The challenges of testing voice-controlled audio systems

Steve F. Temme[1,†]
(1. Listen, Inc., Boston, MA 02118 USA)

Abstract: Smart devices that are voice-controlled such as smart speakers, hearables, and vehicle infotainment systems are notoriously complex to test. They have numerous connections from wired to wireless and contain much signal processing, both on the record and the playback side. This means that their characteristics change according to 'real world' conditions of the environment that they are used in, such as background noise, playback levels, and room acoustics. Furthermore, their multifunctional nature means that there are many aspects of the device that may need to be tested, ranging from voice recognition to music playback, operation as a hands-free telephone, and in the case of hearables, hearing assistance. Due to their complex non-linear use cases, these devices often need to be tested at different levels and different environmental conditions. This paper focuses on tools and techniques to accurately measure the audio performance of such devices under the many various real-world conditions in which they are used.

Keywords: hearables, automotive infotainment, smart speakers, smartphones, test

0 INTRODUCTION

Smart Devices such as smart speakers, hearables and automotive infotainment systems have become increasingly challenging to test. They have many possible interfaces ranging from hardwired to wireless (Bluetooth, cloud-based), smartphone, voice ("Hey Siri", "OK Google", "Alexa"), and in the case of automotive, even USB memory stick and CarPlay/Android Auto. There is usually much signal processing, both on the record side (e.g. beamforming, background noise filtering, voice activity detection, and on the playback side (e.g. loudness, compression, equalization, and active noise cancellation).

This means that their characteristics change according to 'real world' conditions such as the physical environment and background noise. Furthermore, their multifunctional nature means that there are many aspects of the device that may need to be tested, ranging from voice recognition to music playback or

† Email address: stemme@listeninc.com

even operation as a hands-free telephone, telephone headset or hearing aid. These devices often need to be tested at different levels and in different environmental conditions, for example different physical setups and with/without background noise, different signals etc.

Although, there are currently no standards for testing most smart devices, principles and test configurations are borrowed from many other audio devices and use existing standards such as IEC and BS EN for loudspeakers and headphones[1,2], IEEE for headsets, IEEE/TIA/ITU for telephone test [3,4,5,6], ANSI and IEC for hearing aid standards [7,8] , and ETSI [9] for background noise. Flexibility of the test system and experience with testing a wide range of acoustic devices is therefore critical to enable a device to be completely characterized.

This paper explains how to implement both basic acoustic tests and more complex real-world tests along with the techniques and standards that may be used. Most of the tests discussed are relevant to all smart devices including smart speakers, hearables and automotive infotainment, but some hearable-specific additional tests are also detailed. Finally, we present a check list of the test-system functionality you should look for when choosing a system to fully characterize a smart speaker or other smart device.

1 Basic Acoustic Tests

1.1 Wired and Wireless Measurements

Basic acoustic tests measure the device under static conditions to measure one or multiple microphone(s) and speakers or earphones characteristics such as frequency and impulse response, distortion, and for stereo headphones, left/right tracking. It is a good idea to measure both wired and wireless performance for comparison. This requires a calibrated input (e.g. a headphone amplifier or auxiliary input) for wired responses and a calibrated Bluetooth interface for wireless measurements (Fig. 1).

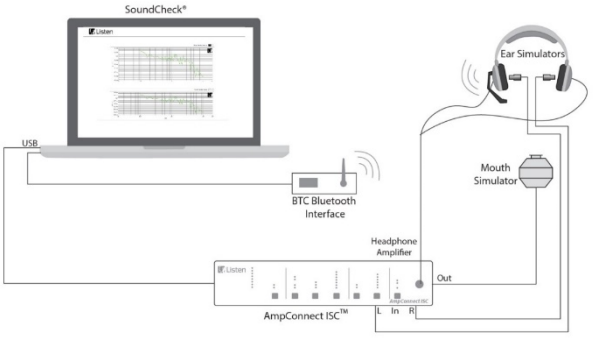

Fig 1. Test configuration for basic wired and wireless (BT) headset measurement. For automotive infotainment systems, substitute a 6-mic array and use the car's speakers via the head unit.

In addition, a calibrated mouth simulator or source speaker, equalized for a constant sound pressure output level versus frequency, and a calibrated measurement microphone or ear simulators are required for both send and receive acoustical measurements. Although a single measurement microphone usually suffices for smart speaker tests, a Head & Torso simulator is the ideal choice for hearables and infotainment systems measurements, since it has both a calibrated mouth and ear

simulators to truly represent a typical listener. However, for hearables, ear couplers are often used as a lower-cost option, and for automotive testing, an array of 6 measurement microphones is often substituted.

1.2 Measurements Via the Cloud

While, in all cases, the measurements themselves are similar to standard loudspeaker and microphone measurements, injecting the stimulus and extracting the recorded signal present challenges. Usually, the only way to do this is like an actual user - via the cloud using voice commands.

Usually, an artificial mouth often contained within the HATS, plays the voice activation commands to the device's microphone(s) which is typically connected via Bluetooth to a smartphone and/or a head unit. In the case of a smart speaker, an artificial mouth or small speaker can be used, and a smartphone is not required as the device itself has the internet connection. The smartphone communicates to the proprietary voice services in the cloud e.g. Alexa Voice Services (AVS) that tells the voice services which uploaded test signal to stream/playback through the smartphone via Bluetooth to the hearable or infotainment system. The test measurement system uses a triggered record to capture the loudspeakers playback signals that can then be analyzed by the test system.

Although the actual physical testing setup is very similar from device to device, for each it is necessary to understand how to wirelessly route the signal. The test signal must be in the cloud to enable playback. Each manufacturer's ecosystem is different in how it plays back from the cloud. Some enable you to upload your own recordings, whereas others require them to be on a media streaming platform such as Spotify. For microphone testing, some systems such as Alexa allow you to access recordings you have made; others prevent this for security and privacy reasons, which makes microphone testing challenging. Each device needs activating with a different wake word, has a different delay compensation, and records for a different amount of time after it hears the wake word. This needs figuring out (largely by trial and error) for each speaker that needs testing. The physical setups can be taken from the test standards for that particular device. For example, for hearables we suggest the physical configuration for headsets recommended by IEEE 269 [4], in which tests are conducted with the hearable worn on a head and torso simulator (HATS). For smart speakers, since their practical use case resembles speakerphones, the physical geometry recommended by IEEE 1329-2010 [3], which involves testing on a table inside an anechoic chamber, is a good starting point. In all 3 cases, the artificial mouth (which may be contained within HATS), plays the voice activation commands to the device's microphone(s) which is typically connected via Bluetooth to a smartphone (or directly to the cloud in the case of smart speakers). The smartphone communicates to the proprietary voice services in the cloud e.g. Alexa Voice Services (AVS) that tells the voice services which uploaded test signal to stream/playback through the smartphone via Bluetooth to the device. The test measurement system uses a triggered record to capture the loudspeakers playback signals that can then be analyzed by the test system. Example test configurations are shown in Fig. 2 below.

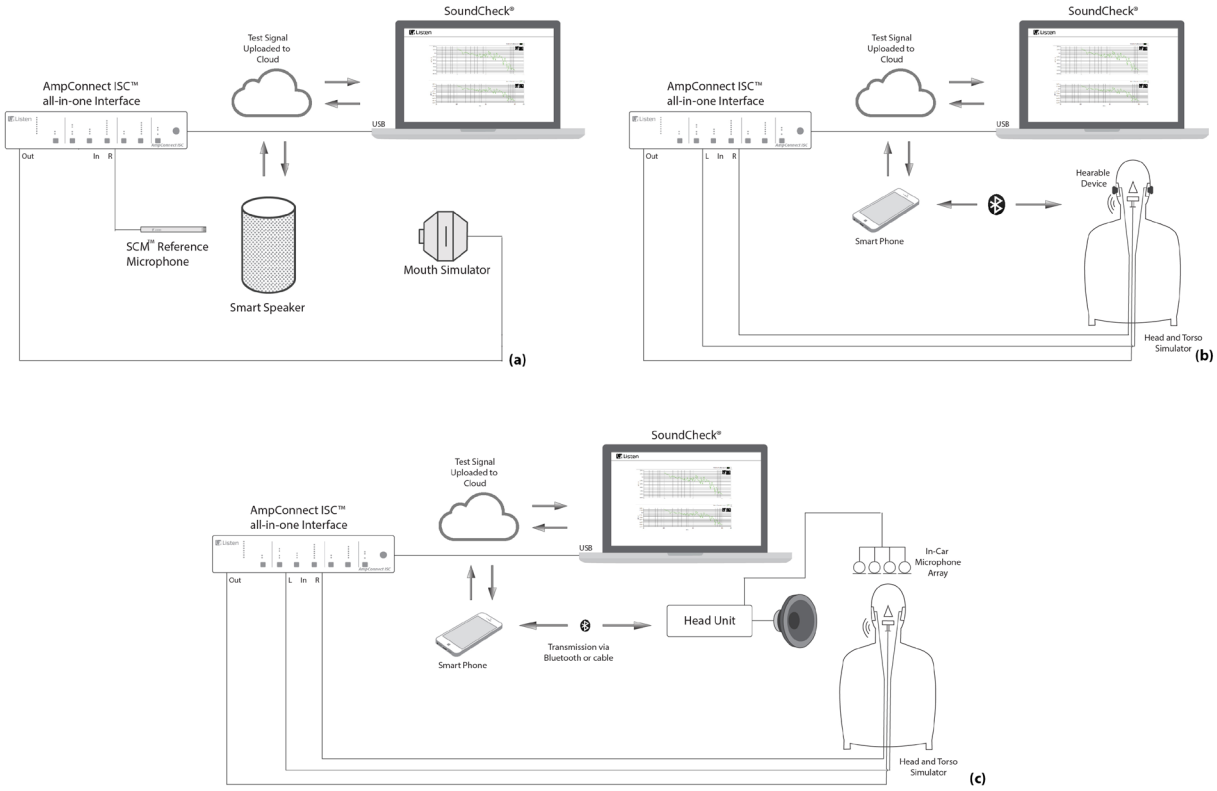

Fig. 2. Test configuration for basic speaker measurements via the cloud for a) smart speaker; b) hearable; and c) automotive infotainment

1.3 Playback

Let's now discuss the steps necessary to create a playback test using voice activation. It is ideal to automate the test as much as possible, so wake words and test stimuli should be pre-recorded and uploaded to the music service in the cloud so that they can be triggered by the test system. First, record the activation signal. This might be something like 'Alexa, play test signal 1'. Naturally the system's appropriate wake word should be used. If a wide range of voice recognition systems from different manufacturers is being tested, it makes sense to record the different voice wake words separately from the command so that any combination can be selected.

The test stimulus also needs to be created and uploaded to the cloud. Bear in mind that it may be compressed when it is uploaded which might introduce some distortion. It can be verified that the MP3 encoding itself does not degrade the test signal by encoding and decoding the test signal, and using transfer function analysis between the original signal and the encoded version. Test signals will be discussed in more detail later, but for basic playback performance testing, a standard stepped sine sweep is preferred.

Once the stimuli and voice commands are created and uploaded, the test sequence simply needs to play the wake word followed by the activation command via the mouth simulator and capture the response with the reference microphone(s) or artificial ears. Stimulus and analysis are then compared as in a traditional measurement to objectively qualify characteristics such as frequency response,

impulse response and distortion.

With the test system taking care of all these interfaces, the frequency output and distortion for an earphone, car infotainment system or smart speaker can be displayed in exactly the same way as for a basic wired speaker, or a speaker connected via a Bluetooth interface.

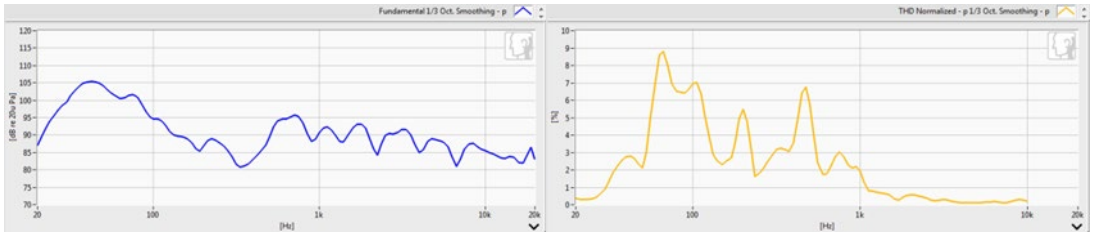

Fig. 3. Frequency and distortion responses of an infotainment multichannel playback system

1.4 Open Loop Measurements

One of the main challenges that test system manufacturers have faced in developing accurate smart system tests, is overcoming sampling rate error. Traditional measurement systems rely on the device under test having a synchronous input and output. Hearables, infotainment systems, and smart speakers are intrinsically open loop, which means that although the device can record or playback a signal there is no synchronous signal path. This introduces the possibility for sampling rate error. In other words, 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 a slightly different sample rate, e.g. 44.09 kHz due to skew in the actual rate of the clock crystal used to drive the sampler (Fig. 4). 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 results 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. It also changes the phase response.

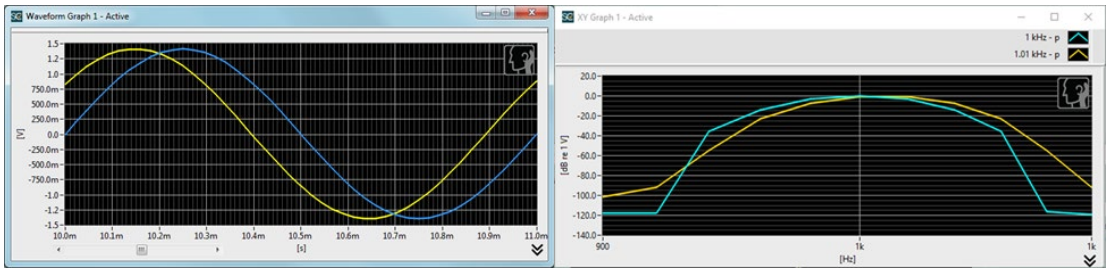

Fig. 4. Open loop measurements with differing sampling rates, and the Resampled Recorded Waveform to Match Sample Rate of Test Signal

To overcome this sampling rate error, an algorithm is applied which searches the beginning of the response waveform for a ‘trigger’ which is used to provide a reference point for alignment and shifting of the response to the stimulus signal. This trigger may be a steady state sinusoid at a pre-set level and frequency (Fig. 5), or for more robust performance that is less susceptible to false triggers,

a log chirp. 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 device and makes testing them straightforward.

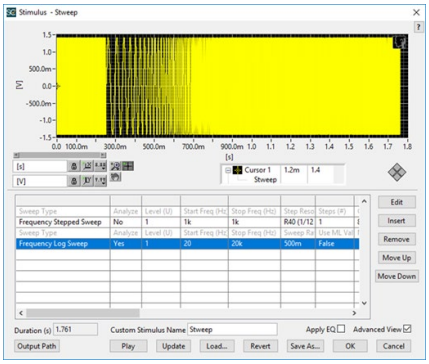

Fig. 5: 1kHz frequency tone prepended to stepped sine stimulus to calculate sampling rate and simplify record triggering

With the test system taking care of these issues, frequency output and distortion can be displayed in exactly the same way as for a smart device connected via a Bluetooth interface, directly via a debug port, or conventional wired.

1.5 Recording

Of course, it is also necessary to measure the microphone(s) or microphone array in the smart device. The set up for this is very similar, except that in this case, the wake word and the test stimulus is played through the mouth simulator or source speaker, captured by the device's microphone(s), then the recorded test stimulus is downloaded from the cloud and routed back into the test system, either by recalling the recorded wave file or using a virtual audio cable (Fig. 6). Once the signal has been brought into the test system this way, as with the playback test, the appropriate calculations are applied to overcome frequency shift and error and it is analyzed as normal.

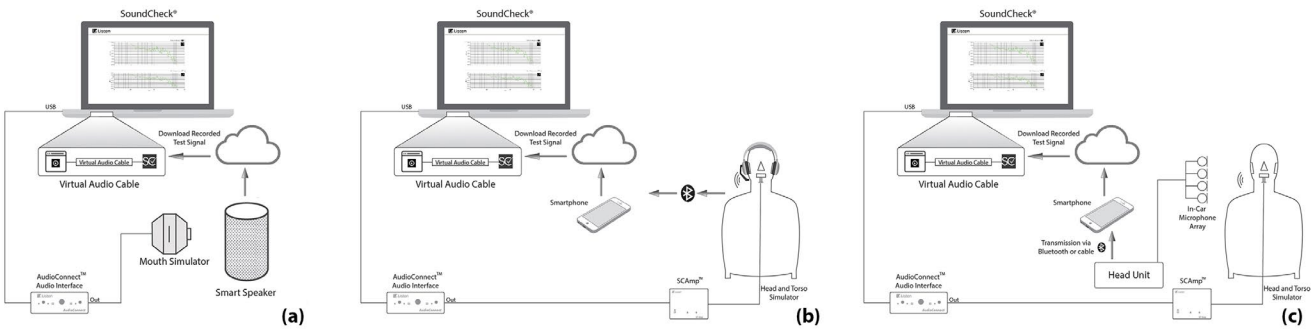

Fig. 6: Test configurations for measurement of the microphone within a) a smart speaker; b) a hearable; and c) an automotive infotainment system

2 Advanced or ‘Real World’ Testing of Smart Devices

2.1 Non-Linear Behavior

While the basic tests described above are ideal for measuring frequency response and distortion, they barely scratch the surface in terms of the breadth and depth of tests that one might want to carry out on a smart device such as a smart speaker, hearable, or infotainment system in the R&D lab. Consider that smart devices and systems are generally non-linear and their response varies according to many factors such as type of background noise, volume of background noise, whether the device is playing music or communicating voice, and more.

These considerations alone bring up a multitude of tests that could yield valuable information about the systems’ performance. While the specific tests that are carried out will vary according to the manufacturer’s design objectives, some of the test methodologies that might be used and some of the industry standards that provide a useful starting point, should be discussed.

A significant difference from basic tests is that advanced tests generally require the use of speech or music as test signals. This means that a test system that can support such test signals is essential and it is also important that the system can measure the active speech level of speech-based signals to ensure calibration and standardization as commonly used in telephony measurements.

2.2 Speech Recognition

Speech recognition tests are important for evaluating the system’s ability to understand voice commands. The physical setup is the same as in the basic test outlined above. One test method is to create a series of specially titled musical tracks and upload them to the particular cloud service e.g. iTunes. The track titles are the words or phrases that you want to test, and the actual content of the tracks is a single, dual or multitone that enables you to identify the track by its audio content. Generally ‘Harvard Sentences’ - a collection of phonetically balanced sentences that use specific phonemes at the same frequency they appear in English – are used for the track titles. These are recommended by ‘IEEE Recommended Practices for Speech Quality Measurements for VoIP, Cellular and other telephone systems’[10] and widely used in telecoms voice recognition. Pre-recorded versions of these can be downloaded online, or your own can be recorded (for example, if you are testing the device with different accents). These track titles are used to request playbacks, and a limit is used to detect whether the correct, or indeed any, signal was played back (Fig. 7).

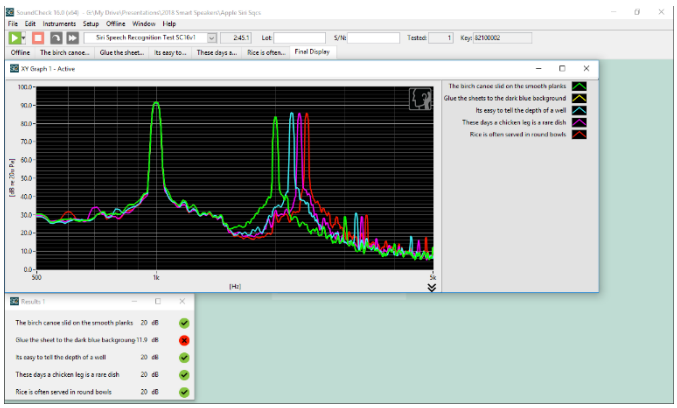

Fig. 7. Graphs showing results of end-to-end speech recognition

There are many variations on this test, for example testing different voices and accents, testing an active talker versus an interfering talker or background noise levels, testing with different types of background noise, etc.

2.3 Background Noise

Background noise is necessary for a wide range of smart system and device tests including voice recognition, SNR optimization of microphone arrays, beamforming directionality studies and more. Where background noise is needed, ETSI standard ES 202 396-1 [9] contains binaural recordings made in different noise environments e.g. café noise, subway, road noise, traffic noise, sirens etc. These can be played using a costly dedicated system, or with sufficient expertise, by programming your own playback method. However, some systems offer integrated ETSI background noise modules, which fully integrate the ETSI standard library with the test system, playing the background noise at calibrated levels which can be controlled and adjusted as part of a pre-programmed test sequence. This test sequence may include loops to incrementally increase the volume or change the noise and repeat the test, making it simple to create a test that increases the background noise by fixed levels until the voice is no longer accurately recognized. In addition to being a very cost-effective approach, it also significantly simplifies the physical test setup and reduces test development time.

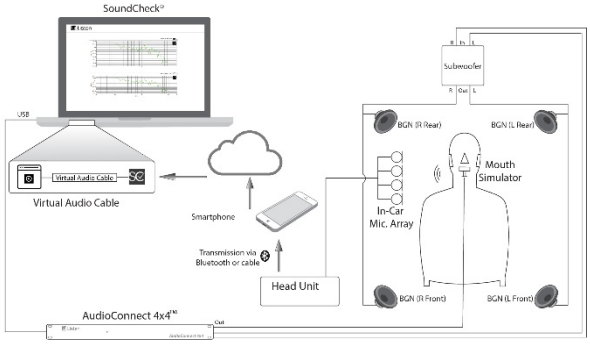

Fig. 8. Test configuration for hands-free measurements with ETSI standard background noises. In a hearable set-up the device's microphone would replace the in-car array, and the signal would go directly from that to the smartphone.

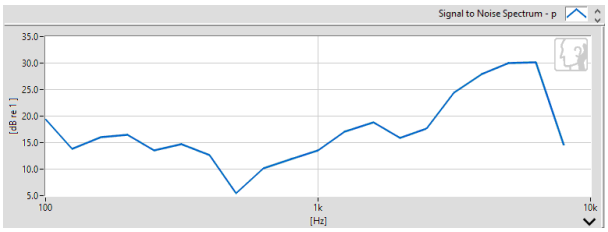

Fig. 9. SNR for a microphone array measured with ETSI standard background noises

2.4 Voice Quality

Smart speakers often function as handsfree communication devices. For such use cases, we can lean heavily on telephone test sequences such as the TIA standard TIA920-B [5] which covers both handsfree devices (as in automotive applications and smart speakers) and headset devices such as hearables. In fact, smart speakers and infotainment systems can be tested as a speakerphone *exactly* as defined in the TIA standard, and a hearable can be tested as a headset *exactly* as defined in the TIA standard.

Although the test setup in each case is identical to the basic acoustic test setups outlined above, the test sequences for measurement to the standard are highly complex and either require considerable expertise in this area to create them, or they need to be purchased as an off-the-shelf package.

2.5 Noise Cancellation

Although most commonly used for hearables, noise-cancellation tests also have applications in automotive infotainment systems for reduction of cabin background noise. The test configuration for measuring noise cancelling headphones is shown in Fig. 10. The test procedure can be broken down into three measurements and three calculations. First, the headphones are removed from the head and torso, the noise signal is played, and the spectrum measured through the open ear. This un-occluded ear spectrum is used as the baseline for the noise. Next, passive isolation is measured by placing the headphones onto the head and torso, playing the noise signal again and measuring the noise through the artificial ear. Finally, the measurement is repeated with the active noise cancellation (ANC) circuit turned on.

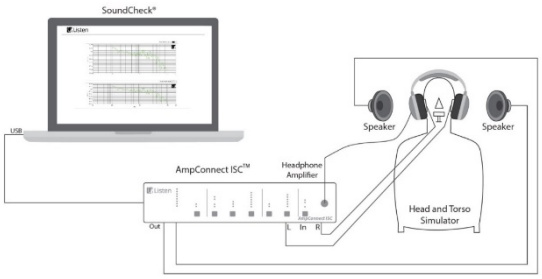

Fig. 10. Test configuration for measuring ANC headphones

These measurements result in three spectra which are used to calculate the three attenuation parameters. Passive Isolation, which quantifies how much noise is attenuated simply by the headphones being worn, is calculated by subtracting the un-occluded baseline curve from the second measurement where the headphones are in place, but noise cancellation is dis-engaged. Passive isolation can be significant across the frequency band but is most prominent at higher frequencies. Circum-aural headphones and in-ear earphones provide a reasonable level of passive isolation, whereas supra-aural headphones and earbuds typically offer much less.

Active Attenuation, which quantifies how much noise is reduced (or sometimes increased) when the active cancellation circuit is engaged, is calculated by subtracting the second measurement

(noise cancellation off) from the third measurement (noise cancellation on). Typically, this will be most prominent in the lower frequencies.

Finally, total attenuation is calculated by subtracting the curve with noise cancellation turned on from the baseline measurement without headphones. This represents the end-user's experience of the device, combining both passive and active components in attenuating background noise (Fig. 15).

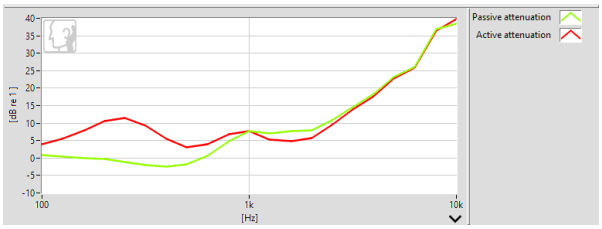

Fig. 11. Typical Noise Cancelling Measurement Results

Often this test is optimized by performing each measurement multiple times and averaging the resulting spectra. This accounts for variations in the fit of the headphones, which for some devices can impact the passive and even the active portions of the noise cancellation performance.

Some ANC headphones allow control of the degree of noise cancelling typically via a smartphone app. This is supposed to account for different listening environments e.g. on a plane with full noise cancelling on versus walking on a sidewalk next to a busy street where one should be more aware of their environment to avoid getting hit by a car, bike or even another person. In this case, a measurement showing the different levels of noise cancellation versus frequency for the different environments is a good idea. Again, this is a case for using a calibrated, simulated background noise system such as specified by ETSI to recreate the different test environments in a very repeatable way.

Also, some noise cancelling headphones will actually alter the frequency response of the music being played when the noise cancelling circuit is engaged. To test for this effect, one can measure the frequency response using the methods outlined above, both with and without the noise cancellation turned on, and compare the resulting response curves. Differences, if they are present, will typically be in the 1 – 3 kHz range.

2.6 Hearing Assistance

Fairly specific to hearables is the integration of hearing enhancement so that they also act as "personal sound amplification products (PSAPs)" or hearing aids, and it is desirable to measure their amplification or compression algorithms along with their corresponding attack and release times. This is usually measured according to the ANSI/IEC standard [7,8] in an anechoic test box (Fig. 12) where the loudspeaker is equalized to have a flat frequency response and the hearable is inserted into a 2cc coupler or "artificial ear".

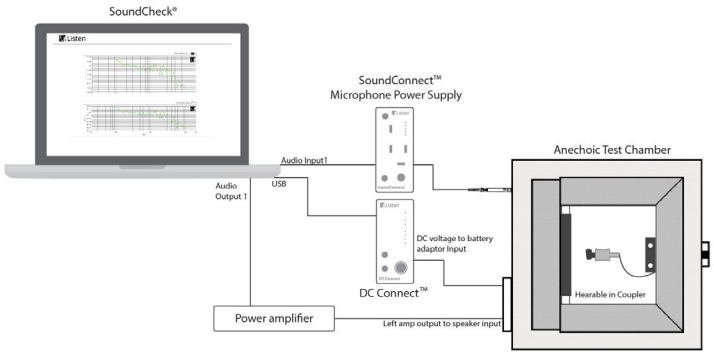

Fig. 12. Test configuration for measuring hearable to hearing aid standards.

The ANSI S3.22 standard [7] describes the method for testing hearing aids with AGC (automatic gain control) and could be applied to hearables that offer this functionality. The device is placed in an anechoic box, and an amplitude sweep from 50-90 dBSPL played through the test box speaker. This sweep is repeated at four frequencies (500 Hz – 4 kHz) in octave increments. The output of the hearing aid is plotted vs. the input on the graph to quantify the gain added for low level signals and compression at high levels for hearing loss compensation (Fig. 13).

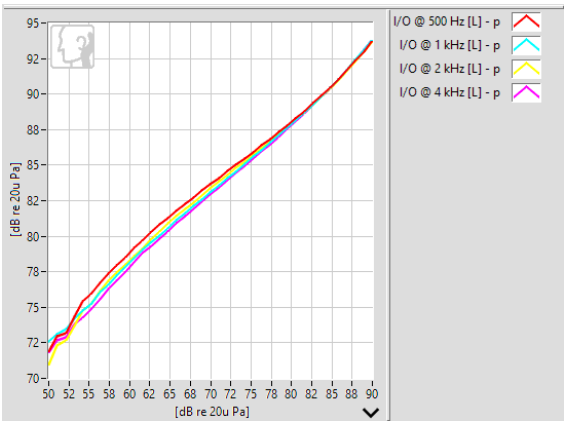

Fig. 13. Hearable measurement of Input vs. Output

3 Conclusions & Test System Recommendations

This paper has described a variety of ways to measure the audio performance of modern-day smart devices including smart speakers, headphones/hearables, and automotive infotainment systems. Since most of these systems have voice control and smartphone integration with algorithms built-in, it is necessary to perform measurements that include these functionalities. These complex functionalities require open-loop measurements that include the entire signal path and use speech and music as test signals under real-world conditions. These measurements represent the convergence of tests from loudspeakers, headphones, microphones, telephones, hearing aids, and more. Between these, a huge number of test stimuli, algorithms, analysis options and signal paths

options are necessary, not to mention many ‘useful-to-have’ features such as integrated background noise and a test configurator that makes it easy to run many iterations of a test continuously. Below is a checklist of the functionalities that should be evaluated before selecting a test system:

- Tests both the input and output components of the device (i.e. loudspeaker or earphones and microphones)
- Can create a compound stimulus with frequency and/or log chirp trigger to enable accurate open-loop testing
- Can be set to analyse only parts of the response signal, so that the trigger tone can be eliminated from the analysis
- Ability to use speech and music as test signals, calibrate levels, and equalize a mouth simulator (requires the inclusion of active speech algorithms)
- Background noise generation – although this can be generated externally, testing is much simpler and faster if it can be created by the test system and integrated with the test sequence
- Can accept signal via virtual audio cable in order to route in signals from the cloud – native windows multimedia interface
- Availability of pre-written TIA test sequences if device needs to be tested to telephony standards (may be manually programmed with sufficient expertise)
- Test programming flexibility – one that allows easy duplication of sub-sequences and loops is extremely helpful when repeating a test at various levels
- Offers a calibrated Bluetooth interface
- Mentor A2B interface compatibility and control from within test system if testing automotive smart devices (e.g. in-car voice control)

REFERENCES

[1] IEC 60268-7: Sound system equipment – Part 7: Headphones and earphones

[2] BS EN 50332-1 Sound system equipment – Headphones and earphones associated with portable audio equipment – Maximum SPL measurement methodology and limit considerations

[3] IEEE 1329-2010, “Standard Method for Measuring Transmission Performance of Speakerphones”, October 2010

[4] IEEE 269: Standard Methods for Measuring Transmission Performance of Analog and Digital Telephone Sets, Handsets, and Headsets

[5] TIA 920.130: Telecommunications Communications Products Transmission Requirements for Digital Interface Communications Devices with Headsets

[6] ITU-T Recommendation P.51, “Artificial Mouth”, August 1996

[7] ANSI S3.22 Specification of hearing aid characteristics

[8] IEC 60118-7: Electroacoustics - Hearing aids - Part 7: Measurement of the performance characteristics of hearing aids for production, supply and delivery quality assurance purposes

[9] ETSI ES 202 396-1: Speech and multimedia Transmission Quality (STQ); Speech quality performance in presence of background noise; Part 1: Background noise simulation technique and background noise database
[10] IEEE Recommended Practice for Speech Quality Measurements
[11] Karlheinz Brandenburg, "MP3 and AAC Explained", presented at AES17th International Conference on High-Quality Audio Coding, August 1999
[12] Steve Temme et al., "The Challenges of MP3 Player Testing", presented at the AES 122nd Convention, May 2007
[13] Glenn Hess et al., "Challenges of IoT Smart Speaker Testing", presented at the AES 143rd Convention, October 2017
[14] Steve Temme, "Testing Audio Performance of Hearables", presented at the International Conference on Headphone Technology, August 2019
[15] ~~[1]~~ IEC 61094-4, "Measurement Microphones Part 4: Specifications for Working Standard Microphones", November 1995

The Author

Steve Temme is founder and President of Listen, Inc., manufacturer of the SoundCheck audio test system. Steve founded the company in 1995, and for the past 24 years the company has remained on the cutting edge of research into audio measurement, regularly introducing new measurement techniques, algorithms and hardware to enable testing of all kinds of devices ranging from basic transducers to complex audio systems. In recent years, Listen's product development focus has been on creating the algorithms, interfaces and test methods to measure today's many and varied wireless and cloud-communicating devices. Prior to founding Listen, Steve worked for many years as an acoustic test and measurement applications engineer at Brüel & Kjær, and also as a loudspeaker design engineer at Apogee Acoustics. He holds a BSME from Tufts University, has authored numerous papers on acoustic testing, and has lectured extensively throughout the world.