



AN02061: Voice Product Design Guidelines

Publication Date: 2026/9/2

Document Number: XM-014887-AN v1.0.0

IN THIS DOCUMENT

1	Introduction	1
2	General comments	1
3	Acoustic Guidelines	2
4	Signal interface	3
5	Other sources	7
6	Further reading	9

1 Introduction

This document provides general guidelines for the design of products implementing XMOS voice pipelines. It will cover loudspeaker and microphone choice as well as enclosure design and handling of digital signals. Following these guidelines should ensure the intended performance of the signal processing algorithms in the XMOS voice products.

Details on the XMOS voice products can be found on the XMOS website:

- ▶ [XCORE-VOICE](#)
- ▶ [XVF3610](#)
- ▶ [XVF3620](#)
- ▶ [XVF3800](#)
- ▶ [lib_voice](#)

2 General comments

Note that the requirements discussed here are only intended as general guidelines to help devise a more specific product specification. The actual specification will depend on the intended application. For example, even within a telecommunications application, a handset, personal speakerphone, or shared space speakerphone will each have very different technical specifications.

Additionally, many certification requirements have multiple available levels of certification. For example, Microsoft Teams Audio Requirements has a basic and a premium certification level, where the premium level features more stringent requirements. A similar situation with basic and premium criteria can be found in Amazon's certification requirements for voice services.

Therefore, we will not cover all the different potential requirements or applications. These must be decided on a case-by-case basis to match with the intended application of the product with consideration to the desired certification requirements.

Instead, this document will cover the technical areas that should be considered when designing the product specification. Where numerical values are given, they should not be considered recommended or required, but instead are provided as a "ball-park" figure for context. These figures have been chosen based on a generic product design for a typical smart speaker or personal speakerphone which aims to pass the basic level of certification requirements.



3 Acoustic Guidelines

The following sections provide guidelines for the physical acoustic design of a product.

3.1 Microphones

The microphones used in the design should have the following characteristics:

- ▶ Digital Micro-Electro-Mechanical System (MEMS) microphones which output Pulse Density Modulation (PDM) or I2S signals are recommended.
- ▶ The noise floor should be much lower than any expected signal. In general manufacturer specifications list the signal to noise ratio (SNR) as the difference in noise floor compared to the level of the signal when a 94 dBSPL acoustic signal is detected. In this case, an appropriate SNR would be at least 60 dB.
- ▶ The sensitivity should be chosen so the digital output is reasonable given the expected sound levels. For far field voice approximately -30 dBFS @ 94 dBSPL would be appropriate. For the SNR of 60 dB listed above this would mean the noise floor is lower than -90 dBFS.
- ▶ The microphone sensitivity should be set so that there is still some headroom when the device loudspeaker is playing at maximum volume. This will ensure that the microphones do not saturate and that the echo canceller can operate effectively.
- ▶ Distortion should be low for the expected levels present. For most applications Total Harmonic Distortion (THD) should be -40 dB or lower. Note that the levels from the playback channel of the device itself will potentially be louder than any speech signals, so the maximum level of the playback channel at the microphone positions should be used as the reference when considering distortion.
- ▶ The frequency response should cover the desired voice band. For a wideband application, a response from 100 Hz to 10 kHz should be easily obtainable.
- ▶ For most applications the directivity should be omnidirectional.

3.2 Loudspeaker

The loudspeaker, power amplifier and Digital-to-Analogue Converter (DAC) are considered together as the playback path of the product. This should have the following characteristics:

- ▶ The frequency response must cover the intended application:
 - ▶ For a product intended for voice applications this should cover the desired voice band, e.g., wideband should cover 250 Hz to 6.3 kHz.
 - ▶ For a product which can also play music etc., the loudspeaker can have a wider frequency response; this should not degrade the performance of the XMOS voice pipeline.
- ▶ The sensitivity should be sufficient for the application:
 - ▶ For a voice product, it should be capable of producing 75 dB SPL at the user's ear location when a -18 dBFS signal is supplied. This can be considered the nominal sensitivity of the product.
 - ▶ For a more general application, the maximum volume produced can be higher. However, much higher volumes will impact the ability to detect voice during playback.
- ▶ The distortion of the playback path should be -30 dB or lower, and preferably as low as possible. Like the microphone distortion, this will depend on level. Assuming the distortion is sufficiently low at nominal playback levels, it could be acceptable to have higher distortion at higher levels. This will degrade performance of the voice pipeline but that is up to the designer. Minimising non-linear distortion is important for the echo canceller to operate effectively.



- ▶ The directivity of the loudspeaker should not affect the voice pipeline performance significantly. However, it may be possible to use a directional loudspeaker to achieve required levels at the user position while minimising the feedback path to the microphones hence improving echo performance.
- ▶ If stereo playback is considered, the cross talk between channels should be minimal, however, this should not affect the performance of the voice pipeline.

The distortion of the echo path can be calculated by measuring the transfer function between the loudspeaker and the microphones. This will include the loudspeaker, the acoustic path and the microphone. The distortion should be -30 dB or lower for the expected levels at the microphones. This can be measured by playing a swept sine wave from the loudspeaker and measuring the output at the microphones using measurement software such as [Room EQ Wizard \(REW\)](#), [Audio Precision audio analysers](#), or similar. This should be lower than -30 dB THD (less than 3%).

3.3 Enclosure and mounting

The mechanical design of the product should account for the following:

- ▶ The mounting of the loudspeaker and microphones should allow sound to propagate to and from the user without significant impediments, such that the performance of the transducers is not degraded.
- ▶ In cases where multiple microphones are used, the microphone spacing must be considered. The spacing requirements will vary depending on which voice pipeline is being designed for, but broadly the microphones should not be too close together or too far apart. A spacing of 3-10 cm between pairs of microphones is typically suitable to avoid spatial aliasing.
- ▶ The feedback path from the loudspeaker to the microphones should be minimised. Several possible paths are present:
 - ▶ The acoustic path outside of the enclosure can be minimised by increasing the distance between loudspeakers and microphones or by choosing a more highly directional loudspeaker.
 - ▶ The acoustic path inside the enclosure can be minimised by sealing the rear chambers behind the loudspeakers and microphones.
 - ▶ The vibration path in the physical structure of the enclosure can be minimised by mechanical design of the housing structure, choice of materials and mounting of the transducers. This can include decoupling the transducers from the enclosure using compliant mounts or damping materials. The vibration path must remain time invariant.
- ▶ For a given nominal level, the closer a microphone is placed to a loudspeaker the lower its gain must be in order to avoid clipping. This will reduce the far-field sensitivity of the device.
- ▶ Vibration of loose parts in the product can cause rattles and buzzes which create non-linearity and will drastically reduce echo performance. Ensure all components, panels and connectors are fixed.
- ▶ Microphones should be prevented from detecting any additional noise sources from the product, e.g., cooling fans. Similar measures should be taken as to the loudspeaker feedback path.

4 Signal interface

The following points must be considered when dealing with the signals passed between the XMOS voice pipeline and the rest of the product.

- ▶ The playback signal which will be sent to the device loudspeakers should also be provided to the XMOS pipeline as a reference signal for the echo canceller algorithms.



There should be minimal delay between the reference signal and the loudspeaker signal.

- ▶ The delay between the reference signal and the loudspeaker signal should remain constant. Although the XMOS pipeline can adapt to small changes in delay, the performance will be much better if the delay is stable. See the [AEC Reference Signal](#) section for more details.
- ▶ No additional processing should be applied to the loudspeaker signal, which is not applied to the reference signal, i.e., all processing must be applied to the playback signal before it is copied out to the loudspeaker and reference signals. This point is particularly important for any non-linear processing, such as dynamic range compression, limiting or bass enhancement, but also for any processing with a long impulse response, e.g., reverberation.
- ▶ If a clock signal is provided over I2S, or similar, the timing of the clock should be stable and the synchronisation of this and any other timing signals must be consistent.
- ▶ Clock drift or jitter can cause the Acoustic Echo Canceller (AEC) filter to continuously adapt and will degrade performance.

4.1 AEC Reference Signal

AEC Delays and Alignment

The time window modelled by the AEC is finite due to the filter tail length. To maximise its performance it is important to ensure that the reference audio is presented to the AEC time aligned to the audio being reproduced by the loudspeakers.

Fig. 1 shows a well aligned AEC filter, where the whole impulse response is captured. Fig. 2 and Fig. 3 show examples of poorly aligned AEC filters where the reference audio arrives either too early or too late with respect to the microphone signal, resulting in a start of the filter filled with zeros or the loss of the tail of the impulse response.

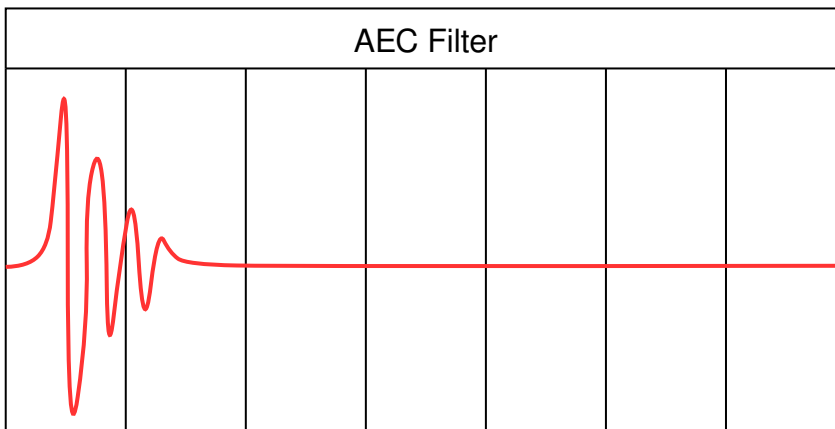


Fig. 1: In this AEC filter, the reference signal and microphone signal are well aligned, and the full impulse response is captured within the filter.

The reference audio path delay and the audio reproduction path delay may be significantly different, requiring additional delay to be inserted into one of the two paths, to correct this delay difference. This can be achieved by either using a fixed delay or Automatic Delay Estimation and Correction (ADEC) (XCORE-VOICE, XVF3610, and lib_voice only). All applications should aim to use a fixed delay where possible, as this will provide the best performance.



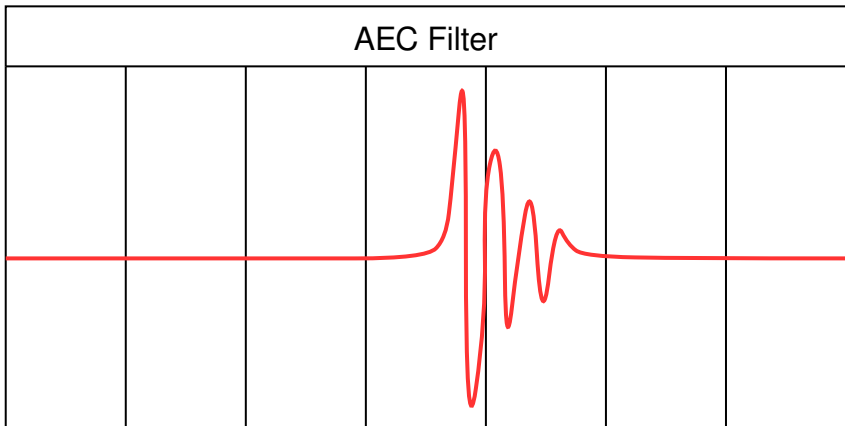


Fig. 2: In this AEC filter, the reference signal arrives before the microphone signal, so the start of the filter is filled with zeros and the tail of the impulse response is lost.

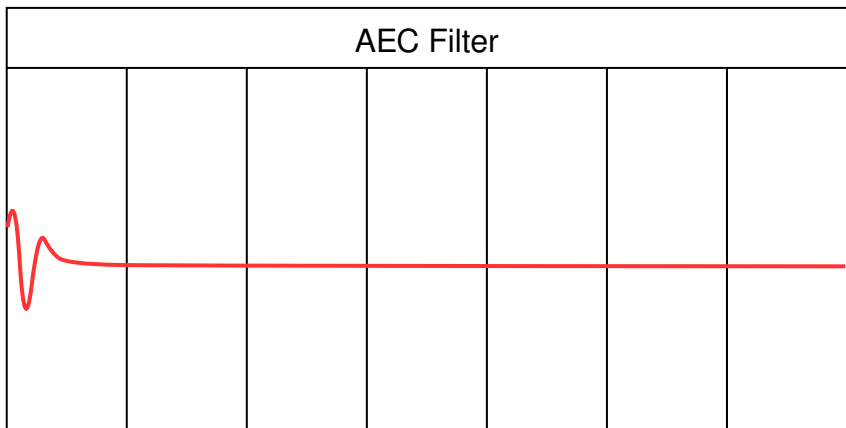


Fig. 3: In this AEC filter, the microphone signal arrives before the reference signal, and the start of the impulse response is lost.

Systems with fixed delay

In an ideal system design, the relative delay between the reference audio and microphones is well known and fixed. In these systems a fixed delay should be applied to either the reference or microphone signals to ensure they are time aligned at the input of the AEC. Examples of these systems are shown in [Fig. 4](#) and [Fig. 5](#). In both these systems, a fixed delay can be used safely.

Systems requiring ADEC

In some applications, it is not possible to know the relative delay between the reference audio and microphone signals, or the delay may vary over time. For example, in a Smart TV, the audio may be played out of the TV's internal loudspeakers or an external sound-bar, and the delay will be different in these two scenarios. An example of this system is shown in [Fig. 6](#). In these applications, ADEC can be used to ensure the reference and microphone signals remain time aligned at the input of the AEC, even when the delay



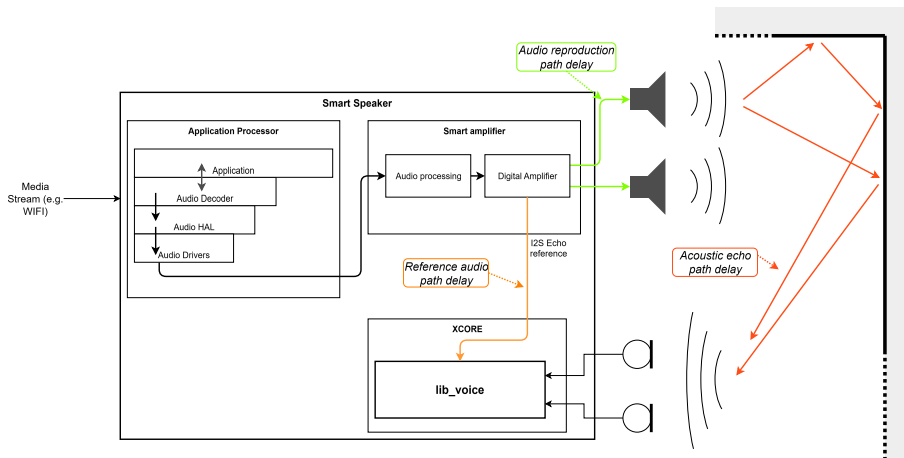


Fig. 4: In this system, the reference audio is sent to the XCORE device over I2S from the smart amplifier chipset. This will result in a well defined and fixed delay between the reference audio and the microphone signals, which can be compensated for by applying a fixed delay to either the reference or microphone signals as required, without the need for ADEC.

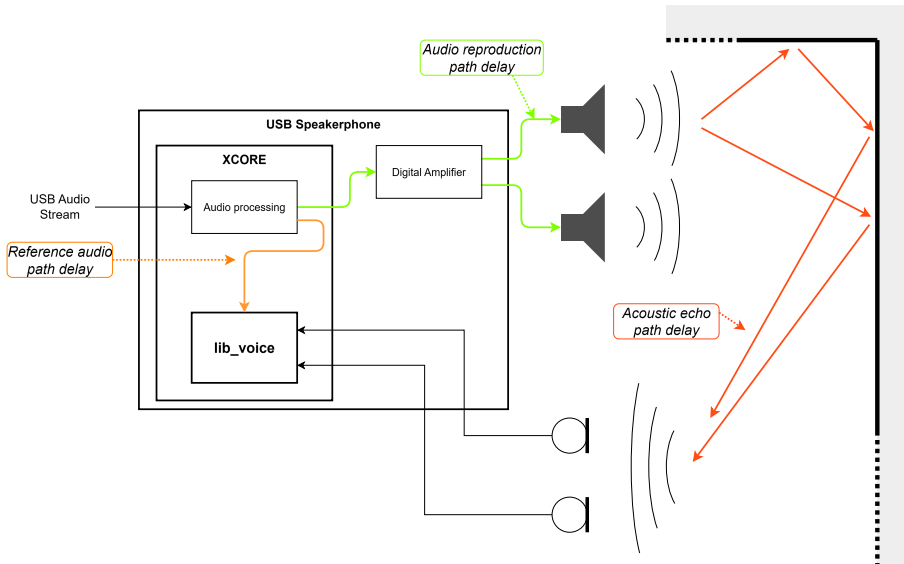


Fig. 5: In this system, the reference audio is sent to the XCORE device over USB from a host device, and the XCORE device sends the signal to the amplifier and loudspeakers. A duplicate of the signal is sent to the AEC. The timing over USB may vary, but the relative delay between the loudspeaker and microphone signals is fixed, so ADEC is not required.



changes. ADEC is only available in XCORE-VOICE, XVF3610, and lib_voice. Further information on ADEC can be found in the [lib_voice documentation](#).

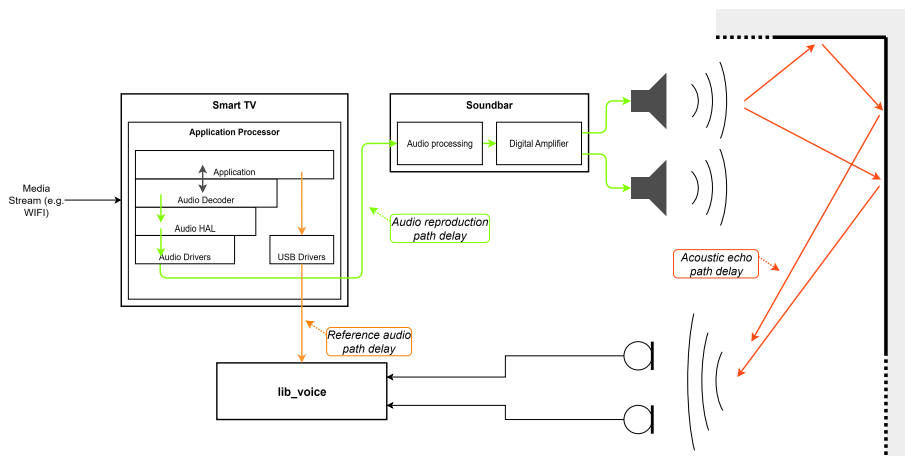


Fig. 6: In this system, the reference audio is sent to the XCORE device over USB from a host device. The audio reproduction path goes through the Smart TV's internal drivers, before being processed in the soundbar. Different models of soundbar would have different latencies, and the latency may also change based on the TV's audio settings, so the relative delay between the reference and microphone signals may be unknown and variable, requiring ADEC to maintain time alignment at the input of the AEC.

5 Other sources

The specifics of enclosure design, in particular relating to the acoustic performance of loudspeakers, are beyond the scope of this document. There is a wealth of literature covering this topic as well as consultancy services for design.

More detailed information on transducer mounting and handling can be found from the component manufacturers. They will also provide guidance on handling during manufacture and other important considerations.

Additionally, specification requirements for various product applications can be found in certification documents. These certifications will allow the product to bear a mark which demonstrates that the product is suitable for use with a particular service.

The following list gives some potentially useful documents. This is not a comprehensive list and does not constitute a recommendation of these suppliers. XMOS does not accept any liability based on these documents.



Supplier	Component	URL
Infineon	MEMS microphones	https://www.infineon.com/dgdl/Infineon-AN557_MEMS_microphone_mechanical_and_acoustical_implementation-AN-v01_01-EN.pdf?fileId=5546d4626102d35a01612d1e3bf96add
In-venSense	MEMS microphones	https://www.invensense.tdk.com/en-us/download-resource/1003-recommendations-mounting-and-connecting-invensense-mems-microphones
Knowles	MEMS microphones	https://www.mouser.com/pdfdocs/Knowles-Acoustic-Interface-Design-Guide.pdf
CUI Devices	Micro loudspeakers	https://www.cuidevices.com/blog/how-to-design-a-micro-speaker-enclosure
Fostex	Loudspeakers	https://www.fostex.jp/en/speaker-unit/
Microsoft	Teams Conferencing requirements	https://learn.microsoft.com/en-us/microsoftteams/devices/certification-specifications
Amazon	Voice services requirements	https://developer.amazon.com/alexa/console/avs/preview/resources/details/AVS%20Acoustic%20Self-test%20Checklist



6 Further reading

- ▶ *XMOS Voice Interface Selection Guide*
<https://www.xmos.com/voice-interface-selection-guide/>
- ▶ *XMOS Voice Interface Knowledge Base*
<https://www.xmos.com/documentation/XM-015450-TM/html/doc/rst/>



Copyright © 2026, 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.

XMOS, XCORE, VocalFusion and the XMOS logo are registered trademarks of XMOS Ltd. in the United Kingdom and other countries and may not be used without written permission. Company and product names mentioned in this document are the trademarks or registered trademarks of their respective owners.

