White Paper

Speech Intelligibility Evaluation Highlights Differences Between LTE Public Safety Devices

Introduction

There is no disagreement that speech intelligibility is a critical requirement for first responder communications devices. In the real world of noise-filled emergency environments, being misunderstood or having to repeat oneself can have catastrophic consequences.

Spirent recently became **the first organization to offer a Speech Intelligibility Evaluation service** based on the ABC-MRT16 algorithm developed by the U.S. National Telecommunications and Information Administration (NTIA) and used by the National Institute of Standards and Technology's (NIST) Public Safety Communications Research Division (PSCR).

As we were putting the finishing touches on our automation and reporting capabilities, we put four currently available commercial devices through their paces.

The results were eye opening: in some typical first responder noisy environments, phrase misunderstanding occurred almost 50% of the time for some devices as compared to less than 10% of the time for others.

Read on for the details, and for what it all means.

Speech Intelligibility Evaluation Highlights Differences Between LTE Public Safety Devices

Criticality in PS-LTE

First responder networks are beginning the transition from legacy TETRA and P.25 networks to PS-LTE – Public Safety-over-LTE – commercial cellular technologies based on 4G LTE. The cellular standards organization, 3GPP, has incorporated mission-critical requirements such as push-to-talk, priority preemption, and proximity (device-to-device) operation. Networks are coming online (such as FirstNet in the U.S.) and are being planned and trialed around the world (UK, Korea, France, Australia and others).

PS-LTE makes **four compelling promises** for emergency services agencies:

1. First responder networks, largely voice-centric today, will be able to leverage what the average smartphone user has had for a decade: high speed data, live video, location tracking, group messaging, etc.
2. LTE's enormous device vendor ecosystem and access to an app-enabled environment will provide unparalleled choices and innovation.
3. The economics of leveraging commercial LTE network components, existing commercial networks and the vendor ecosystem will result in a favorable cost structure.
4. Interoperability between agencies (a driving force behind the funding of FirstNet) will be enabled by software within a unified network.

But these benefits can't come at the price of speech intelligibility. The difficulty of delivering understandable speech in the midst of active emergency situations is far more technically challenging than for delivering good consumer-grade cellphone audio. "A lot of work has been done in the [Land Mobile Radio] world to investigate and come up with solutions," says Ken Rehbehn, founder and principal analyst at CritComm Insights, and a former 911 dispatcher and active firefighter and EMT in Montgomery County, MD. "That progress needs to be sustained as LTE devices come on the market."

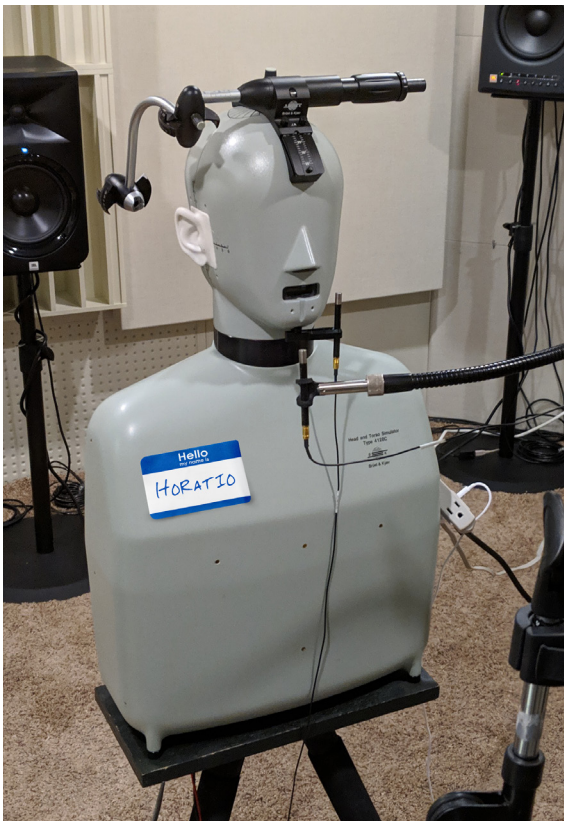
What is Speech Intelligibility?

Testing for voice quality began in the Bell System decades ago, with human listeners and the development of Mean Opinion Score (MOS). Today's MOS algorithms (such as POLQA) are excellent at evaluating the quality of audio transmission, but they are not sufficient for Public Safety scenarios. That's because of background noise.

First responders need to communicate in acoustically challenging environments. They need to be understood while there are alarms sounding in a building, when outside with nearby vehicle sirens and helicopters circling overhead, and when next to firetrucks with pumps running.

MOS algorithms, generally speaking, evaluate the ability of a device and/or a communications channel to accurately reproduce audio. Speech intelligibility, on the other hand, is about the ability of a device to transmit all the critical components of speech in the presence of these background noises so that words and phrases can be easily understood.

How Spirent Evaluates Intelligibility



We start by creating a controlled lab environment that accurately recreates challenges in the field. Horatio, our head and torso simulator, sits in the center of an acoustic isolation chamber. A series of voice phrases are played out of Horatio's mouth and captured by the device-under-test. Simultaneously, real-world recordings of background noise are played through speakers surrounding Horatio for 360° reproduction.

While on an active call, the device under test operates in its normal mode, working to distinguish between the spoken word and various background noises. We record the upstream audio at the receiving end.

Until fairly recently, the only way to judge intelligibility has been via "subjective testing," where panels of listeners (humans, that is) score the test results. Today the scoring is handled by the state-of-the-art ABC-MRT16 algorithm. The methodology employs "rhyme testing." Sets of short words (consonant-vowel-consonant) are used that only differ in the initial or final consonant. One sequence, for example, is "sane, name, game, tame, cane, fame." The words are spoken multiple times while a particular type of background noise is played, and the test is repeated for several noise profiles.

The ABC-MRT16 algorithm models the human auditory system to produce a score between zero and one. A score of < 0.5 is considered a failure, meaning that the phrase was not understood and would need to be repeated. ABC-MRT16 has been validated by comparison to human subjective tests for a wide range of conditions. It achieves a very high Pearson correlation of over 95%.

Benchmark Test Methodology

For our benchmark test, we use both speech and noise from the NIST/NTIA database. 1200 phrase samples are played for each test run, consisting of 300-word samples spoken by four different talkers. We first run the voice samples in a quiet room test, and then compare that to a series of runs with noise profiles including Nightclub, K12 Fire Rescue Saw, Firetruck Pump Panel, and others. You can listen here and here to a few speech-plus-noise examples.

A full sweep of one device against all of the NIST noise profiles requires upwards of 10,000 test samples to be played, and their corresponding audio results to be recorded. We run each device in speakerphone mode and in non-speakerphone mode. For our benchmark test of 4 devices, that is about 80,000 distinct speech test points. We've automated the test execution including network setup, device calling, and sample sequencing to ensure error-free test runs and logging of data (and for our own sanity). We've also automated the back-end storage, scoring and reporting of the results.

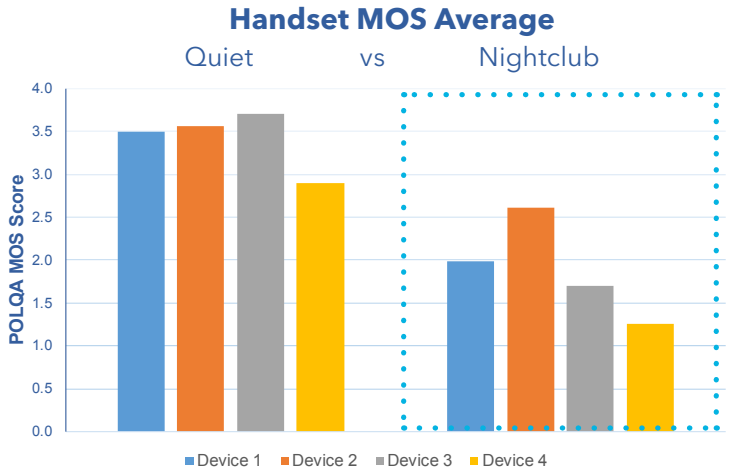
Isn't POLQA Good Enough?

The ITU-T P.863 standard adopted the POLQA algorithm for voice quality testing in 2011. It is the go-to standard used in the consumer cellular industry for assessing narrowband and "HD Voice." And it's great for that use case. But in noisy environments, POLQA penalizes a device's MOS if the background noise makes the received audio significantly different than the reference audio file. As a result, relying on POLQA results alone are likely to lead to incorrect conclusions.

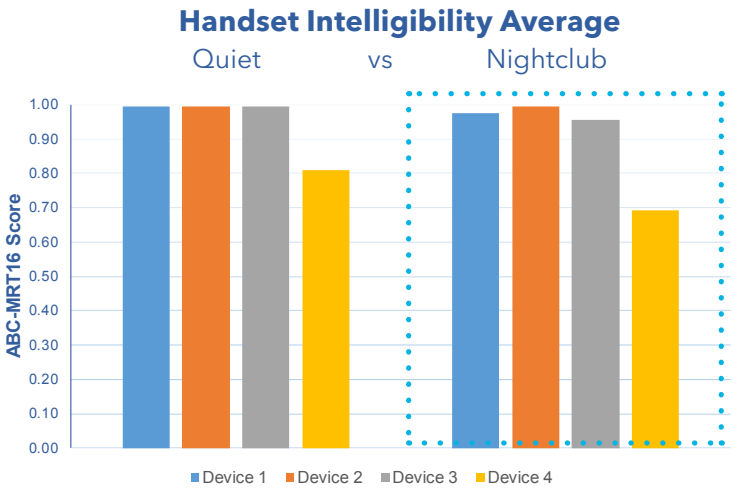
When we ran our benchmark ABC-MRT16 tests on the four commercial devices, we also executed POLQA testing in parallel. At first glance the results of the two algorithms are directionally similar – when you eye-ball it, it looks like a device that performs lower with one algorithm also performs lower on the other. That is a misleading result.

Here's where the difference matters. Take a look at this comparison of the same four devices, for one particular noise environment, tested with POLQA and tested with ABC-MRT16.

We start, at the right, using 0.0-to-4.0 range POLQA scoring for a Quiet environment vs the Nightclub noise profile. With no background noise Device 4 is a bit sub-par (a MOS of under 3.0) but the others do quite well. With the Nightclub noise present, Device 2 struggles, and Devices 1, 3 and 4 fall under 2.0 (poor). You might be tempted to conclude that none of these devices is particularly good in the presence of this noise environment (and you would be wrong!).



Now we repeat the test run, this time using the 0.0-to-1.0 range ABC-MRT16 Intelligibility scoring algorithm. In the quiet room, Device 4 is again a bit sub-par, and the others are excellent. When we're in the Nightclub, things sound very different this time. Devices 1, 2 and 3 all perform very well, scoring over 0.95. It is only Device 4 that is in trouble.



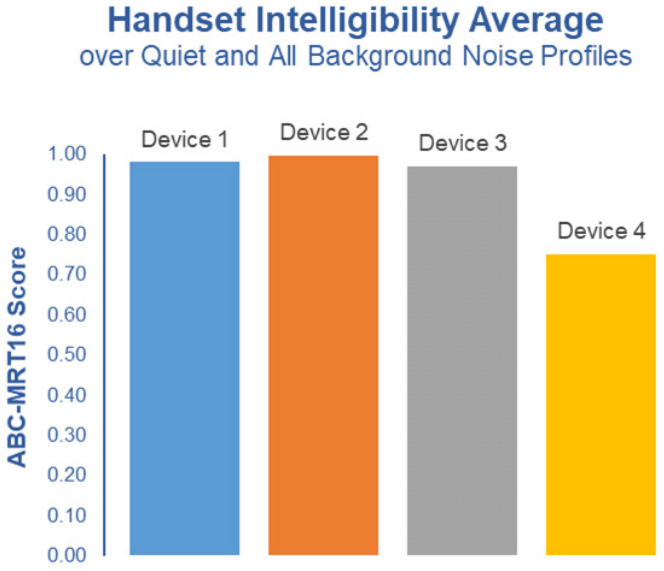
With POLQA alone, we would have incorrectly implicated two devices. It's not the right tool to use when evaluating speech intelligibility.

Real Devices with Real Differences

We ran four commercially available LTE devices through the benchmark noise test. We cannot name the devices (or even hint at the identity of the manufacturers), but we can share some of the key findings for comparison.

Starting at the top level KPI, the broad average of all scores for a device (all speech samples, no noise and all noise profiles, speakerphone and non-speakerphone) reveals that one of the devices doesn't perform as well as the other three.

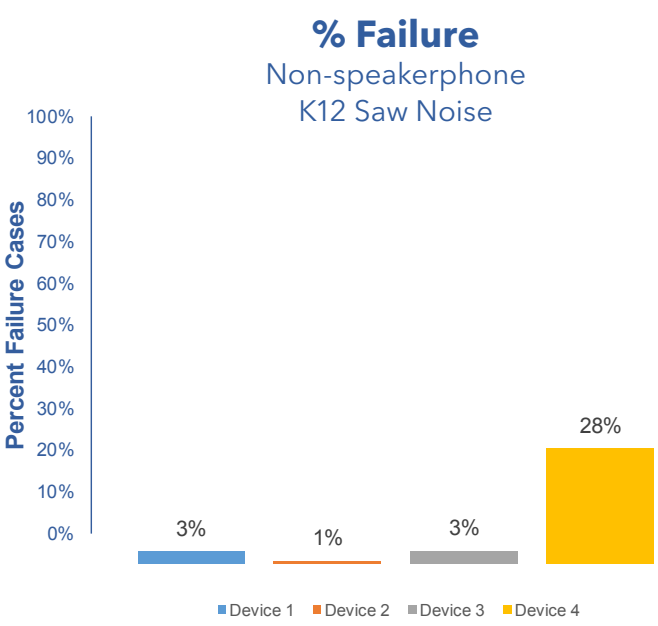
There is, though, **a more meaningful KPI** than the average score. Recall that for an individual speech sample, a score of under 0.5 indicates that the listener could not correctly identify the word. In real life, that means that either they would ask for a re-transmission, or worse, they would misunderstand what was said. Then, the critical question becomes, how often would that happen?



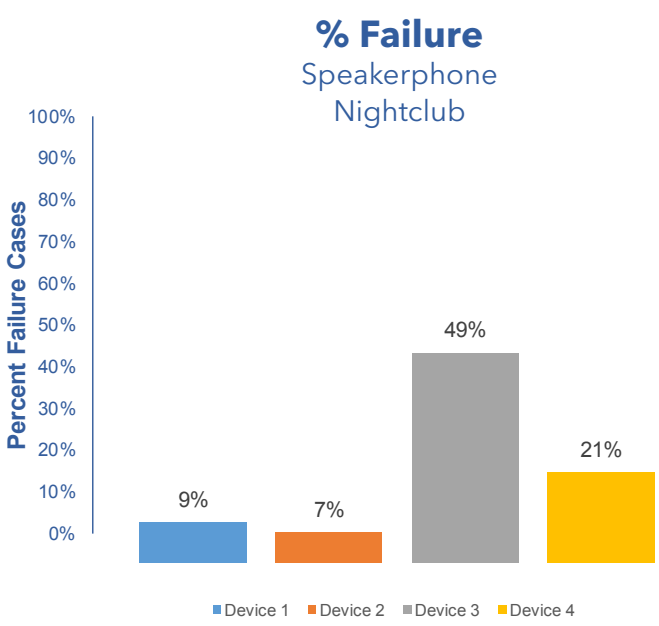
Speech Intelligibility Evaluation Highlights Differences Between LTE Public Safety Devices

Let's zero-in on the K12 Saw noise scenario, and count failures. In this run, we're operating the device in non-speakerphone mode. Now we can see where one of the problem areas is. And more importantly, we start to understand how this device will perform in the field. Device four failed over 25% of the time. One in four times the message was misunderstood or needed to be repeated.

If all of the devices were performing similarly, we might write it off as too-tough of an environment. But each of the other three devices worked quite well. The listener would only have missed between 12 and 36 out of 1200 words.



The story shifts when we look at another case. Now we're in the Nightclub, and the device is operating in its speakerphone mode. This is a more challenging scenario for all of the devices. In this case, Device 3's performance falls off dramatically, failing almost 50% of the time. Meanwhile, the best-performing device in the test manages to deliver intelligible speech all but 7% of the time.



Speech Intelligibility Evaluation Highlights Differences Between LTE Public Safety Devices

About Spirent Communications

Spirent Communications (LSE: SPT) is a global leader with deep expertise and decades of experience in testing, assurance, analytics and security, serving developers, service providers, and enterprise networks.

We help bring clarity to increasingly complex technological and business challenges.

Spirent's customers have made a promise to their customers to deliver superior performance. Spirent assures that those promises are fulfilled.

For more information, visit: www.spirent.com

What to Make of the Results?

It's not so much that one of these particular devices is better than the other. Or even that some do better in one case vs another. There are **two key takeaways** that we see here:

The first is that there is proof that **devices can be built that perform well in the extreme audio environments** in which first responders operate. More noise environments and devices need to be studied, but in this initial benchmark test, looking across all of the scenarios and talkers, the best performing device outperformed the worst by at least a six to one margin. It's a solvable problem.

The second conclusion is that **the bar has been raised by device manufacturers**. As Ken Rehbehn said, "It is in the best interest of device designers, operators and buyers to make sure that the equipment that they are certifying is up to standard in terms of usability." Public safety agencies, network operators, and device and equipment manufacturers should consider how well the gear that they are buying and building will perform in the field.

Recommendation: Assure the Promise of Speech Intelligibility

When device vendors and mobile operators choose to serve the public safety market, they have a vested interest in ensuring that they're doing everything possible to maximize speech intelligibility. A key part of this assurance is selecting the right testing algorithm along with highly specialized test equipment, environment and expertise. Spirent Communications provides all of that, without the upfront and ongoing costs of owning and staffing a lab capable of rigorous PS-LTE intelligibility testing.

For more information, visit us at www.spirent.com/solutions/public-safety-over-lte-testing. If you would like to talk with us about speech intelligibility evaluation, please reach out at www.spirent.com/contactspirent (select "Public Safety over LTE").

spirent™
Promise. Assured.

Contact Us

For more information, call your Spirent sales representative or visit us on the web at www.spirent.com/ContactSpirent.

www.spirent.com

Americas 1-800-SPIRENT
+1-800-774-7368 | sales@spirent.com

Europe and the Middle East
+44 (0) 1293 767979 | emeainfo@spirent.com

Asia and the Pacific
+86-10-8518-2539 | salesasia@spirent.com

Rev A | 04/19