Electronic Design

CHIN BECKMANN, CEO, DSP Concepts,
http://www.dspconcepts.com

Navigating the Challenges of Embedded Voice-Control for Smart TVs

Embedded voice-control for smart TVs is here, but technical issues remain. What needs to be done with the audio front end and far-field voice implementation to deliver a high-performance, consistent solution?

With the pandemic forcing us to spend more time at home, analysts predict the average American will spend 2 hours and 46 minutes watching TV every day this year. Knowing American consumers are spending so much time watching TV, we can also assume they spend quite a bit of time searching for new content.

Historically, most people would use a remote as their search mechanism, but anyone who's ever manually searched for a new show using an on-screen keyboard knows how clunky and slow it is. Push-to-talk (PTT) remotes have emerged as a savior in the last decade by providing a better user experience than keyboards. However, PTT on remotes still requires that the user actually finds the remote and presses the microphone to effectively control the TV by voice.

This lackluster experience fails to deliver on the real promise of voice control for TVs—the ability to seamlessly switch the channel or change the volume through voice commands when our hands are full or we're working far from the TV in the kitchen. With Americans staying home and watching more TV than ever, there's a clear need for voice-enabled content discovery—but only if it actually works.

Smart TVs have a complex ecosystem with multiple accessories, such as remote controls, sound bars, set-top boxes, and in some cases a streaming device like a Roku or Apple TV, all of which can be connected to the TV. Any of these components could be used to host the voice user interface (VUI). In fact, some companies have already dipped their toes into voice control for TV by integrating the technology into some of these components, like Samsung has with its sound bars.

You might be thinking, "Why would brands need to add voice to their TVs if they've already added it to accessories like sound bars and remote controls?" The answer is that they want to make their way into our living rooms through as many devices as possible. But at the same time, not all accessories can provide a superior user experience when it comes to far-field voice control.

The *table* illustrates those limitations with a comparison of the factors that affect the performance of voice-activation smart TVs and their connected accessories. For instance, remote controls can only go up to the mid-field voice-activation range given the form factor, battery operation, and bill-of-materials (BOM) limitations on these devices.

Going Full Far-Field

For brands that want to provide a full far-field experience where users can control their TVs from across the living room or kitchen, the switch from PTT to mid-field voice activation may not deliver enough value to justify the investment. Sound bars can provide far-field voice control, but for brands that sell both TVs and sound bars, adding far-field voice to

	Smart TV	Remote control	Sound bar
Form factor	Large	Small	Medium
Performance	Far-field (6m+)	Mid-field (3-6m)	Far-field (6m+)
# of microphones	2 or 4	2	At Least 4
Relative BOM cost of always-listening	Low	High	Medium
Power budget constraints	Low	High	Low

Different requirements for voice-activated design.

both devices can lead to a fragmented and confusing user experience.

Now let's talk about *why* adding far-field voice to the TV itself is a viable option. To start, a TV's large form factor allows for integrating more powerful systems-on-chip (SoCs) to add complex algorithms for always-listening implementation, which improves its far-field voice-detection capabilities. Secondly, a TV is usually the largest "hub" in the living room that can control multiple entertainment devices using only voice if designed correctly.

Thanks to advances in technology we're starting to see more TVs offer embedded voice control, which promises a substantially better user experience compared to other accessories. However, delivering a high-performance VUI for smart TVs, where a user can sit on their couch and start watching a new program just by saying "Alexa, play The Last Dance" or "Hey Bixby, play Tiger King," is easier said than done. Adding voice to the TV itself is an entirely different process and there are several key differences to consider—even for a giant like Samsung.

Building Blocks for Embedded Voice Control in TVs

Now that we've covered why brands are looking to add voice directly to TVs, and why these devices are well-suited for it, let's take a closer look at the various technical elements needed to do it successfully *(Fig. 1)*.

The All-Important Audio Front End

The audio front end (AFE) includes the microphone array and signal-processing blocks required to achieve far-field voice performance on smart TVs. It processes the multichannel microphone array signal to cancel out any interfering background noises or the device's own playback signal to produce a clean voice signal. The resulting clean signal is then sent to the wake-word detection engine. This engine is what actually recognizes the wake word, such as "Alexa" or "OK Google," which has been pre-programmed on the device.

The AFE often uses multiple signal-processing algorithms to effectively cancel out unwanted interference signals while preserving user speech. Here, we'll break down what those algorithms are and the purpose they serve:

Microphone array

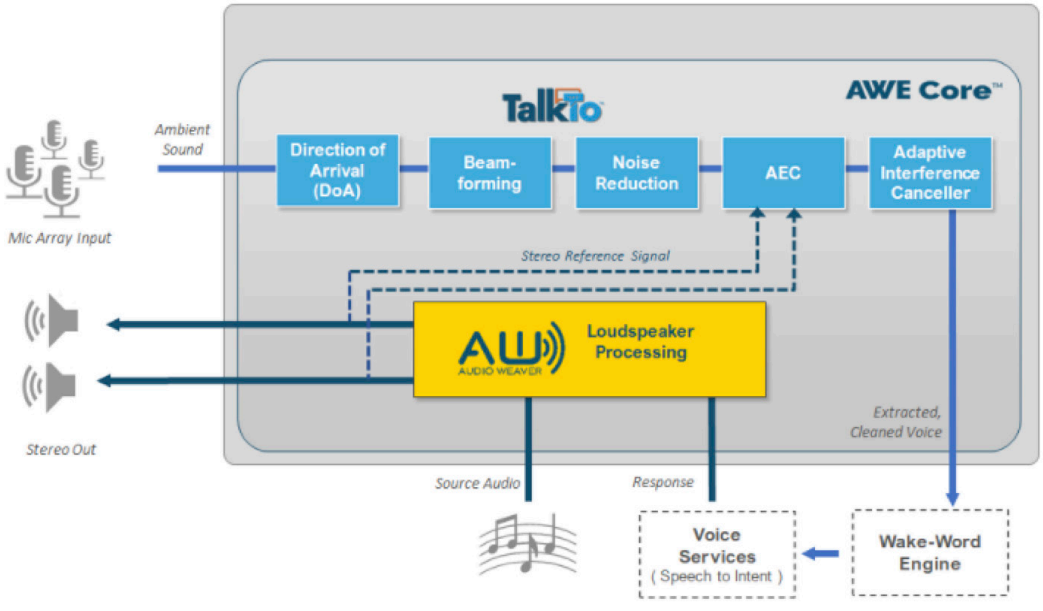
Voice-activation systems require one or more microphones working in tandem to capture the incoming signal that's passed on to the AFE. Size, cost, performance, and reliability are the primary factors to consider when selecting the microphone array; the performance of the array hinges on the number of microphones, geometry, and placement on the device.

Given TVs are typically located against a wall, the TV's microphones must be on a horizontal strip. The strip can be on the front with the microphones facing in the direction of the user, or on top (bottom) with the microphones facing up (down). With flat-screen designs, the geometry is also limited to linear arrays with a minimum of two microphones in the array.

The combined multi-mic signal from the microphone array improves the effective SNR available for signal processing in the rest of the audio signal chain. It's also good practice to separate the microphones from the loudspeakers to minimize direct coupling of sound.

Direction-of-arrival (DoA) detection

DoA detection determines the position of the user relative to the product so that the microphone array can steer the beam in the direction of the user's voice.



1. Shown is a typical voice-activated system.

Beamformer

The beamformer accepts sounds coming from the determined DoA while rejecting sounds from other directions. The beamformer's performance depends heavily on the microphone array geometry, beam width, background noise level, and the effective SNR of the microphone array. TVs are often located in the corner of a room and, therefore, the beamformer algorithms must steer the beam within a 180-deg. field of view.

Single-channel noise reduction (SCNR)

The SCNR algorithm can suppress up to 10 decibels of stationary noise from things like the sound of a fan blowing or a microwave running, and serves as the first step in the noise-cancellation algorithms.

Acoustic echo canceller (AEC)

The AEC rejects the playback signal on the device's speaker to pick up the user's voice command. The known playback signal is fed as a reference signal to the AEC and essentially "cancels" the speaker signal from the microphone. Current state-of-the-art AEC designs are able to cancel out 30-35 decibels of playback signal; with low distortion speakers, it's possible to cancel up to 40 decibels. The more echo cancellation, the better voice-activation performance.

The AEC's primary purpose on a smart TVs is to cancel out the content playing on the TV while the user is giving voice commands. Because TVs usually have stereo playback systems, a stereo echo canceller is required to achieve optimal echo cancellation. The echo attenuation performance is cut in half when the stereo playback channels are downmixed to mono to cancel out the playback content.

Adaptive interference canceller (AIC)

Apart from cancelling its own playback content using a multichannel echo canceller, the AFE also must cancel out external noise from other sound sources in the room. That's why you must build with an AFE that has a robust interference canceller algorithm, which rejects interfering sounds that are difficult to cancel out with a traditional beamformer (e.g., loud music playing on the sound bar in the living room or microwave noise in the kitchen).

The AIC algorithm we've developed at DSP Concepts doesn't require a reference signal to cancel out the interfering noises. Instead, it uses a combination of beamforming, adaptive signal processing, and machine learning to cancel out up to 30 dB of interference noise while also preserving the desired speech signal.

Wake-word detection

The resulting clean voice from the AFE is compared to a wake-word utterance, such as "Alexa," to detect the presence of a wake word. A wake-word detection algorithm is usually a machine-learning model, the size of which also impacts its accuracy. For example, a 1-MB model that's been trained on a significantly large amount of training data will be more accurate than a 64-kB model, but it also consumes more CPU resources. A 1-MB Alexa wake-word engine running on an Arm Cortex-A53 instance will typically consume 150 MHz of CPU resource. A large wake-word model is usually required to reduce the false alarms and accurately detect the wake word.

Smart-TV OEMs traditionally used their own custom wake words such as Bixby for Samsung TVs' wake word. With the proliferation of Amazon Alexa and the Google Voice Assistant, OEMs are increasingly providing users with the option to select between multiple wake words. Typically, these dual wake words are required to run in parallel to provide the best detection performance, which will also double the CPU resource usage.

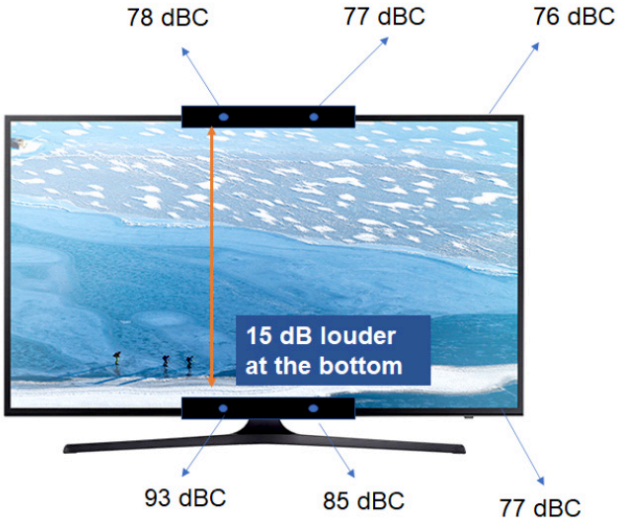
Playback Processing Joins the Party

Playback processing is another critical element of voice control for TVs, with major TV OEMs often including stereo playback speakers within the TV itself as well as a subwoofer for mid- to high-end TVs. The speaker output obtained from the playback-processing blocks can substantially impact the voice-activation performance. For example, speech-recognition performance can be substantially affected if the playback content on the speakers introduces significant nonlinear distortions that are hard to cancel out by the AEC algorithm.

Some basic playback modules, such as speaker equalizers, can compensate for non-flat frequency response, and crossover filters are able to split audio into low- and high-frequency signals for woofer and tweeter playback on smart TVs. Aside from those, the following processing blocks are the key to delivering a sophisticated playback experience:

- *Multiband compressor/Peak limiter:* A compressor and peak limiter allow the speakers to play loudly without distorting or compromising voice-recognition performance.
- *Volume management:* Have you ever been watching a show at a comfortable volume, only to have the sound suddenly start blaring when it cuts to a commercial? You're not hearing things—this happens because the playback content on the TV can have a wide dynamic range in sound pressure levels and can thus vary greatly between different types of content. In addition, TVs also provide multiple modes of operation, such as music, movie, and voice mode, where the permitted dynamic range is preset. This requires an intelligent volume-management algorithm that can normalize the dynamic range of the content to provide a uniform listening experience.
- *Dialog enhancement:* Apart from volume management, spoken content on the TV needs to be intelligible across different playback modes. Dialog enhancement works with mixed content like movies and TV programs to improve the perceived intelligibility of speech content.

Playback processing and Voice UI must be tuned in parallel to ensure optimum performance on voice-activated TVs. It's recommended to monitor the AEC performance while tuning the playback path to achieve the best of both playback and voice-activation performance.



2. Sound levels are measured at the top and bottom of this TV screen. A microphone array at the bottom center measures 15 dB more sound level compared to the top.

Overcoming Far-Field Voice-Implementation Constraints

Despite the large form factor of TVs, implementing voice activation on these devices comes with its own challenges. The flat-panel display and the thin screen size of today's TVs limit the amount of components that can go inside, as well as how they can be configured. Some of these challenges and ways to mitigate the performance degradation of Voice UI on TVs include:

Optimum placement of microphone array for far-field voice in TVs

Placement of the microphone array on flat-screen TVs is a tradeoff between performance and cost. The stereo speakers are typically located at the bottom of the TV. Therefore, a microphone array mounted on the top of the flat screen provides the best echo-canceller performance given that direct coupling of speaker-to-microphone signals is minimized.

Figure 2 shows a measurement taken at DSP Concepts with a sound-level meter. We measured the sound levels to be 15 dB more at the bottom of the screen compared to the top. Therefore, the AEC must operate 15 dB better with the bottom array to provide the same barge-in performance as at the top array. The quieter the sound, the better the AEC performance.

However, the top array requires longer cables to run from the main SoC closer to the bottom of the screen, which increases the BOM costs. We recommend considering this cost vs. performance tradeoff in the microphone-array design and location by accurately measuring the sound levels at each desired mic location.

Eliminating external interference noises

The area around the TV in a typical living room is crowded with other home audio entertainment devices, such as sound bars and external surround speakers that are still connected to the TV. However, they act as interfering noise sources for far-field voice activation.

For example, the playback content on the sound bar could be originating from the TV via the HDMI connector, but the TV has no control over the latency of the playback path to provide efficient echo cancellation. Thus, the playback content becomes a noise distractor to the TV that must be efficiently cancelled out by noise-suppression algorithms. In this case, a device such as DSP Concepts' Adaptive Interference canceller could be employed to cancel the interfering noise without impacting the wake-word detection performance.

Energy Star requirements

Smart TVs with the latest must-have features like ultra-high resolution, internet connectivity for streaming applications, and high dynamic range displays are required to meet Energy Star standards for power consumption. Energy Star-certified TVs must consume less than 0.5 W in sleep mode, and "on" mode power requirements vary according to screen area. The product must consume no more than 3 W while in standby-active, low mode.[1]

The standby-power requirement is particularly important for voice-control implementations and can cause some design limitations. In the presence of an external DSP that can perform voice-activation detection (VAD), the main SoC can be turned off in standby mode and activated by an interrupt signal upon voice detection. In designs that include a single multicore SoC, the chip can switch between different power modes.

For example, DSP Concepts' implementation of far-field algorithms on the NXP RT685 SoC consume less than 30 mW while in always-listening mode. This low -ower implementation includes 2-mic processing together with AIC, SCNR, and a wake-word engine.

What to Expect Next

Voice control is quickly becoming expected in the living room, but it has to work well and consistently for consumers to truly embrace it. Understanding the nuances of adding voice to TVs is critical for brands hoping to drive widespread adoption and increase their market share—even for the most experienced companies in the space. We expect to see a range of TVs with embedded voice-control hit the market within the next year, but which will deliver an experience that consumers consider worth buying remains to be seen.

Chin Beckmann co-founded and has led DSP Concepts since its inception in 2003. Her leadership has resulted in Audio Weaver, a platform-disrupting audio product development, and she has raised more than $25 million of venture financing. Under her direction, Audio Weaver has continued to solve the most

difficult audio-processing problems, resulting in design wins among Tier 1 Automotive and Consumer OEMs globally.

Prior to DSP Concepts, Chin held software engineering roles at Bose, Proteon, and Data General. Chin holds a BS in Electrical Engineering from Boston University and an MBA from Northeastern University. She is also pianist for the California Pops Orchestra and fluent in English, Spanish, and Chinese.

Reference

1. https://www.energystar.gov/products/electronics/televisions/key_product_criteria