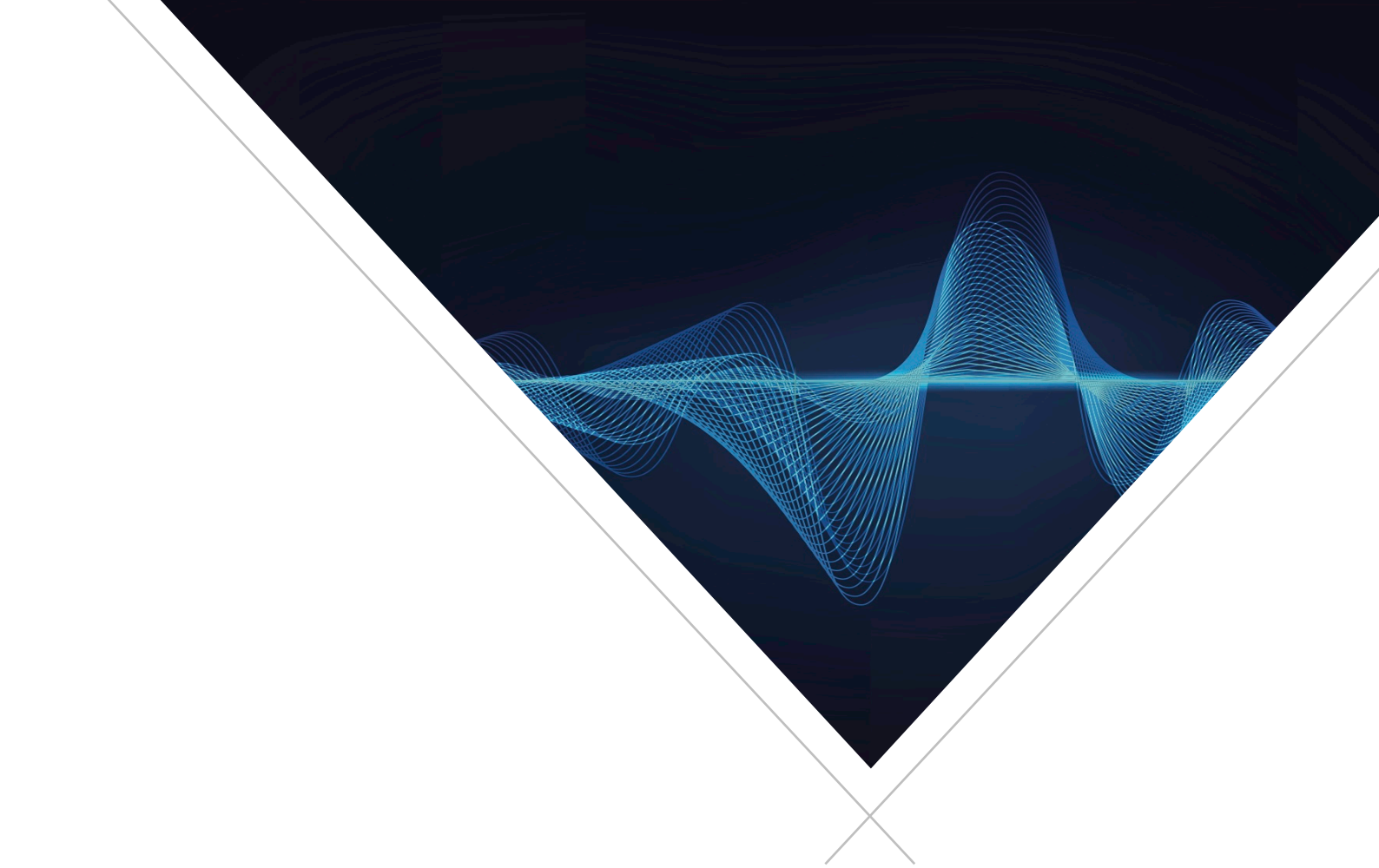

RECOMMENDATIONS AND CONSIDERATIONS FOR TESTING AND OPTIMISING FAR-FIELD VOICE APPLICATIONS

WHITEPAPER

JUNE 2018

XMOS

1. OVERVIEW

XMOS VocalFusion™ voice processors capture voice interactions and commands from across the room. These high-performance devices are designed for use in conference calling devices (human-to-human), or devices which connect with automatic speech recognition systems (human-to-machine) systems, either on the cloud or local application processor. This document applies to all far-field human to machine (ASR) use cases.

The VocalFusion portfolio includes a range of dev kits to rapidly enable prototypes, offering:

- linear mic arrays for “edge-of-room” devices with stereo-AEC, optimised for smart TVs, soundbars, set-top boxes and digital media adaptors
- linear mic arrays with mono-AEC for other smart home devices such as control panels and washing machines
- circular mic arrays for conference calling, smart speaker and “centre-of-room” implementations

After initial evaluation work using a VocalFusion dev kit, the silicon and software can be easily deployed in prototypes for integration and optimisation in acoustic enclosures, before moving to volume production and market launch.

This document aims to help engineers and developers in the planning of tests during development. VocalFusion can be deployed in a diverse range of applications and this document provides general guidelines, recommendations, and areas to consider when developing, testing and optimising the deployment of VocalFusion in devices with ASR connectivity.

2. TEST PLAN

When voice-enabling a product, testing needs to occur at multiple stages to ensure quality of the end-implementation. During each test stage, there are a large number of parameters which can be changed to optimise the user experience (see the Test parameter variables section) and these should be documented in a test plan.

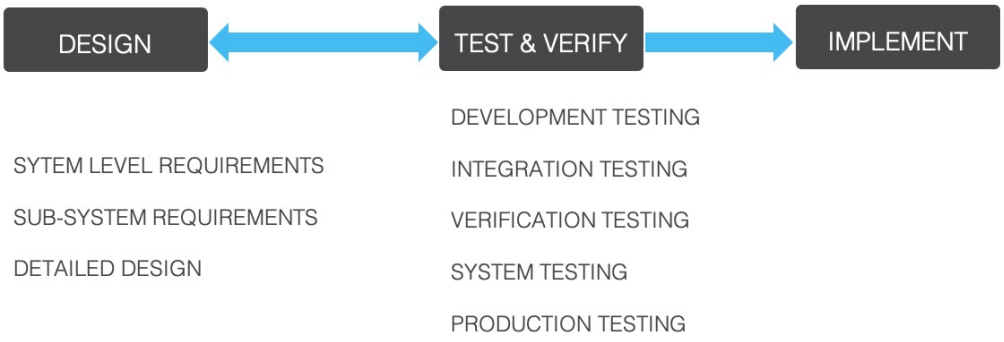

FIGURE 1: MODEL TO ILLUSTRATE THE TEST STAGES FOR A VOICE ENABLED PRODUCT

Typically, the initial development testing phase, will be undertaken using a development laptop and the XMOS VocalFusion dev kit; at this stage the DUT is typically a PCB with no external case.

During integration testing, the XMOS VocalFusion silicon with voice DSP algorithms is integrated into the external case (acoustic enclosure or housing), complete with the microphone(s) and loudspeaker(s) of the final product. The acoustic enclosure of the product will have a significant impact on performance, and it is strongly recommended that tuning of the VocalFusion voice algorithms within the acoustic enclosure is undertaken as early as possible. The tuning of a device will typically require multiple test runs; ideally all configuration parameters, results and audio tracks should be carefully recorded for future reference.

The operation of any other significant physical feature (for example motors) within the final product will also impact results, and consideration should be given to the point at which tests should be undertaken with those physical features running and integrated inside the final acoustic enclosure. Depending on the amount of change implemented at each stage consideration should be given to which new tests need to be undertaken, and which previous tests re-run.

3. TEST SCENARIOS

It is recommended that tests should:

- focus on those scenarios which best represent the environment in which the final product will be used
- reveal the performance limits of the DUT
- assist in optimising the performance of the complete system

Considerations when developing test scenarios:

- the size of the room
- the minimum and maximum distance between the talker and the device
- the expected output level from the microphones to the ASR
- how mobile and active the person talking is likely to be
- the ambient noise level
- the amount of echo dampening provided by soft furnishings (sofas, curtains, carpets) or reverberation (RT60) caused by reflective surfaces (windows, glass walls and mirrors, shiny hard-surfaced furniture, solid stone floors)
- diverse operational environments: for example, a voice-enabled TV in the large open hallway, of a large house with lots of hard marble and also the living room of a small apartment with a sofa, curtains and carpets

4. VOCALFUSION: OPTIMISED FOR DOMESTIC ENVIRONMENTS

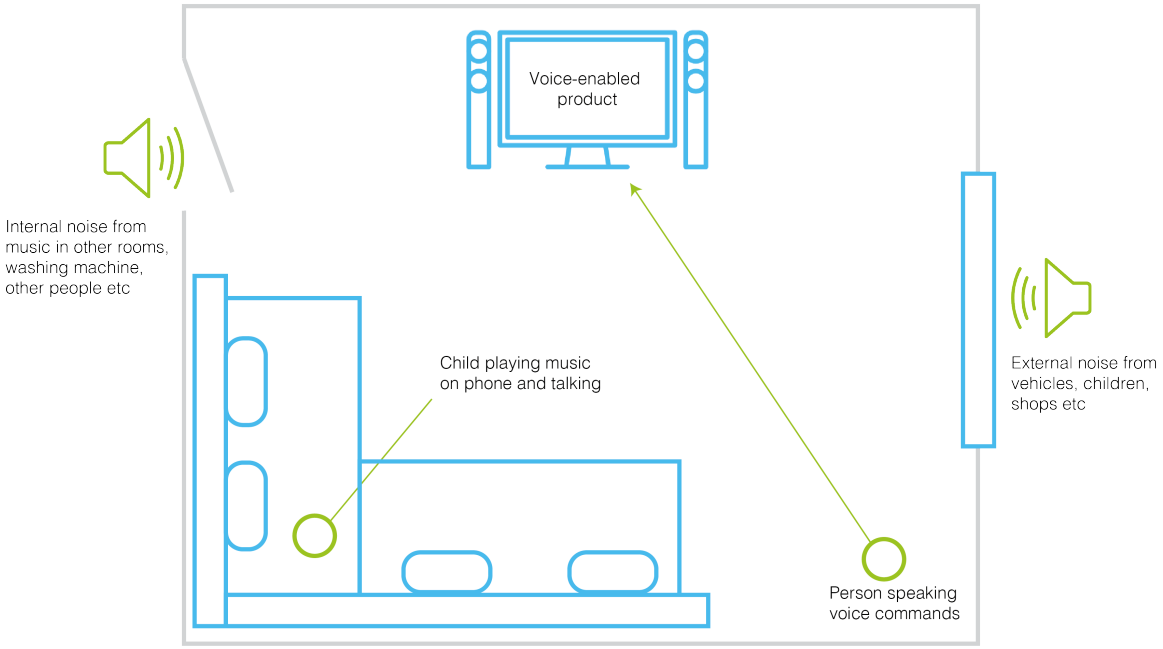

FIGURE 2: EXAMPLE OF A DOMESTIC LIVING ROOM WITH VOICE ENABLED STEREO TV, WITH EXAMPLES OF NOISE SOURCES

In this example, four omnidirectional digital microphones (in green above) are integrated into the TV, facing up towards the ceiling or out into the room. The TV stereo speakers are in the vertical edges of the device, and the XVF3500 device is used to enable far-field voice interactions.

VocalFusion devices provide an integrated voice DSP solution (set out in Figure 3). Your test plan and criteria needs to consider each feature highlighted below.

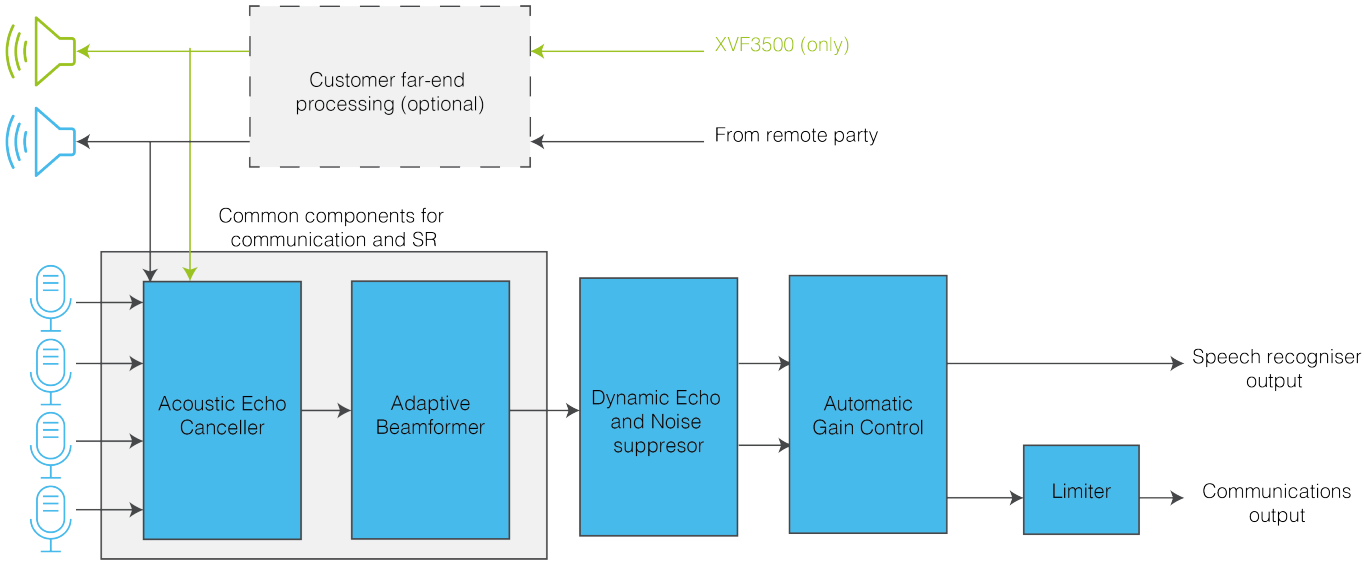

FIGURE 3: VOCALFUSION VOICE PROCESSING DSP PIPELINE, WITH ADDITIONAL PATH IN BLUE FOR XVF3500 STEREO-AEC PRODUCT. NUMBERS ARE REFERENCED IN THE PARAGRAPHS WHICH FOLLOW.

A. STEREO ACOUSTIC ECHO CANCELLATION (AEC)

The stereo sound from the TV is picked up by the microphones; our algorithms remove this playback audio from the captured signal. This enables barge-in: so, in a smart speaker the user can interrupt whatever is playing and issue a voice command and in conference-calling, the user can talk-over other parties, naturally.

B. ADAPTIVE BEAMFORMING

This identifies and isolates the voice command / voice of interest and applies the DSP algorithms to capture and intelligently track a clear audio signal as it moves around the room. For more information on beamforming, please see our VocalFusion DSP Databrief.

C. DYNAMIC DE-REVERBERATION

Our algorithm removes room echoes (eg voice bouncing off the window or TV screen).

In the home, furniture and soft-furnishings absorb typical reverberation (someone talking, the sound from the TV, and low-level background noise), so we expect an RT60 of between 400-600ms. In contrast, if the stereo-TV is to be placed in an open hallway of a large house with high ceilings and solid floors, we'd expect significant amounts of reverberation, and an RT60 of between 600-1000ms.

D. NOISE SUPPRESSION

Our algorithm suppresses background noise, like external street noise and other noise from around the house or room.

E. AUTOMATIC GAIN CONTROL

VocalFusion provides two outputs: voice data optimised for ASR or voice data optimised for conference call applications. Tuning of the device is undertaken on one output.

For ASR applications, AGC is switched off and Gain Control set to an appropriate value.

For conferencing applications, the voice signal is passed through AGC, so the quietest voice in the room is heard equally.

Together, all of these features deliver excellent far-field voice capture – (distances of >1m).

In some cases, the voice-enabled unit does not emit sound – for example, consider a voice enabled digital media adaptor plugged into the (non-voice enabled) stereo TV. In this case VocalFusion can provide a sixth function - configurable AEC latency, to adjust for and eliminate any delay in the audio reference signal.

5. THE INCREASING IMPORTANCE OF FAR-FIELD PERFORMANCE

Test scenarios should reflect the environment in which the product will be used, including room size, to ensure the range of performance is assessed adequately.

In many countries, new homes are getting larger, increasing the need for high-performance far-field voice capture. In other countries new builds are getting smaller, but previously sub-divided rooms are giving way “open plan” living spaces, particularly in the kitchen, dining and lounge areas. Again, this increases the need for excellent far-field voice capture over significant distances.

Open plan living spaces create challenging acoustic environments (see Figure 5); they can:

- be more reverberant, as a single hard floor may carry throughout the space
- be noisier, with appliances like kitchen extractor fans, dishwashers and TVs operating at the same time
- involve multiple people talking in different parts of the space
- have difficult corners, where multiple previous separate rooms have been joined together, creating complex acoustic reverberations and room echoes
- incorporate large doors opening onto open outdoor spaces, and therefore more likely to encounter higher levels of background noise

Larger rooms and open plan spaces increase customer expectations of the distance that far-field voice commands will be supported.

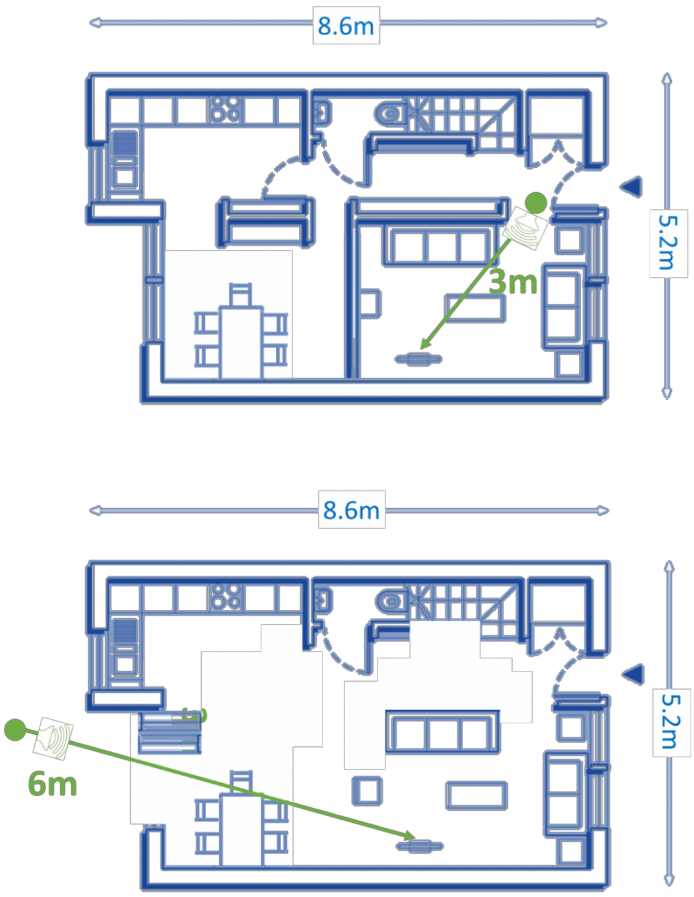

FIGURE 5: ILLUSTRATION FROM RIBA SPACE STANDARDS FOR HOMES, SHOWING HOW OPEN-PLAN LIVING WITHIN THE SAME FOOTPRINT LEADS TO SIGNIFICANTLY DIFFERENT ACOUSTIC CHALLENGES.

In the first layout, the separate living room avoids interference from kitchen noise. In the second layout, the open plan arrangement puts greater demands on far-field DSP, which now needs to handle additional noise from the kitchen, dining room and garden areas.

6. TESTING CONSIDERATIONS

The acoustic enclosure of the device under test (DUT) will have a significant impact on performance, and it is recommended that tests and tuning within the acoustic enclosure are undertaken as early as possible within the development and testing process.

6.1. TRANSLATING TEST SCENARIOS INTO TEST ENVIRONMENTS

Once you have considered the typical intended environment for the DUT, you can create suitable test scenarios.

For example, Figure 3 shows a domestic living room with a voice-enabled TV and various noise sources. We used this to develop the test scenario shown in Figure 6. In this ‘edge-of-room’ scenario, the design intent of the TV speakers is considered to be the hemisphere in front of the screen.

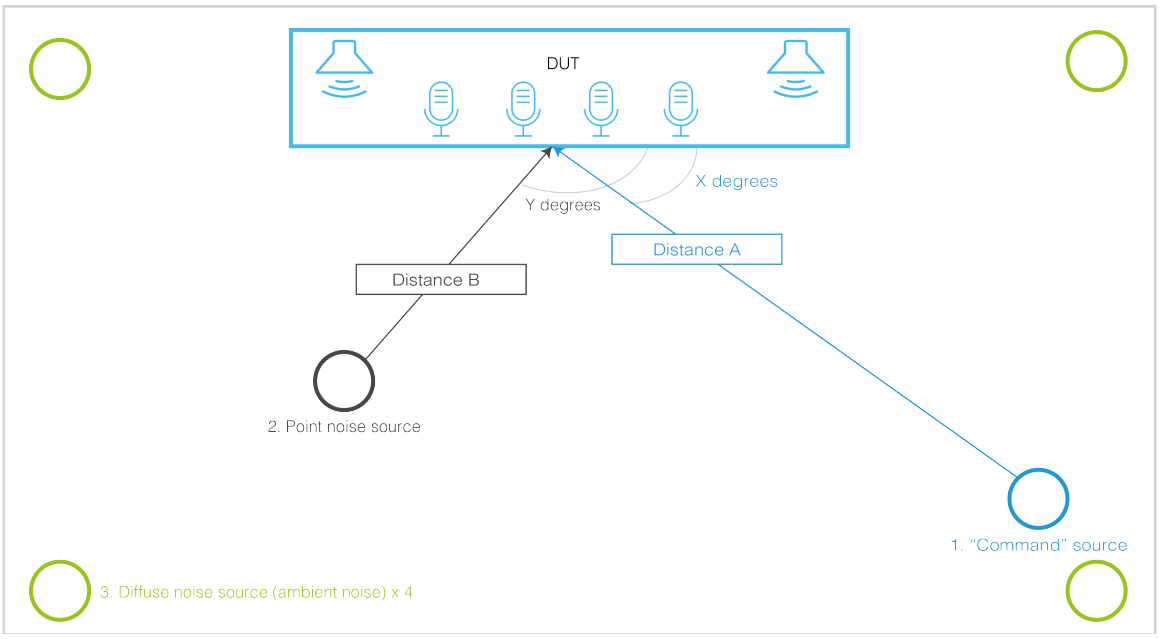

FIGURE 6: TEST SCENARIO BASED ON FIGURE 2 ILLUSTRATION OF DOMESTIC LIVING ROOM WITH TV

However, if the DUT is an omni-directional (cylindrical) smart speaker, then testing both "edge-of-room" and "center-of-room" scenarios should be considered. Greater attention should be paid to the scenario where the smart speaker is placed near the edge or corner of a room. This is a fairly typical given the wall-placement of domestic power sockets.

Placing an omni-directional speaker at the edge of the room creates complex echoes that the voice DSP algorithms then have to address.

6.2. PHYSICAL TEST ENVIRONMENT

Ideally, tests should be undertaken in multiple controlled environments, to ensure repeatability:

- A room with "normal" levels of reverberation (RT60), comparable to the final environment the device will typically be used in
- A second room with "less than typical" reverberation enables AEC and noise suppression to be more accurately tested

Many test labs enable different RT60 environments by the removal or addition of panels with different acoustic properties. If tests are undertaken across multiple labs care should be taken to ensure that such panels are consistent and comparable.

The ETSI recommendations (ETSI EG 202 396-1 V1.2.2) are frequently used as a basis for defining ideal test environments. It contains recommendations for room size, RT60 and ambient noise level and can be used to assess the suitably of test environments.

6.3. TEST PARAMETER VARIABLES

Test plans should consider adjustment of multiple parameters, and these will vary depending on the DUT, typical uses cases, and design intent. It is recommended that the following elements and parameters are considered in the design of the test plan:

6.3.1. NOISE SOURCES

A. Voice command ("command") speaker – wake word level (SPL)

- Tests should be undertaken both using constant wake word level, and increasing wake word level (ramped)
- For some ASRs it may be relevant to test the performance of 2nd and further voice commands
- The voice commands should be recorded and played back, to increase repeatability
- Tests using multiple voices, both male and female, are advised
- For some scenarios it may be relevant to test,
    - With a moving (mobile) “command” speaker
    - With multiple “command” speakers, to simulate multiple people in the room talking and issuing voice commands

B. For each of the 1. point noise, 2. DUT (ie in Figure 3, this would be the TV (DUT) output volume) and 3. ambient (diffuse background) noise sources,

- Tests should be undertaken with constant level, and increasing level (ramped)
- The noise used for each source should be recorded and played back, to reduce variability
- Depending on the use case, spoken words, music, white and pink noise should be considered. (White and pink noise: both broad-band, with energy in all frequencies are more likely to represent a real-life noise. Pink noise could be used to represent air conditioning for example, and compared with a recording of air conditioning has the advantage of being consistently repeatable)

6.3.2. AMBIENT (BACKGROUND) NOISE

This can be simulated by using a diffuse noise source so that the noise is bounced around the room. Angles, distances and heights of the noise sources should be considered, for example:

A. Angles between point noise source, “command” speakers and DUT
B. Distances between point noise source, “command” speakers and DUT
C. Relative height of point noise source, “command” speakers and DUT

When measuring distances and angles it is useful to physically mark the reference measuring point on the speakers and DUT.

6.3.3. AEC / BARGE-IN PERFORMANCE

This is assessed by running multiple tests of varying combinations of volume of DUT output, wake word volume and point noise volume, and assessing the False Accept Rate (FAR), and False Reject Rate (FRR) of wake word command.

6.3.4. SPEAKERS

It is important to use the same speakers for the source of the point noise, “command” speaker, ambient noise, and those integrated into the DUT. The "command" speaker(s) should be of high quality to provide accurate reproduction of the recorded test phrases.

6.3.5. AUDIO LEVEL

This should be recorded by a calibrated SPL meter in dB at the DUT, or 1m away from the audio source. The measurement microphone should be in very close proximity and in the same plane as the DUT microphone. The choice of SPL weighting should be consistent for all measurements and appropriate for the testing.

6.3.6. INTERNET CONNECTIVITY

A wired internet connection, and not WIFI, is recommended for connection to cloud-based ASRs. This aids reliability and reproducibility as a poor WIFI connection can look like poor microphone performance and thereby corrupt the test results.

6.3.7. VERSION HISTORY

Care should be taken to note the version of any algorithms, keyword models or ASR interfaces being used. This is particularly important when tests are undertaken over a protracted period of time, and to ensure comparable results when benchmarking against other voice DSP solutions.

Additionally, parameters within some models, for example some 3rd party keyword models, can be adjusted to optimise for specific implementations.

The method to check the VocalFusion firmware version can be found in the VocalFusion Software Design Guides.

6.3.8. EQUIPMENT

The XMOS VocalFusion Tuning Guides provide recommendations for speakers, SPL meters and other test equipment,

- HEADPHONES: It is recommended that a good quality (closed) pair of headphones is used in combination with the audio card on the development PC. An example is the Beyerdynamic DT990.
- REFERENCE LOUDSPEAKERS: It is recommended that a good quality loudspeaker (with amplifier) is available for simulating a near-end speech source. An example is the Genelec 8020D. Please ensure that the loudspeaker does not have any additional processing activated, such as bass boost etc.
- SPL METER: For calibrating the sound levels it is convenient to have a Sound Pressure Level (SPL) meter available. An example is the RION NL-5 Sound Level Meter.

6.4. MIC-ARRAY GEOMETRIES

During the development testing phase (Figure 1), the mic array on the VocalFusion dev kit will probably be used. For optimum product performance, we recommend you use the same microphone geometries as our dev kits:

- XMOS VocalFusion linear dev kits feature 4 Infineon IM69d130 MEMS microphones in a 100mm linear microphone array, with equally spaced microphones.
- XMOS VocalFusion circular dev kits feature 4 Infineon IM69d130 MEMS microphones in a 75 x 43mm microphone array of 4 microphones configured in a rectangle.
- In linear mic arrays, the VocalFusion voice DSP enables mics to be asymmetrically distributed, but it is highly recommended that they should all be in a single straight line. However, if variations to the layout used in the XMOS mic arrays are implemented then the performance impact must be assessed by the customer for their specific implementation.

Our VocalFusion Tuning Guides and VocalFusion DSP Databriefs, are available to XMOS customers under NDA. These provide further details on how the physical geometry of the microphone array in the DUT must be defined and supplied to the VocalFusion algorithms.

6.5. PHYSICALLY OPTIMISING THE MICROPHONES AND SPEAKERS IN THE DUT

In the early stages of testing the mic array on the VocalFusion development kit will probably be used. However, in the final product the customer may choose to use different microphones. XMOS suggests the following general guidelines on microphone selection and placement.

Microphones should:

- be carefully selected to optimise far-field performance. This will be improved with high SNR microphones; our dev kits use high-SNR Infineon IM69d130 MEMS microphones.

  For further details and best practice please refer to the Infineon MEMS microphone app notes, for example AN557: MEMS microphone mechanical and acoustical implementation.

  VocalFusion algorithms can support other mics with digital PDM interfaces within a 4-mic linear or circular array, but the performance must be assessed by the customer for their specific implementation.

- always be carefully mounted so that they are acoustically sealed and physically isolated from the product case and PCB.

  If the device needs to be touched in normal use, eg if it has a keyboard, or USB sticks and cables need to be inserted and removed, then avoid mounting microphones on the base of the product as they are likely to pick up vibrations during use.

  If the device has a touch control or buttons, please ensure that the microphones are not part of the same component within the product.

- be as far away as possible from mechanical noise sources, for example fans, keyboards and speakers. Vibration from those mechanical sound sources must be minimised; this can be achieved via appropriate dampening and designing chambers within the industrial enclosure to isolate the noise source and or the microphones.

- not be near or in contact with the speakers.

- be as far away as possible from heat sources, for example thermal vents, and wireless antenna.

- be protected from dust, liquids and impact during device production, assembly and customer use.

  If touch controls are implemented, then thought should be given to the likely “accidental touchpoints” that fingers my touch when approaching those controls, to ensure that dirt from fingertips does not accidentally get rubbed into microphones.

  The product chassis and speakers should be acoustically insulated to avoid sound/mechanical vibration coupling. Speakers should be sealed from the microphones.

7. FURTHER DOCUMENTATION AND SUPPORT

A comprehensive range of documents provide further support for the implementation and optimisation of our VocalFusion technology. In this guide, reference is made the following documents (some are under licence):

- VocalFusion Fine Tuning Guide for stereo-AEC development kits
- VocalFusion DSP Databriefs for stereo-AEC development kits
- VocalFusion Software Design Guides

These documents, plus Tuning Scripts are also included in the VocalFusion software download. The examples given are for our XVF3500 VocalFusion dev kit. Similar documentation is available for our other VocalFusion dev kits at xmos.ai.

Our Field Application Engineers also work with customers to advise on best practice and offer guidance. For more information on the support available, please see xmos.ai/support.

ACRONYMS

AEC: Acoustic Echo Cancellation, which removes a known reference from a microphone signal. For example, in a soundbar system, removes the music that originated from the soundbar speaker that has bounced around the room and been heard by the microphones.

AGC: Automatic Gain Control, which works to set the output volume at a desired level and maintain it there regardless of the level of the microphones. As a person walks around the room the signal received by the microphones changes, so the AGC tries to keep it consistent, and in doing so keeps the apparent volume of the person speaking constant.

ASR: Automatic Speech Recognition, which comprises keyword (wake word) and natural language processing (NLP). Often has a local keyword detector, which has different properties to generic speech recognition, and is highly optimised for the keyword(s); typically have small dictionary of trigger words, which may include short phrases of commonly used commands, highly optimised to trigger in noisy and adverse conditions (eg music being played).

DSP: Digital Signal Processing. Mathematical manipulation of an information signal.

DSP: Digital Signal Processor. Microprocessor optimised for digital signal processing, for example, xCORE.

DUT: Device Under Test (the product being tested.)

FAR: False Accept Rate, which is often used in keyword detection assessment. In this case, it measures the observed conditional probability of a keyword being detected given an input which does not contain a keyword.

FRR: False Reject Rate, which is often used in keyword detection assessment. In this case, it measures the observed conditional probability of a keyword not being detected given an input which contains a keyword.

MEMS: Micro-Electro Mechanical Systems. A technology applied to microphones, which has enabled the miniaturisation and use of microphones in many more objects.

SNR: Signal to Noise Ratio, which is the ratio of signal power to the noise power, often expressed in decibels. Decibels (dB) are a measure of a relative and logarithmic scale. Negative values indicate that the signal you are measuring is quieter than the noise you are measuring. 0dB indicates equivalence. For example, if 2 speakers are at 0dB to each other, they are at the same volume; if one is turned up by 6dB, then that speaker is twice as loud as the other; if one speaker is turned up by 12dB it will be twice as loud.

RT60: The average time for the SPL of an impulse to decay by 60dB. This measured in seconds, but typically quoted in ms (milli-seconds).

SPL: Sound Pressure Level is the measure of the pressure of a sound wave relative to the air around it. It is measured in decibels (dB).

IEC 61672:2003 defines a set of frequency weightings which help relate absolute SPL measurements to real world scenarios. The two most commonly used weightings are dBA which has a response similar to the human ear, and dBC which has a flatter response.

Copyright © 2020, All Rights Reserved.

Xmos Ltd. is the owner or licensee of this design, code, or Information (collectively, the “Information”) and is providing it to you “AS IS” with no warranty of any kind, express or implied and shall have no liability in relation to its use. Xmos Ltd. makes no representation that the Information, or any particular implementation thereof, is or will be free from any claims of infringement and again, shall have no liability in relation to any such claims.