ALANGO

Technologies and solutions

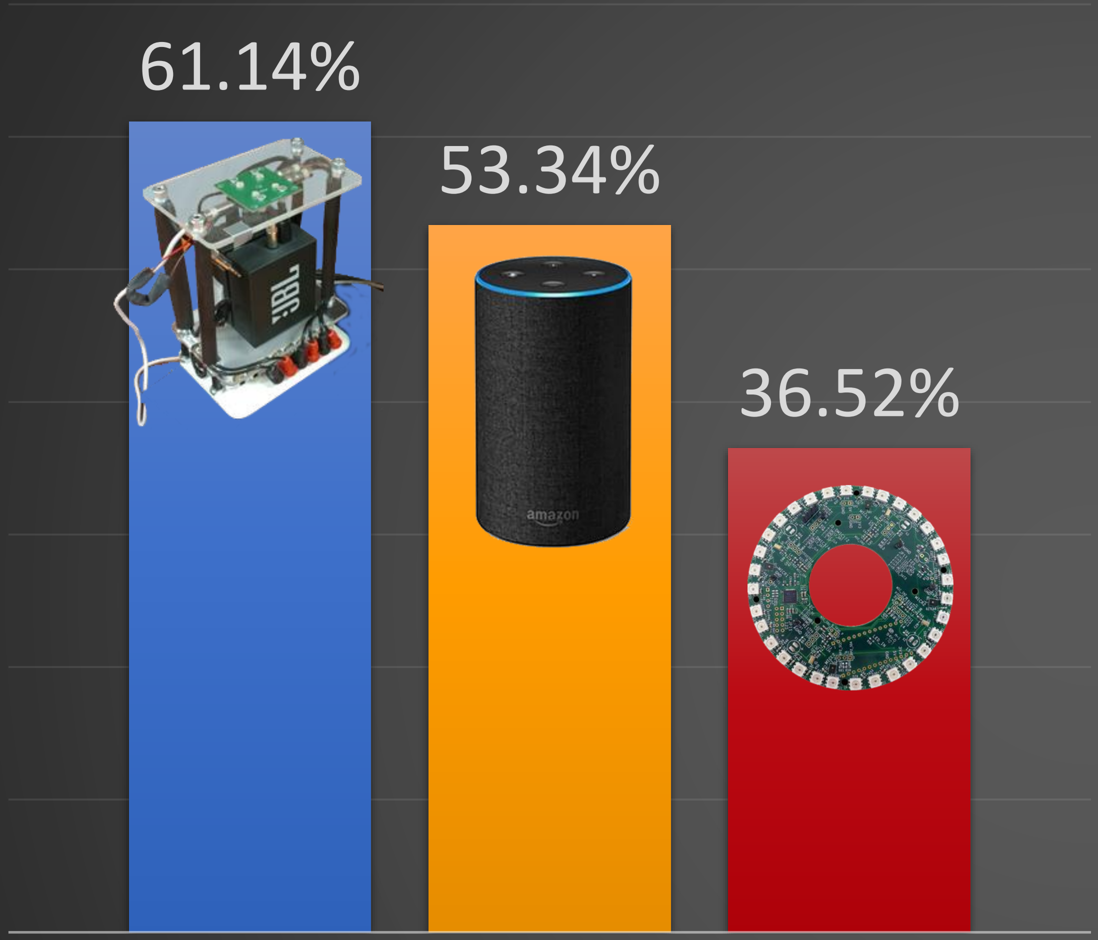
Keyword Recognition Performance with Alango Voice Enhancement Package (VEP)

DSP software solution
for multi-microphone voice-controlled devices

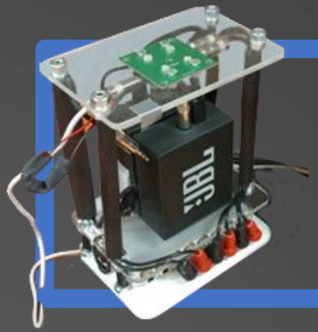
V1.19, 2018-12-25

ALANGO
Technologies and solutions

Executive Summary

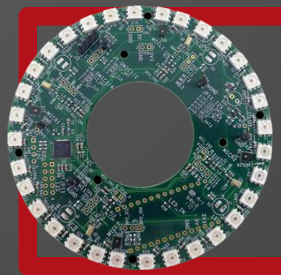


Keyword Recognition Rate
Average Over Tested Scenarios



Alango Duplexa
4-mic

Amazon Echo 2
7-mic

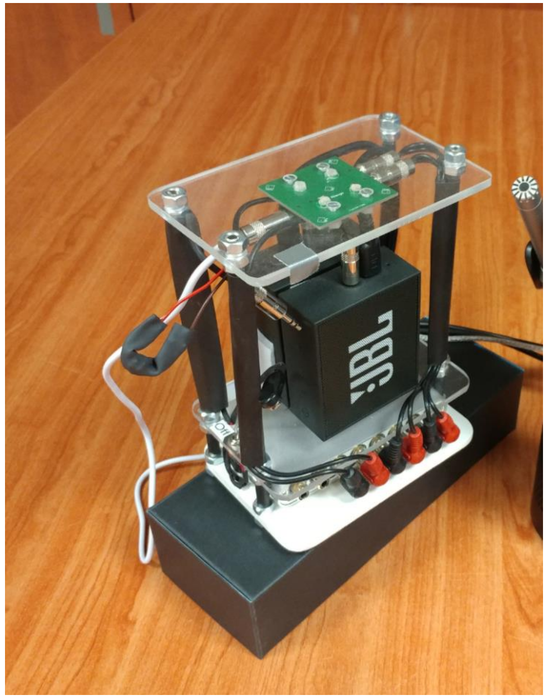


Synaptics
4-mic

Comparison Goal and Devices Under Test

The goal of the test detailed in this document is to **compare keyword recognition rates in various noise environments** using the following devices:

- **Alango Duplexa -** Alango VEP-powered 4-microphone smart-speaker
- **Amazon Echo 2**
- **Synaptics AudioSmart** 4-Mic Development Kit



Alango 4-mic Duplexa

Amazon Echo 2

Synaptics AudioSmart 4-mic Dev Kit

Test Summary and Intent

Summary of devices tested:

Alango "Duplexa" – a 4 microphone smart-speaker demo/test kit powered by
- Alango VEP ("Voice Enhancement Package") multi-microphone beamforming DSP library and
- Sensory TrulyHandsfree Keyword recognition technology 6.0 with high resolution speech features and on-device wake word post-qualification (in Sensory's THF SDK v6.5.2).

The device is capable of:

- recognizing "Alexa" keyword followed by a user command
- showing direction of arrival to the user
- performing corresponding actions, such as controlling music playback, showing (on screen) transcribed text recognized from user speech

- **Amazon Echo 2** – a 7 microphone consumer device
- **Synaptics AudioSmart** – a 4-microphone development kit for Amazon AVS

Intent:

Test and compare keyword recognition rate of three devices in absolutely equivalent, controllable and reproducible acoustic environments.

The test is completed by placing the devices in identical acoustic conditions, pronouncing a trigger word, and registering device reaction.

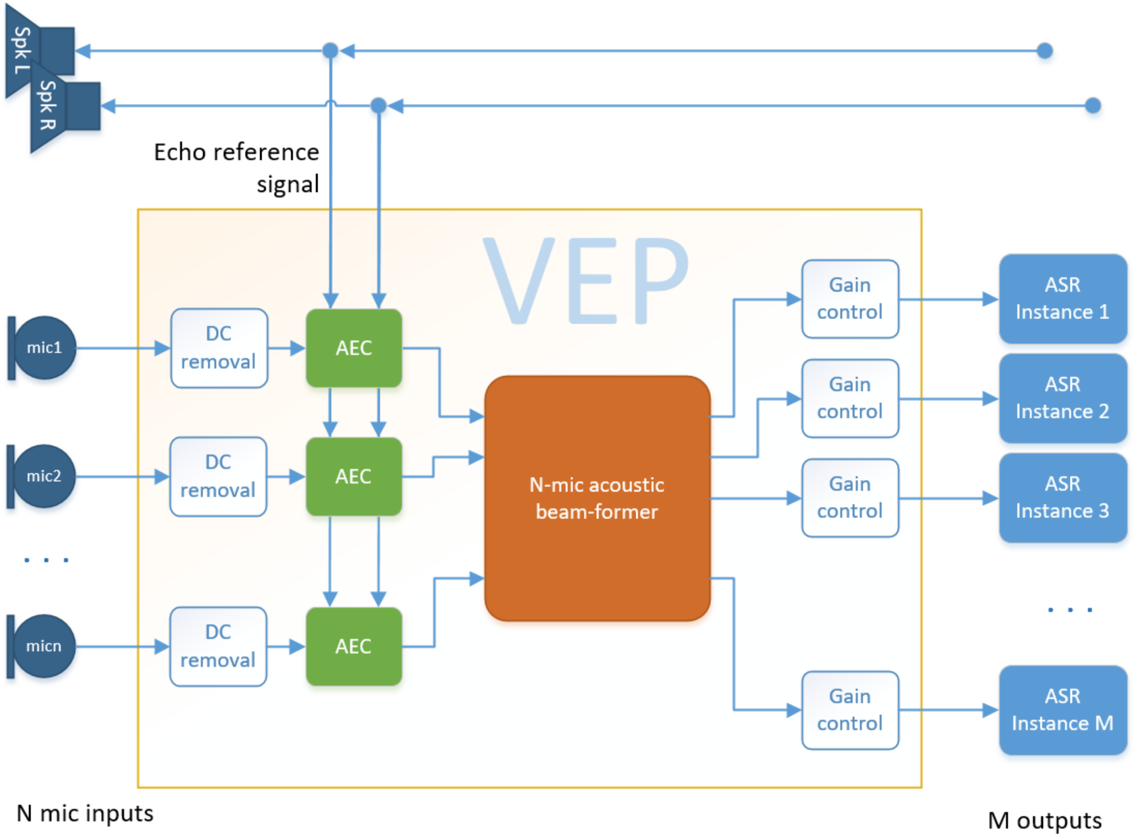
About Alango VEP Library

Alango Voice Enhancement Package (VEP) is a suite of real-time software DSP technologies designed for improving speech recognition performance in multi-microphone, voice-controlled multimedia devices. The VEP audio signal processing solution is designed to run on embedded platforms and is intended for use as a front-end to Automatic Speech Recognition (ASR) and Keyword Recognition (KWR) engines. The VEP DSP library is highly configurable, flexible, and universal. Typical VEP applications include voice-controlled multimedia and smart devices, such as home speakers, automotive infotainment systems, etc.

The technologies included in VEP are universal and do not interfere with the signal processing performed by modern ASR engines. VEP algorithms are designed to never spoil speech recognition rate, but only improve it. Because of this, VEP is ASR engine-agnostic; it can be used in conjunction with any ASR engine.

VEP includes the following high-level DSP blocks:

- DC removal
- Acoustic echo canceller (mono or stereo)
- Multi-microphone beam-former
- Gain control
- Single-channel noise reduction (per channel)



About Alango VEP Library (cont.)

VEP receives N signals from the N physical microphone sensors (N >=2).

- VEP supports single-, two- and three-dimensional microphone arrays.

VEP is universally configurable to support **multiple microphone configurations and geometries.**

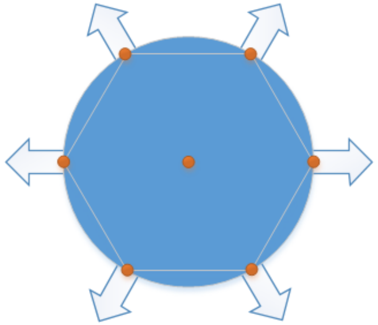
- VEP supports circular, linear and non-symmetric microphone configurations.

VEP output consists of M+1 signals:

- Each of the M outputs (“beams”) enhances the voice coming from a specific spatial direction.
- The additional “Mux” output combines all the enhanced beams into a single audio channel based on internal logic.

Thus, the integrator may select the most suitable **VEP integration strategy:**

- a “multi-beam integration approach” in which each of VEP’s M output beams is analyzed by a dedicated ASR/KWR instance,
- a “single-beam integration approach” in which on only the VEP’s Mux channel is used by ASR/KWR, with beam switching and selection managed internally by VEP



Circular 7-microphone array for smart speakers with 360-degree voice pickup

Linear 6-microphone array for flat panels with 180-degree voice pickup

ALANGO
Technologies and solutions

About Alango VEP Library (cont.)

VEP integration approaches compared

Multi-beam integration approach:

Advantages:

- no need to search for the “right” direction to listen to based on some heuristic == **the best possible recognition results!**
- very low delay VEP processing (only 24 ms algorithmic delay)

Disadvantages:

- increased resource consumption, since M ASR instances must be run simultaneously to listen to M beams (at least while waiting for the trigger)

Single-beam integration approach:

Advantages:

- significantly reduced computational resources -- single ASR instance on Mux beam is enough

Disadvantages:

- possibly imperfect performance in difficult noisy situations
- increased audio latency (configurable, > 100 ms) for Mux beam selection and decision making

Test Description and Prerequisites

To exclude any possible deviations in the test flow and to ensure repeatability, the entire test is done in a fully automated environment without any human interaction or manual operations.

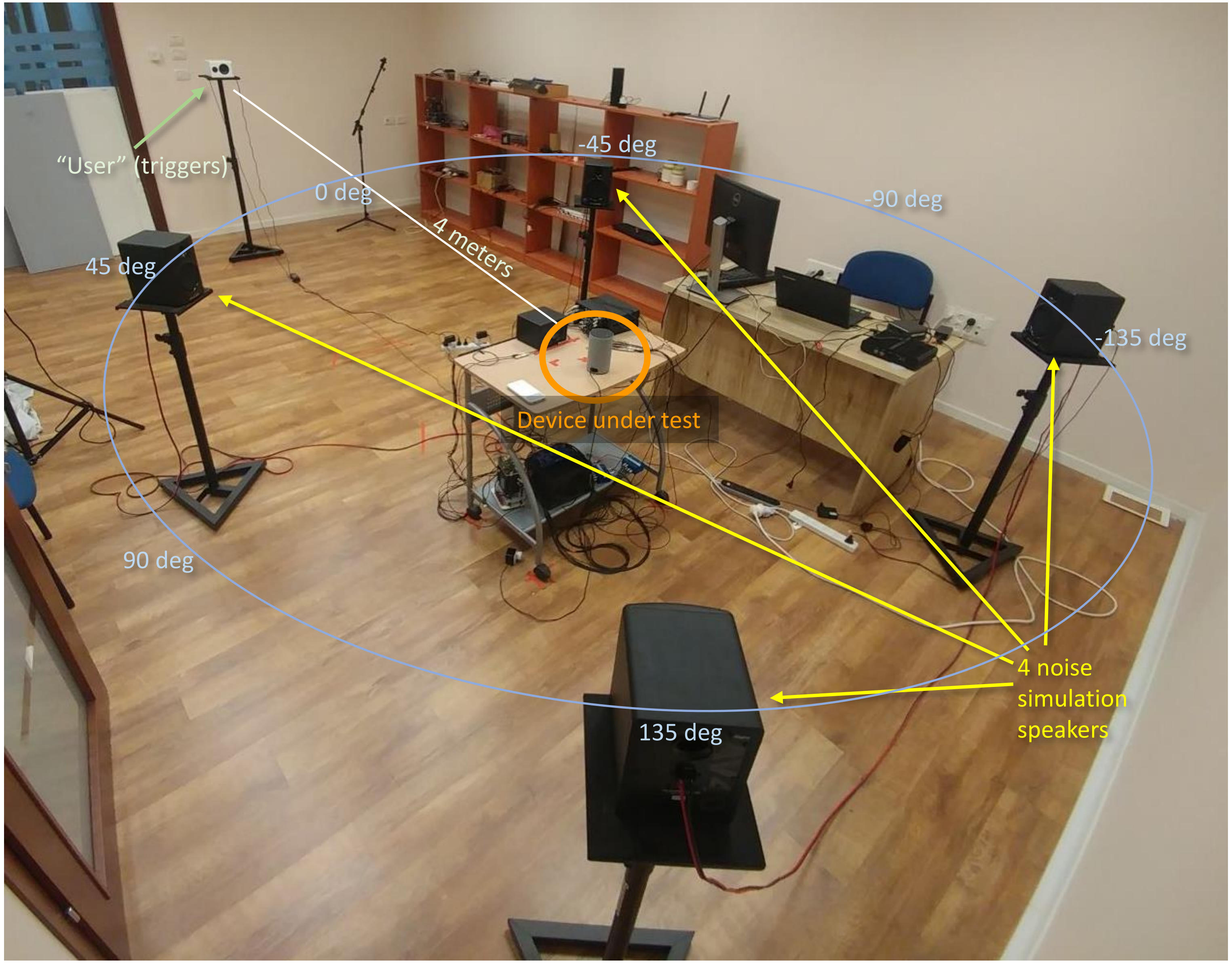
Test setup and components:

- Real meeting room
- Devices under test (DUT): Alango Duplexa (4 mic), Amazon Echo (7 mic), Synaptics Audio-Smart (4 mic)
  - All devices interface with Amazon AVS, either natively or via Raspberry PI running AVS firmware image
- A mouth simulator generating the voice trigger “Alexa” using different voices and voice levels
- Surround noise simulation system consisting of 4 speakers and test noises

Test flow:

- Place the tested devices near each other at the same location of the table, then turn on the devices
- Start noise playback using 4 loudspeakers. The loudspeakers play uncorrelated sounds and noises to simulate the real acoustic environment in which different sounds arrive from different directions and sound sources
  - Pre-defined noise levels: ~55-60 dB SPL @ DUT
- Place mouth simulator at a fixed position (~4 meters away from the DUT) to play “Alexa” trigger word using
  - 3 different voice levels
  - 10 different real human voices (5 male and 5 female)
- Start playback of trigger word “Alexa” followed by playback of “Stop it” (to suppress device reaction/answer on the trigger) using different pre-recorded real human voices
- At the end of the test, collect “Alexa” recognition statistics via a log appearing in the Amazon AVS app

Test Setup



"User" (triggers)
0 deg
4 meters
-45 deg
-90 deg
45 deg
-135 deg
Device under test
90 deg
4 noise simulation speakers
135 deg

What Is Compared

The following charts display comparisons of "Alexa" keyword recognition rate for:

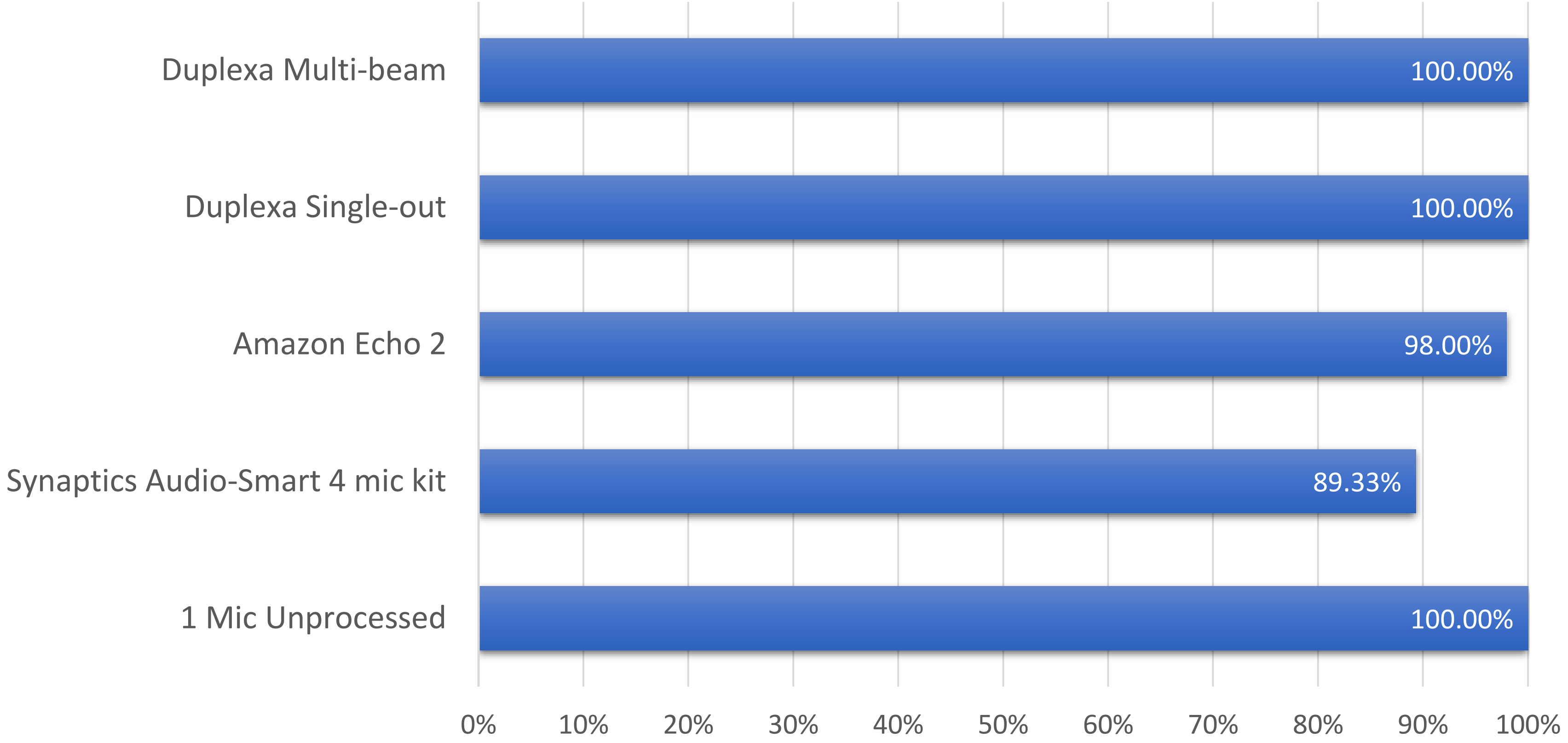
- Duplexa 4-mic using VEP multi-beam
- Duplexa 4-mic using VEP single-output
- Amazon Echo 2
- Synaptics Audio-Smart 4mic development kit
- Duplexa "no processing" (effectively, Sensory KWR working with unprocessed single mic raw audio)

The tests are performed under various simulated acoustic noise conditions:

- Not noisy, silent room, normal speech level
- Not noisy, silent room, weak speech level
- Surround cafeteria noise
- Uncorrelated white noise
- Distractor (TV) at 45 degrees
- Distractor (TV) at 135 degrees (behind)

Test Results

Not noisy, silent room, normal speech level



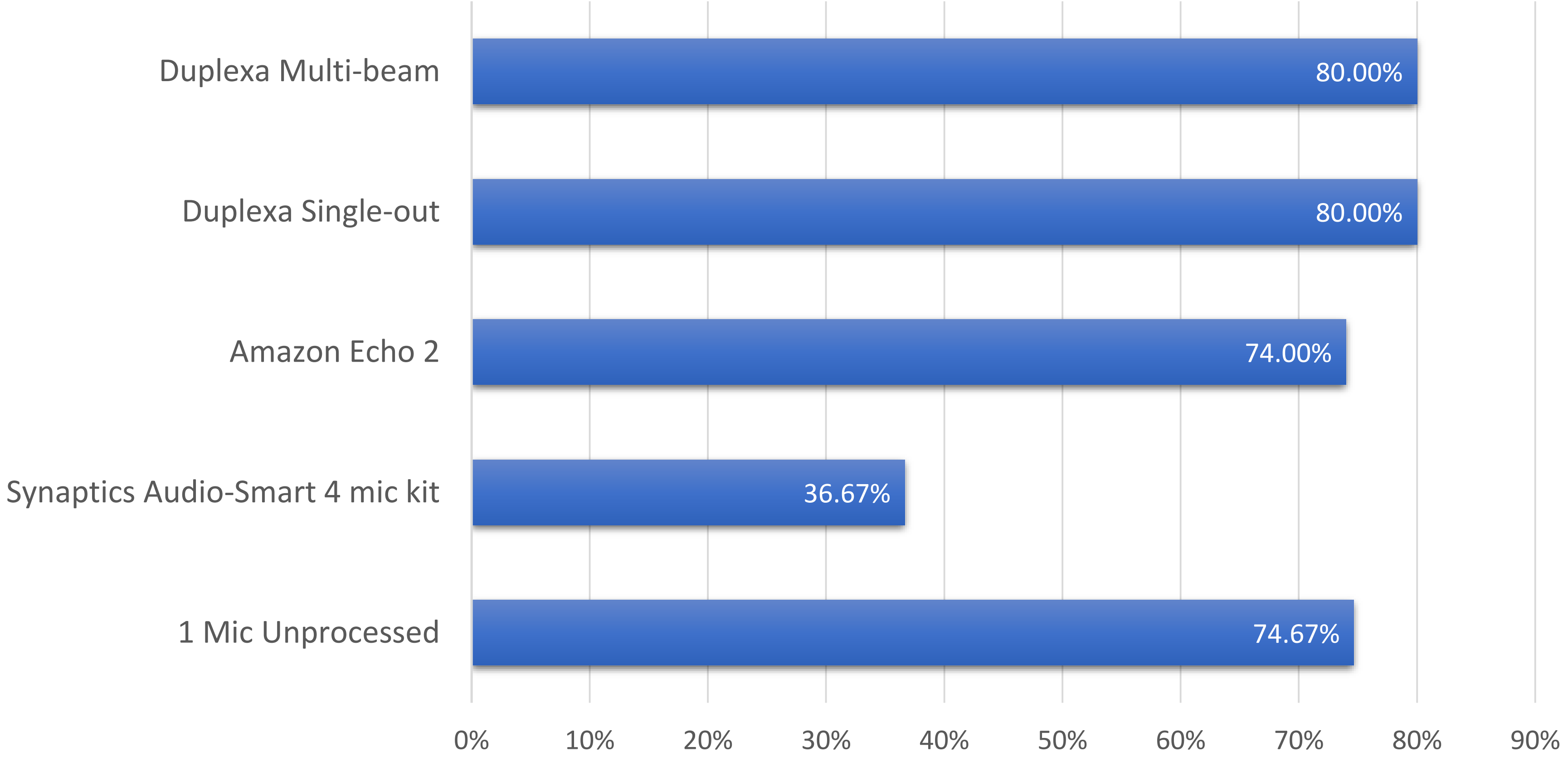
Exemplary audio snippets as recorded @ DUT (normalized for listening convenience)

Note: the imperfect recognition of Amazon Echo 2, and especially, of Synaptics Audio Smart is not accidental -- it is consistent and reproducible.

The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~35 dB SPL @ DUT; user speech level and distance: 94,88,82 dB SPL @ MRP (65,59,53 @ DUT), distance to DUT is 4m.

Test Results

Not noisy, silent room, weak speech level



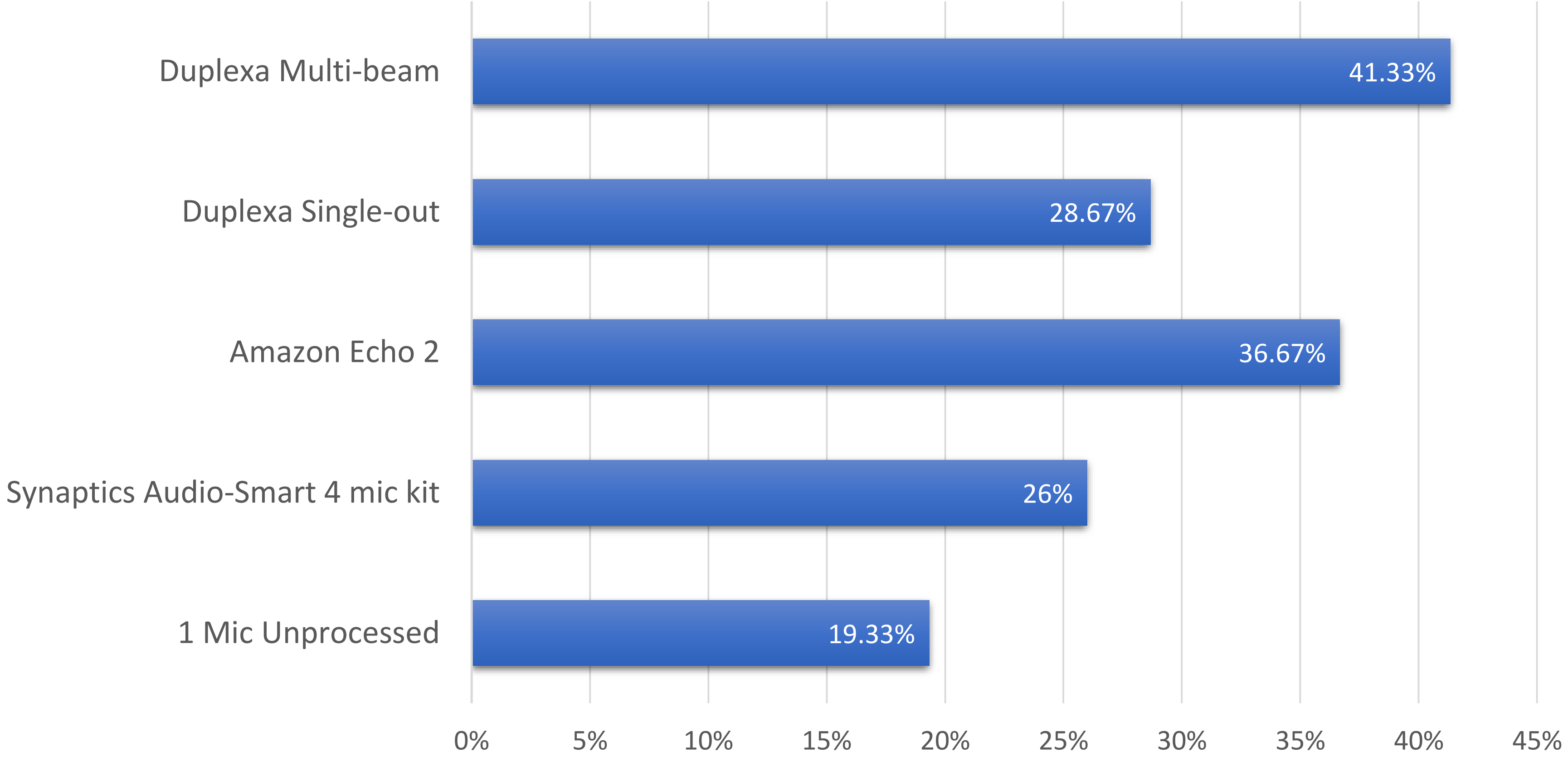
Exemplary audio snippets as recorded @ DUT (normalized for listening convenience)

Note: according to our observations, Synaptics suffers from a high level of electrical self-noises, most likely produced by LEDs operation which significantly spoils the results in silent conditions.

The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~35 dB SPL @ DUT; user speech level: 50,44,38 dB SPL @ DUT

Test Results

Surround cafeteria noise



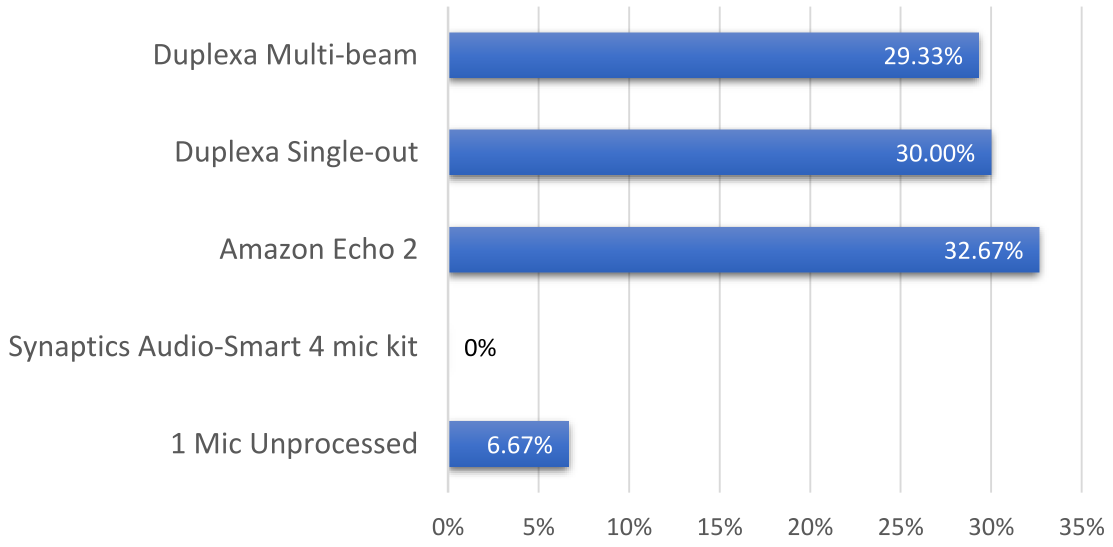
Exemplary audio snippets as recorded @ DUT (normalized for listening convenience)

The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~55 dB SPL @ DUT; user speech level: 60,54,48 dB SPL @ DUT.

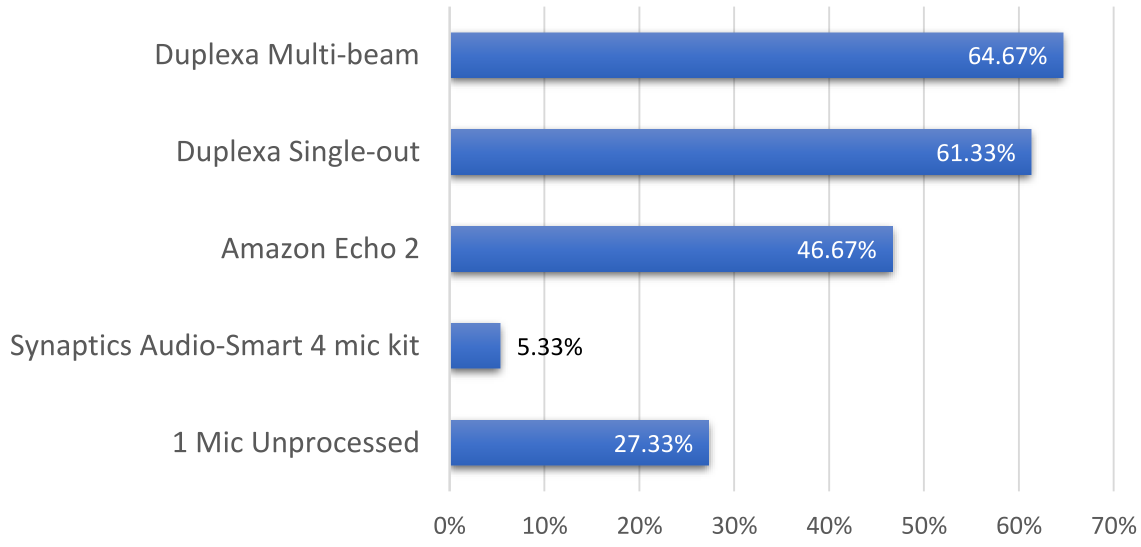
ALANGO Technologies and solutions

Test Results

Uncorrelated white noise



Surround uncorrelated white noise from all sides (4 speakers)



Uncorrelated White noise from behind (2 speakers)

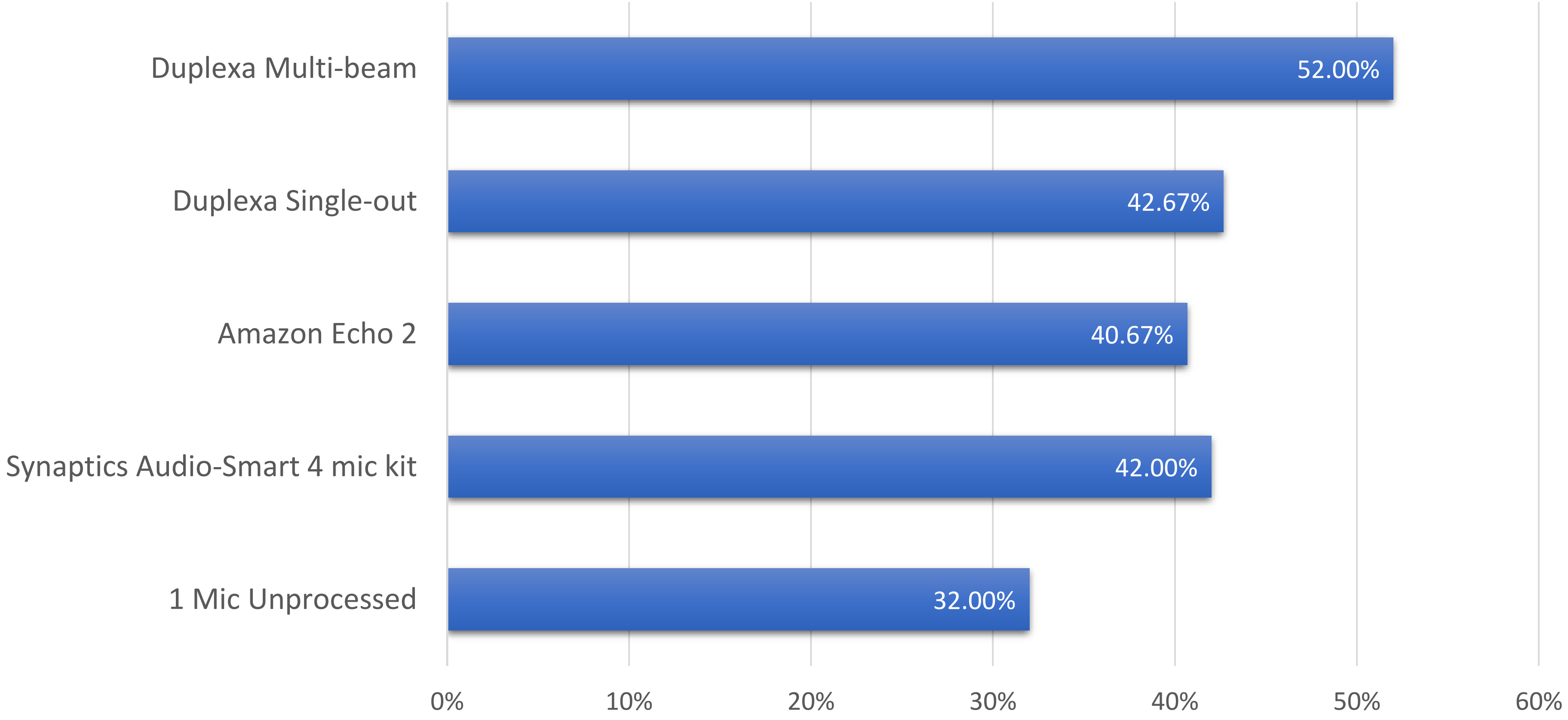
Note: there is no mistake; the Synaptics kit completely fails on this test. The test has been repeated several times to exclude any mistakes and re-validate correctness.

The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~60 dB SPL @ DUT; user speech level: 60,54,48 dB SPL @ DUT.

ALANGO
Technologies and solutions

Test Results

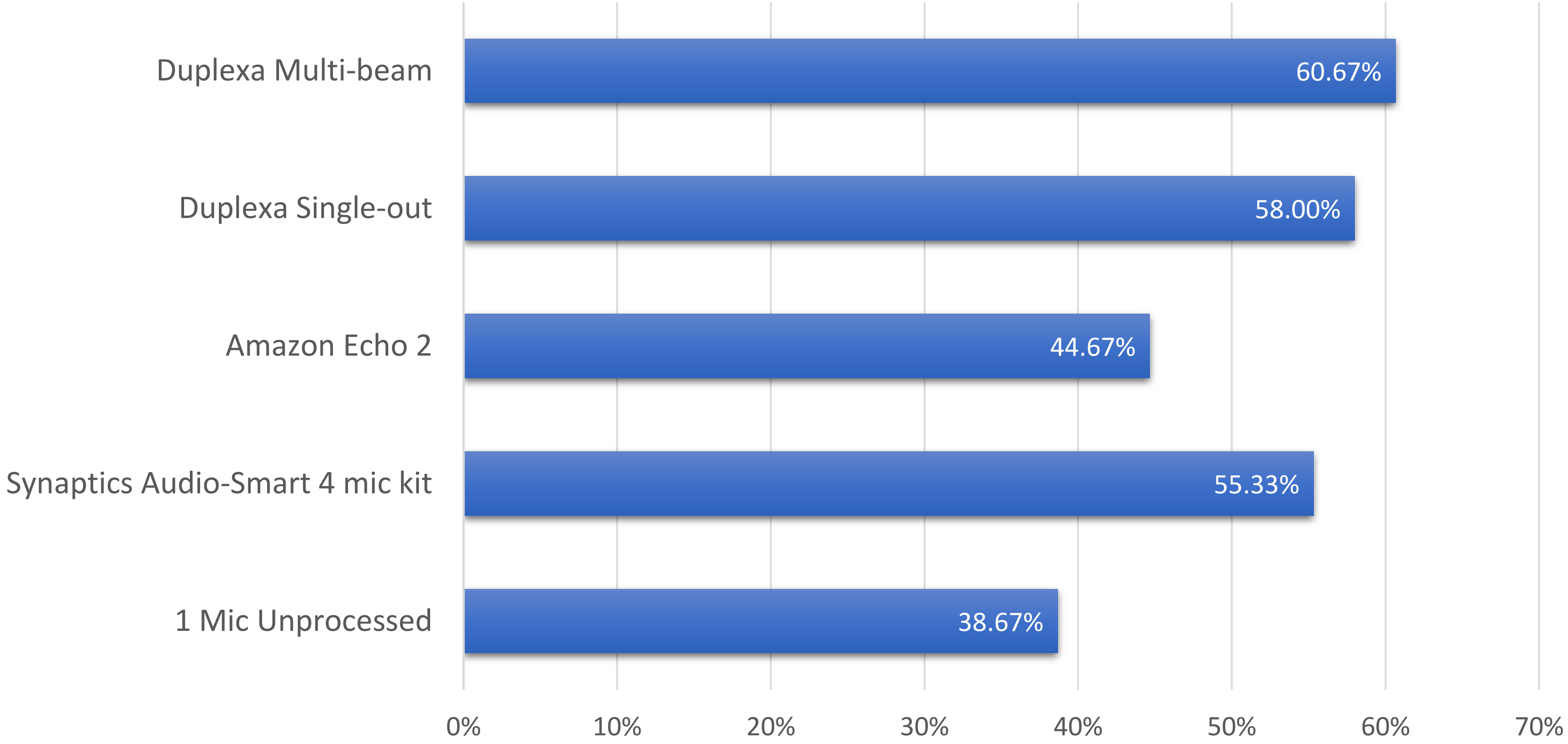
Distractor (TV) at 45 degrees



The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~55 dB SPL @ DUT; user speech level and distance: 64,58,52 dB SPL @ DUT.

Test Results

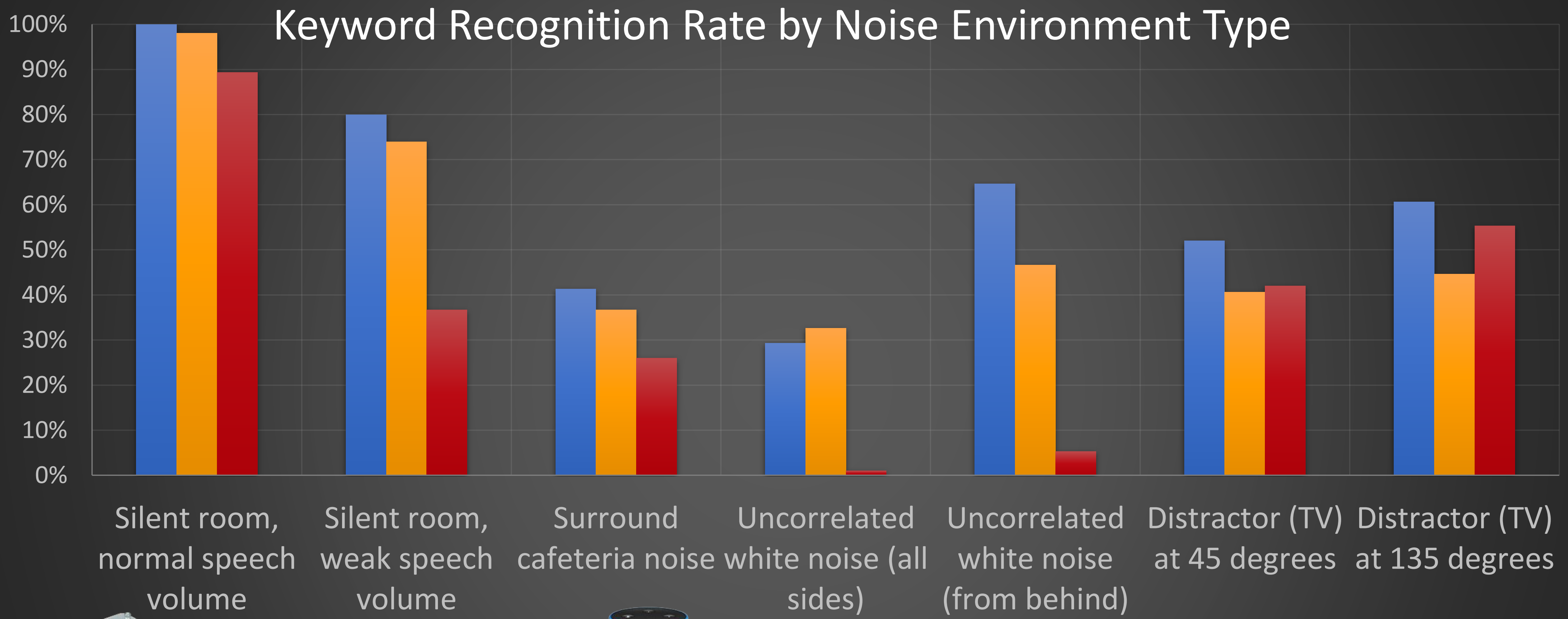
Distractor (TV) at 135 degrees (behind)

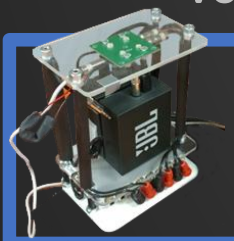


Exemplary audio snippets as recorded @ DUT (normalized for listening convenience)

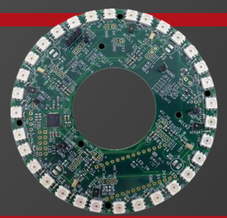
The results above are the hit-rate average over 150 “Alexa” triggers spoken by 10 different voices (5 male + 5 female) at 3 different SPL levels, 5 times for each trigger. Average noise level: ~55 dB SPL @ DUT; user speech level and distance: 64,58,52 dB SPL @ DUT.

Summary of Test Results


Keyword Recognition Rate by Noise Environment Type


Alango Duplexa
4-mic

Amazon Echo 2
7-mic


Synaptics
4-mic

Conclusions

Alango Duplexa demo and test kit, powered by VEP DSP library and equipped with 4 microphone array, outperforms (in the “multi-beam” approach) or works on par with (in the single-output approach) the 7-microphone Amazon Echo 2.

Duplexa demonstrates an even larger performance advantage compared to the Synaptics Audio-Smart 4mic dev kit.

Alango beamforming significantly improves ASR results in all acoustic conditions compared to a single-microphone device having no beamforming capabilities.

Both Duplexa and Echo 2 are shown to never spoil KWR performance compared to a single mic, unprocessed reference signal.

Contact information

Don’t hesitate to contact us if you want to be our customer or just have some comments. We are looking forward to hearing from you!

Please, send your questions, comments, thoughts, proposals to info-il@alango.com or specifically to:

Mr. Robert Schrager (Sales enquiries):	sales-il@alango.com
Mr. Alex Radzishevsky (Technical enquiries):	tech-il@alango.com
Dr. Alexander Goldin (CEO):	ceo-il@alango.com

Alango Technologies Ltd.
2 Etgar St. PO Box 62
Tirat Carmel 39100 Israel
Phone: +972 4 8580 743
Fax: +972 4 8580 621