Value of high-SNR microphones in Voice User Interface

About this document

Scope and purpose

This document highlights the benefits of microphones in a Voice User Interface (VUI), with the focus on challenging use cases, and recommendations for evaluation standards and the need for system optimization for improved overall VUI system performance.

Intended audience

This document helps guide system evaluation and testing teams tasked to evaluate microphones, microphone arrays or complete VUI systems.

Table of contents

1 What is a VUI?

A VUI enables interaction between people and devices using voice as the means of communication. It enables the transfer of information in the form of commands and questions to an electronic system with or without cloud connectivity. VUIs are implemented in many consumer applications such as smartphones, smart TVs and smart-home devices (e.g. Amazon Echo or Google Home). The concept of VUI is based on capturing audio signals using a single microphone or an array – see Figure 1.

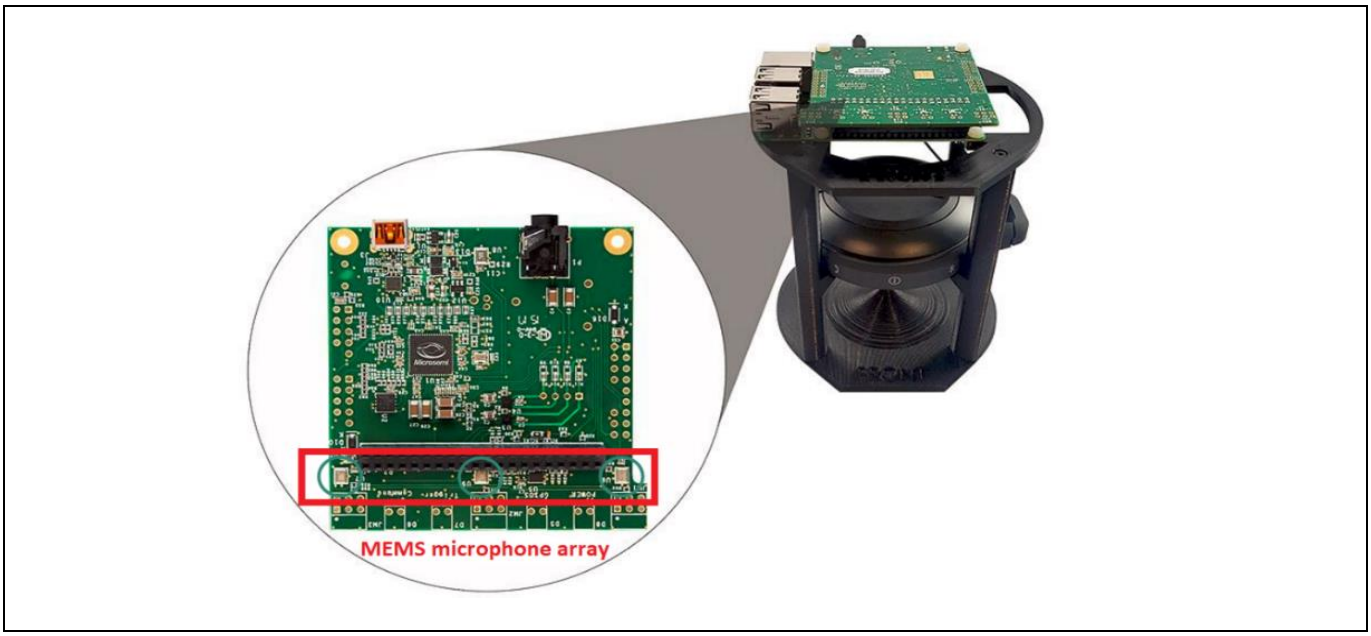

Figure 1 Interior of a smart speaker with a linear microphone array

The recorded voice command is processed by an application processor to improve the signal quality by means of beam-forming, noise cancelation and other speech-enhancement algorithms. The improved signal is sent to the cloud (e.g. Amazon Web Services/AWS or Google Cloud) for keyword and command recognition. The corresponding output signal (e.g. an answer to a question or a command) is finally aired or executed by the VUI or a secondary integrated device component – see Figure 2.

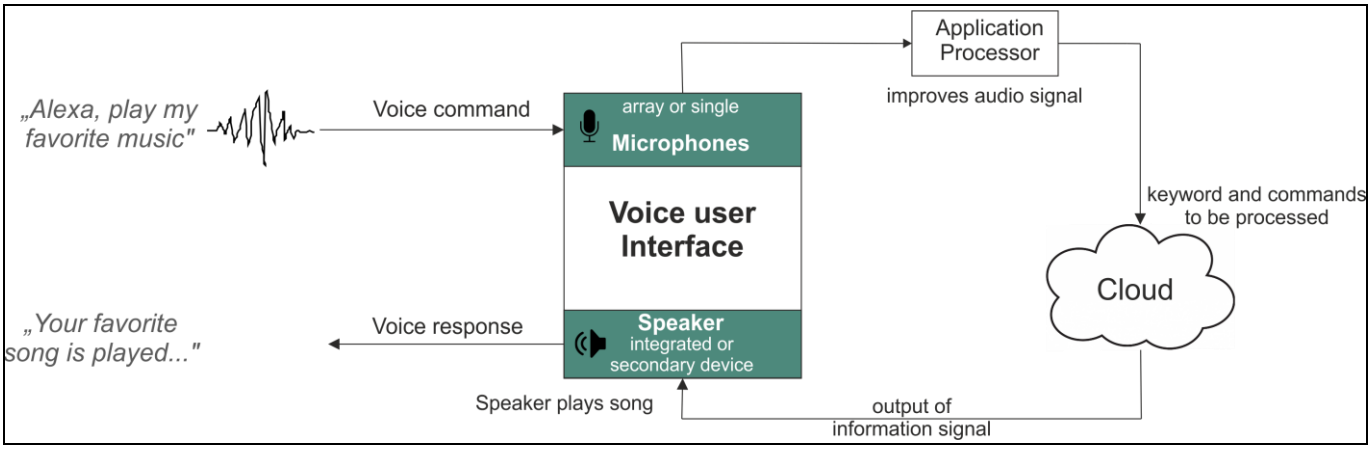

Figure 2 Open-loop signal flow diagram of a VUI

2 Benefits of high SNR

Most VUI interfaces use MEMS microphone arrays for blind source separation and speaker localization, and to detect commands in the presence of background noise (known as the cocktail party effect). The performance of a microphone array is defined by the performance of its individual microphones. Microphone performance is often characterized by self-noise and dynamic range – see Figure 3.

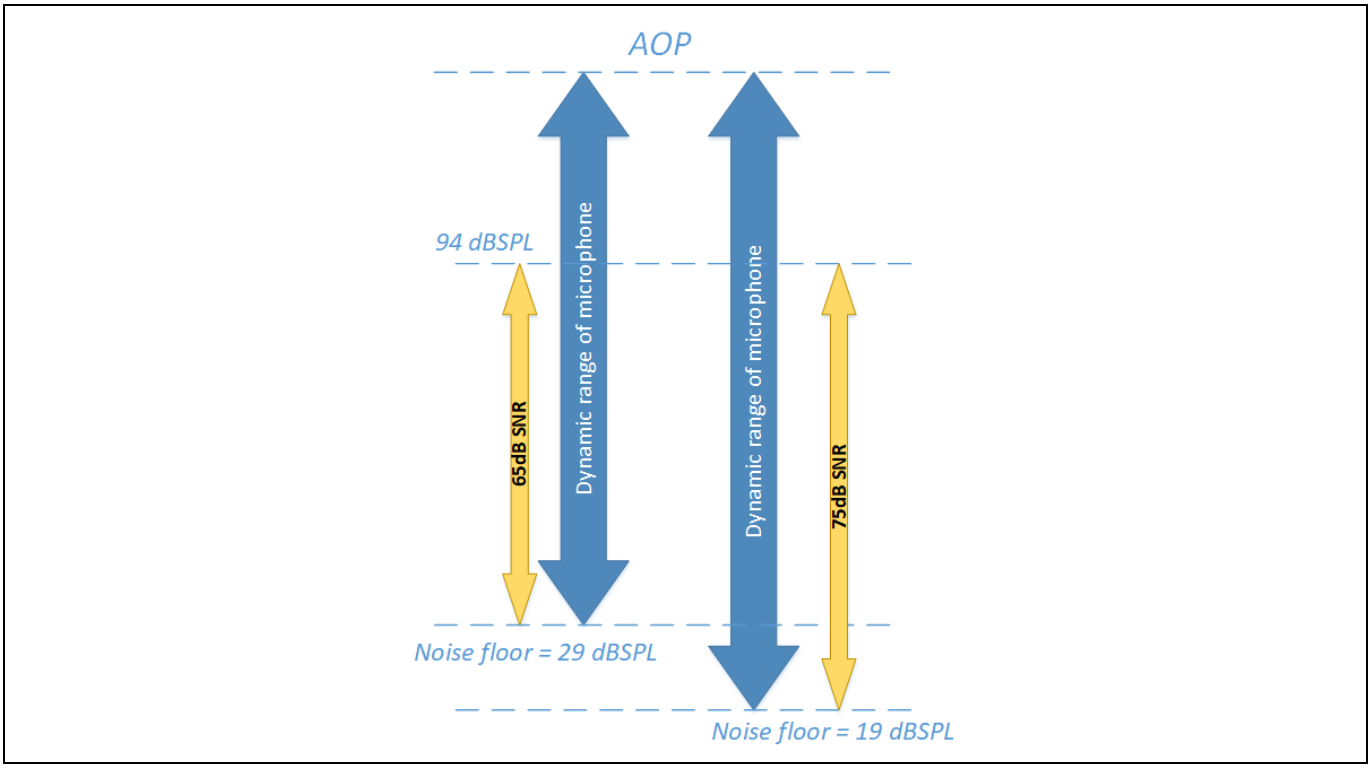

Figure 3 Relation between SNR and self-noise (noise floor) of the microphone

The upper limit of the dynamic range is defined by the Acoustic Overload Point (AOP). The lower limit is defined by the Signal-to-Noise Ratio (SNR). The SNR describes the self-noise of a microphone. A microphone can only pick up signals with a Sound Pressure Level (SPL) above its self-noise floor. Microphones with a high SNR can therefore work with lower audio sound pressure levels than microphones with lower SNR. Likewise, VUIs with an array of microphones provide higher-quality audio raw data as input for the application processor. As the raw data input contains more information and less self-induced noise, the subsequent processing in the cloud (compare with Figure 2) becomes easier and more efficient. For example: if keywords like “Hey Siri” must be confirmed for system wake-up, better input audio data results in a higher hit-rate, lower false acceptance and therefore a reduced error rate for system wake-up.

2.1 Challenging use cases

The current generation of VUI devices are focused on providing optimal performance for use cases that involve normal speech (60 dB_{SPL}) at a distance of 1 to 3 m. In the laboratory, these ideal conditions can be easily realized. However, real-world conditions provide many use cases where the performance levels drop below the 60 dB_{SPL} limit. Reasons for this could include a larger distance between user and VUI, whispered voices or commands spoken with varying sound pressure levels. Under such challenging conditions, the microphones with lower SNR would find it difficult to capture the audio signals correctly. Therefore, in challenging or simply realistic scenarios, the use of microphones in VUI systems results in superior performance, as shown by the examples in Figure 4.

Benefits of high SNR

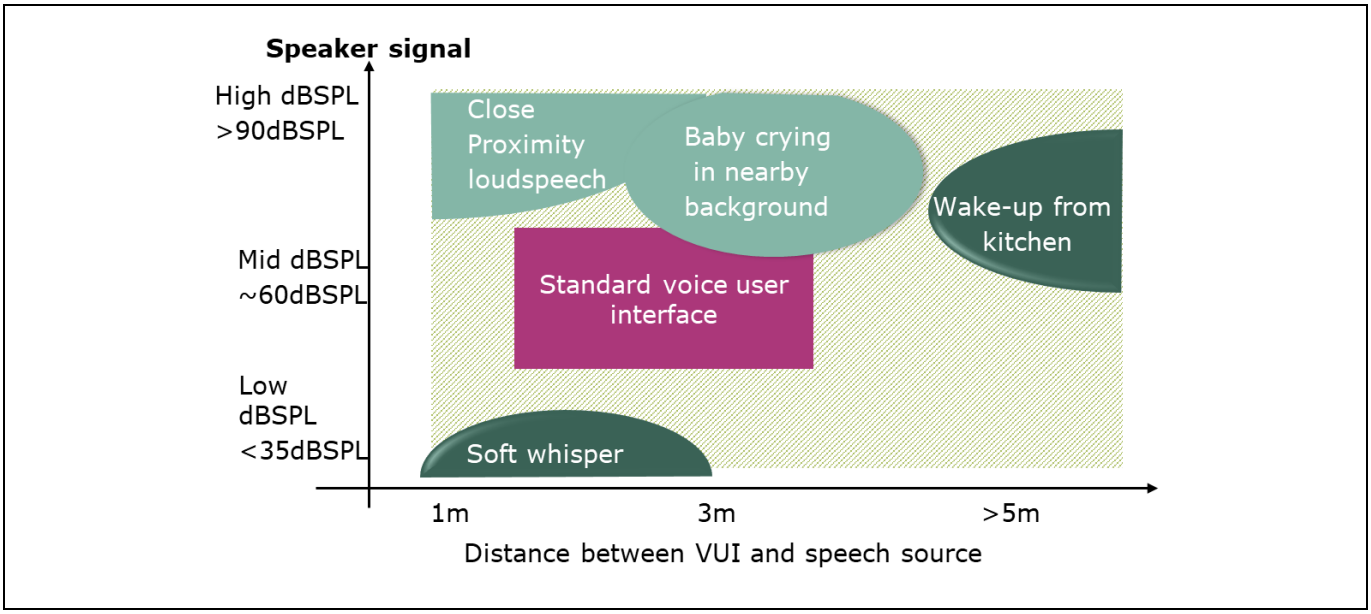

Figure 4 VUI use cases varying in speech signal levels and distance to the device

2.2 Whisper/soft-voice scenario

There are multiple scenarios where people would prefer to communicate with a VUI device in a soft voice. Examples include where there are people sleeping in the same room, parents not wanting to wake up their sleeping child, or simply to avoid disturbing someone else reading a book. Speech with a low SPL will lead to erroneous command recognition if the captured sound level is close to the self-noise level of the VUI microphone. Even the most sophisticated signal processing won't be able to succeed if the microphone's raw data contains too much noise. However, low self-noise microphones leave enough headroom to ensure that even low-SPL speech signals can be amplified, processed and recognized.

2.3 Cross-room scenario

The user and the VUI device may not be in the same room. The user may want to lower the sound of a VUI device (e.g. a smart TV) in the living room while working in the kitchen, or increase the volume of a newsfeed while getting ready in the bathroom. In both far-field scenarios (7 to 10 m distance) the user's voice will be attenuated by the distance and physical barriers such as walls, and therefore it can easily drop below the normal speech level of 60 dB_{SPL}. Similar to a soft or whispered voice, the self-noise floor of the microphones used will determine the overall VUI performance. The lower the microphones' self-noise, the bigger the distance between the user and the VUI device where speech recognition will remain possible.

3 Performance evaluation

Currently, there are no specific testing standards defined for VUI devices. In order to generate meaningful and repeatable evaluation results, the following test assumptions were used:

1. The test measures speech intelligibility and not speech quality (intelligibility signifies how well it can distinguish between similar-sounding words for better command understanding)
2. The test works with all standard speech bandwidths (4 kHz, 8 kHz or 20 kHz)
3. The test should not be susceptible to voice codecs and noise-suppression systems
4. The test should be applicable for realistic (noisy) environments
5. The test should be able to evaluate the performance of the VUI microphones as well as VUI systems such as speakers, headsets, handsets, etc.

After studying different testing standards like PESQ/POLQA and the Speech Transmission Index (STI), Audio Precision's "Articulation-Band Correlation Modified Rhyme Test" (ABC-MRT) based on traditional modified rhyme testing was chosen as the evaluation tool. ABC-MRT was the test that came closest to the above-mentioned requirements for a VUI device performance evaluation.

3.1 Test set-up

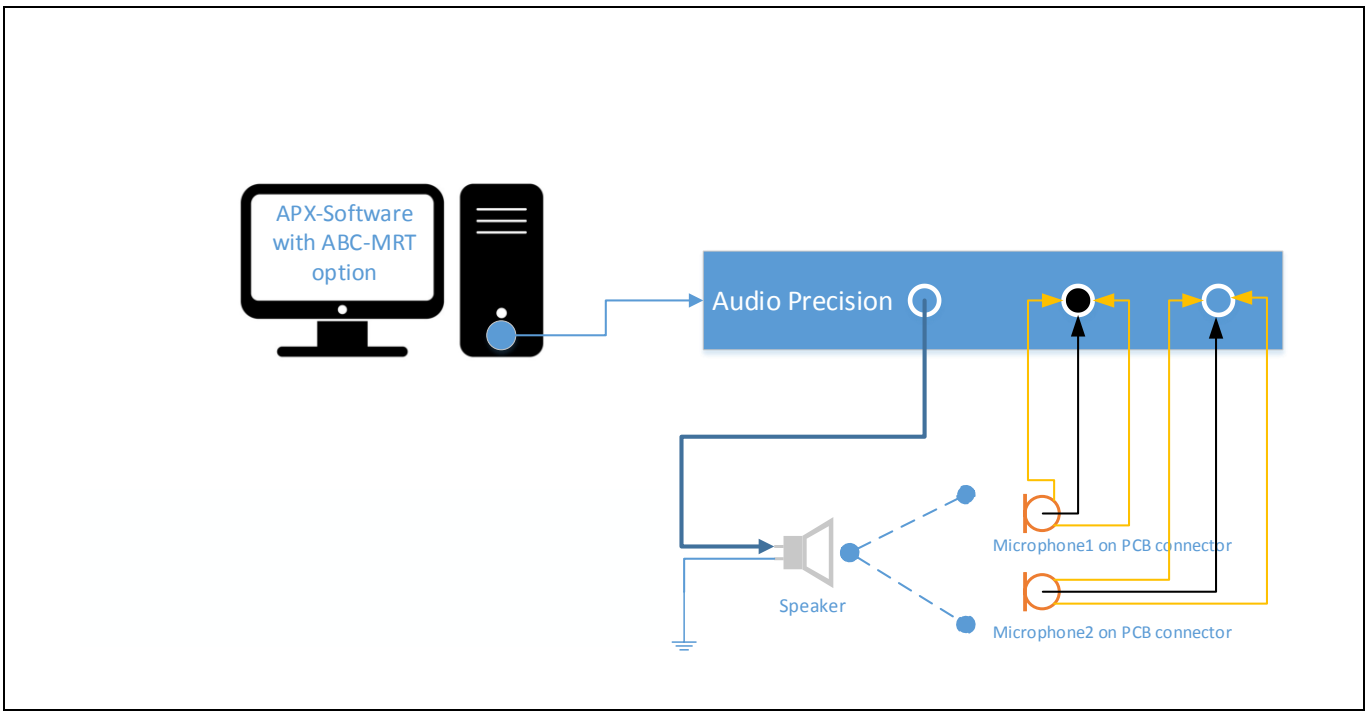

Figure 5 Overview of the Audio Precision interface

Figure 5 gives an overview of the component-level test set-up. The microphones (in yellow) are mounted on flexible PCBs. The PCBs are connected to Audio Precision with a PC interface to the Audio Precision-based ABC-MRT software extension. During the evaluation, an external speaker plays a set of keywords from the ABC-MRT database via Audio Precision hardware. The recorded audio from the microphone is fed to the Audio Precision hardware to perform a speech intelligibility evaluation. The ABC-MRT software provides an intelligibility score (ABC-MRT score) from 0 to 1, where 0 signifies that the audio stream fed from the VUI device/microphone interface has no match with the actual speech signal, while 1 represents 100 percent matching. For more information on how ABC-MRT works, see Technote 134: ABC-MRT in APx Audio Analyzers.[1]

3.2 Test environment

The test environment consists of the Device Under Test (DUT) placed in the center of the testing room. Three noise-signal speakers and one speech-signal speaker are located around the DUT at a distance of 1 m.

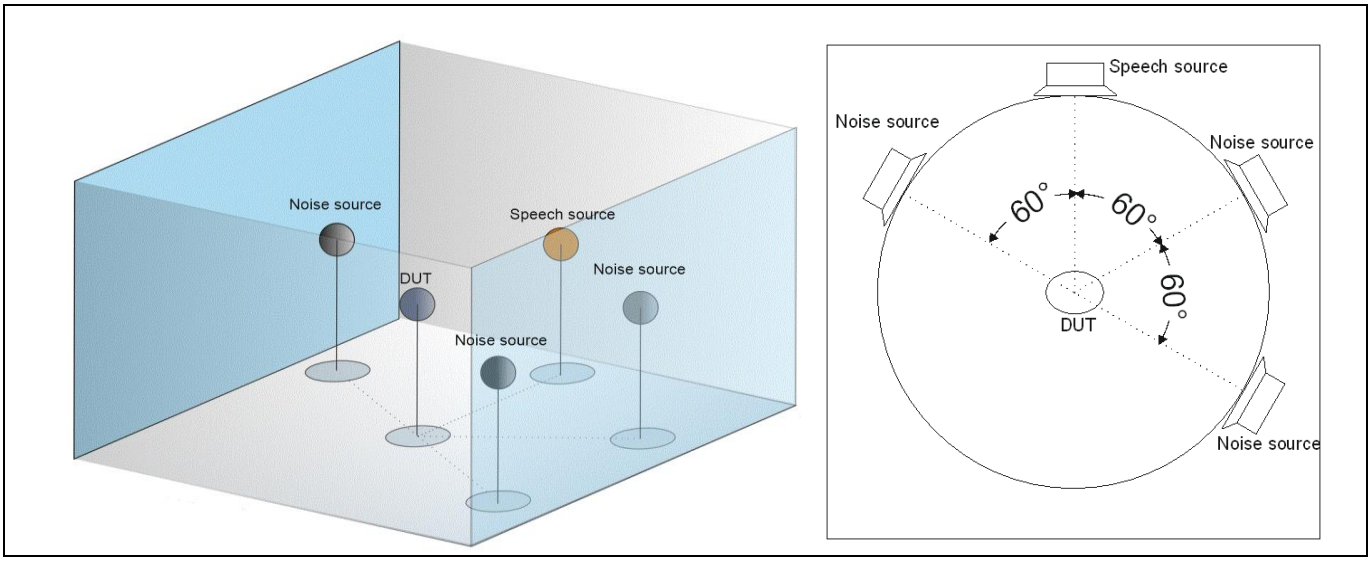

Figure 6 Test environment

The test environment is based on the European Telecommunication Standards Institute (ETSI) guidelines, ETSI EG 202 396-1,[4] with some customizations. The specifications of the test environment used for the evaluation of challenging use cases are listed below:

1. Room treatment: semi-anechoic room with dimensions approximately 4.4 m x 3 m
2. Reverb time: very low
3. Acoustical damping with an absorption factor around 95 percent
4. Noise floor of the room below 10 $dB_{SPL}(A)$
5. Speakers used for speech signal and noise sources
6. Height of the set-up from the ground is 1.4 m

4 Evaluation results

4.1 Whisper/soft-voice scenario

Evaluated at microphone level with an SPL of 25 to 30 dB averaged over the test signal, 75 dB SNR microphones showed up to 40 percent better speech intelligibility compared to 65 dB SNR microphones. The evaluations were made in the presence of babble noise (also known as office noise). The x-axis in Figure 3 refers to the difference between the background noise and speech signal (e.g. x-axis value of -5 refers to background noise being 5 dB lower than the speech signal).

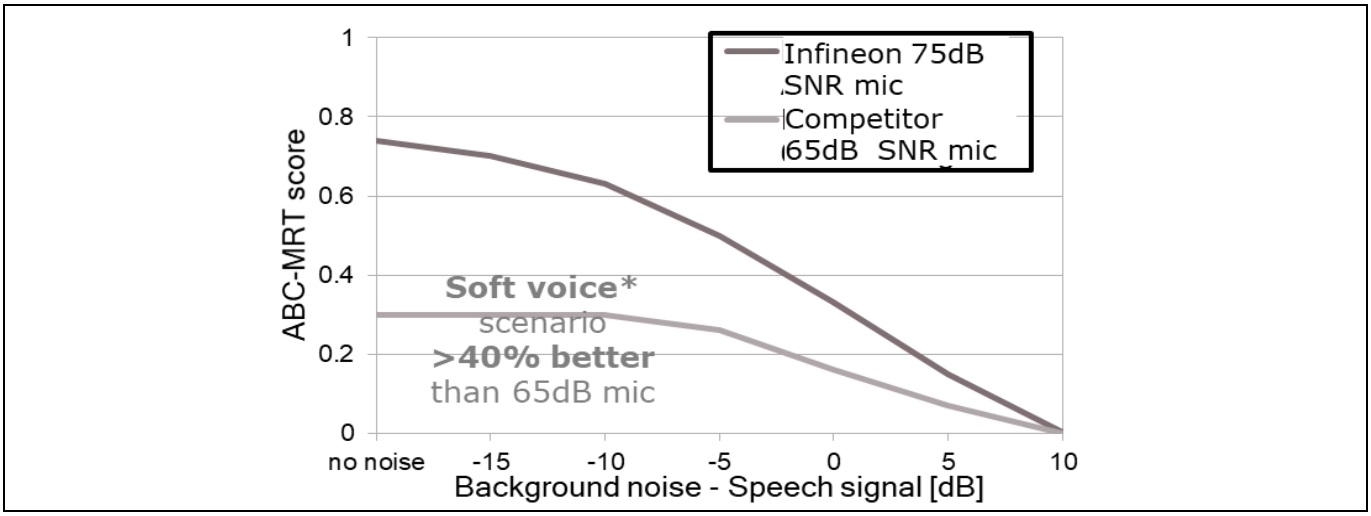

Figure 7 ABC-MRT score for whisper/soft-voice scenario

4.2 Cross-room scenario

Evaluated at microphone level with an SPL of 30 to 35 dB averaged over the test signal, 75 dB SNR microphones showed up to 25 percent better speech intelligibility compared to 65 dB SNR microphones. The evaluations were also made in the presence of babble noise. The axes in Figure 8 correspond to those in Figure 7.

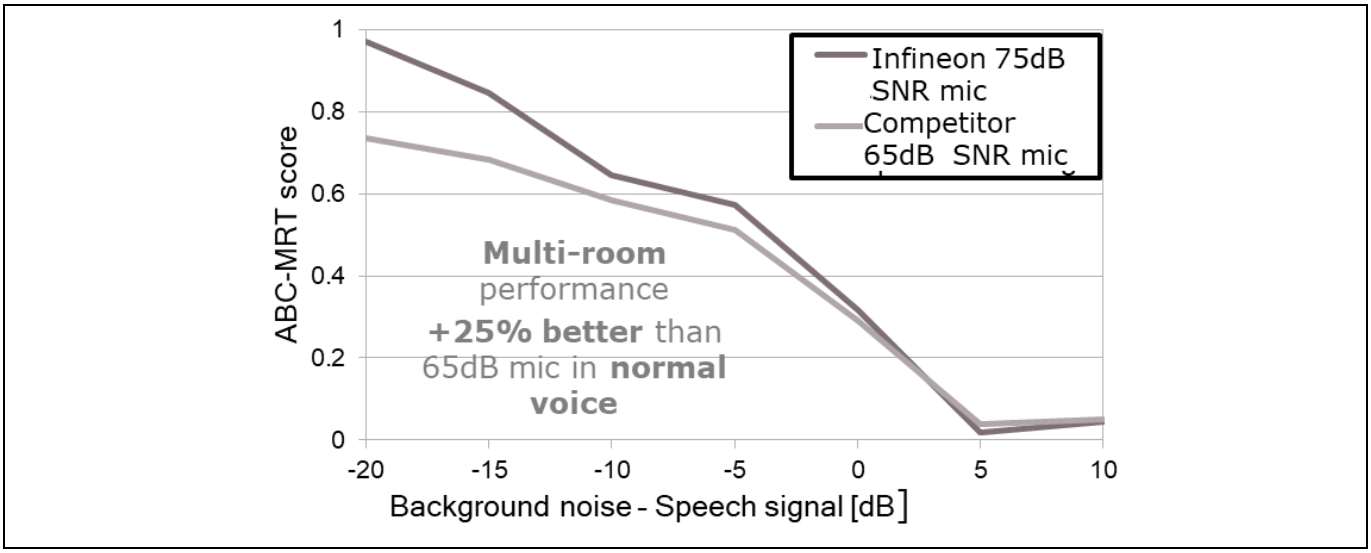

Figure 8 ABC-MRT score for cross-room scenario

5 System test set-up

A similar approach can be chosen for system-level tests. In the corresponding set-up (in comparison to Figure 9), the captured audio from a microphone array is processed by means of beam-forming, noise suppression and other methods before it is sent to Audio Precision.

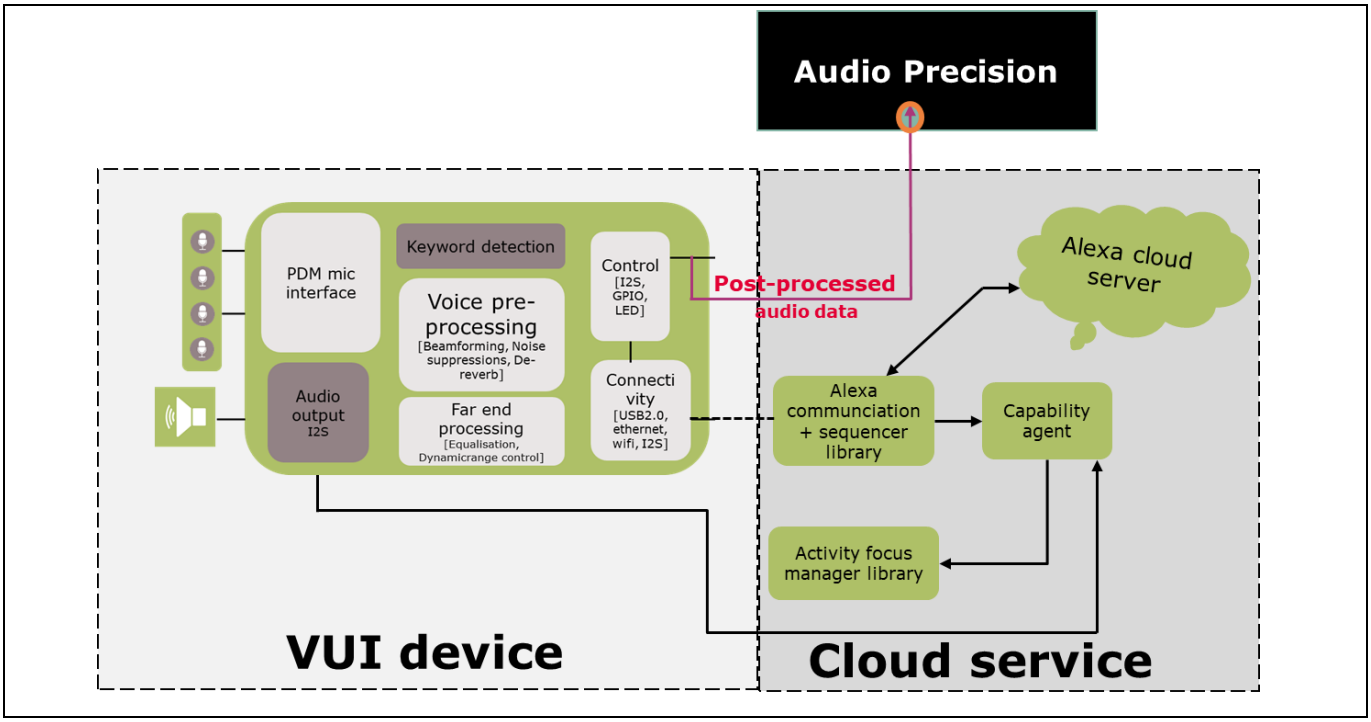

Figure 9 System-level evaluation set-up

6 System optimization and evaluation standards for microphones

6.1 Algorithms

Currently, algorithm providers focus on making algorithms that are hardware agnostic. Instead of being optimized for given hardware, current algorithms try to give a decent performance with every type of hardware. For example, if a certain VUI device implementation has to allocate an array of microphones, the algorithm should be tailored to the specifications of the microphone, so that the benefits of the superior hardware component/s can be fully utilized.

This will allow a VUI system to achieve better performance in challenging environments.

Acoustic noise suppression in VUI devices is used to capture the background noise and correspondingly generate an anti-noise version of the same, which is added to the captured audio to negate the noise and have a clearer speech audio. The working principle is explained well in Application Note AN538: Infineon Microphone in Noise-Canceling Headsets. VUI devices incorporate similar features to perform speech enhancement. High-SNR microphones enable the user to capture raw signals that are high in intelligibility, and therefore better post-processing of signals can enable better and cleaner (less noisy) speech for a higher hit-rate and lower false acceptance rate.

Other benefits of implementing microphones with low self-noise include:[4]

- High-resolution spatial information from microphone array: Content-rich signal information from the microphone array helps in gathering well-defined spatial information of the acoustic VUI environment for more accurate beam-steering and blind source separation.
- Increased Interference Reduction (IR): It evaluates the suppression of the interfering speech signal achieved by the filter. It is the average difference between the segmental power of the interfering clean speech signal and the segmental power of the filtered version of it.
- Higher Signal Distortion Index (SDI): The SDI measures the amount of distortion in the filtered version of the desired source signal with respect to the clean desired source signal at a reference microphone.

6.2 Testing standards

The testing methods used to evaluate VUI systems are not standardized and vary between device manufacturers. As a result of this, a lot of use cases reflecting real-life scenarios such as whispered voice and cross-room voice are overlooked and not evaluated. Although some standards are proposed by popular voice service providers, these are very loosely defined and do not cover the entire spectrum of use cases in the environment setting.

This application note has highlighted the following topics:

a. Challenging use cases like whispered voice and cross-room voice are evaluated for speech intelligibility
b. Benefits of microphones to capture audio signals with higher intelligibility are outlined
c. There is a brief overview of system/algorithm improvements necessary for microphone benefits
d. The need for a well-defined testing standard for better evaluation of VUI systems is covered

7 References

[1] Technote 134: ABC-MRT in APx Audio Analyzers

[2] Infineon Microphone in Noise-Canceling Headsets

[3] Speech Enhancement Using Microphone Arrays

[4] European Telecommunication Standards Institute (ETSI) Guidelines ETSI EG 202 396-1

Revision history

Document version	Date of release	Description of changes
V0.7	2019-02-13	First draft release
V0.8	2019-04-12	Second draft version
V1.0	2019-05-01	Final release

Trademarks
All referenced product or service names and trademarks are the property of their respective owners.

Edition 2019-04-30

Published by

Infineon Technologies AG

81726 Munich, Germany

Do you have a question about this document?

Email: erratum@infineon.com

Document reference
AppNote Number
AN 1904 PL38 1904 101645

IMPORTANT NOTICE

The information contained in this application note is given as a hint for the implementation of the product only and shall in no event be regarded as a description or warranty of a certain functionality, condition or quality of the product. Before implementation of the product, the recipient of this application note must verify any function and other technical information given herein in the real application. Infineon Technologies hereby disclaims any and all warranties and liabilities of any kind (including without limitation warranties of non-infringement of intellectual property rights of any third party) with respect to any and all information given in this application note.

The data contained in this document is exclusively intended for technically trained staff. It is the responsibility of customer's technical departments to evaluate the suitability of the product for the intended application and the completeness of the product information given in this document with respect to such application.

For further information on the product, technology, delivery terms and conditions and prices please contact your nearest Infineon Technologies office (**www.infineon.com**).

WARNINGS

Due to technical requirements products may contain dangerous substances. For information on the types in question please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by Infineon Technologies in a written document signed by authorized representatives of Infineon Technologies, Infineon Technologies' products may not be used in any applications where a failure of the product or any consequences of the use thereof can reasonably be expected to result in personal injury.