Audio Engineering Society

Convention Paper

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Challenges of IoT Smart Speaker Testing

Glenn Hess[1], and Daniel Knighten[2]

[1] *Indy Acoustic Research, 6602 E 75th St., Suite 105, Indianapolis, IN 46250*

[2] *Listen, Inc., 580 Harrison Ave. Suite 3W, Boston, MA 02118*

Correspondence should be addressed to Glenn Hess or Daniel Knighten (glenn.hess@indyacousticresearch.com or dknighten@listeninc.com)

ABSTRACT

Quantitatively measuring the audio characteristics of IoT (Internet of Things) smart speakers presents several novel challenges. We discuss overcoming the practical challenges of testing such devices and demonstrate how to measure frequency response, distortion, and other common audio characteristics. In order to make these measurements, several measurement techniques and algorithms are presented that allow us to move past the practical difficulties presented by this class of emerging audio devices. We discuss test equipment requirements, selection of test signals and especially overcoming the challenges around injecting and extracting test signals from the device.

1 Introduction

A relatively new class of IoT products has emerged that are classified as smart speakers. Examples of such devices include the Amazon Echo™, Google Home™, Apple HomePod™, and Harman Kardon Invoke™. All these devices share certain key characteristics. First and foremost, rather than having physical controls these devices use voice control to respond to command words and short phrases. Second, rather than playing back and recording signals to local storage media, they store and playback signals from Internet based cloud services. These devices are typically used for entertainment, personal assistance, and home control.

Compared to earlier generations of consumer audio playback devices this class of products introduces new challenges in performing objective measurements. These devices typically do not provide any direct path to inject or extract response signals from the embedded acoustic transducers, speakers and microphones. In addition, their use of voice recognition for control makes even activating these devices challenging. For this paper we chose a representative and popular device and describe how we measured its audio performance.

2 Audio Performance

To provide an objective evaluation of the device's audio performance, we will describe techniques to characterize the frequency response, output level, and distortion of the device under test. This will allow direct comparison between IoT smart speakers and conventional speakers.

When measuring conventional speakers and microphones, the device under test is most commonly stimulated with a sinusoidal test signal, often swept across a frequency band. The measurements can be derived directly from the response.

With an IoT smart speaker, there are three main challenges. First, IoT smart speakers require verbal interaction to operate. That is, they listen for and then respond to a specific activation word or phrase. Second, there is typically no direct path to inject or extract a response signal. Finally, all IoT smart speakers use various types of active signal processing and may not respond linearly to all types of signals.

Below we will discuss how we overcome each of these challenges.

3 Activating Smart Speakers

Trial and error was required to learn what activation phrase was required to get the device to record, establish how long it would record for, and then to get it to playback a specific signal. Once we established these phrases, we could get the device to either play back the desired stimulus for speaker testing, or record a signal for microphone testing. In our analysis, we parse the response and ignore the activation phrase.

In general, these devices are always listening with their internal microphone(s) active. They respond to an activation word or wake-up phrase followed by a command string. The challenge in testing audio performance is activating the device followed by applying a suitable test signal to keep the smart speaker active and operating in a normal state.

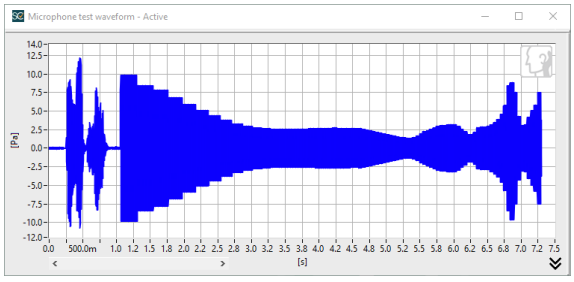

Figure 1. Activation word and test signal waveform for microphone measurement

4 Retrieval of Response Signal

In the case of our sample device, we discovered that all audio recorded by the device is available for download via the account to which the device is registered. The only significant limitation is that it only records for approximately 8 seconds. However, we adjusted our test signal so that it would fit within the recording limitations. We were then able to play a composite signal consisting of a voice recording of the activation word followed by a stepped sine sweep. After the recording was captured by the device, we accessed the associated website and downloaded the audio recording. This captured recording was then analysed to determine the microphone's characteristics.

5 Storage and Playback of the Stimulus

In order to test the speaker function, we had to upload a test signal to our account with the associated Internet service as if it was a musical recording. We then dictated to the device under test to playback the test signal. From there we captured the playback of the stored test signal.

The service requires that stored audio signals be encoded in the MP3 format. Test signals were encoded to 128 kbps mono fixed bit rate using the ACM MP3 encoder which introduced minimal degradation to the source signal.

6 Measurement Equipment

The following equipment was used for measurements:

- RME Fireface UCX Audio Interface
  Recommended for use with Listen, Inc.'s SoundCheck software
- Bruel & Kjaer 2690 NEXUS Conditioning Amplifier
  Used in conjunction with preamp and microphone
- Bruel & Kjaer 2669 Microphone Preamp
  Used in conjunction with conditioning amp and microphone
- Bruel & Kjaer 4190 Condenser Microphone
  A 12.5 mm laboratory standard free-field microphone per IEC 61094-4

- Bruel & Kjaer 4227 Mouth Simulator
  Complies with the specifications given in ITU-T Recommendation P.51
- Crown D-75 Power Amplifier
  Used in conjunction with the mouth simulator
- Listen, Inc. SoundCheck ver. 15
  Used to generate and analyse all signals

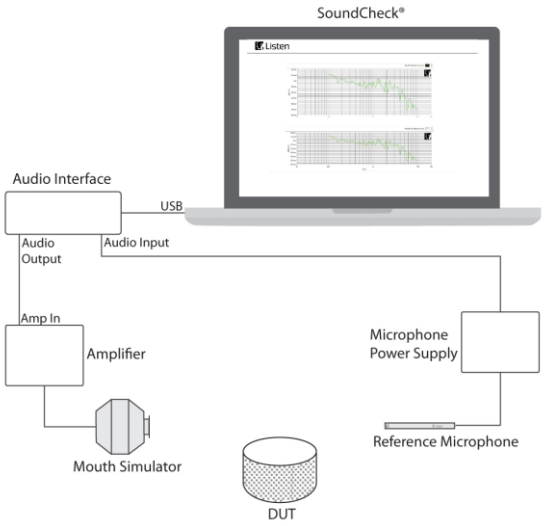

Figure 2. Test system connection diagram

7 Sampling Rate Error

Traditional measurement systems rely on the device under test being essentially a filter, which is a device with a synchronous input and output. As a category, smart speakers are intrinsically open loop, to mean that there is no synchronous input and output. Instead as described above, the device can record a signal or playback a signal but it does not have a synchronous signal path.

This introduces the possibility for sampling rate error. That is, the device may record a signal to a file with a sample rate of 44.1 kHz, but it may have in-fact been recorded at 44.09 kHz or another similar rate due to skew in the actual rate of the clock crystal used to drive the sampler. A similar error can occur when playing back test signals. That is, the test stimulus may be sampled at 44.1 kHz, but due to error in the playback sample rate the file is actually played back at a slightly faster or slower rate.

This sampling rate error will result in the component tones of the test stimulus being shifted to either a higher or lower frequency. This shift can then lead to measurement errors due to loss of coherence between the stimulus and response signals.

To overcome this sampling rate error, we apply an algorithm which searches the beginning of the response waveform for a steady state sinusoid at a pre-set frequency. The signal is then shifted to DC using a heterodyne filter and all other frequencies are filtered out. The output of the heterodyne filter includes the phase information which is ultimately used to estimate actual playback or recording sample rate of the response signal. With this information, the entire response waveform is resampled to the correct stimulus sample rate prior to analysis. This frequency shift step corrects for sampling rate error in the DUT and makes testing these devices straightforward.

The algorithm search for the sinusoidal signal described above is triggered by a 1 sec white noise signal. This trigger signal is inserted prior to the sinusoidal test signal as shown in the figure below. The trigger step is configured with a time offset to exclude the white noise from the algorithm search and heterodyne analysis.

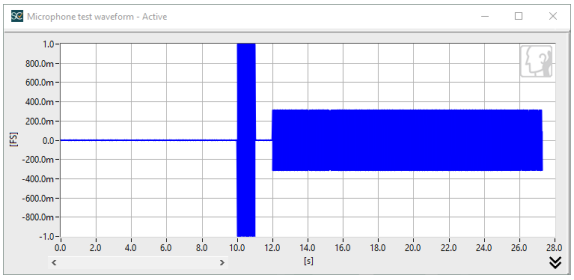

Figure 3. White noise and test signal waveform

8 Stimulus and Analysis

The actual signal used to measure the frequency response, distortion, and other audio characteristics of the device for these tests was a stepped sine sweep. In the case of the microphone measurements,

the amplitude at each frequency step of the sweep was equalized to produce an equal pressure response of 89 dBSPL (-5 dBPa) at the mouth reference point of a mouth simulator, per IEEE 1329-2010. The sweep provided 1/12 octave resolution from 100 Hz to 10 kHz. For the speaker test, the stimulus signal had an amplitude of –13 dBFS. A 1/3 octave resolution sweep was used from 20 to 100 Hz and 1/12 octave resolution from 100 Hz to 20 kHz.

Analysis was performed using a heterodyne filter which extracted the fundamental and harmonic distortion products from the response signal.

9 Physical Setup

IEEE 1329-2010 was referred to for the physical test setup. In many ways, smart speakers most closely resemble speakerphones as covered in this IEEE standard. All tests were conducted inside an anechoic chamber with the device under test placed on a table as defined in the diagram below and shown in the figure that follows.

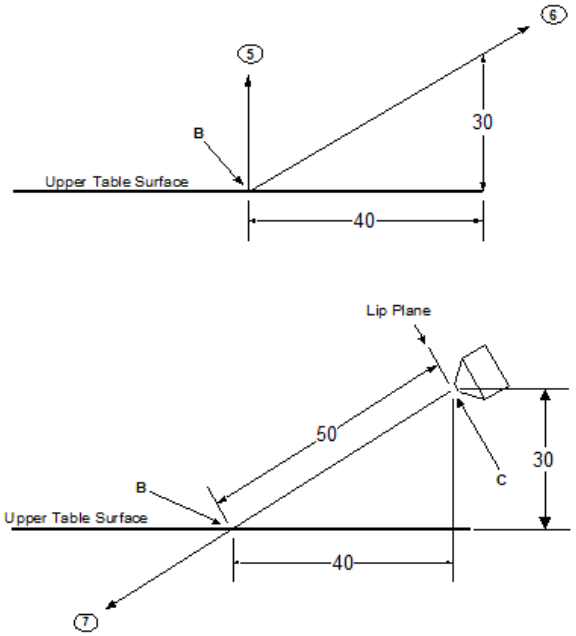

Figure 4. IEEE 1329-2010 test setup with table (dimensions in cm)

Figure 5. IEEE 1329-2010 test setup used (full anechoic chamber)

10 Device Microphone Measurements

Microphone system performance includes measurements of frequency response, sensitivity, and distortion.

Measurement steps included the following:

1. Apply the test signal using a calibrated mouth simulator
2. Retrieval of the voice interaction recording from the associated website
3. Transfer of the recording to software analysis system

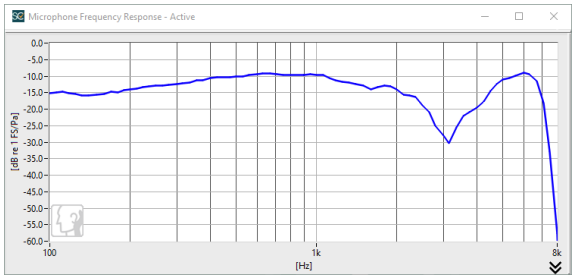

Figure 6. Microphone frequency response

The microphone frequency response is relatively flat except for the dip at 3 kHz. This was due to the device's microphone proximity to the table surface causing a "table bounce effect".

Freq. (Hz)	100	500	1k	5k	10k
Sens. (dB re 1 FS/Pa)	-15.2	-10.1	-9.6	-11.0	-164

Table 1. Microphone sensitivity levels

The sensitivity is measured in dB FS/Pa, where FS is relative to digital full scale in the waveform recorded by the DUT, and relative to the sound pressure in Pascals at the mouth reference point.

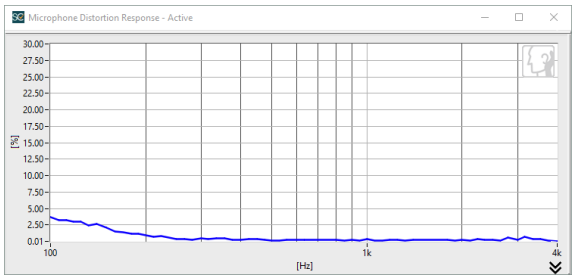

Figure 7. Microphone distortion response

The total harmonic distortion is relatively low above 200 Hz, but increases at lower frequencies due to residual distortion of mouth simulator.

Freq. (Hz)	100	500	1k	4k
THD (%)	3.7	0.2	0.3	0.01

Table 2. Microphone distortion percentages

11 Device Speaker Measurements

Speaker system performance includes measurements of frequency response, sensitivity, and distortion.

Measurement steps included the following:

1. Upload test signal to the associated online music library in MP3 format
2. Test operator dictates activation phrase to play test signal
3. Capture of the response signal for software analysis system

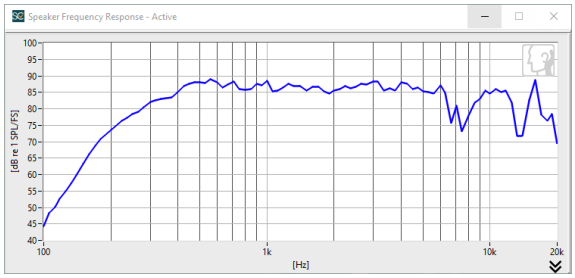

Figure 8. Speaker frequency response

The frequency response is relatively flat in the passband from 400 Hz to 6 kHz. The 400 Hz roll-off is due to physical limitations of the device's loudspeaker.

Freq. (Hz)	100	500	1k	5k	10k
Sens. (dB re 20u Pa/FS)	44.4	88.0	88.6	85.3	84.6

Table 3. Speaker sensitivity levels

The sensitivity is measured in dB re 20 uPa/FS, where output is sound pressure (SPL) at the reference mic and input is relative to digital full scale in the signal played back by the device.

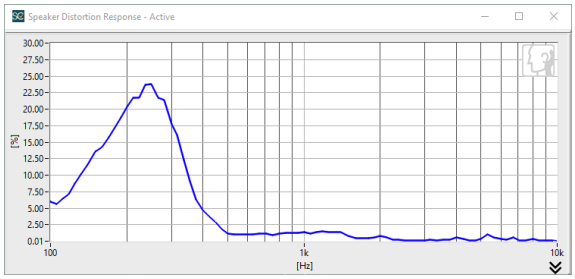

Figure 9. Speaker distortion response

Total harmonic distortion relatively low above 500 Hz, but increases at lower frequencies due to limitations of device's loudspeaker.

Freq. (Hz)	100	500	1k	5k
THD (%)	5.9	1.2	1.4	0.3

Table 4. Speaker distortion percentages

12 Conclusions

Smart speakers present two main testing challenges. The practical challenge of playing back test signals and recording response signals and the technical challenge presented by sampling rate error. However, as we demonstrate in this paper, both challenges can be overcome and conventional results can be extracted.

In the future, the authors hope to explore tests involving non-stationary signals, including noise, modulated noise, and actual speech and music. In addition, we would like to explore how well the devices can detect their activation phrases while playing back music, in the presence of noise, and with multiple interfering talkers.

References

[1] IEC 61094-4, "Measurement Microphones Part 4: Specifications for Working Standard Microphones", November 1995

[2] IEEE 1329-2010, "Standard Method for Measuring Transmission Performance of Speakerphones", October 2010

[3] ITU-T Recommendation P.51, "Artificial Mouth", August 1996

[4] Karlheinz Brandenburg, "MP3 and AAC Explained", presented at AES17th International Conference on High-Quality Audio Coding, August 1999

[5] Steve Temme et al., "The Challenges of MP3 Player Testing", presented at the AES 122nd Convention, May 2007