Smart Speaker Acoustic Measurements

Audio Precision and GRAS Sound & Vibration
Application Note // By Joe Begin

APPLICATION NOTE

Introduction

Smart speakers are a relatively new class of consumer audio device with unique characteristics that make testing their audio performance difficult. In this application note, we provide an overview of smart speaker acoustic measurements with a focus on frequency response—the most important objective measurement of a device's audio quality.

Background

A smart speaker is an Internet-connected (usually wireless) powered speaker with built-in microphones that enables you to interact with an intelligent virtual assistant (IVA) using voice commands. Using voice only, you can direct it to perform tasks such as play audio content (news, music or podcasts, etc.) from the internet or a connected device, control home automation devices, or even order items from a connected online shopping service. Amazon was the first major company to release a smart speaker, called Echo, with an IVA known as Alexa, and it still has the dominant market share. Other significant entrants in this space include Alphabet (Google Assistant™ on Google Home™ smart speakers), Apple (Siri on HomePod mini speakers), Microsoft® (Cortana® on 3rd party speakers), Samsung® (Bixby® on Samsung Galaxy Home® speakers) and a few others from China, Japan and South Korea.

From humble beginnings when first introduced, smart speakers quickly skyrocketed in popularity. In November 2014, shortly after the release of the first-generation Amazon Echo, a popular technology review site said of the event, "… *the online retailer took another strange turn in the world of hardware by unveiling a weird wireless speaker with, Siri-like ability to recognize speech and answer questions.*" [1] But since then, the proliferation of smart speakers has exploded, and strong growth is projected to continue for at least the next five years. Within three years of the 2014 introduction, the USA alone had an installed base of 67 million smart speakers in households, and that number grew by 78% in 2018 to 118 million units. [2] The global smart speaker market was valued at USD 4.4 billion in 2017 and is projected to reach USD 23.3 billion by 2025. [3] Several companies with IVAs license the technology to other manufacturers—for example, both Bose and SONOS offer Alexa-enabled and Google Assistant-enabled smart speakers. And not only speakers are getting "smart"; IVA technology, with microphones and loudspeakers to support it, is being added to all sorts of devices like refrigerators, microwave ovens and set top boxes, enabling voice control of those devices. Additionally, most smartphones can also play the role of a smart speaker.

Smart Speaker IVAs

An interaction with a smart speaker begins with a specific "wake word" or phrase, for example, "Alexa" for Amazon, "Hey Siri" for Apple, etc., followed by a command. In their normal operating mode, smart speakers are in a semi-dormant state, but are always "listening" for the wake word, which triggers them to acquire and process a spoken command.

In terms of speech recognition, smart speakers themselves are only capable of recognizing the wake word (or phrase). The more computationally intensive speech recognition and subsequent processing is done by the IVA on a connected server. [4] The IVA converts the user's speech to text and attempts to interpret the command. To invoke the requested response from the device, the spoken command must contain a sequence of keywords recognizable by the IVA. A successful interaction may result in the requested action being taken by the IVA (e.g., "Set a timer for 10 minutes.") or by a connected Web service (e.g., "Play an Internet radio station.").

Audio Subsystems

Smart speakers contain several distinct audio subsystems, including:

- A Microphone array
- A powered (active) loudspeaker system
- Front-end signal processing algorithms for tasks such as beamforming, acoustic echo cancellation and noise suppression

An array of microphones is used instead of a single microphone to enable the device to take advantage of beamforming, a signal processing technique that can effectively increase the signal-to-noise ratio of the speech signal sent to the IVA for processing. [5] Based on correlations among signals received at different microphones in the array, a beamforming algorithm can detect the most likely direction of the talker in a room and, in a sense, focus on that direction by combining the various microphone signals in a way that attenuates signals coming from other directions. This can effectively reduce the level of ambient noise and room reverberation in the speech signal sent to the IVA. Noise suppression may also be used to reduce the level of non-speech-like signals.

Ideally, a smart speaker will be able to respond to spoken commands (by first recognizing the wake word) even while playing audio content such as music or speech in a room. Acoustic echo cancellation (AEC) is essential for preventing the loudspeaker output from completely masking the microphone input for this task. The signal being played on the loudspeaker system can be used as a reference signal for the AEC algorithm, enabling it to ignore the content being played and to recognize the wake word. Typically, playback is paused after the wake word is detected in order to help improve command recognition.

Audio Signal Paths

The primary audio paths for a smart speaker are between the device and the IVA or a network server using the Internet with a Wi-Fi® or wired connection. On the input side, a speech signal containing a spoken command is sensed with the device's microphone array, digitized and then uploaded to the IVA for signal processing and command interpretation. On the output side, digital audio content is transmitted from a Web server to the device, where it is converted from digital to analog, then finally to an acoustic signal as it is played over the device's loudspeaker system.

In addition to the two primary paths above, smart speakers may have several other audio paths, including:

- An analog output jack for connecting to an external powered speaker system
- An analog input jack for using the smart speaker as a simple powered speaker
- Bluetooth® connection for playing audio content on an external Bluetooth speaker, streaming content from a smartphone or tablet as a music source, or in some cases acting as a handsfree device for telephone calls
- Network connections to other smart speakers for multi-room music, stereo pairing or intercom functionality
- Connections to home automation devices, e.g., for two-way intercom connection to a security device, or audible status messages

Audio Testing

The audio subsystems of smart speakers have a multitude of components that contribute to overall performance and audio quality, including microphones and microphone arrays, A/D and D/A converters, power amplifiers, loudspeaker drivers, digital signal processors, audio codecs, etc. In addition, several system-level functions such as beamforming, echo cancellation, wake word recognition, etc., contribute to overall quality. At some stage, each of these components and systems must be tested. Testing end-to-end performance of an overall smart speaker system is also desirable.

Different test contexts—R&D, validation, production test, quality assurance—have different goals and different levels of access to subsystems and components. For example, during product design, R&D engineers might well be able to isolate the active crossover functionality of a system on a chip (SOC) by physically tapping into chip level connections (and have the first-hand product knowledge to be able to use the resulting signals). Similarly, for production test, manufacturers have the option of temporarily loading special test-specific firmware into the device to enable functional tests which are not available in off-the-shelf units. For example, noise reduction could be disabled allowing the microphone input system to be tested with sinusoidal signals instead of speech.

Testing the overall end-to-end performance of a smart speaker's primary input and output audio paths can be quite challenging for the following reasons:

1. Input to and output from a smart speaker are both acoustic, and acoustic test is by its nature more complex than electronic (analog or digital) audio test. Acoustic tests require calibrated microphones, usually an anechoic test chamber, and a quality loudspeaker system to stimulate DUT microphones.
2. Smart speakers are inherently open-loop devices. On the input side, a signal (typically speech) is captured, digitized and then transmitted to a server somewhere as a digital audio file. To assess the input path performance, the audio file must be retrieved from the server and analyzed in comparison to the signal that was generated in the first place. On the output side, audio content that originates as an audio file on a server is streamed to the device where it is converted to analog and played on the device's loudspeaker system. To assess the output path performance, the device's loudspeaker output must be measured with a measurement microphone and compared with the original signal from the server. The original signal is often in the form of an encoded audio signal (e.g., MP3 or AAC), which requires that it be decoded before analysis.

3. The A/D and D/A converters in the device will invariably have different sample rates than the audio analyzer, requiring some form of compensation during analysis. [6]

Measuring Frequency Response

The most important aspect of the performance of any audio device is its frequency response. [7] Frequency response is a type of "transfer function" measurement. For a device under test (DUT), it represents the magnitude and phase of the output from the DUT per unit input, as a function of frequency. Devices are often compared in terms of the "shape" of their frequency response curves, which typically refers to the magnitude response only (not phase), and in addition normalizes the magnitude to a reference value. For example, the response magnitude might be normalized to its value at some reference frequency, say 1 kHz, such that the normalized curve passes through 0 dB at 1 kHz. Usually, a flat frequency response (constant response magnitude versus frequency) is desirable in audio systems to ensure that source material is faithfully recorded and reproduced without spectral coloration. Flat frequency response is quite achievable in electronic audio systems, but much more difficult in acoustic devices, especially loudspeakers.

For loudspeakers, frequency response is also the basis from which another important metric, sensitivity, is derived. Sensitivity is a measure of the output from the device per unit input. For example, a loudspeaker system with a specification of "94 dBSPL, 1 W/1 m" indicates that for an input of 1 W, it will output a sound pressure level of 94 dB at a distance of 1 m. For loudspeakers, sensitivity is calculated by averaging the frequency response magnitude over a range of frequencies, because they typically do not have flat frequency response.

Because frequency response is such an important audio quality metric, audio analyzers have several different ways of measuring it. For example, some of the measurements available in Audio Precision audio analyzers that could be considered for smart speaker testing include:

- Stepped sine sweep
- Logarithmically-swept sine (chirp)
- Multitone
- Transfer function

Each measurement technique has certain advantages and disadvantages with respect to measuring the primary input and output paths of a smart speaker.

Stepped Sine Sweep

Stepped sine testing is the classic means of measuring frequency response that has been in use since the earliest days of audio test, even before audio analyzers existed. It involves testing a DUT by "sweeping" over a series of discrete frequency steps within the frequency range of interest—usually some portion of the audible frequency range, considered to be 20 Hz to 20 kHz. At each frequency step, the device is stimulated with a sine wave, and its output is analyzed to determine metrics such as level and harmonic distortion.

In general, one advantage of stepped sine testing over other techniques is that because of its long history of use, it is often considered to be a type of "gold standard" against which other techniques may be compared if there is any doubt about measurement integrity. Another advantage is that other audio quality metrics such as total harmonic distortion (THD) and inter-channel phase can be measured at the same time as frequency response.

A general disadvantage of stepped sine testing is test time. Devices often have a transient response when the stimulus frequency is changed abruptly, requiring more time at each step for the device to settle to its steady-state behavior. Another related disadvantage is poor frequency resolution. Because the frequency steps are discrete, higher resolution requires more steps, which increases test time.

For acoustic devices, one of the biggest disadvantages of stepped sine testing is that a test environment free of reflections (usually an anechoic chamber) is required. An ordinary room with reflective surfaces can't be used because reflected sound waves will combine (either constructively or destructively) with the direct sound waves of interest, causing severe measurement errors.

For testing the input path of a smart speaker, stepped sine testing has two additional disadvantages:

1. Smart speakers are designed to capture and process speech signals. It is therefore quite likely that they will have digital signal processing (DSP) designed to attenuate sinusoidal signals. It might be difficult to effectively test this path with a stepped sine signal unless this feature can be disabled for the test.
2. When conducting an end-to-end test, there is a limit to the length of signal the device will record when attempting to process speech commands (e.g., approximately 7 seconds, including the keyword, in the case of the Alexa IVA service). In this case, a stepped sine stimulus would have to be less than about 6 seconds.

Stepped sine testing works well for testing the output path of a smart speaker, provided the test environment is anechoic within the frequency range of interest.

Logarithmically Swept Sine

The log-swept sine technique (sometimes called chirp or continuous sweep) was introduced in 2000. [8] It uses a type of sinusoidal stimulus in which the frequency is continuously swept logarithmically from low to high frequency in a short time (from a fraction of a second to a few seconds). Logarithmically swept sine testing is very popular in audio test because it can be used to derive a number of audio quality metrics, including THD and individual components of harmonic distortion with a short test time.

Log-swept sine testing has some additional advantages for acoustic measurements:

1. Because the frequency response is derived from the impulse response, the analysis can make use of what is known as a "time selective" or "quasi-anechoic" technique. Before calculating the frequency response, the end of the impulse response is windowed to remove the effect of room reflections. Although this reduces the accuracy of the measurement at lower frequencies, if used carefully, it can enable accurate measurements of the mid to upper frequency range response of an acoustic device without the need for an expensive anechoic chamber.
2. In Audio Precision's APx500 audio analyzers, the log swept sine measurement designed for acoustic testing, (known as "acoustic response") can be used to detect rub and buzz defects in loudspeakers.

The log swept sine technique can be used to test the input path of a smart speaker. However, the signal is still a type of sinusoid, so care should be taken when testing devices in which the DSP intended to block pure tones can't be disabled. For example, it might be necessary to use a short sweep, so the system does not have enough time to recognize and react to the fact that a sinusoidal signal is present.

With its quasi-anechoic capability and optional ability to detect rub and buzz, the log swept sine is an ideal stimulus for testing the output path of a smart speaker.

Multitone

As the name implies, a multitone signal is comprised of a series of sine waves (tones) at discrete frequencies combined together. The phase of each tone is often adjusted to reduce the signal's crest factor (the ratio of its peak to rms level). To measure frequency response, the tones are usually spaced logarithmically in frequency. In AP audio analyzers, the multitone analysis measurement uses a DSP technique which enables it to measure distortion* and noise simultaneously with frequency response. One of the unique features of multitone analysis is that it enables measuring noise in the presence of signal.

Advantages of multitone analysis include:

1. Multitone testing is fast and delivers a large number of audio quality metrics with a single measurement.
2. Each multitone has a unique "signature", enabling the analyzer to trigger accurately on the signal when noise or other signals are present.
3. When a reasonable number of tones are included, the signal becomes different enough from a pure tone that it is unlikely to be blocked by a noise suppression algorithm.

Some disadvantages to multitone analysis are:

1. Because all the frequencies are stimulated simultaneously, the signal energy level at each frequency is lower than it would be for a pure sine or chirp signal. Stated another way, the crest factor of the signal is much higher† than a sinusoid.
2. Similar to a stepped sine, a multitone typically has limited frequency resolution, and an anechoic chamber is required for acoustic tests.

In a free-field environment, such as an anechoic chamber, multitone analysis works well for testing both the main input path and output path of a smart speaker. One caveat is that on the input side, enough tones should be used that a noise cancellation algorithm will not block the signal. On the output side, from the perspective of the IVA, a multitone signal is just another music track for the device to play.

* ***Note:*** *The distortion metric derived from multitone analysis is referred to as "Total Distortion" rather than THD, because there is no way to distinguish between harmonic distortion and intermodulation distortion.*

† ***Editor's note:*** *The initial publication stated "lower"; it has been corrected to "higher".*

Transfer Function

Transfer function analysis, sometimes referred to as "dual-channel FFT analysis" or "dynamic signal analysis", involves stimulating a DUT with a broadband* signal (such as noise, music or speech) and acquiring the output from the DUT. The transfer function (i.e., frequency response) is then derived from the input and output signals using a mathematical technique known as the complex discrete Fourier transform (DFT). The term "complex" in this context refers to the fact that the frequency response includes both magnitude and phase. A byproduct of the analysis is a result known as the Coherence function—a function with a value between 0 and 1 at each frequency that indicates the degree to which the output signal is coherent with (i.e., related to, or caused by) the input signal.

Most analyzers require that you use one of the analyzer input channels to measure the input to the DUT as well as its output. In AP's APx500 audio analyzers, the stimulus signal can be taken from the generator signal or from a file on disk, freeing up an input channel for additional measurements. The APx measurement has another feature that is especially useful for open loop measurements: the ability to trigger on the signal of interest using a signal matching algorithm based on cross correlation.

The main advantage of transfer function analysis over other frequency response measurement techniques is that it can use any broadband signal. This makes it an excellent choice for testing the input path of a smart speaker, because a speech or speech-like signal can be used as the stimulus, virtually ensuring that the signal will be acquired and processed unaltered by any noise cancellation algorithms. Transfer function analysis can also be used for the output path of a smart speaker, using music or noise as a stimulus. An anechoic chamber should also be used when testing both the input path and output path of a smart speaker, to avoid reflections.

* *We use the term broadband to refer to a signal that contains energy at all frequencies within a certain frequency range.*

Practical Considerations for end-to-end Smart Speaker Testing

Next, we consider some of the practical aspects of end-to-end testing smart speakers. The two primary audio paths are quite different and will be covered separately.

When conducting audio tests, it's generally a good practice to adhere to industry standards or to at least use them as a guideline. Smart speakers are so new that there are currently no industry standards for testing them. However, in terms of form and function, a smart speaker is similar to a speakerphone and there are several national and international standards which focus on speakerphone tests such as IEEE 1329. [9] These standards can be used as a guideline when testing smart speakers.

Smart Speaker Input Path

As mentioned above, for the primary input path of a smart speaker, a speech (or speech-like) signal is sensed with the device's microphone array, digitized and uploaded to the IVA for signal processing and command interpretation. An audio test of this path involves "tricking" the device to acquire and save a test signal, retrieving the recorded signal from the back-end server, and comparing it to the original test signal.

Physical Test Setup

Figure 1 shows a test setup which is based on IEEE standard 1329. The standard specifies that speakerphones be tested on a tabletop approximately 1 m wide by 1 m long (39″ x 39″) in an anechoic chamber or in a simulated free field (using time-selective measurement techniques). For the send direction (the equivalent of the smart speaker input path), the device is positioned 40 cm (15.7″) from the edge of the table, with a mouth simulator* located at the table edge and 30 cm (11.8″) above the table surface. In Figure 1, the mouth simulator is 30 cm (11.8″) above the top of the smart speaker rather than the table surface. This is because a smart speaker's microphone array is typically located on the top surface of the device, as opposed to speakerphones, which typically have microphone(s) located at the base of the device.

* *A mouth simulator is a special type of loudspeaker used to simulate speech in a precise and repeatable manner. For measurements where adherence to standards is not critical, a small, full-range (100 Hz to 8 kHz) single driver loudspeaker could be used instead of a mouth simulator.*

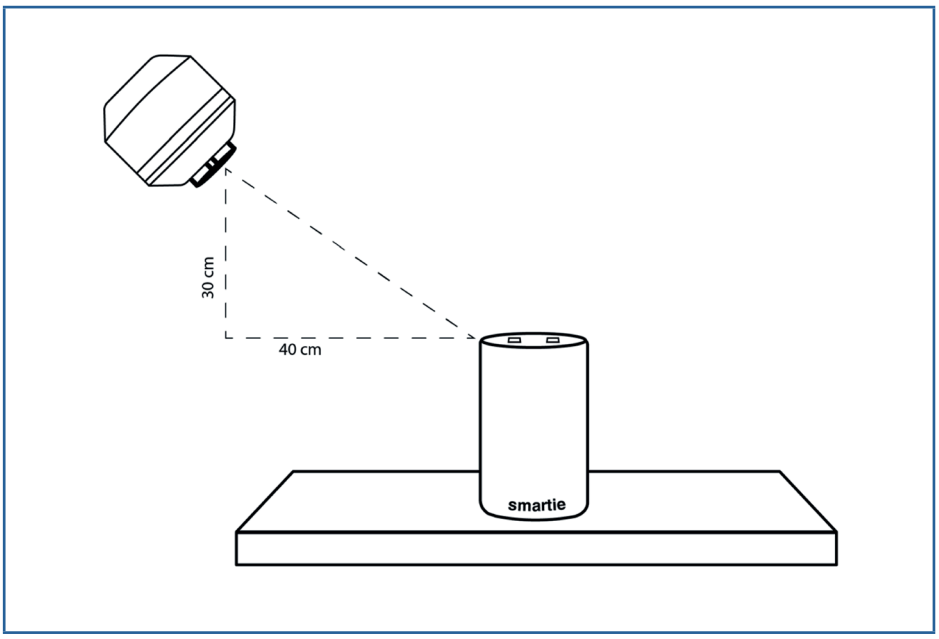

FIGURE 1.
Setup for smart speaker input path test.

Mouth Simulator Calibration

Like any loudspeaker, a mouth simulator does not have a flat frequency response; typical mouth simulators have a deviation from flatness of ±10 dB or more within the frequency range from 100 Hz to 8 kHz (the frequency range of male speech). For a mouth simulator to faithfully reproduce an audio signal such as speech, it must be equalized to have a flat response within this range. Measurements to equalize (and calibrate) a mouth simulator are typically made with a ¼" measurement microphone at a point called the Mouth Reference Point (MRP), centered on the mouth opening and located 25 mm (1") in front of the lip ring. To equalize a mouth simulator, its frequency response is measured, then inverted and applied as an EQ curve (a feature built into the audio analyzer). This results in the mouth simulator having a frequency response magnitude that is flat within a specified tolerance (e.g., ±0.5 dB from 100 Hz to 8 kHz).

The test signal should also be presented at a known level. IEEE 1329 specifies a level of −5 dBPa (89 dBSPL) at the MRP, which corresponds to a normal or nominal level of speech. If a speech signal is used, the level should be set using the "active speech level"—a level metric that ignores the silent periods between speech phrases and sentences. The input path of a smart speaker will likely have nonlinear processing such as automatic gain control (AGC). As such, it might be prudent to repeat the test at a few different levels. For speakerphone loudness ratings, IEEE 1329 specifies tests at levels from 79 to 99 dBSPL at the MRP in 5 dB steps.

Test Signal Sample Rate

In the analysis phase, the signal acquired by the smart speaker and uploaded to the IVA will be compared to the stimulus signal. This will require that the two signals have the same audio sample rate. For the input path, smart speakers typically use a sample rate of 16 kHz, which enables a bandwidth of just less than 8 kHz. This sample rate is used in "wideband speech" applications such as Voice over Internet Protocol (VoIP) and newer versions of the Bluetooth Hands-free Profile (HFP). It is referred to as wideband, because it has twice the bandwidth of ordinary digital telephone lines (just under 4 kHz), which enables clearer, more natural sounding speech.

If the test stimulus does not already exist, it makes sense to create it at the same sample rate (typically 16 kHz) as the DUT output. Audio analyzers can usually create a stimulus signal at a variety of sample rates. If the stimulus signal already exists, but is at a different sample rate, at some point it will need to be converted to the sample rate of the DUT output. This is easily accomplished with an audio waveform editing software package like Audacity® [10] (open source) or GoldWave® [11] (inexpensive).

Using the Wake Word

When testing the input path, the wake word must be used to activate the smart speaker, causing it to anticipate a spoken command and record a few seconds of audio. One option is for the test operator to just say the wake word, as if giving the device a command, and then immediately generate the test signal from the mouth simulator. This works, but the timing between the wake word and the test signal may vary from test to test.

Another option is to record a person saying the wake word and then prepend it to the stimulus signal using an audio waveform editor (i.e., insert it at the beginning of an audio file containing the stimulus). In this case, to conduct a test, the wake word with stimulus is simply played through the mouth simulator at the required level. With this method, the time between the wake word and the stimulus is constant from test to test. This option also works well when the DUT is inside a test chamber and the operator and audio analyzer are located outside the chamber.

Retrieving and Analyzing the DUT Output Signal

The process of activating the smart speaker with the wake word followed by the audio stimulus signal will trigger the DUT to record several seconds of audio and upload it to the IVA for speech recognition processing. To complete the analysis requires retrieving this recorded audio file from the back-end server, converting it to the .WAV format, and analyzing it with the audio analyzer control software. Details of this process will vary, depending on which IVA is used. For example, the popular Alexa service has a Web portal where users can log into their

account to manage interactions with connected smart devices. In the Settings menu under History, there is a record of each interaction with a connected smart device, including a date/time stamp, a transcription of the interpreted command and a tool to play the recorded audio on the PC's speakers. Using the Web developer mode features built into Web browsers, it is possible to retrieve a hyperlink that will enable you to download the recorded .WAV file directly. You can then open it in the audio analyzer software for analysis.

Example—Transfer Function Measurement of a Smart Speaker Input Path

This example illustrates a test of the input path of a smart speaker using the transfer function measurement with a stimulus waveform consisting of speech chatter. The signal, which was recorded at a night club with many people talking at the same time, is essentially random speech noise. As shown in Figure 2, the smart speaker wake word was prepended to the stimulus waveform. For the transfer function measurement, the system triggers on the wake word and includes it in the analysis with the rest of the stimulus signal.

The frequency response magnitude derived by retrieving the recorded .WAV file from the IVA's server and analyzing it with respect to the stimulus signal is shown in Figure 3. In this case the mouth simulator was driven at a level such that the rms level measured at the MRP was 89 dBSPL.

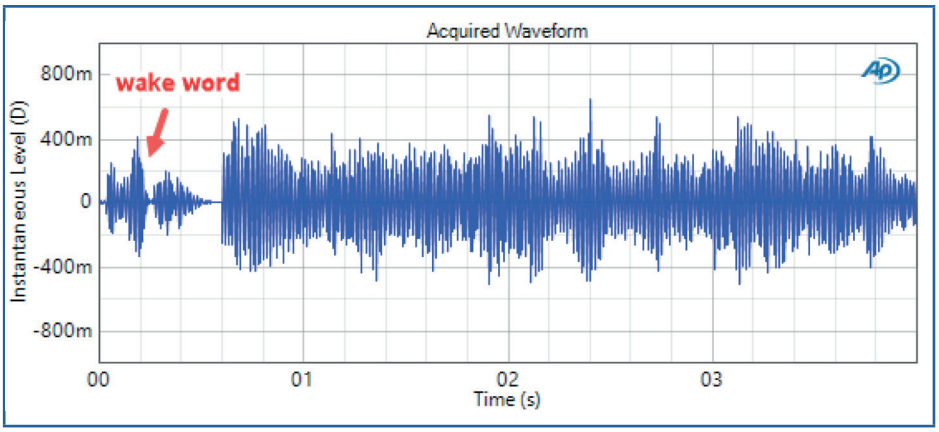

FIGURE 2.
Initial portion of the stimulus waveform (speech chatter with pre-pended wake word).

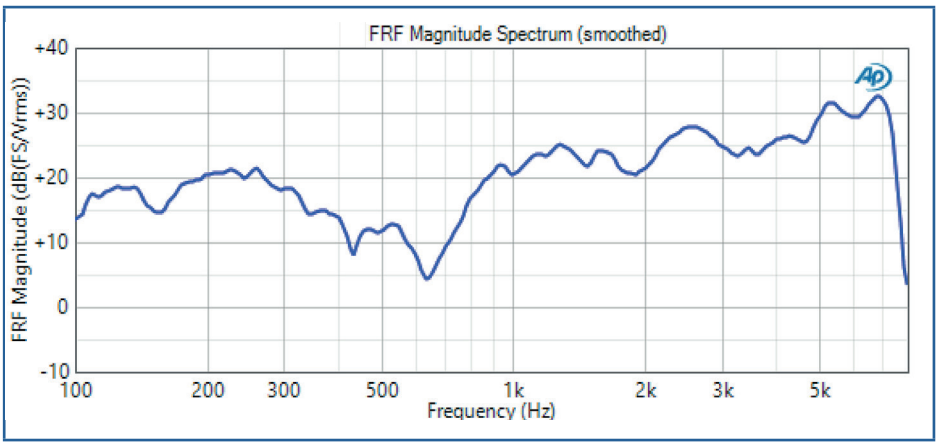

FIGURE 3.
Smart speaker measured frequency response magnitude using the stimulus of Figure 2.

Example—Log Swept Sine Measurement of Smart Speaker Input Path

This example features a test of the input path of the same smart speaker using a log swept sine stimulus. Figure 4 shows the stimulus waveform for a 0.35 s long sweep from 50 Hz to 8 kHz, with a pilot tone and the wake word inserted before the chirp signal. The analyzer triggers on the pilot tone and uses the pilot tone frequency to correct for the slight sample clock difference between the DUT and the analyzer. In this case, the wake word and pilot tone are not included in the analysis of the frequency response.

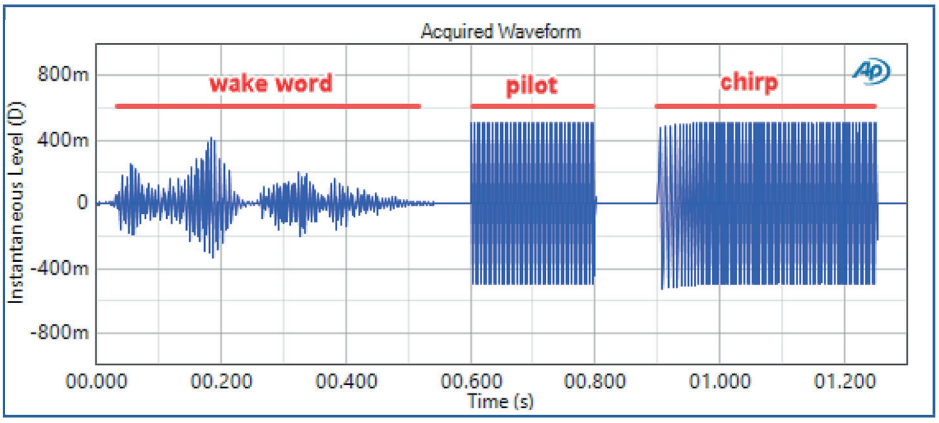

FIGURE 4.
Stimulus waveform for log-swept sine test of smart speaker (wake word, 200 ms pilot tone and 350 ms chirp, 50 Hz to 8 kHz).

Figure 5 shows the acoustic level response of the smart speaker input path for sweep lengths of 0.35, 1.0, and 4.0 s. Some interesting observations:

1. The response curves from the chirp analysis and the transfer function measurement with a speech signal, above, are remarkably similar in shape.
2. The response to the chirp stimulus does not vary much with the length of the sweep.

A second smart speaker we have tested behaves differently; the response measured with a speech signal is quite different than when measured with a chirp, and the chirp response changes significantly with sweep length. [12] This is most likely due to the second speaker having a more aggressive noise suppression algorithm than the one tested in this example.

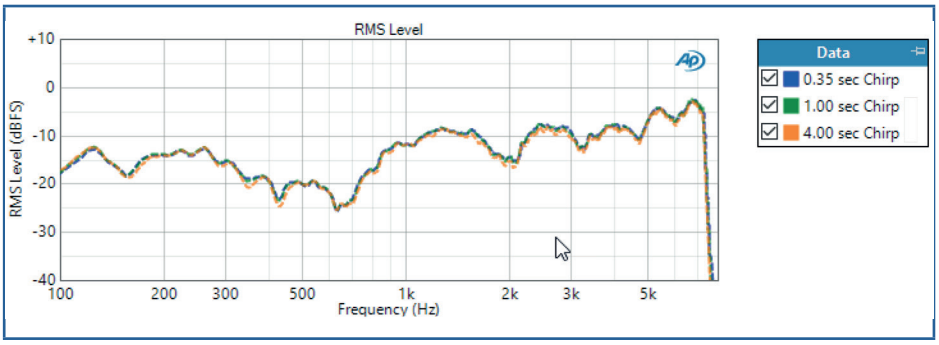

FIGURE 5.
Smart speaker input path level response due to chirp signals varying in length from 0.35 to 4.0 s.

Example—Multitone Measurement of Smart Speaker Input Path

The input path of the first smart speaker above was again measured with a multitone consisting of tones logarithmically spaced at standard 1/6-ocatve frequencies from 100 Hz to 8 kHz, with the overall signal length (the period of the lowest frequency tone) being 200 ms. The wake word was again prepended to the stimulus signal to activate the smart speaker. In the case of multitone measurements, the highly selective triggering mechanism of the measurement is used to trigger on the signal itself, and the wake word is excluded from the analysis.

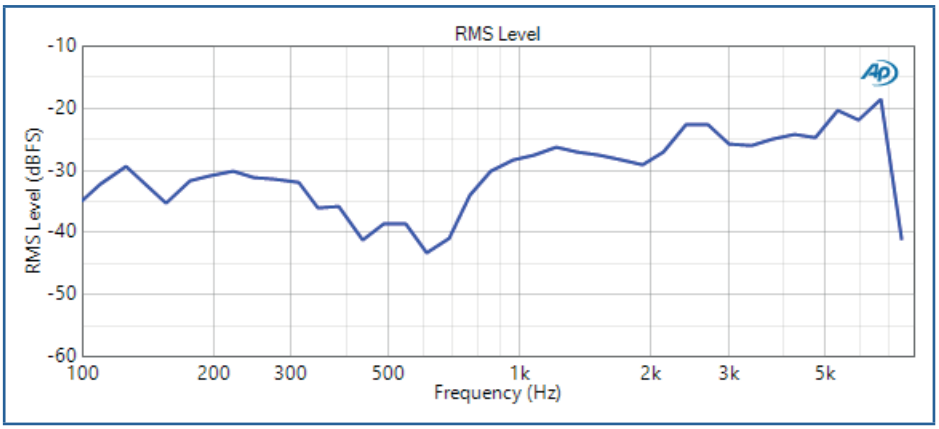

FIGURE 6.
Smart speaker input path level response measured with a multitone.

Figure 6 shows the response of the smart speaker input system to the multitone signal. Despite the lower resolution, the shape of the curve closely matches the curves from the transfer function and chirp measurements. Note, however, that the absolute level is about 17 dB lower than the level of the chirp response. This is due to the multitone signal having a much lower crest factor than the chirp signal.

Smart Speaker Output Path

The smart speaker primary output path involves digital audio content being transmitted from a Web server to the device, where it is converted from digital to analog, then finally to an acoustic signal as it is played over the device's loudspeaker system. An audio test of this path requires capturing the acoustic output signal from the device with a microphone connected to the input of an audio analyzer, and comparing this to the unaltered original audio file.

Physical Test Setup

Figure 2 shows the setup for a smart speaker output path, which is also based on using the IEEE 1329 standard for speakerphones as a guideline. In this case the mouth simulator is replaced with a measurement microphone. This test should also be conducted inside an anechoic chamber, unless a quasi-anechoic or (time-selective) measurement technique is used.

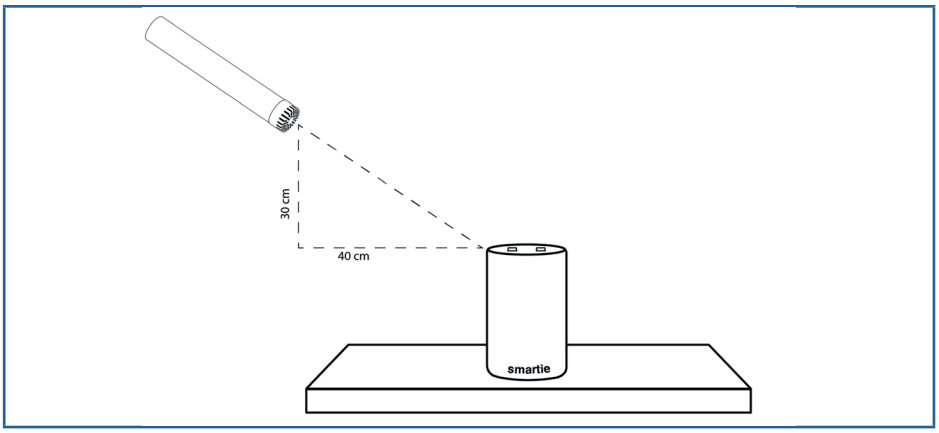

FIGURE 7.
Setup for smart speaker output path test.

Preparing the Test Signal

Smart speakers can typically play audio content from an online music service or internet radio station, or from a user's own music repository located either in "the cloud" or on a user's device. Usually, it is not possible to get an exact copy of the audio file streamed from a third-party streaming service. Therefore, the best approach is to prepare an audio file containing the stimulus signal and upload it to the music repository. In the case of the popular Alexa service, Amazon initially allowed users to upload their own content for streaming to smart devices, but later changed policies to disallow it. An effective workaround is to use a third-party streaming service for which there is an Alexa "skill" (e.g., Plex Media Server™). With this approach, it is possible to create a media server on your own PC and tell the IVA to stream a file from your server to a connected smart speaker.

Here again, the analysis requires that the sample rate of the signal acquired by the audio analyzer match the sample rate of the original signal. However, the music service from which the file is streamed will likely have its own constraints that have to be met before the file can be streamed. For example, some services require that it be encoded to .MP3 format at a sample rate of 44.1 kHz, and assigned multimedia tags such as artist, album and track. Once the audio file containing the stimulus signal has been prepared, it must be uploaded to the server.

Capturing and Analyzing the DUT Output

Once the file containing the stimulus signal has been prepared and uploaded to the connected music service, the test can be conducted. To get the smart speaker to play the test signal usually requires a spoken command. For example, for a smart speaker whose wake word is "Smartie", the command might be "Smartie: Play the song Test Signal by artist Audio Precision." If the IVA successfully interprets the command and finds the music track, a few seconds later it will respond with a "spoken" reply, such as "Okay, playing the song Test Signal by Audio Precision," followed by the test signal. To capture the test signal, the measurement in the audio analyzer must be started before the test signal is played.

To ensure that the test signal is analyzed (not the IVA's spoken reply), the audio analyzer must have a means to trigger on the desired signal. The trigger mechanism may vary by the type of measurement/stimulus. For example, in AP audio analyzers the following trigger mechanisms are used:

- A multitone signal has a unique signature that enables the measurement to trigger on the signal itself with a high degree of success.
- For open-loop measurement, log swept sine (or chirp) measurements and stepped sine measurements use a sinusoidal pilot tone. The analyzer uses a frequency-selective threshold trigger to locate the pilot tone and the test signal.
- The transfer function measurement has a "match" feature that enables it to trigger on the test signal itself. For signals such as music with repetitive phrasing, the user can prepend a short section of a unique signal, such as a maximum length sequence (MLS) to the stimulus, for improved triggering.

Example—Log Swept Sine Measurement of Smart Speaker Output Path
Figure 8 shows the stimulus and response waveforms from a smart speaker test using a log swept sine measurement from 50 Hz to 20 kHz with a 2.0 s sweep length. The 0.2 s pilot tone is visible at the beginning of the waveforms.

The pilot tone is used not only for triggering, but also to measure the difference in sample rates between the audio analyzer and the DUT. As noted above, the sample clock of the DUT will never exactly match the sample clock of the audio analyzer. For example, the original signal may have been created at a sample rate of 48 kHz, but the DUT sample rate is slightly different—say 47.990 kHz. The sample rates of the two signals must be identical, otherwise artifacts will be introduced in the results. [6] When chirp measurements are conducted in an open-loop configuration in APx500 audio analyzers, the system measures the difference in sample rates by analyzing the frequency of the pilot tone. It then re-creates the stimulus signal such that the stimulus and response have identical sample clocks.

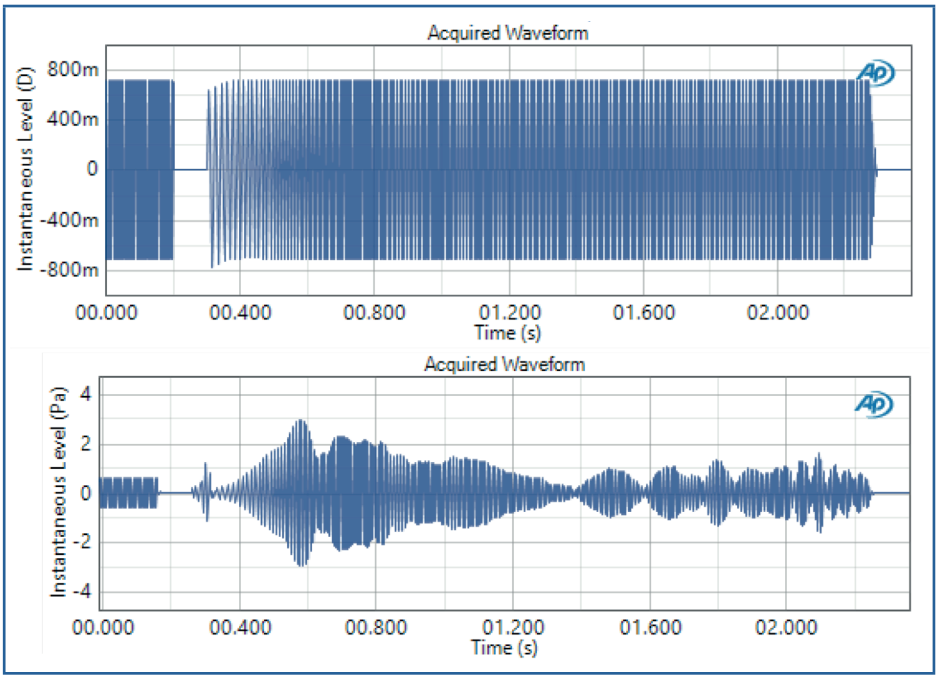

FIGURE 8.
Stimulus waveform (upper) and acoustic pressure response waveform (lower) from a smart speaker test. Log swept sine chirp from 50 Hz to 20 kHz in 2.0 s with a 200 ms pilot tone.

Figure 9 shows the acoustic level response derived from the stimulus and response waveforms in Figure 8. Dips in the curve occurring at regular frequency intervals starting at about 1.5 kHz are likely due to acoustic reflections from the table combining destructively with sound waves propagating directly from the speaker to the microphone.

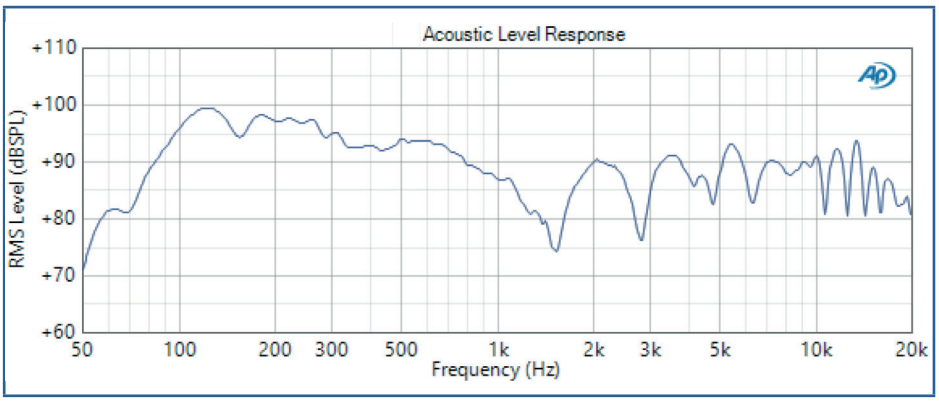

FIGURE 9.
Smart speaker frequency response level measured via log swept sine test using the waveforms shown in Figure 3.

Example—Transfer Function Measurement of Smart Speaker Output Path
Figure 10 shows the first few seconds of a music signal used to test the output path of the same smart speaker. The brief 0.1 s burst at the very beginning of the waveform is a short MLS signal prepended to the music signal to improve signal matching for triggering.

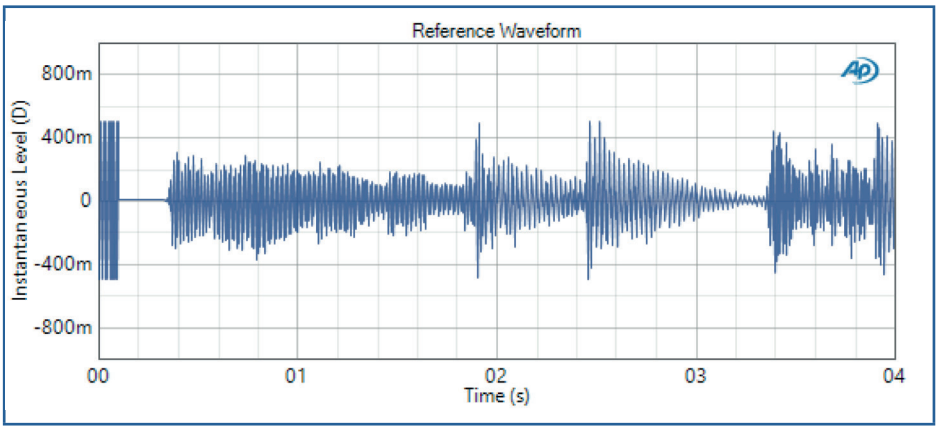

FIGURE 10.
Initial portion of a music signal used to stimulate a smart speaker.

The frequency response magnitude calculated from analyzing 20 s of the music signal is shown in Figure 11. Note that this result is a direct measure of the smart speaker's sensitivity versus frequency, with units of dB (Pa/FS).

Note the similarity between Figure 9 and Figure 11. Slight differences in shape between the two curves are likely due to the device responding differently to the music signal than a chirp signal at certain frequencies.

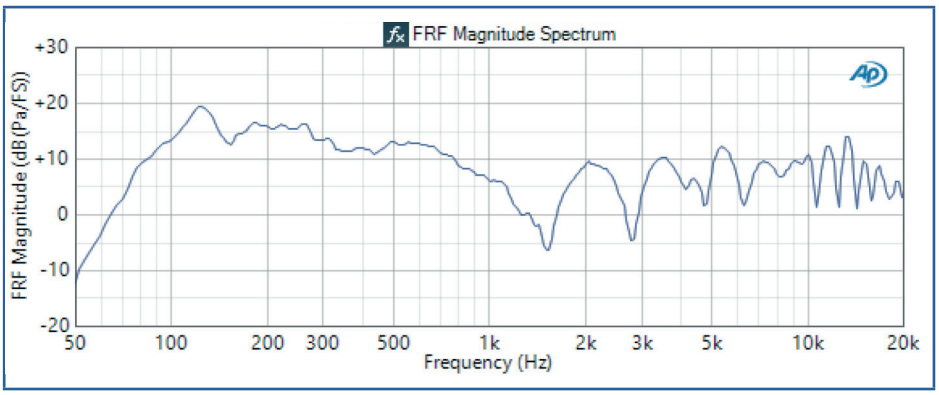

FIGURE 11.
Frequency Response Magnitude of the smart speaker measured using the Transfer Function measurement with a music signal.

Conclusion

This concludes our high-level overview of smart speaker acoustic measurements. We've discussed some of the inherent challenges involved in testing smart speakers—both their input and output are acoustic, they must be controlled by spoken commands, and testing them involves interacting with an IVA service and a back-end server. Despite these challenges, armed with a good audio analyzer and some acoustic accessories, plus a little inventiveness to overcome some of the obstacles, it is relatively straightforward to make meaningful audio quality measurements of smart speaker devices.

References

1. Wired.com. Amazon's weird Siri-like speaker is another way to get you to shop. (2014).
2. The Smart Audio Report, Winter 2018, NPR and Edison Research (2019).
3. Smart Speaker Market, Allied Market Research https://www.alliedmarketresearch.com/smart-speaker-market (2019).
4. The Anatomy, Physiology, and Diagnostics of Smart Audio Devices. AES Convention e-Brief 426 (2018).
5. Microphone Array Beamforming. InvenSense Application Note AN-1140 (2013).
6. Measuring Audio when Clocks Differ. AES Convention Paper 10055, NY (2018).
7. Toole, Floyd (2009). Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms. Taylor and Francis.
8. A. Farina, "Simultaneous measurement of impulse response and distortion with a swept sine technique," Presented at the 108th AES Convention, Paris, France, 2000.
9. IEEE 1329-2010 - IEEE Standard Method for Measuring Transmission Performance of Speakerphones.
10. Audacity - Free, open source, cross-platform audio software. https://www.audacityteam.org/
11. GoldWave - Digital Audio Editing Software. https://www.goldwave.com/
12. Audio Precision Technote 138 – Transfer Function Measurements with APx500 Audio Analyzers. 2019.

Amazon, Echo, Alexa and all related logos are trademarks of Amazon.com, Inc. or its affiliates • Google Assistant and Google Home are trademarks of Google LLC • Apple, HomePod mini and Siri are trademarks of Apple Inc., registered in the US and other countries and regions • Microsoft and Cortana are trademarks of the Microsoft group of companies • Samsung, Bixby and Samsung Galaxy Home are trademarks or registered trademarks of Samsung Electronics Co., Ltd. • Bose is a trademark of the Bose Corporation • SONOS is a registered trademark of Sonos, Inc. • Blutooth is a registered trademark of Blutooth SIG, Inc. • Audacity is a registered trademark of Dominica Mazzoni • GoldWave is a registered trademark of GoldWave Inc. • Plex Media Server is a trademark of Plex, Inc. • Wi-FI is a registered trademark of Wi-Fi Alliance®

© 2021 Audio Precision, Inc and GRAS Sound & Vibration / All Rights Reserved. XXVI0622145923 www.audioprecision.com / www.greaacoustics.com