VocalFusion: recommendations and considerations for testing and optimising far-field voice applications

Contents

Overview

XMOS VocalFusion™ captures voice interactions and commands from across the room, for use in either conference calling devices (human to human), or devices which connect with automatic speech recognition (ASR – human to machine) systems (on the cloud or local application processor) to resolve and action the user's intent. VocalFusion is designed to be the most accurate, most integrated, and easiest to use voice interface.

The VocalFusion portfolio includes a range of development kits to quickly and easily voice enable prototypes, with

- circular mic arrays for conference calling, smart speaker and "centre of room" implementations
- 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.

This document applies to all far-field human to machine (ASR) use cases.

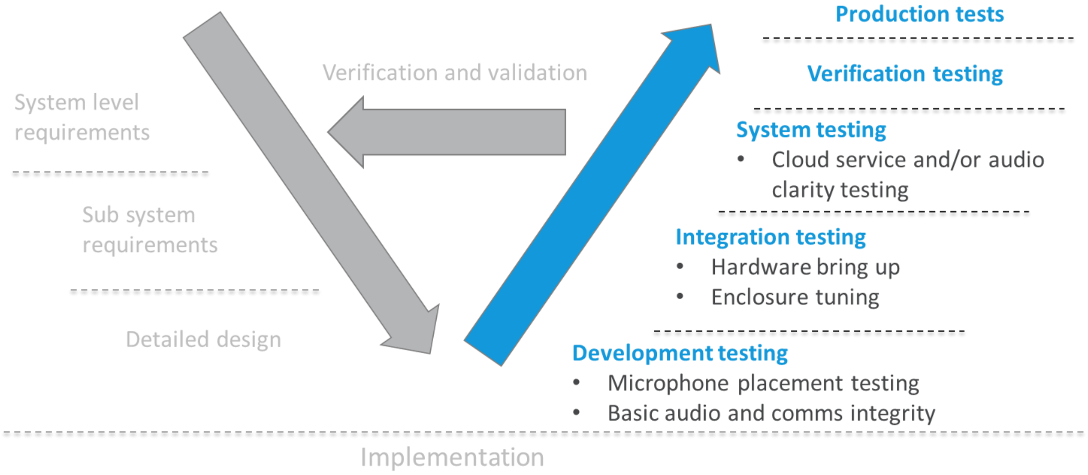
After using the XMOS VocalFusion development kits for initial evaluation activity, the VocalFusion silicon and software can be easily deployed in prototypes for integration and optimisation in acoustic enclosures, before moving to commercial volume in the final product.

This document is **intended for use by engineers and developers** to aid in the planning of tests during the development process. VocalFusion can be deployed in a diverse range of applications; this document provides **general guidelines, recommendations, and areas to consider when developing, testing and optimising the deployment of VocalFusion in devices with ASR connectivity**.

Test plan

When voice-enabling a product, testing needs to occur at multiple stages to ensure integrity and optimisation of the 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: V-Model for systems development, as an illustration of the test stages which need to be considered for a voice enabled product



Typically, the initial **development testing** phase, Figure 1**,** will be undertaken using a development laptop and the XMOS VocalFusion development 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.

Test scenarios

It is recommended that tests should,

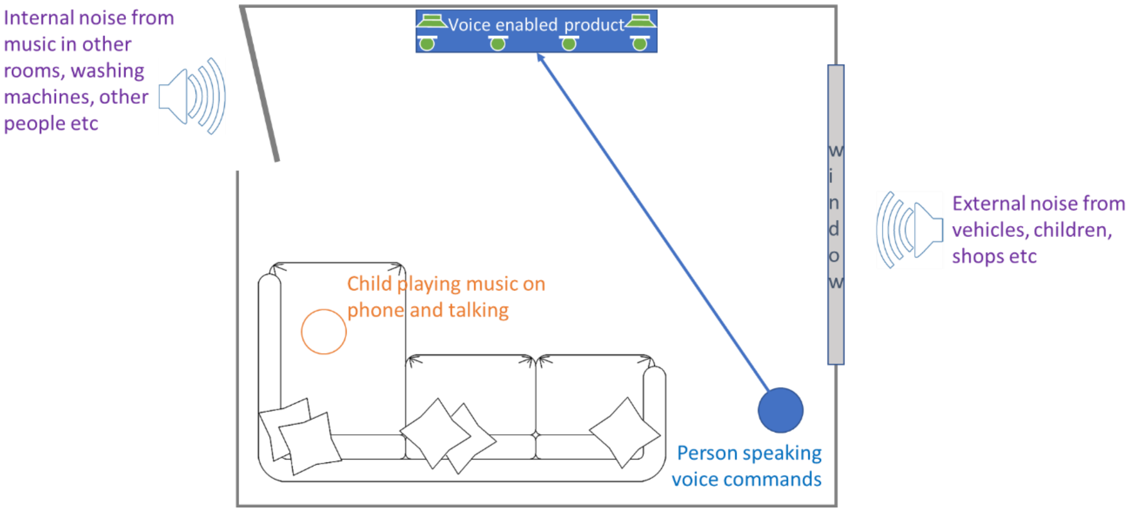
- focus on those scenarios which best represent the environment in which the final product will be used
- reveal the performance limits of the DUT
- assist in optimising the performance of the complete system.

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.

VocalFusion – optimised for domestic environments

Figure 2: example of a domestic living room with voice enabled stereo TV, with examples of noise sources

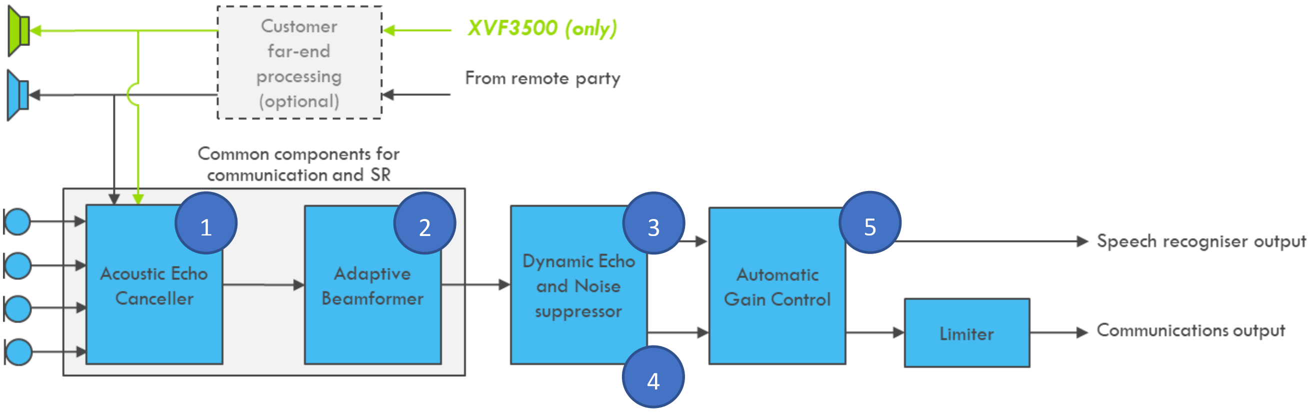

Internal noise from music in other rooms, washing machines, other people etc
Voice enabled product
w i n d o w
External noise from vehicles, children, shops etc
Child playing music on phone and talking
Person speaking voice commands

Let’s assume in this example that four omnidirectional digital microphones are integrated into the TV (in green above), facing up towards the ceiling or out into the room. The TV has stereo integrated speakers in the vertical edges of the device, so the **VocalFusion stereo-AEC device** with a linear mic array is used to enable far-field voice interactions.

Considering the example of a voice-enabled stereo-TV, Figure 2, in a domestic room when a person is speaking commands, and review how each of the VocalFusion features shown in Figure 3 works in practice.

VocalFusion devices provide an integrated voice DSP solution, which performs full duplex acoustic echo cancellation (AEC), beamforming, stationary and non-stationary noise suppression, and automatic gain control (AGC),Figure 3.

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.



The VocalFusion stereo-AEC device provides six key features, each of which should be considered when developing a test plan and test criteria,

1. **Stereo acoustic echo cancellation**: the stereo sound from the TV as heard by the microphones is removed from the signal that they capture (that is, stereo-AEC algorithms remove the playback signal).
    - In use these features support **full-duplex** operation, ensuring that the user can talk-over and **barge-in** over the sound from the TV.
        - In a smart speaker implementation, barge-in ensures that the user can talk over music. In a conference calling device, barge-in ensures that the user can talk over other parties to maintain a natural conversation flow.
2. The **adaptive beamformer** identifies and isolates any voice content in the listening space by focussing the microphones on to the person speaking and keeping their voice clear. If the person issuing voice commands stands up and walks around then the adaptive beamformer will use signals from the four microphones to track the talker as they move.
    - Refer to the **XMOS VocalFusion DSP Databrief**, available under NDA, for more information on the **Region of Interest** supported for linear mic arrays.
3. Dynamic **de-reverberation** removes **room echoes**, for example, any echo caused by the person's voice bouncing off the window or glass TV screen.
    - In the domestic setting of our example we can anticipate that the furniture and soft-furnishings will absorb some of the reverberation of the person speaking, the sound from the TV, and the background noise. This environment would typically have an RT60 of between 400-600ms.
    - In contrast, if the stereo-TV is to be typically used in large open hallway of a large house with lots of hard marble then there will be very significant amounts of reverberation. This environment would typically have an RT60 of between 600-1000ms.
4. **Noise suppression** addresses background noise. Looking at Figure 2, external noise from the street passing through the window, internal noise from other rooms in the home, and from other people or equipment in the same room.
5. VocalFusion provides the voice data as one of **two outputs**: one optimised for ASR, and a second optimised for humans to hear, for example for conference call applications. Tuning of the device is undertaken on one output.
For conferencing applications, the voice signal is passed through **Automatic Gain Control**, to ensure that even if the person talks quietly the system can hear the person speaking.

For ASR applications, AGC is typically switched off. the gain set to an appropriate **Gain Control** value.

- Together, all of these features deliver excellent voice capture performance over the **far-field**, that is distances over 1m and more typically far beyond 3m in a domestic environment.

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**.

6. Consumer electronics products can be voice enabled using devices such as digital media adaptors, and other plug-in after-market far-field voice accessories. In these solutions the latency between the audio output and the AEC reference signal arriving at the **VocalFusion stereo-AEC device** is unknown. During the development and tuning of the device, the configurable AEC latency features supported in the VocalFusion stereo-AEC solution enable the AEC reference signals to be accurately calibrated, and the latency adjusted to ensure that the sound leaving the TV is accurately cancelled.

The increasing importance of far-field performance

Test scenarios should be developed to reflect the environment in which the product will be used. In the case of our stereo-TV example, the device will typically be used in a domestic setting. The size of the typical room should be considered when reviewing the expected operating distance that the far-field voice device, so that appropriate tests can be developed to understand the full range of performance.

In many countries around the world, new detached homes are getting larger in size[i], Figure 4, increasing the need for excellent far-field performance.

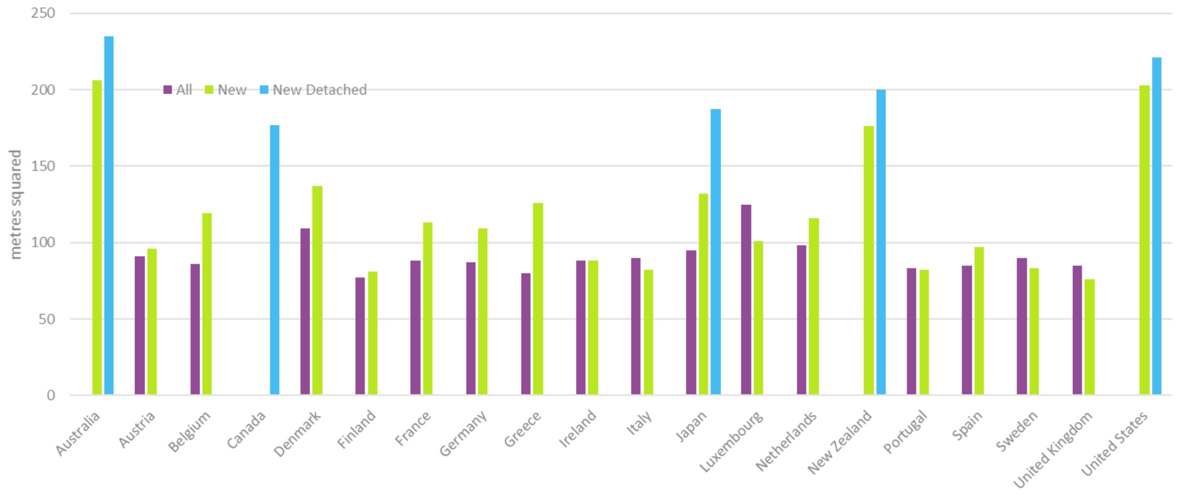


Figure 4: International house sizes

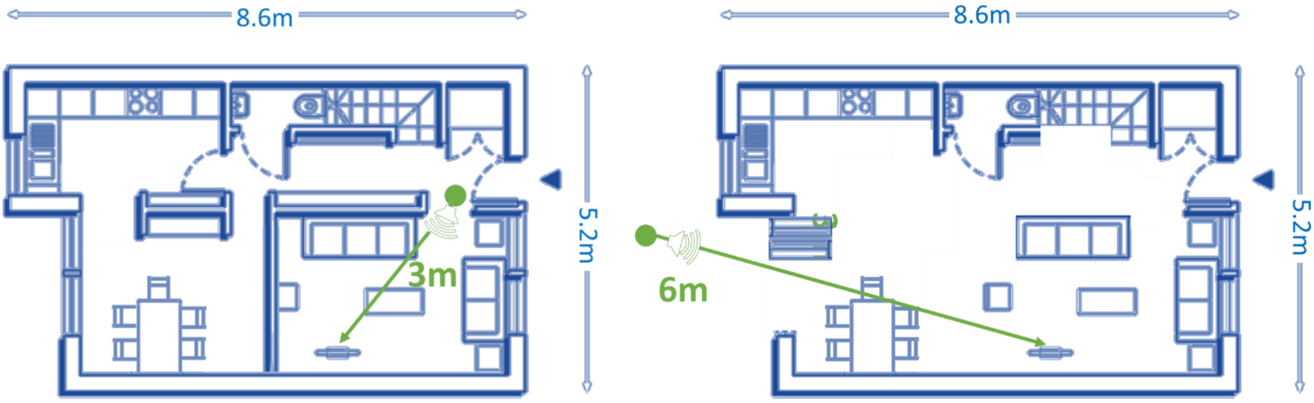
In other countries new builds are getting smaller[ii], however, the internal space is often no longer sub-divided by solid walls into separate rooms, and instead “open plan” living spaces combining kitchen and living space are very common in some countries, again increasing the need for excellent far-field performance over greater distances.

Open plan living spaces create quite challenging acoustic environments, Figure 5. They can often:

- be more reverberant, as a single hard floor surface may be used throughout the space
- include noise from kitchen extractor fans, dishwashers and TVs in use at the same time, within the same open space
- include 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

Additionally, larger rooms and open plan designs create a single vista; this “single room” increases customer expectations of the distance that far-field voice commands will be supported.

Figure 5: taking an illustration from RIBA Space Standards for Homes, to show how moving to open-plan living within the same footprint can lead to significantly different acoustic environments and far greater distances over which the far-field voice control may be expected to work.



Separate living room, isolating noise from kitchen, dining room and garden.

Open plan space: far-field DSP needs to handle background noise from kitchen, talkers at the dining table, and external garden noise all now in the “same space”.

Testing considerations

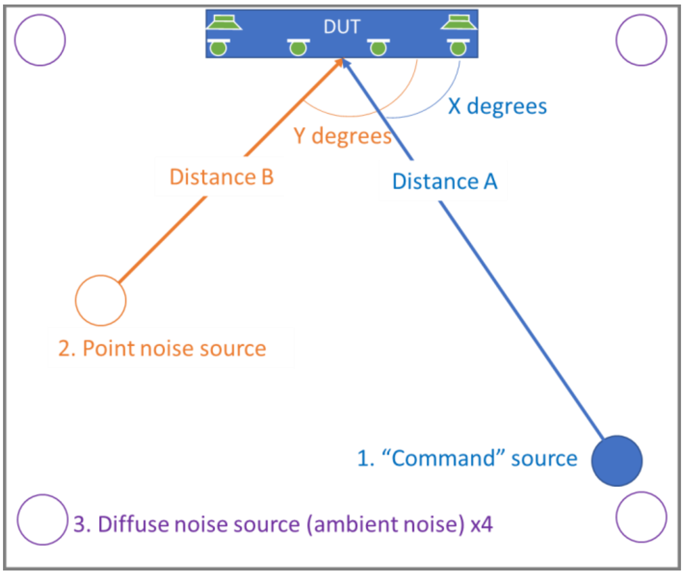
The acoustic enclosure of the DUT will have a significant impact on performance, and it is recommended that tests and tuning within the acoustic enclosure are undertaken as early as possible within the development and testing process.

Translating test scenarios into test environments

After consideration has been given to the typical environment where the device under test (DUT) is likely to be used, then suitable test scenarios can be created.

For example, taking the scenario illustrated in Figure 3 of a domestic living room with a voice enabled stereo TV and various noise sources, then the test scenario shown in Figure 6 could be developed. In this “edge of room” scenario, the design intent of the external speakers in the stereo TV should be considered to be the hemisphere in front of the screen.

Figure 6 Test scenario, based on illustration in Figure 2 of domestic living room with stereo TV



In contrast, if the DUT is an omni-directional cylindrical smart speaker, then testing at both "edge of room" and "centre of room" scenarios should be considered.

- In this case, greater attention should be paid to the scenario where the cylindrical speaker is placed near the edge or corner of a room (typically due to the location of domestic power sockets). Placing an omni-directional speaker near the edge of the room will create complex echoes for the voice DSP algorithms to address.

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.

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.

1. Noise sources
    a. Voice command speaker ("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 (i.e. in Figure 3 where the TV is the DUT, this would be the TV 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; 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.

2. Ambient (background) noise 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" speaker and DUT
    b. Distances between point noise source, "command" speaker and DUT
    c. Relative height of point noise source, "command" speaker and DUT

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

3. AEC / barge-in performance 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.

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

5. The audio level should be recorded by a calibrated SPL meter in dB at the DUT, or 1m away from the audio source. The measurement mic should be in very close proximity and in the same plane as the DUT mic. The choice of SPL weighting should be consistent for all measurements and appropriate for the testing (see Acronyms section for further information).

6. A wired internet connection, not WIFI, is recommended for connection to cloud-based ASRs, to aid reliability and reproducibility (poor WiFi connections can look like poor microphone performance – corrupting the results of tests).

7. Care should be taken to note the version of any algorithms, keyword models or ASR interfaces which are being used. This is particularly important when tests are undertaken over a protracted period of time, or 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 **XMOS VocalFusion Software Design Guides.**

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.
- 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.

Mic-array geometries

During the development testing phase (Figure 1**),** the mic array on the XMOS development kit will probably be used. For optimum product performance, XMOS recommends use of the same microphone geometries as the development 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.**

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

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
  - Far-field performance will be improved with high SNR microphones - XMOS VocalFusion development 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; we recommend that **the performance 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, for example, 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, then 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.

Further documentation and support

A comprehensive set of VocalFusion documents provide further detailed support for implementation optimisation and are available to XMOS VocalFusion customers under NDA, for example, documentation for the VocalFusion Stereo Dev Kits is available here. XMOS Field Application Engineers work with customers to advise them on best practice and guide them on specific issues to consider.

In particular reference is made in this document to,

- VocalFusion Fine Tuning Guide for stereo-AEC development kits
- VocalFusion DSP Databriefs for stereo-AEC development kits
- VocalFusion Software Design Guides
- And these three documents, amongst others, plus **Tuning Scripts** are also included in the **VocalFusion software download**.

The examples given are for our VocalFusion stereo-AEC dev kit. Similar documentation for our other 4 VocalFusion dev kits is also available via www.xmos.com.

Acronyms

AEC Acoustic Echo Cancellation: 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: works to set the output volume at a desired level, and maintain it there regardless of the level of the microphones. A 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. 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 (e.g. 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. Often used in keyword detection assessment, and in this case measures the observed conditional probability of a keyword being detected given an input which does not contain a keyword.

FRR False Reject Rate. Often used in keyword detection assessment, and in this case 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. 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.

About XMOS

XMOS is a leading supplier of voice and audio solutions to the consumer electronics market. Unique silicon architecture and highly differentiated software positions XMOS at the interface between voice processing, biometrics and artificial intelligence.

See more at www.xmos.com. Follow them on LinkedIn, Twitter and Facebook, and view their video library on YouTube.

[i] Source: demographia.com/db-intlhouse.htm

[ii] Source: UK RIBA Space Standards for Homes.