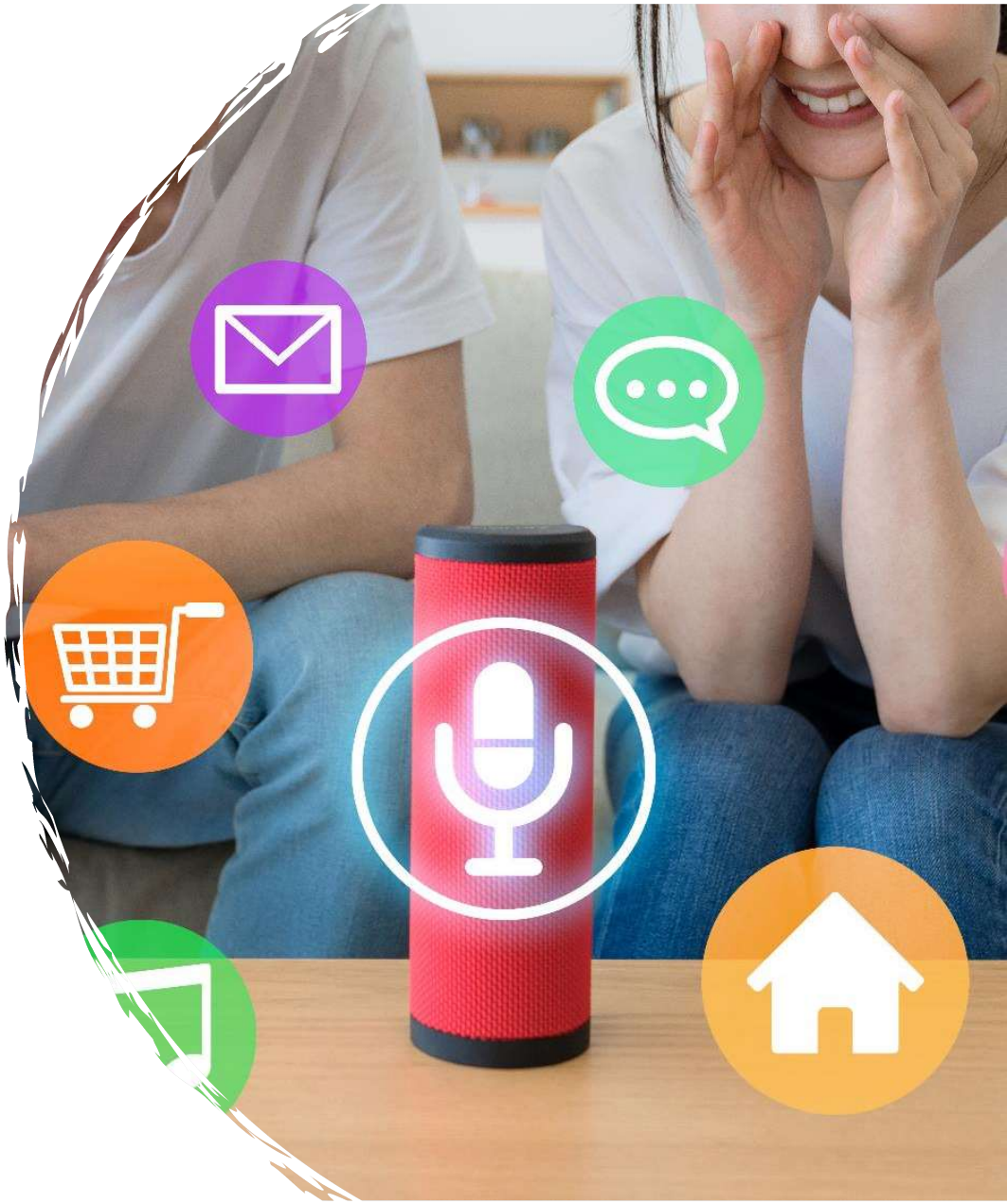
STQ Workshop

Measuring Acoustic Performance of Voice Assistance – In Practice, Alexa

David Berol



20/11/2022

Agenda

- My background with Alexa
- Introduction to Alexa / AVS Certification
- Different Audio Related Test with AVS
  - Wake Word Detection Delay (WWDD)
  - User Perceived Latency (UPL)
  - False Rejection Rate (FRR) / False Alarm Rate (FAR)
    - Test Room Requirements
    - Introduction to Detection Theory
  - Response Accuracy Rate (RAR)
- Approach to testing device in production

Alexa and me

- Joined Alexa Voice Service (AVS) in 2017 as first Field Applications Engineer
  - Previous background in product marketing and microphones for Akustica (Bosch), Audience (Knowles) and Wolfson Microelectronics (Cirrus)
- Helped write the first draft of the Acoustic Certification criteria and defined room requirements
- Supported as a solution architect for several years and maintained assistance on improvement projects impacting testing and certification.
- Until recently (July 2021-October 2022) was the manager of the AVS Certification.

knowles

Audience

wolfson® microelectronics

AKUSTICA Bosch Group

Introduction to Alexa Voice Service (AVS)

- Amazon released the 1st generation Echo in 2014 and in effect introducing Alexa at the same time
- In June 2015, Amazon announced the Alexa Voice Service (AVS) to Third Party (3P) hardware makers. This enabled 3P developer to incorporate Alexa into their products.
- This framing and branding has changed over the years but today AVS is framed as:

Alexa Built-in devices are powered by **Alexa Voice Service (AVS).** AVS manages the services and infrastructure required for Alexa experiences, and provides a suite of device APIs, SDKs, hardware kits, and documentation. There is no charge for using AVS APIs and SDKs.

- This presentation is going to describe the testing associated with 3P devices and focus on the acoustic certification criteria and tradeoffs. Amazon built devices (Echo) have a different set of tests and procedures but aren't as publicly available to reference.

AVS Certification

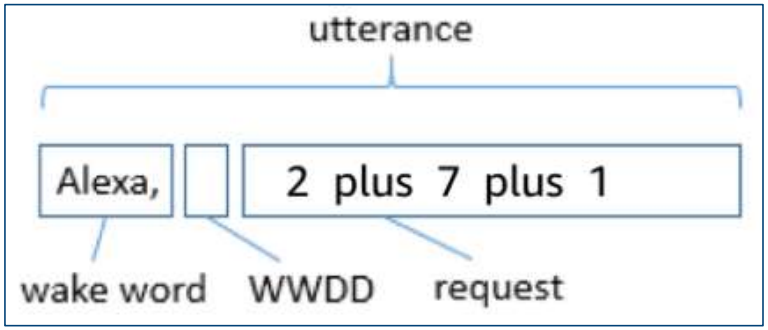
- After the integration is complete for an AVS Device (speaker, soundbar, TV, headphone, thermostat, etc.) the device is certified against a set of design areas. These are:
  - Functional
  - User Experience (UX)
  - Music
  - Acoustic
  - Security
- Not all areas apply for all devices (for example music) but for our discussion I am going to focus on the Acoustic testing

AVS Acoustic Certification / WWDD

- Several different criteria are used to evaluate a device under test:
  - Wake Word Detection Delay (WWDD)
  - User Perceived Latency (UPL)
  - False Rejection Rate (FRR)
  - False Alarm Rate (FAR)
  - Response Accuracy Rate (RAR)

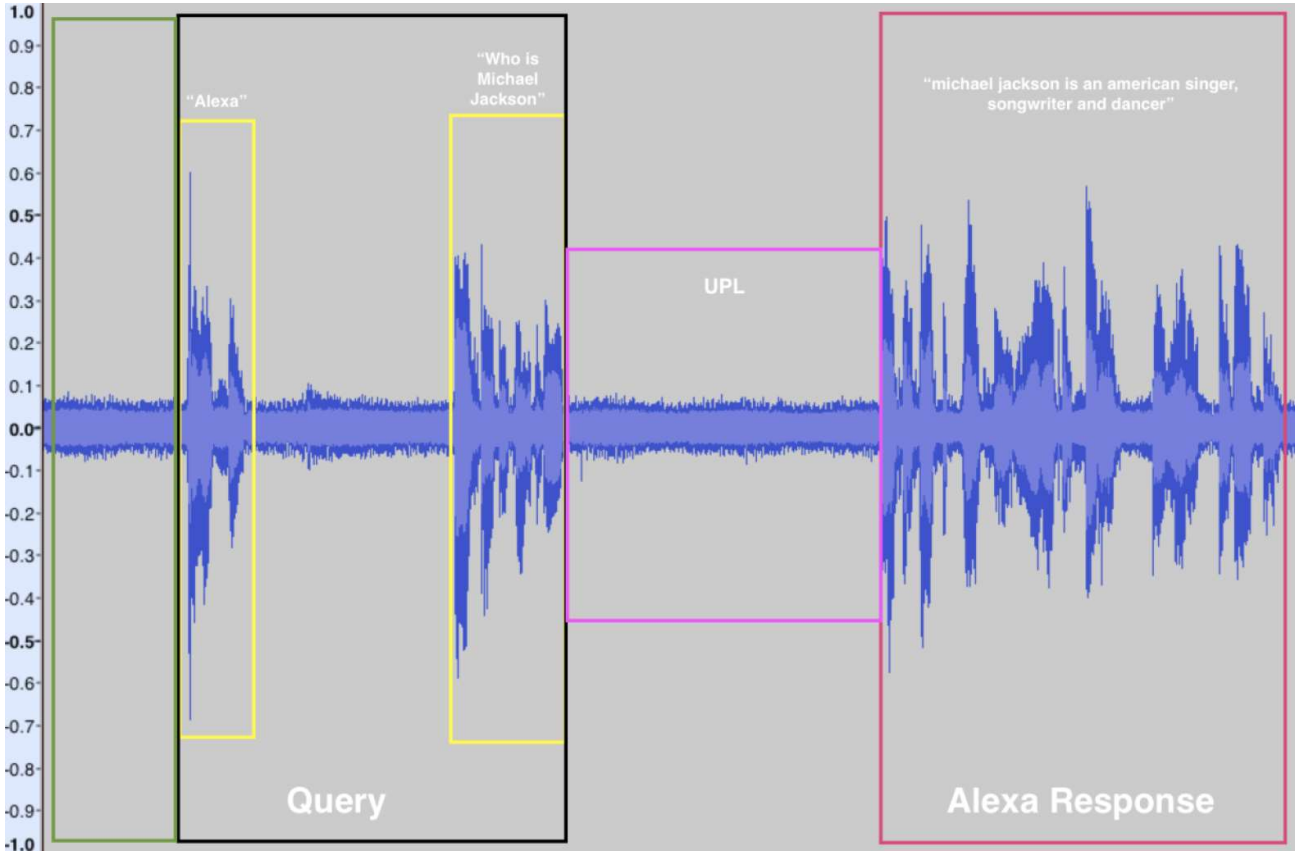
- Wake Word Detection Delay (WWDD)

  The minimum time that a particular device requires a user to pause between saying "Alexa" and the remaining request in order for the Alexa Voice Service to reliably receive the entire request



User Perceived Latency

- User Perceived Latency (UPL) is the human perceived delay in the time it takes for an Alexa device to respond to voice. It is calculated as the duration from the end of the user's speech request to the start of Alexa's response audio playback. An Alexa device is tested for UPL to ensure it does not cause excessive delay in response.



Acoustic Setup / Room Requirement

- Before going deeper into the rest of the Alexa testing, let's review how Amazon settled on a room requirement.
- We wanted to leverage state of the industry and have a low barrier to entry so not full anechoic. Most device in the start of AVS were inside the home and typically the living room.
- What those guiding principles ETSI ES 202 396-1 (v1.7.1) was used.

6.1 Test Room Requirements

The reproduction technique chosen does not require specific types of rooms such as anechoic rooms. The technique is applicable in typical office rooms as well as in anechoic or semianechoic rooms. The playback room should meet the following requirements:

- **Room Size:**

  The room size should be in a range between 2,5 × 3 m and 3,5 × 4 m. The room height should be between 2,20 m and 2,50 m.

- **Treatment of The Room:**

  Office type rooms should be equipped with a carpet on the floor and some acoustical damping in the ceiling as typically found in office rooms. A curtain should cover one or two walls in order to avoid strong reflections by hard surfaces in the room. The reverberation time of the room should be less than 0,7 s but higher than 0,2 s between 200 Hz and 8 kHz.

  For anechoic or semianechoic no additional treatment is needed.

- **Noise Floor:**

  In order to reduce the influence of external noise the noise floor measured in a room should be less than 30 $dB_{SPL}(A)$.

Room Requirement cont'd

- After years of supported various types of companies on the supply chain of making consumer electronics and testing Alexa devices it became clear that the spec some values up for interpretation. There was a work item open to hopefully bring clarity around the spec, especially for users only looking at this section and not the full document.
- This yielded ETSI ES 202 396-1 (v1.8.1):

6.1 Test Room Requirements

The reproduction technique chosen does not require specific types of rooms such as anechoic rooms. The technique is applicable in typical office rooms as well as in anechoic or semianechoic rooms. The playback room should meet the following requirements:

- **Room Size:**

  The room size shall be at least 3,0 × 3,0 m to accommodate the speaker placement in clause 6.2. The room should be larger to avoid placement of speakers too close to the walls. This implies the room size should be at least 4,0 × 4,0 m. The room height should be made in consideration of the reverberation time. This leads to a typical range between 2,20 m and 2,50 m. Equal dimensions in length and width are not recommended as are being a multiple of the height dimension.

- **Treatment of The Room:**

  Office type rooms should be equipped with a carpet on the floor and some acoustical damping in the ceiling as typically found in office rooms. A curtain should cover one or two walls in order to avoid strong reflections by hard surfaces in the room. The reverberation time of the room should be less than 0,7 s but higher than 0,2 s for each of the octave bands of 250 Hz and 8 kHz. Explicitly that would be: 250, 500, 1 000, 2 000, 4 000, 8 000 Hz.

  For anechoic or semianechoic rooms, diffuse treatment may be needed to meet reverberation time minimums.

- **Noise Floor:**

  In order to reduce the influence of external noise the noise floor measured in a room should be less than 30 $dB_{SPL}(A)$.

AVS Certification – Speaker Placement

- Speaker layout within “ETSI Room” for AVS Acoustic Certification

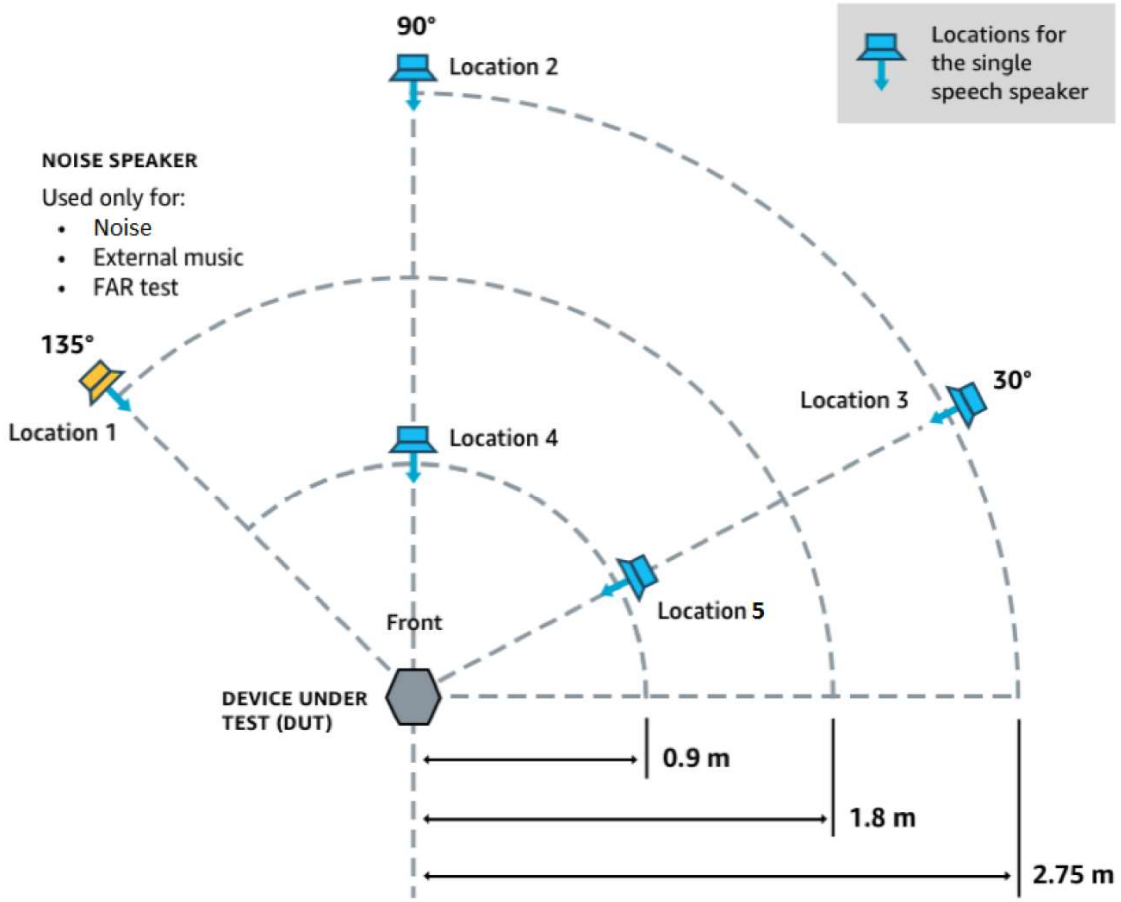


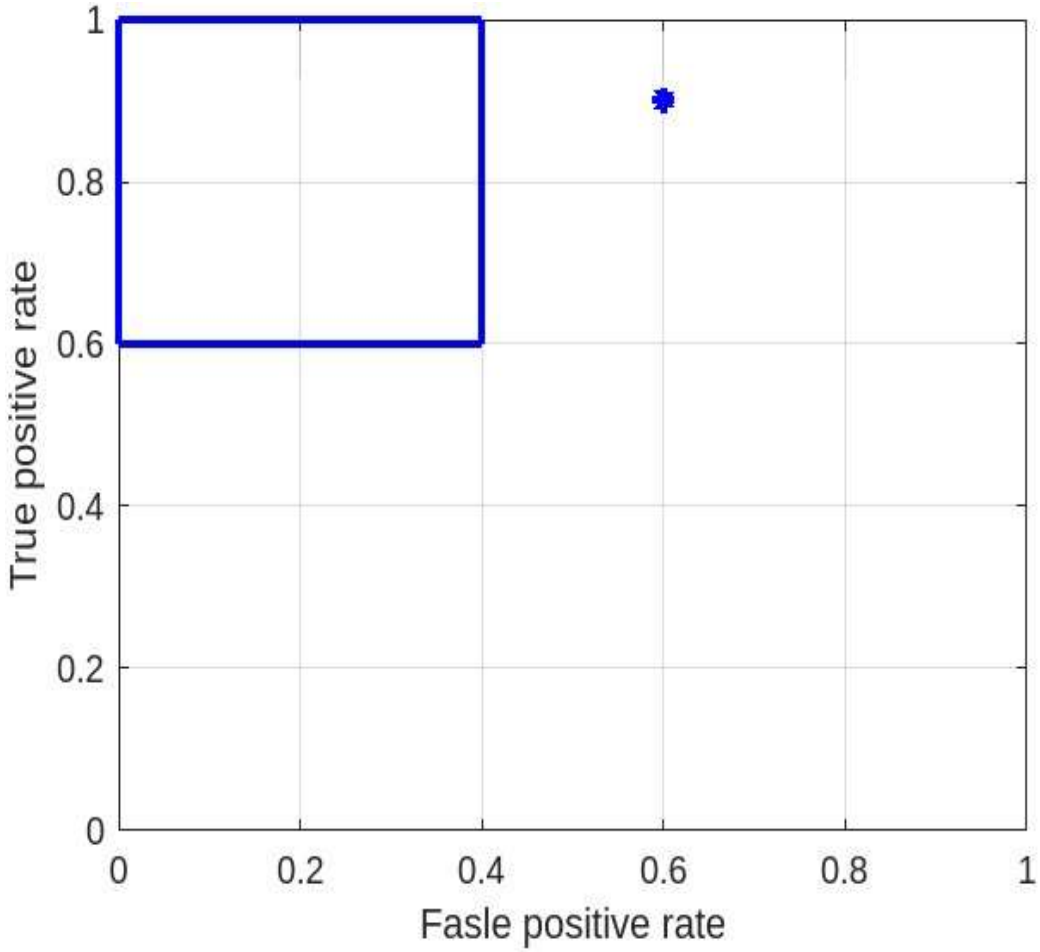
Figure 15: Far-field and Near-field test environment layout

False Rejection Rate / False Accept Rate

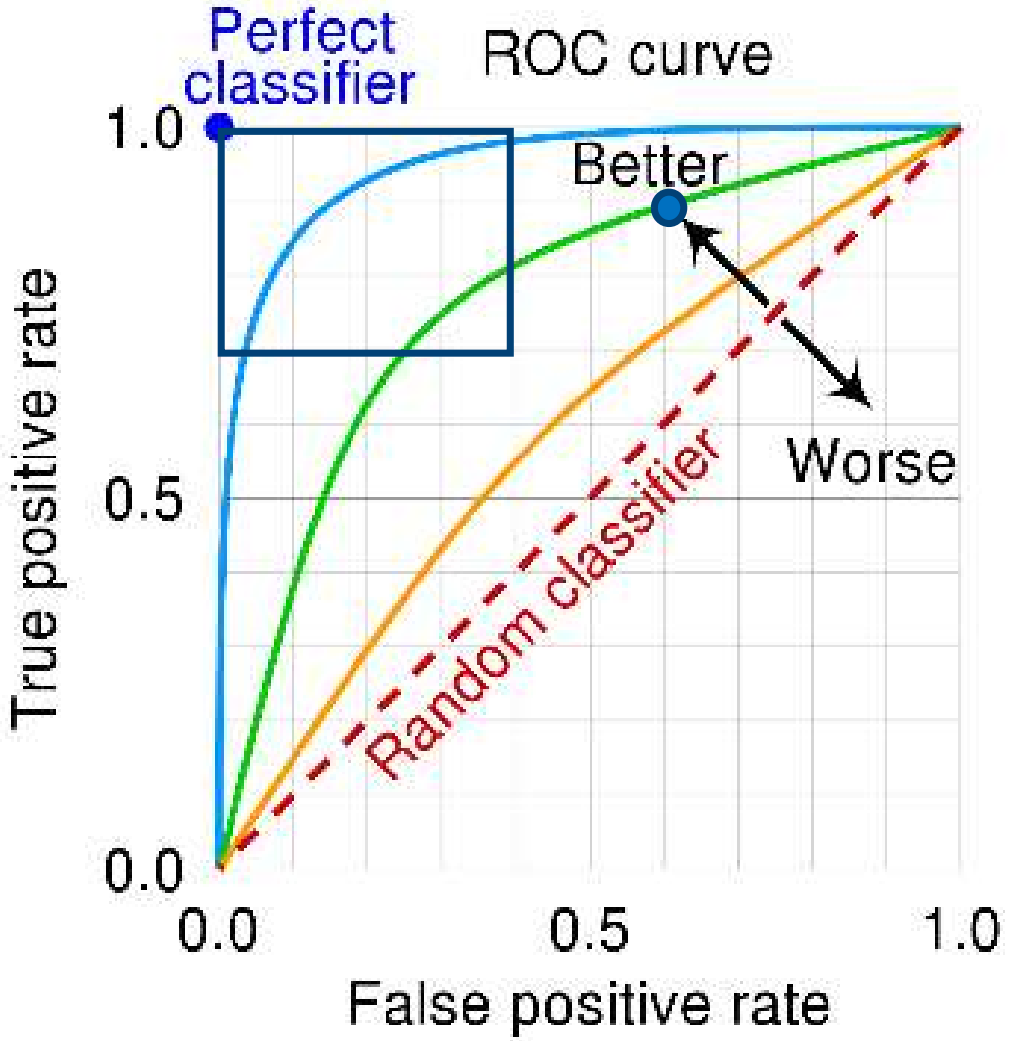
- The Wake Word False Rejection Rate (FRR) is the number of missed wake words per wake words spoken. For example, if ten requests are given and the device wakes up nine out of ten times, the FRR is (1 missed wake word / 10 wake words spoken), or 0.1 (10%).

- In the False Accept Rate (FAR) test (aka False Alarm Rate), you assess the number of times a device incorrectly awakes during the test period.
  - Originally, this was a getting 3 or less wake events on a 24-hour file. This time period was extrapolated from the expected wake permitted in a given week.
  - To help with test time, the team evaluated short windows of time in the 24-hour file for a detection and parsed that down into a much shorter file giving the same coverage.

Brief Introduction to Detection Theory

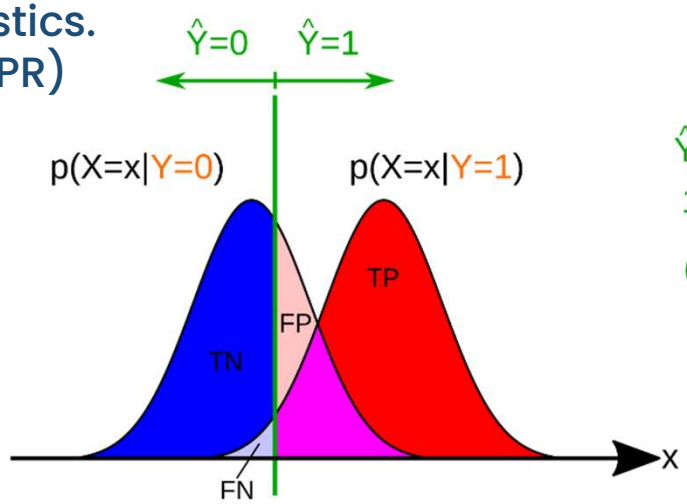


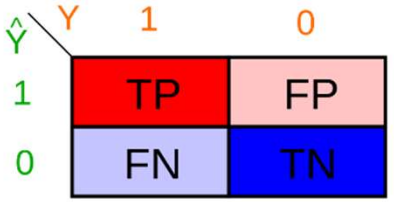
How to improve performance

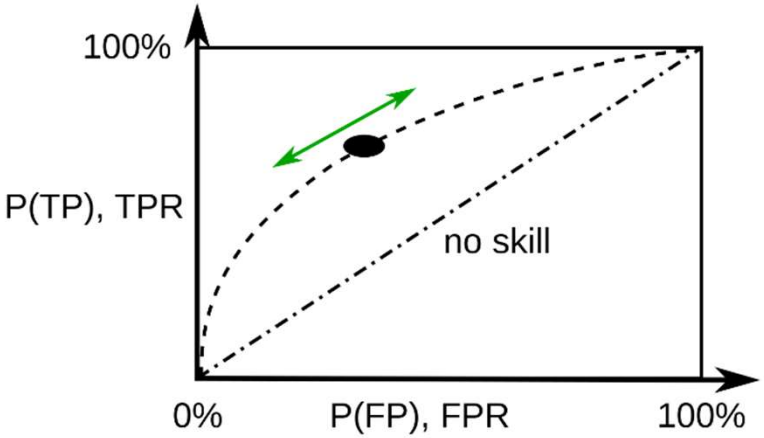


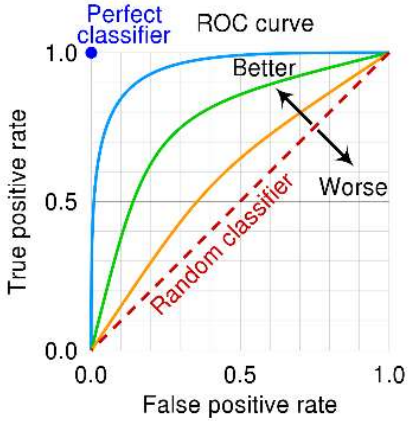
Receiver Operating Curve (ROC) and Classifiers

- The balance between FAR / FRR performance and the wake word models leverage null hypothesis statistics.
- True Positive Rate (TPR)
- TPR = FRR – 1









https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Response Accuracy Rate

- The Response Accuracy Rate (RAR) of a device is the number of right responses per corresponding requests. For example, if ten requests are given and the response from Alexa is good nine out of ten times, the RAR is (9 correct responses / 10 requests), or 0.9 (90%).

- Originally, the performance was evaluated against the response from the device. “What time is it?”. It was confirmed that the “Time Is …” was in the response.

- Now, in addition to that, the test case will be marked correct if the phrase passed an ASR matches the given transcription. “What time is it? = What time is it?”

- This test is viewed as most closely aligned with the device’s acoustic front end (AFE) which is almost always externally supplied (not by Amazon).

Considerations for Testing in Production

- The nature of testing a live voice assistant in production is that things will change.

- Specifically for Alexa, an example of this was a test utterance: “Remind me to buy Panasonic ear bud headphones” which resulted in “Adding Buy Panasonic Ear Bud headphones to your to do list”
- This worked until Alexa added a reminders feature which changed the behavior on devices when invoked with “Remind me”. Additionally, 3rd party devices or AVS devices didn’t yet support reminders.
    - The resulting response was “Your device doesn’t support reminders” which doesn’t give enough information if it parse the whole phrase which is the goal of the test.

- In practice, QA and certification testing should be done assuming the test will go a planned but have the infrastructure to capture or record (camera, logs, etc.) anything we there is a repeated difference from the publish expectation.

Questions

Backup