre-processing received from multi-microphone systems. The ultra-low-power C5000™ DSP family is great for battery-powered applications and capable of performing routines like beamforming, array signal-to-noise ratio (ASNR) and dynamic range compression (DRC) signal processing, acoustic echo cancellation and de-reverberation, and some level of speech recognition. The floating-point-capable C6000™ family handles more complex applications that require more microphones, more sophisticated noise-reduction algorithms and even audio post-processing functions for a playback path. These DSP families are supported by a processor software development kit (SDK) that provides a common software basis for all products and enables easier movement across the portfolio.

TI also offers field-proven software libraries that include beamforming, ASNR, DRC, voice activity detection (VAD) and acoustic echo removal (AER) for incorporating and tuning to your solutions as needed.

Collectively, these devices and corresponding software can enable audio pre-processing solutions optimized for your particular application. This hardware/software suite offers a scalable product line that adds voice as the user interface. Finally, the TI Designs reference design library, including the 66AK2G02 based reference design, can help you quickly establish an evaluation and demonstration platform.

Summary

All types of embedded systems are considering interaction with electronic systems with voice as the user interface, from consumer audio devices, to home automation systems, to industrial automation equipment. Services like Siri, Cortana, Google Now and Alexa have shown that the user experience can be dramatically improved relative to voice response systems we have used in the past (consider an airline's voice response enabled customer service line) to be very natural and seamless. Cloud computing and speech recognition services are easily accessible, powerful and constantly improving. Internet connectivity is nearly ubiquitous and very easy and cost-effective to embed in systems. Signal-processing techniques that have been used effectively in traditional communication systems for decades are getting new life cleaning and enhancing input voice signals for speech recognition engines.

These facts have encouraged device manufacturers to add voice as the user interface and have shown consumers that the products can work well and reliably. In the coming years, we can expect to start interacting with more systems via voice and observing the integration and interaction of systems based on patterns that we determine or are automatically learned. Playing music or retrieving trivia answers are only the beginning; it will be exciting to see how voice as the user interface unlocks innovation across the industry.

For more information please visit: www.ti.com/audio

IMPORTANT NOTICE FOR TI DESIGN INFORMATION AND RESOURCES

Texas Instruments Incorporated ('TI") technical, application or other design advice, services or information, including, but not limited to, reference designs and materials relating to evaluation modules, (collectively, "TI Resources") are intended to assist designers who are developing applications that incorporate TI products; by downloading, accessing or using any particular TI Resource in any way, you (individually or, if you are acting on behalf of a company, your company) agree to use it solely for this purpose and subject to the terms of this Notice.

TI's provision of TI Resources does not expand or otherwise alter TI's applicable published warranties or warranty disclaimers for TI products, and no additional obligations or liabilities arise from TI providing such TI Resources. TI reserves the right to make corrections, enhancements, improvements and other changes to its TI Resources.

You understand and agree that you remain responsible for using your independent analysis, evaluation and judgment in designing your applications and that you have full and exclusive responsibility to assure the safety of your applications and compliance of your applications (and of all TI products used in or for your applications) with all applicable regulations, laws and other applicable requirements. You represent that, with respect to your applications, you have all the necessary expertise to create and implement safeguards that (1) anticipate dangerous consequences of failures, (2) monitor failures and their consequences, and (3) lessen the likelihood of failures that might cause harm and take appropriate actions. You agree that prior to using or distributing any applications that include TI products, you will thoroughly test such applications and the functionality of such TI products as used in such applications. TI has not conducted any testing other than that specifically described in the published documentation for a particular TI Resource.

You are authorized to use, copy and modify any individual TI Resource only in connection with the development of applications that include the TI product(s) identified in such TI Resource. NO OTHER LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE TO ANY OTHER TI INTELLECTUAL PROPERTY RIGHT, AND NO LICENSE TO ANY TECHNOLOGY OR INTELLECTUAL PROPERTY RIGHT OF TI OR ANY THIRD PARTY IS GRANTED HEREIN, including but not limited to any patent right, copyright, mask work right, or other intellectual property right relating to any combination, machine, or process in which TI products or services are used. Information regarding or referencing third-party products or services does not constitute a license to use such products or services, or a warranty or endorsement thereof. Use of TI Resources may require a license from a third party under the patents or other intellectual property of the third party, or a license from TI under the patents or other intellectual property of TI.

TI RESOURCES ARE PROVIDED "AS IS" AND WITH ALL FAULTS. TI DISCLAIMS ALL OTHER WARRANTIES OR REPRESENTATIONS, EXPRESS OR IMPLIED, REGARDING TI RESOURCES OR USE THEREOF, INCLUDING BUT NOT LIMITED TO ACCURACY OR COMPLETENESS, TITLE, ANY EPIDEMIC FAILURE WARRANTY AND ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT OF ANY THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

TI SHALL NOT BE LIABLE FOR AND SHALL NOT DEFEND OR INDEMNIFY YOU AGAINST ANY CLAIM, INCLUDING BUT NOT LIMITED TO ANY INFRINGEMENT CLAIM THAT RELATES TO OR IS BASED ON ANY COMBINATION OF PRODUCTS EVEN IF DESCRIBED IN TI RESOURCES OR OTHERWISE. IN NO EVENT SHALL TI BE LIABLE FOR ANY ACTUAL, DIRECT, SPECIAL, COLLATERAL, INDIRECT, PUNITIVE, INCIDENTAL, CONSEQUENTIAL OR EXEMPLARY DAMAGES IN CONNECTION WITH OR ARISING OUT OF TI RESOURCES OR USE THEREOF, AND REGARDLESS OF WHETHER TI HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

You agree to fully indemnify TI and its representatives against any damages, costs, losses, and/or liabilities arising out of your non-compliance with the terms and provisions of this Notice.

This Notice applies to TI Resources. Additional terms apply to the use and purchase of certain types of materials, TI products and services. These include; without limitation, TI's standard terms for semiconductor products http://www.ti.com/sc/docs/stdterms.htm), evaluation modules, and samples (http://www.ti.com/sc/docs/sampterms.htm).

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265