Improving Barge-in Performance on Smart Speakers with Ultra High Dynamic Range Microphone

Application Note
Rev1.0

Author: Tung Shen Chew
Udaynag Pisipati

Vesper Technologies Inc.
77 Summer St, 8th floor
Boston, MA - 02110

AN7-Rev1.0
10/14/2019

Table of Contents

Table of Figures

Introduction

Smart Speakers, as one would expect, are meant to be smart enough to respond to user queries while also delivering high quality audio on the speaker reproducing the entire audible spectrum including the low frequency bass response. But, why is the smart speaker market fragmented between products that offer the smarts such as Echo Dot and Google Home Mini and those that offer high fidelity audio such as the recently launched Echo Studio, Sonos One or Apple Homepod? For a smart speaker to be able to respond accurately to a voice command during high fidelity audio playback, a user scenario called music barge-in, it must completely cancel out the music playing back from its own speakers. If the speaker does not cancel out the music completely, its ability to hear the voice commands will suffer. Playback quality and voice recognition accuracy of smart speaker is a tradeoff between the following factors –

1) Form-factor of the device that defines the size and number of the woofers and tweeters that can be embedded inside the device - a woofer/tweeter in a small enclosure can produce more non-linear distortions in low frequencies due to enclosure reflections.
2) Quality of the Acoustic Echo Cancellation (AEC) algorithms to cancel out the playback signal from the speaker. If the AEC algorithms cannot completely cancel out the non-linear distortions, the voice response accuracy during playback is degraded.

For good barge-in performance, the full duplex Acoustic Echo canceller used in the system is required to adaptively cancel out the loud music playing on the device while preserving the speech signal in the voice command. AEC task becomes more complex as the numbers of speakers and microphones increases. Choosing high quality transducers can reduce the nonlinear distortions in the speaker output increasing the effectiveness of AEC algorithms. However, this also increases the Bill of Materials (BOM) of the system. Therefore, most smart speaker designs tradeoff the bass response on the speakers to

enable good user experience with voice commands. This tradeoff is more prominent when the speakers are playing music at loud volume levels and high bass, simultaneously with voice commands. Fortunately, the known playback signal from the speaker can be subtracted from the signal captured by the microphone before passing it on to the echo canceller. An example of such a system is shown in Figure 1 below.

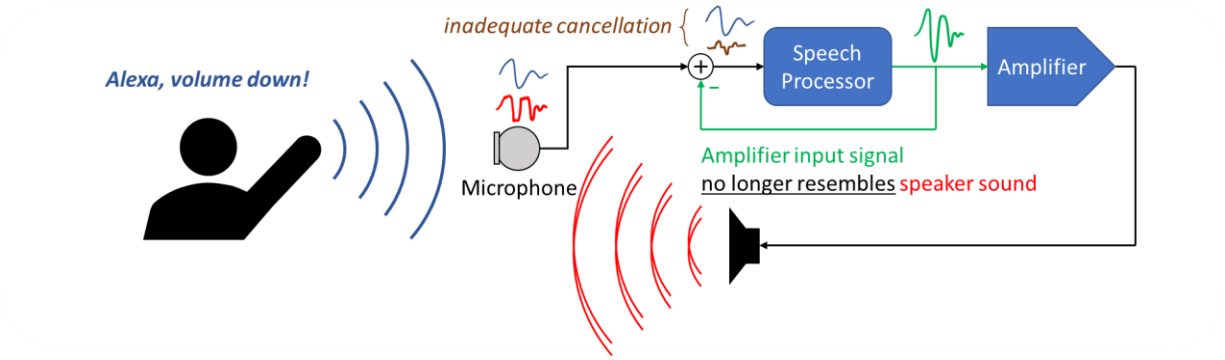

Figure 1: Block diagram of a typical smart speaker

At lower volumes where the speaker distortion is low, this method works well. However, at loud volume levels, the amplifier input signal no longer resembles speaker sound due to the non-linearities in the playback signal. Strong low frequency components saturate the microphones used in proximity of the speaker. Therefore, current implementations must compromise between playback sound quality and voice accuracy or vice versa. To achieve better far-field accuracy in music barge-in scenario, speaker gains or the low frequency bass response is traded-off.

What if we place a reference microphone directly in front of the speaker inside the enclosure and subtract its signal from the main mics to improve barge-in performance. This should allow the speaker hardware to playback the music at loud volume levels while the signal from the microphone array on the device capture the voice command. The reference microphone creates an acoustic feedback path that helps to monitor the envelope of the amplifier output signal. As the amplitude of this envelope increases at

high volume settings, the adaptive echo cancellation algorithms can efficiently cancel out the loud playback sounds. A typical block diagram of such a system is shown below in Figure 2.

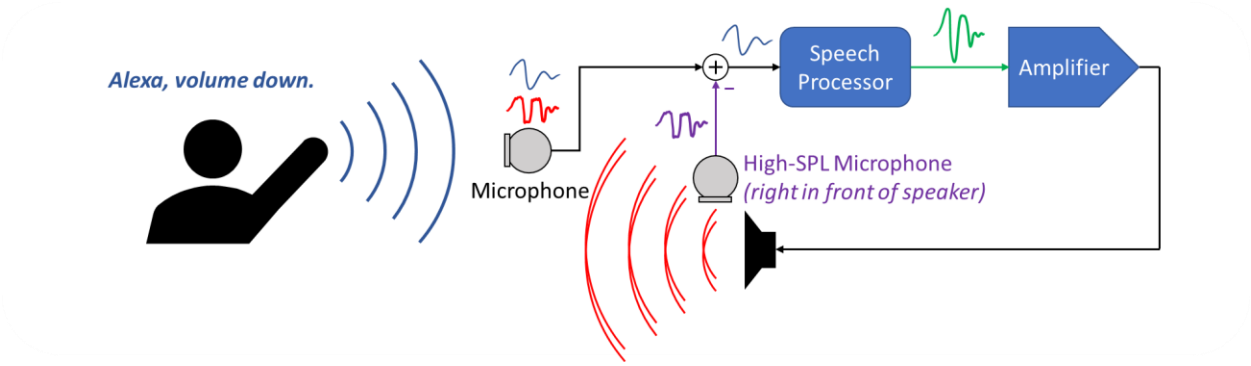

Figure 2: Block diagram of a smart speaker with reference microphone

The barge-in performance of the above solution is primarily dependent on the microphone that is listening to the speaker. Here are some of the characteristics of such a reference microphone

- The microphone must withstand high sound pressure levels, typically around 150dBSPL measured at the front of the speaker. In other words, the microphone should have a high Acoustic Overload Point (AOP) defined as the sound pressure level at which the microphone hits 10% Total Harmonic Distortion (THD)
- Microphone should also recover from the high SPL event fast enough that it can continuously monitor the playback signal
- Signal to Noise Ratio (SNR) of such a reference microphone can be traded off for higher dynamic range given the playback signal is loud enough at the location of the microphone.

Are there any MEMS microphones that can meet the above criteria? Capacitive MEMS microphones have the 10% distortion at an AOP of around 130 dB, which means the signals feeding from the speaker into the microphone will be clipped at these sound

pressure levels. Any changes to the stiffness of the diaphragm to achieve a higher AOP on a capacitive MEMS element will also significantly tradeoff SNR of the microphone. A piezoelectric MEMS element on the other hand can withstand sound pressure levels as high as 170 dBSPL. The linearity of piezoelectric material, therefore, is only limited by the ASIC voltage rail. Rest of the document describes the latest Piezoelectric MEMS microphone, VM2020, with ultra-high AOP and demonstrates the impact of lower distortion in the speaker design on the audio playback levels and music barge-in performance with the example of a reference speaker implementation from Vesper.

Vesper VM2020

VM2020 is an omni-directional differential piezoelectric MEMS microphone with durable construction and robustness to dust, water and other environmental particles while also resisting high SPL events. The differential output of the microphone allows for common mode noise rejection providing a low-noise signal to the ADC. A simple application circuit of VM2020 is shown in Figure 3. The differential outputs on the microphone can be AC coupled to the Analog to Digital Converter (ADC).

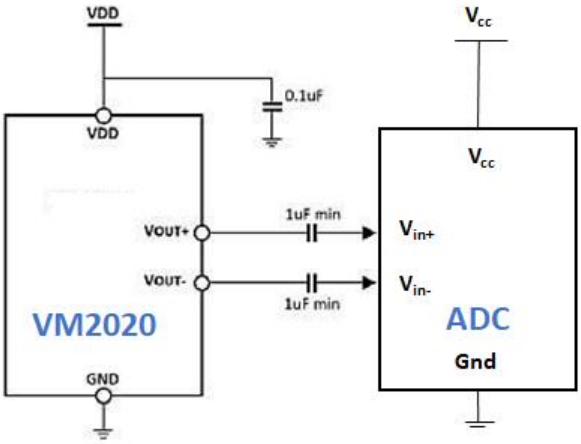

Figure 3: Application circuit of Vesper VM2020 ultra-high AOP microphone

The microphone has a high AOP with a 10% Total Harmonic Distortion (THD) point at 152 dB and a 1% THD at 149 dB. To put that in perspective, a microphone used directly in front of the speaker will experience sound levels in the range 130 – 150 dB depending on the volume setting. This means that the AOP on VM2020 can withstand as much sound level as the speaker can play at maximum loudness setting. This high dynamic range on the microphone captures the full range of sounds at high playback setting. In addition, Piezo microphones are immune to stiction and recover within 10 milliseconds after a high SPL event. Figure 4 shows Vesper Vs. Capacitive MEMS microphones when exposed to a 150 dBSPL event. It is evident that the Piezo microphone is linear at high sound levels and recovers back to normal state immediately after the high SPL event.

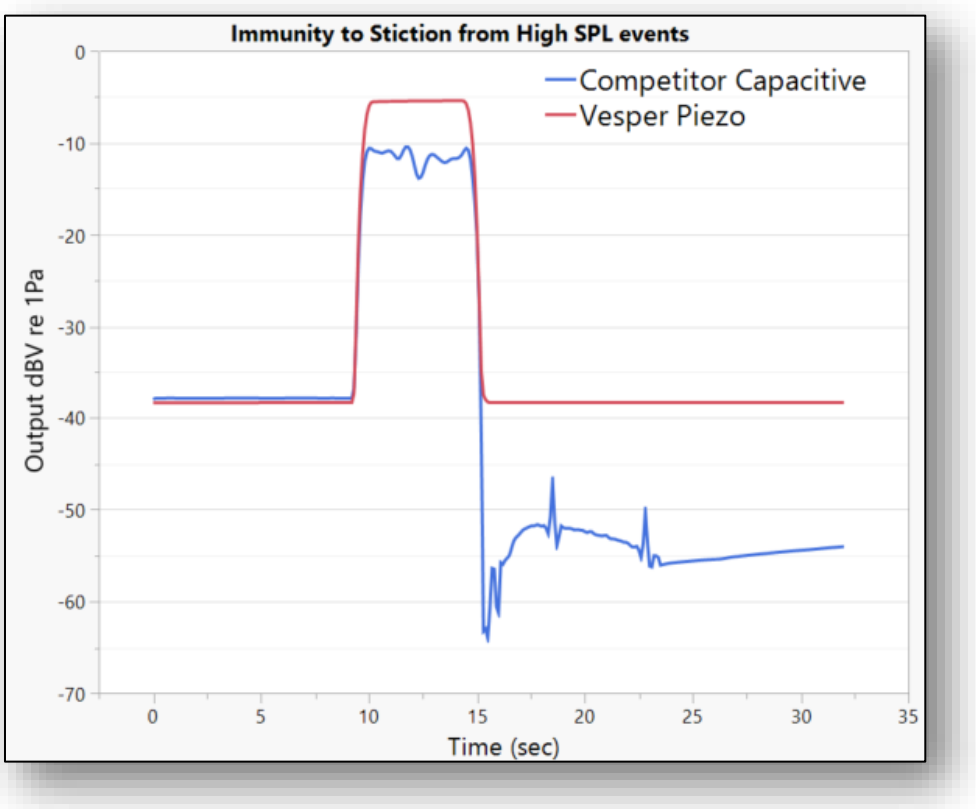

Figure 4: Stiction immunity of Piezoelectric MEMS microphone

Vesper Speaker Design

To help understand how VM2020 improves barge-in performance, Vesper built a reference smart speaker design that can be used in combination with any off-the shelf Raspberry Pi form factor evaluation kits. A picture of the speaker along with the components used in the design is shown in Figure 5.

- Speaker is designed in 2 variants - a) low excursion speaker optimized for cost with a short throw woofer and b) high excursion speaker optimized for bass response with a long throw woofer. High excursion speaker produces higher sound levels for a constant THD level when compared to low excursion design.
- VM2020 microphone is placed < 10mm in front of the woofer in the design
- The evaluation kit can be mounted on the top of the speaker attached to the microphone array board to drive the audio playback on the speaker
- The input capacitor to the ADC should not be too high that it filters out the low frequency components from the reference microphone. The value of the capacitor can be obtained by the formula below. For a 10 KΩ minimum input impedance (Z) of the PCM1865 ADC, to obtain a roll-off frequency of 50 Hz, a 0.15 uF capacitor should be used.

$$\text{Input Capacitor } (C_0) = 1 / (2*pi* f_{-3dB} * Z)$$

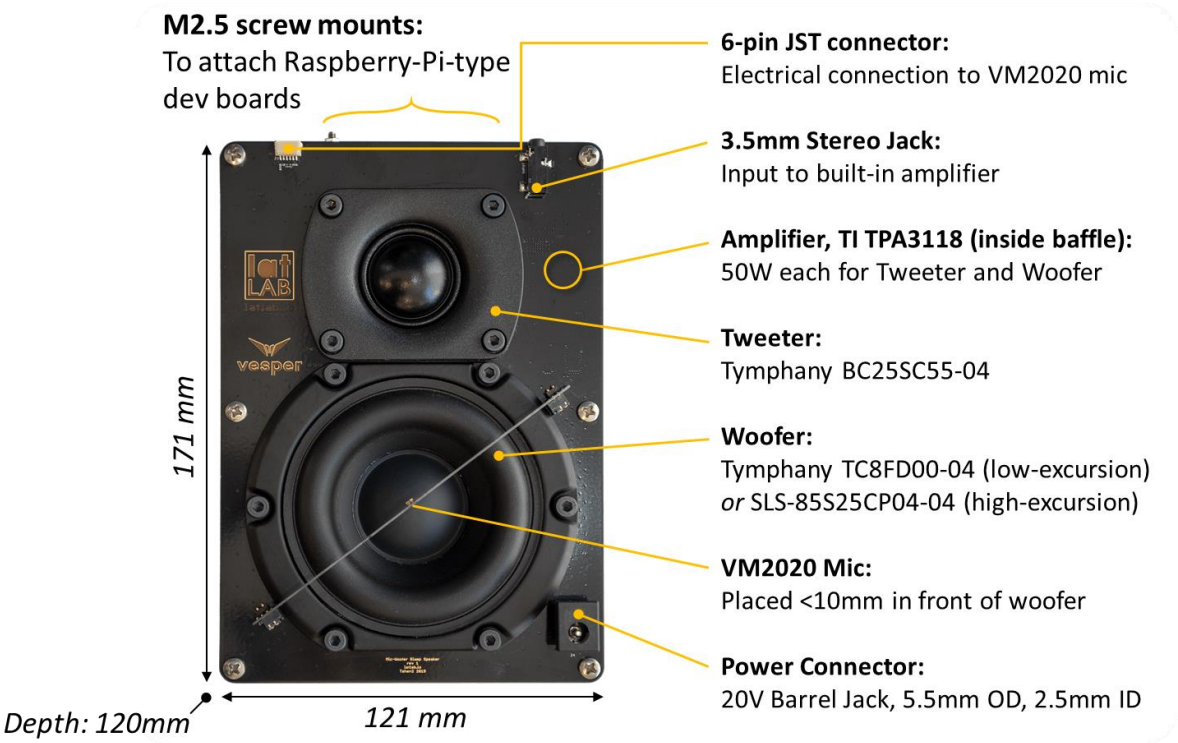

Figure 5: Vesper's reference smart speaker design

A simplified block diagram of the smart speaker design is also shown in Figure 6.

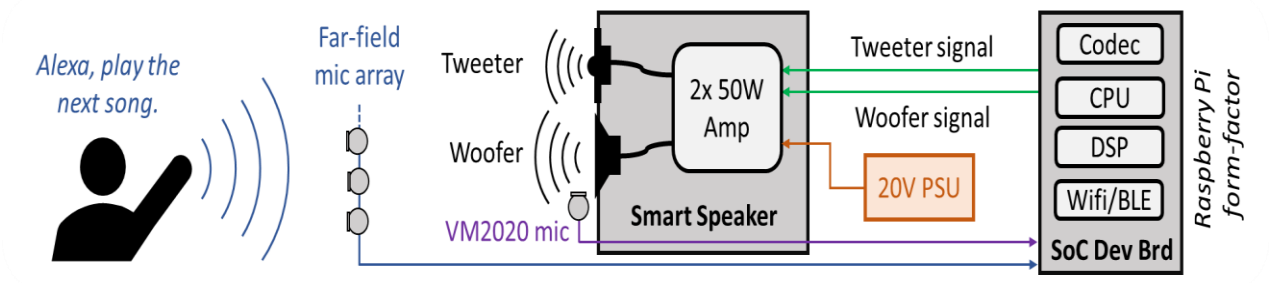

Figure 6: Simplified block diagram of the smart speaker system with VM2020

Smart Speakers must limit their sound output to keep distortion below **1%**. Otherwise, their barge-in performance will suffer. In order to demonstrate the impact of speaker

distortion on the maximum SPL that can be achieved at the output of the speaker, the max SPL Vs frequency at two distortion points - 1% and 10% THD for both designs are measured. Measurements are performed with sine tones at each frequency increasing the sound levels until the required distortion points are reached. The maximum SPL shown at each datapoint is the SPL at which the corresponding 1% or 10% distortion occurs for that frequency.

Figure 7 shows the max SPL curves for high excursion speaker.

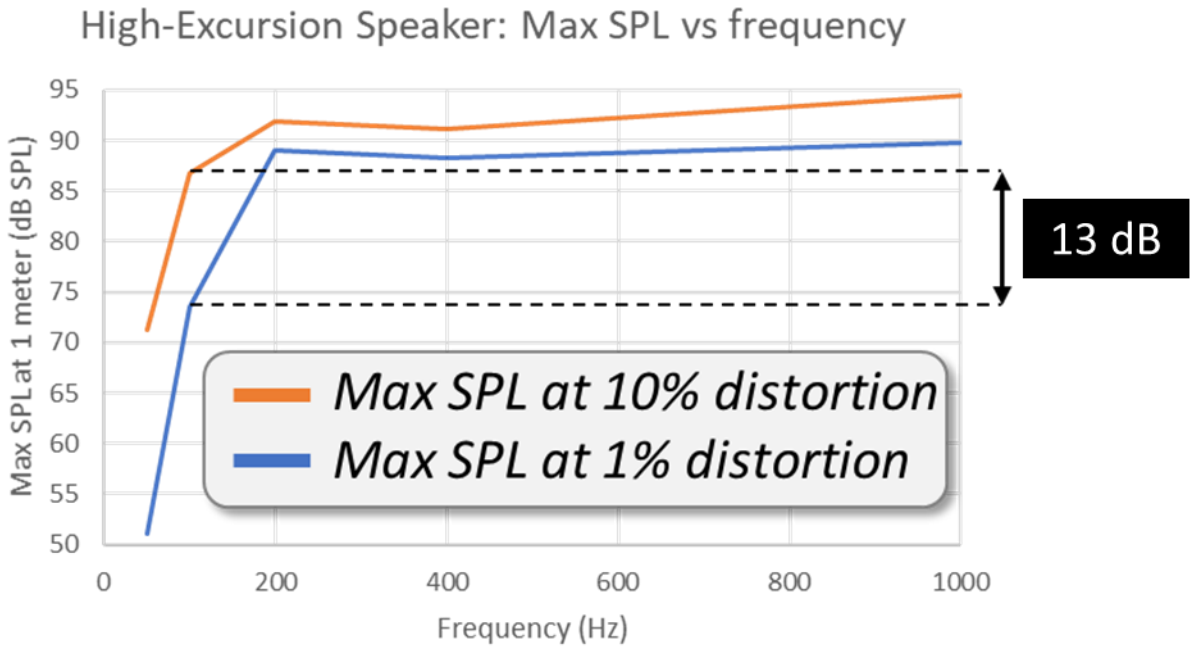

Figure 7: Max SPL Vs. Frequency for a high-excursion speaker design

At 100 Hz, 1% distortion occurs at 74 dB and 10% distortion is at 87 dB. If the speaker must maintain 1% distortion in the playback signal, it must attenuate 13 dB of the signal at low frequencies. Therefore, if a VM2020 were used in such a design, the high AOP on the microphone allows the smart speaker to play 13dB louder at these low frequencies

thereby preserving audio playback quality. Similarly, the results from low excursion design in

Figure 8 shows a difference of 20 dB in maximum SPL between 1% and 10% THD points. In this case, using VM2020 in the design, the speaker can preserve 20 dB of low frequency signal.

Note that the maximum SPL points shown above are specific to our component selection. In a typical smart speaker design, these SPL limits would be much higher and the difference between the 1% and 10% distortion point also depends on the quality and the number of woofers/tweeter used in the design.

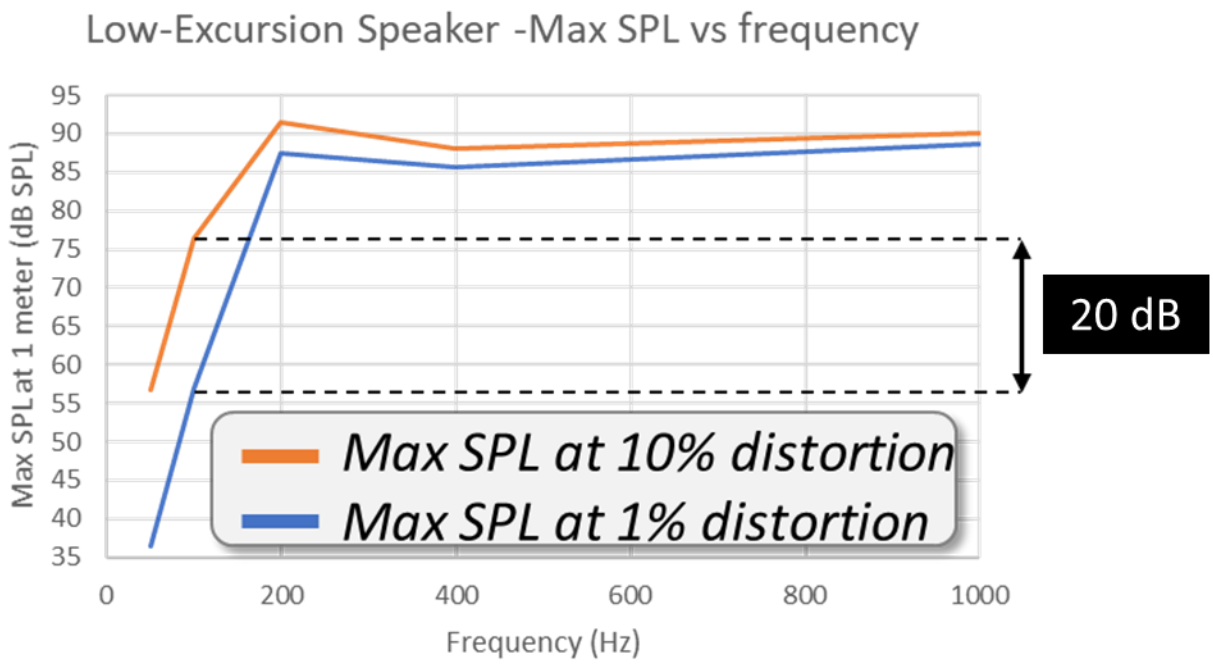

Figure 8: Max SPL Vs. Frequency for a low-excursion speaker design

These results indicate that higher tolerance in distortion of the playback signal leads to higher sound levels that can be produced out of the same hardware design. High SPL playback signals from the speaker can be captured by VM2020 without clipping the

microphone and the AEC algorithms then cancel out this playback signal to hear the wake word utterance from the user. Thus, VM2020 enables better audio fidelity and music barge-in performance.

A few additional insights on the speaker design are provided below

- The responses shown above are dependent on the selection of woofer, tweeter, amplifier as well as the acoustic enclosure design of the speaker.
- Different speaker designs can produce different curves but irrespective of the speaker design, it holds true that higher tolerance on distortion allows higher SPL levels to be produced out of the speaker.
- The low frequency response varies with the resonance frequency of the woofer. At frequencies below resonant frequency, the woofer will produce higher distortions as the sound pressure levels are increased. The woofers selected for this design have a resonance peak between 100 – 200 Hz.
- The sound levels at the location of the reference microphone vary based on the microphone placement. The ideal microphone position for VM2020 is in the front of the woofer where the sound levels are in the range 130-150 dB
- The analog signal captured by VM2020 is digitized using a Texas Instruments TI PCM1865 ADC with an SNR of 112 dB in this design. In general, at high volume settings, the playback signal is loud enough that a higher dynamic range ADC is not required.
- It is recommended to have the same ADCs processing the signals from high AOP microphone as well as the regular array microphones so there is no sampling offset between the regular microphone array and the reference microphone.

MediaTek Smart Speaker Reference Design

Vesper also partnered with MediaTek to build a reference design with VM2020 to expedite time to market for designers. MediaTek's Pumpkin Rich IoT Evaluation Kit with MT8362x chipset is a Raspberry pi form factor design with an add on daughter board for Vesper microphone array including VM2020. The reference design is also integrated with DSP Concepts Voice UI algorithms providing high end AEC and Interference cancellation algorithms for improved music barge-in performance. Additional details on the Rich IoT evaluation kit can be obtained from the MediaTek Website.

Conclusion

Music streaming is the most frequent as well as the most valued use case for voice assistants in smart speaker. But, the design tradeoff between playback quality and music barge-in is an implementation bottleneck resulting in poor user experience and market fragmentation. Vesper's VM2020 enables developers with a high dynamic range microphone that can be used to create products with great sound quality as well as reliable responses to user queries. Product designers no longer need to tradeoff between good voice interface and audio playback quality. The ultra-high AOP on the microphone combined with environmental robustness also offers unique advantages to products that require sound monitoring including smart speakers, soundbars as well as industrial machine monitoring products.

For additional information on Vesper's latest roadmap of microphone products, reach out to sales@vespermems.com.