Cerence SSE Audio Requirements

For Hands-Free Telephony, Barge-In and Wake-up-Word Applications

Version 4.0

Table of contents

1 Purpose

This document specifies the requirements for the audio environment in hands-free telephony, barge-in, and wake-up-word applications. To achieve the best performance of the Cerence SSE speech enhancement solution — in terms of noise suppression, echo cancellation, and multi-microphone processing — these requirements must be fulfilled.

The bidirectional audio environment for hands-free telephony includes the complete path from the network access device (NAD) over the Cerence SSE processing block to the loudspeakers as well as from the microphone over the Cerence SSE block to the NAD.

For barge-in and wake-up-word, the NAD is replaced by the speech recognizer engine and the TTS/prompter.

The software interface of the Cerence SSE solution is described in [1].

1.1 History

Rev.	Section	Main Modifications	Date	Name
1.3	1	Adding history and referenced document tables at beginning on document and delete old *Related Documents* section	2016-07-20	MR
		Add section about microphones with integrated signal processing.		
		Rework section and add more information about the constant relative delay requirement between the microphone and reference channel.		
1.4		Introducing RCG module, updating architecture pictures, add information about volume control.	2016-12-12	MR
		Reorganize ADC + level requirements for microphone audio path.		
2.0		New structure of document	2018-01-19	GK
		Adding delay requirement		

		New definition of level in reference channel		
		Adding frequency response requirement		
		Adding tolerance schemes for SRC		
		Adding requirements for wake-up-word		
2.0.1		Minor fixes in text, no new requirements	2018-01-22	GK
3.0	3.2.2	Added requirement on tolerances for microphone arrays	2018-09-25	GK
	3.2.3	Clarification of requirement for frequency response for microphones		
	3.5.3	New requirement that Cerence must have access to weighting in RCG from loudspeaker signals; recommendation for initial RCG modified; generation of an initial setup for reference channels		
	3.5.4.4	Technical background information on use-case “Barge-In with interfering Media Sources”		
	3.5.5	New requirement on reference channel for combination of WuW and ICC (In-Car-Communication)		
	4.2	Added information on DTMF tones during a CarPlay phone call.		
	All	Minor fixes in text		
3.0.1	All	Minor fixes in text, add citations	2018-10-11	YK
3.0.2	All	Switch to Cerence design	2019-10-31	GK
4.0	3.2.2	Additional requirements on microphones for beamforming (tolerance for phase response)	2019-11-08	GK
	3.2.3	Recommendation for high pass filtering of microphone signals		
	3.3.2	Minimal distances between microphones for ULA reduced		

	3.5.6	New requirement on reference channel if system uses engine sound simulation or active noise cancellation		
	4.4	Recommendation for eCall (GOST 33468)		

1.2 Referenced Documents

Ref. ID	Document name	Version	File	Author
[1]	Cerence SSE API Documentation	4.x	SW_API_documentation.pdf	Cerence
[2]	ITU-T P.1100 Narrow-band hands-free communication in motor vehicles	01/2019	https://www.itu.int/rec/T-REC-P.1100/en	ITU
[3]	ITU-T P.1110 Wideband hands-free communication in motor vehicles	01/2019	https://www.itu.int/rec/T-REC-P.1110/en	ITU
[4]	ITU-T P.1130 Subsystem requirements for automotive speech services	06/2015	https://www.itu.int/rec/T-REC-P.1130/en	ITU
[5]	Accessory Interface Specification	R28	Accessory Interface Specification R28.pdf	Apple

1.3 Abbreviations

Term	Definition
A2B	Automotive Audio Bus (digital audio transfer bus)
ADC	Analog Digital Converter
AEC	Acoustic Echo Cancellation
ASRC	Asynchronous Sample Rate Converter
AuSS	Audio Sub-System (sound processing, amplifier)
BT	Bluetooth
Cerence SSE	Cerence's Speech Signal Enhancement Library
DAC	Digital Analog Converter
dBFSpk	Decibel relative fulls scale (measured as peak level of signal)
HATS	Head and Torso Simulator
HFP	Hands-free Profile (BT profile for telephony)
ICC	In-Car-Communication
IoT	Internet of Things
ITU	International Telecommunication Union
LSG	Loudspeaker Group
LTI	Linear and Time Invariant
MEMS	Micro-Electro-Mechanical Systems
NAD	Network Access Device
NLTV	Non-Linear and Time Variant
PIC	Passenger Interference Cancellation
RCG	Reference Channel Generation
SNR	Signal to Noise Ratio

SPL	Sound Pressure Level
SRC	Sample Rate Converter
SSRC	Synchronous Sample Rate Converter
THD	Total Harmonic Distortion
THD+N	Total Harmonic Distortion + Noise
TTS	Text To Speech
UCA	Uniform Circular Array
UCA_C	Uniform Circular Array with Center Microphone
ULA	Uniform Linear Array

2 Signal Flow Architecture

The following section describes a possible implementation of a hands-free application and a speech recognition application. The respective requirements, which emerge from the setup, must be fulfilled to guarantee a properly working system.

Please note that for a high quality hands-free telephony system, not only requirements regarding the input and output signals of Cerence SSE, but also further requirements for the complete system must be respected as well. These requirements (listed below) refer to the characteristics and integration of microphones, amplifiers, and loudspeakers:

- [2] describes the system requirements for a complete (narrow-band) hands-free system according to ITU-T P.1100
- [3] describes the system requirements for a complete (wide-band) hands-free system according to ITU-T P.1110
- [4] describes requirements for all components of a hands-free system according to ITU-T P.1130

From the integration aspect, Apple CarPlay Siri and Android Auto VR also belong to the telephony applications. In the CarPlay specification [5], Apple makes explicit references to specific versions of ITU-T P.1100 [2] and ITU-T P.1110 [3].

In the following section, the number of microphones and reference channels are limited to one for the sake of simplicity. For sections where the number of these components is relevant, it is mentioned explicitly.

2.1 Generic Audio Architecture

Figure 1 shows a typical setup for an application using Cerence SSE. Cerence SSE processes audio data in two directions:

- “Sending direction”: The signal path from the microphone to the NAD (Uplink) or Speech recognition engine.
- “Receiving direction”: The signal path from the NAD to the loudspeakers (Downlink, for telephony use-cases only).

A **microphone signal** is acquired and, after conversion into the digital domain and synchronous resampling to the operating sampling frequency of the Cerence SSE block, is passed to this block. The Cerence SSE block processes the audio signal and passes it either to

the network access device (NAD, e.g. a BT-chip) in **uplink** direction or to the VoCon Hybrid speech recognition engine.

For telephony, it is recommended to route the **downlink** signal from the NAD through Cerence SSE. Cerence SSE can only then be used to adjust the signal coming from the NAD to improve the performance of the AEC and the speech quality. The processed signal is then passed to the **audio sub-system** (AuSS, see section 3.7) after proper sample rate conversion. Audio signals from TTS or Media are routed directly to the audio sub-system.

After signal processing in the audio sub-system, which can be non-linear and time-variant, its output signal is routed back (optionally together with navigation prompts) to the Cerence SSE as reference signal for echo cancellation. In the **loudspeaker path**, the remote signal may be processed further, but exclusively in a linear and time-invariant manner (see section 3.7.2), before finally being played back over the loudspeaker(s).

The **reference signal** is needed by the AEC component to recognize the signal being played back over the loudspeakers and subsequently remove it from the microphone signal. After sample rate conversion the reference signal is passed to the Cerence SSE.

In general, it will be necessary to employ **sample rate converters** (SRC) in the system because some components might not operate with the same sample rate as the rest of the system. However, in the microphone, reference channel, and loudspeaker paths, only synchronous SRCs (SSRC) are allowed. These three paths are relevant for AEC (see section 2.2) and therefore they must be time-invariant.

NOTE:
*In **synchronous SRCs** (SSRC) input signal and output signal are in the same clock domain. **Asynchronous SRCs** (ASRC), where the input signal and output signal are in different clock domains are time-variant and therefore not allowed in the paths which are relevant for AEC.*

For practical reasons, the SRC in the AuSS Input Path for TTS should also be a SSRC: SSE and AuSS must run in the same clock domain and TTS has no internal real-time clock – this means that there is no need to use a complex ASRC instead of a more efficient SSRC. The SRCs between Cerence SSE and NAD are often realized as ASRCs because the NAD is running in a separate clock-domain (even if the nominal sample rate is the same for both). This is no problem because the audio paths between Cerence SSE and NAD are not relevant for AEC.

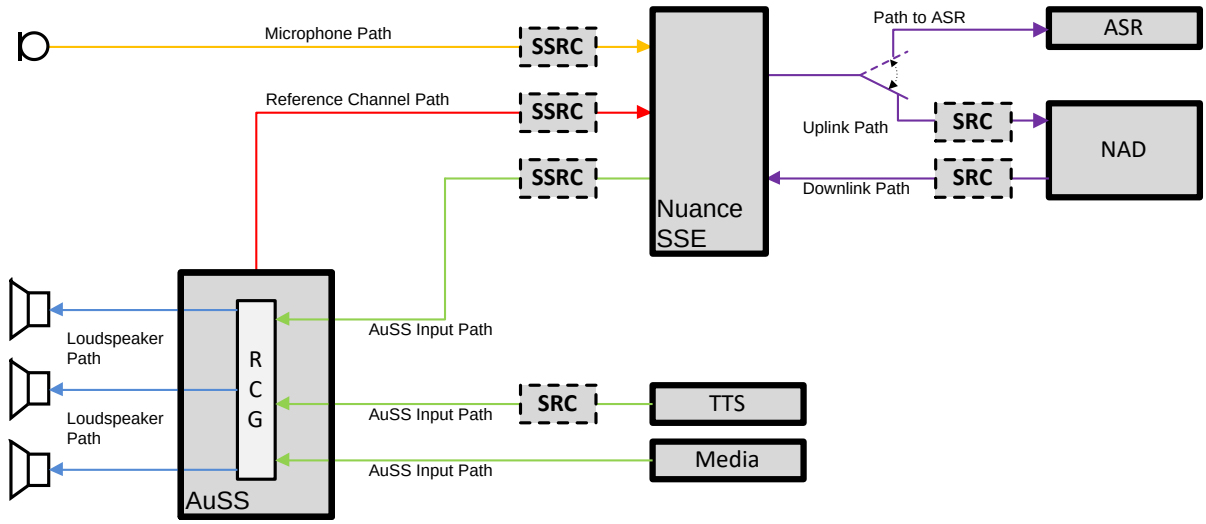

Figure 1: Audio flow for Cerence SSE applications; sample rate converters optional / as required

2.2 Audio Paths Relevant for AEC

Three segments of the audio paths shown in **Figure 1** are relevant for AEC:

- Microphone Path (marked in orange) from microphone to Cerence SSE
- Reference Channel Path (marked in red) from Reference Channel Generation (RCG) to Cerence SSE
- Loudspeaker Path (marked in blue) from RCG to the loudspeakers

The combination of Loudspeaker Path, acoustic path from loudspeaker(s) to microphone(s), and Microphone Path is also called “Echo Path”.

To have a reliably functioning echo cancelling system, it is necessary that all components in both the echo path and the reference channel path are linear and time invariant:

Linear means that the signal processing must not depend on the signal level and the signal processing must not generate additional frequency components (like harmonic distortion). This particularly excludes distortions due to saturation, clipping or quantization effects. All modifications which can be modeled by an FIR (finite impulse response) filter like volume setting or equalizing are allowed.

Time-invariant means that the echo path must not change over time. Especially critical is a time varying delay in the echo or the reference path. Also, time-variant processing in the microphone (e.g., noise suppression by a Wiener filter) is not allowed.

Examples for typical non-linear or time variant components are given in section 3.7.1. The most obvious component that does time variant processing is the volume control. A volume control can also be present in the system as a dynamic volume control (depending on driving speed or background noise level in the car). To be able to avoid echoes due to such time variant behavior the volume control must be applied before the reference channel generation (see also section 3.7.5).

A time-variant part of the echo path is the acoustic path between loudspeaker and microphone (people inside a room can move), but this cannot be avoided and will be compensated by the automatic adaptation control of the acoustic echo cancelling system.

3 Requirements

This chapter describes requirements for all modules and signal paths, which must be fulfilled to ensure a good performance for telephony and speech recognition with a single microphone or a multi-microphone configuration. If a system supports more than one use-case, the requirements must be fulfilled for each use-case. While the basic audio architecture might be the same for each use-case, there might be some differences—like different sample rates or the use of different microphones—that must be considered in the audio architecture.

3.1 Maximum Expected Sound Pressure Level

An important characteristic of a system that uses acoustic echo cancellation (AEC) is the maximum expected sound pressure level (SPLMAX) at the microphone. Audio signals generated by the system (e.g., entertainment signals, TTS prompts or telephone downlink), which must be cancelled by AEC, should not exceed SPLMAX.

NOTE:
For hands-free telephony the "Receiving Loudness Rating (RLR) at maximum volume" according ITU-T P.1100/P.1110 [2][3] *may be used to describe SPLMAX.*

SPLMAX must be defined at the beginning of a project in cooperation with Cerence. To achieve good AEC performance, SPLMAX should not be higher than needed for the corresponding use-cases. For audio levels above SPLMAX the AEC performance might be degraded.

3.2 Microphone Requirements

The Cerence SSE library supports single- and multi-microphone speech enhancement with beamforming and spatial postfiltering as well as acoustic echo cancellation (AEC) and other features. To achieve optimal performance in telephony or speech recognition applications some requirements must be met.

3.2.1 General Requirements

The microphone shall operate linearly (THD+N < 3%) up to the maximum expected sound pressure level (SPLMAX) at the microphone. This is especially important to avoid

degradation of the AEC performance. Clipping due to saturation of the analogue microphone signal or due to saturation of the analogue-to-digital conversion (ADC) at SPL up to SPLMAX is not allowed.

Both omni-directional and directional (e.g., cardioid) microphone types are supported. Regarding microphones with internal signal processing, please refer to section 3.2.4.

The equivalent noise level of the microphone self-noise must not exceed 30 dBSPL(A) (A-weighted, following DIN EN 61672-1:2014-07, Appx. E3). If the system supports telephony applications as well, the frequency response needs to be compliant with ITU-T P.1130(06/2015) [4], Class 1 or Class 2 (see section 3.2.3).

3.2.2 Multi-Microphone Configurations

If a multi-microphone configuration is used, all microphones must have the same characteristics (especially in a microphone array for beamforming). Larger deviations among the microphone characteristics may degrade performance:

- The frequency response of all microphones shall be within a tolerance scheme according ITU-T P.1130 (06/2015) [4], section 8.3.2.1.2, Class 2 (see section 3.2.3). Also, the sensitivity of the microphones at 1 kHz shall be within ±2dB relative to the nominal sensitivity (ITU-T P.1130 (06/2015) [4], Annex A.1, Class 2).
- All microphone signals must have the same polarity. It is not allowed to have an inverted microphone signal (e.g. if an eCall system is looped into one of the microphone signals).
- If the microphones are used as an array, the difference in the phase response shall be less than the following tolerance scheme:

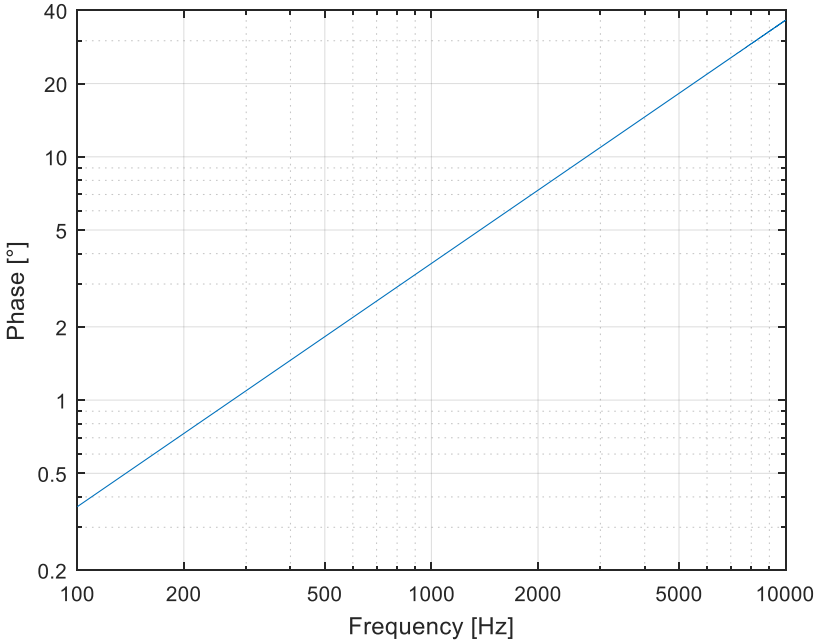

Figure 2: Upper limit for variation of phase response between microphones in an array (beam former)

3.2.3 Microphone Specification for Automotive

A typical maximum expected sound pressure level is 106 dBSPL (according to ITU-T P.1110 [3]). A sensitivity of 300 mV/Pa (at 8 V power supply) is recommended (according to ITU-T P.1110 [3]).

The microphone frequency response shall be compliant to Class 1 (**Figure 3**) or Class 2 (**Figure 4**) of ITU-T P.1130 (06/2015) [4], section 8.3.2.1.2 (test point S1a):

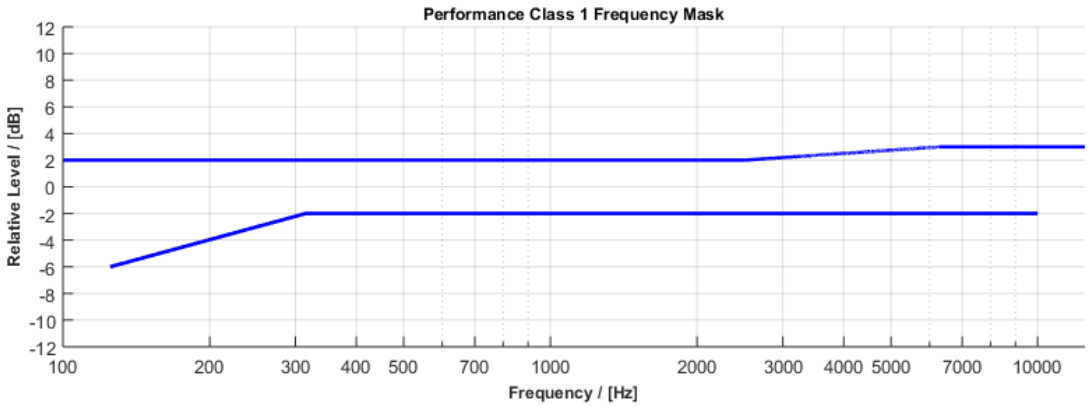

Figure 3: Microphone frequency mask **Class 1** (according to ITU-T P.1130 (06/2015))

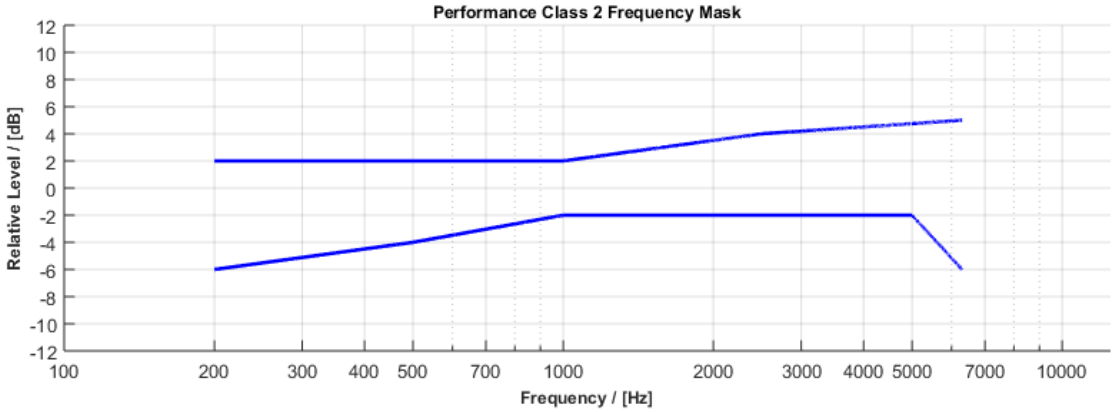

Figure 4: Microphone frequency mask **Class 2** (according to ITU-T P.1130 (06/2015))

NOTE:
For telephony systems supporting wideband or higher audio bandwidth, ITU-T P.1130(06/2015) [4] *frequency response mask Class 3 (**Figure 5**) is not sufficient as qualification against ITU-T P.1110* [3]*. A system with Class 3 microphone is very likely to fail against ITU-T P.1110* [3] *due to missing low-frequency audio content.*

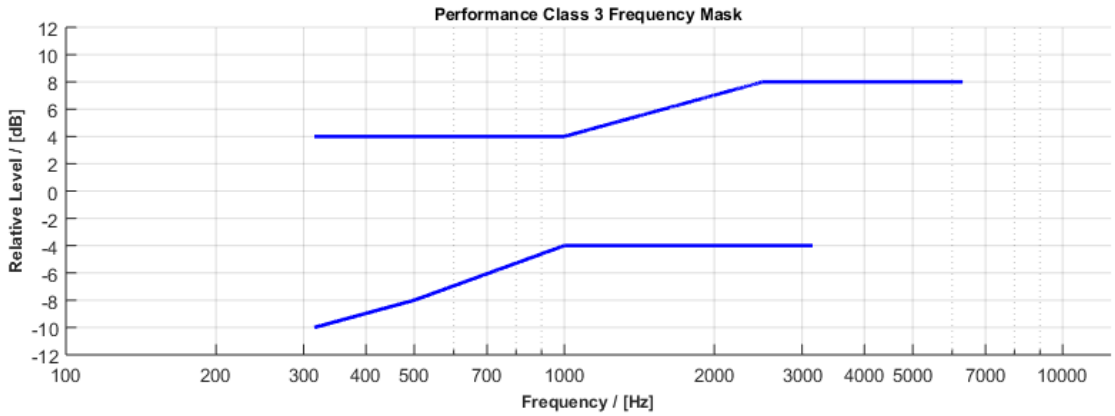

Figure 5: Microphone frequency mask **Class 3** (according to ITU-T P.1130 (06/2015))

Note 1: In accordance with ITU-T P.1130 (06/2015) [4], section 8.3.2.1.2.4, “the [frequency] response characteristics of the microphone should be flat in the frequency range of wideband transmission (100 Hz-7 kHz).” Deviations of the frequency response can partially be compensated by equalizing the signal in the Cerence SSE. However, this might increase the effective self-noise of the system and can cause problems during certification under ITU-T P.1100 [2] or ITU-T P.1110 [3].

Note 2: In accordance with ITU-T P.1130 (06/2015) [4], section 8.3.2.1.2.4, “[…] especially in the presence of background noise, a bandwidth limitation may be desirable." Cerence highly recommends at least an additional high pass characteristic as shown in the following figure:

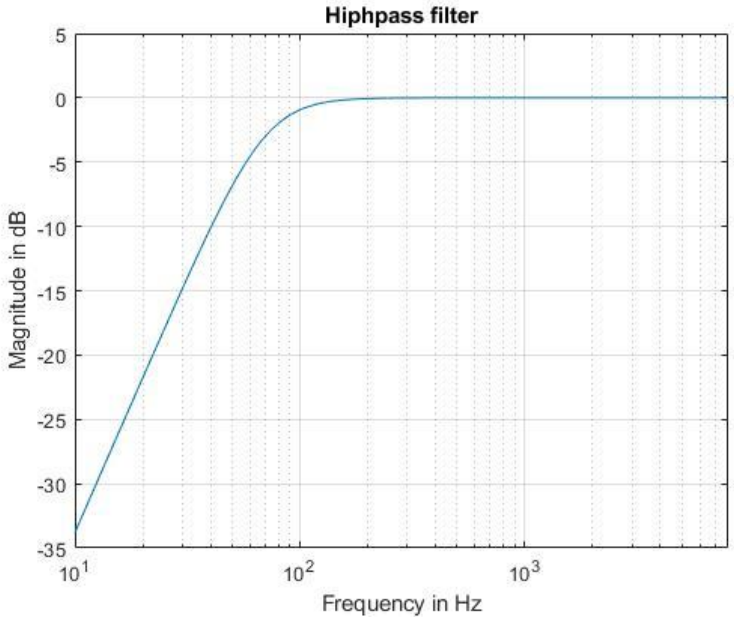

Figure 6: Recommended high pass filtering of the microphone signal to prevent saturation due to background noise

This type of filter prevents an overload of the microphone in situations with high background noise (e.g. open window). The high pass should be realized directly within the microphone, e.g. by the HW design of the microphone itself. This prevents a saturation of the microphone signal and ensures a scaling of the microphone signal which doesn't require a large headroom (which might lead to quantization effects)

Note 3: In ITU-T P.1130 (06/2015) [4], Annex A, a different tolerance scheme is shown (to be measured in an anechoic chamber), but this tolerance scheme is less strict, not suitable for wide-band applications and therefore not referenced by Cerence.

3.2.4 Microphones with Internal Signal Processing

Cerence strongly advises against microphone systems with integrated signal processing—so called DSP microphones—especially if the microphone signal processing includes any non-linear or time variant algorithms (e.g., noise-reduction, automatic gain control or spatial filtering). The reason is that the microphone processing can highly impair the performance of an AEC as it permanently modifies the transfer function of the echo-path which is modelled in the AEC. As a result, echoes or speech distortion can occur. Besides the AEC, also the performance of other algorithms can be impaired by non-linear and time variant

microphone signal processing. Using microphones with internal non-linear or time-variant processing puts a successful certification according to ITU-T P.1100/1110 [2][3] or third-party certification requirements at a high risk.

3.2.5 Microphones with Digital Interface

Microphones with a digital interface are a combination of an analog microphone, an ADC and potentially a sample rate converter. All individual requirements for the acoustic part of the microphone, the ADC, and the sample rate conversion must be considered accordingly.

3.3 Microphone Arrangement

The microphones should be placed such that the SNR of the desired signal is maximized. This ensures that the speech signal is captured in an optimal manner while the interference signals are kept at a low level.

If a microphone array is used, it is mandatory to arrange the microphones in a defined topology. The following array configurations are supported by Cerence SSE: uniform linear arrays (ULA), uniform circular arrays (UCA), and uniform circular arrays with a center microphone (UCA_C).

For microphone arrays, it is essential that the microphones are in free line-of-sight to all desired speaker positions.

3.3.1 Seat-Dedicated Microphones

In the “seat-dedicated” configuration, a separate microphone is installed for each relevant seat in a vehicle. All microphones need to be of the same type. It is important that **acoustic cross-talk** is attenuated by at least 6 dB (under consideration of the microphone tolerances). This means that for **each seat position, given only one speaker in the car and at that position**, the recorded level in the assigned microphone is at least 6 dB higher than in the microphones assigned to other seats.

NOTE:
This can be achieved by spatial separation (e.g., the driver’s microphone is further away from the passenger and vice versa) or by deploying directional microphones with appropriate orientations (the driver’s microphone has its spatial “null” in the direction of the passenger).

Seat-dedicated microphones can be used for PIC (passenger interference cancellation) and for signal mixing configurations.

3.3.2 Uniform Linear Array (ULA)

For a ULA configuration, the microphones are aligned in a straight line with equal distance between adjacent microphones. In case of directional microphones, all elements shall have the same orientation as the steering direction of the beamformer (often broadside direction i.e. perpendicular to the array axis). The recommended distance between the microphones depends on the number of microphones.

Number of microphones	Recommended distance	Optimal distance
2	2.0-10.0 cm	6.0 cm
3	2.0-7.0 cm	4.0 cm
4	2.0-6.0 cm	3.0 cm
5 to 8	2.0-5.0 cm	2.5 cm

3.3.3 Uniform Circular Array (UCA)

In a UCA configuration, all microphones are located on a circle with equal distance between adjacent microphones. Omnidirectional microphones must be used. At least three microphones are required for the UCA configuration. The UCA geometry is specified by the radius of the circle and the number of microphones.

Number of microphones	Recommended radius	Optimal radius
3 to 8	2.0-6.0 cm	4.0 cm

3.3.4 Uniform Circular Array with Center Microphone (UCA_C)

For a UCA_C configuration, the microphones are located on a circle with equal distance between adjacent microphones. Additionally, there is one microphone located at the center of the circle. Omnidirectional microphones must be used. At least four microphones are required for an UCA_C configuration. The UCA_C geometry is specified by the radius of the circle and the number of microphones.

Number of microphones	Recommended radius	Optimal radius
4 to 8	2.0-6.0 cm	4.0 cm

3.3.5 Arbitrary Array Configuration

If a microphone array with arbitrary microphone positions (neither ULA nor UCA nor UCA_C) shall be used, it is highly recommended to contact Cerence to ensure that this configuration will be supported by Cerence SSE as well.

3.4 Microphone Path

In **Figure 1** the Microphone Path is marked in orange. As mentioned in section 2.2, the microphone audio path is part of the echo path and therefore it is necessary that the microphone audio path is linear and time invariant. Saturation and/or clipping must not occur anywhere in the complete microphone audio path.

The complete microphone path (including ADC and SRC) must fulfill the following requirements:

- No additional gain between ADC and Cerence SSE.
- THD+N < 1% for all relevant signal levels corresponding to audio levels up to SPLMAX.
- Frequency response of SRC as required for passband frequencies > 50 Hz (see section3.15.3).

3.4.1 AD-Converter

It has to be ensured that the microphone signals do not clip. This is especially important for the application of AEC. The AD converter should reach digital full scale at the SPLMAX in the

microphone. No non-linear distortions (such as saturation effects) are allowed in the electrical microphone signal or after the AD-converter up to this signal level. With a programmable analog gain stage within the ADC it is possible to adjust the signal range in a flexible way.

Further requirements regarding the ADC:

1. Self-noise: SNR > 80 dB (relative to digital full scale)
2. Crosstalk between ADC channels in multi-microphone configurations for all frequencies < -60 dB
3. Resolution: ≥16 bit
4. Linear operation, THD+N ≤ 1%
5. Preferably a programmable analog input gain
6. No two's complement overflow shall occur.

In multi-microphone configurations the temporal relation between the microphone signals is important. All ADCs must work clock-synchronously (same clock) and sample simultaneously (without relative delay to each other).

3.4.2 Sample Rate conversion

The microphone audio path is relevant for the operation of AEC and therefore it must be linear and time invariant. If a sample rate converter is needed in the microphone audio path, it is important that only synchronous sample rate converters (SSRC) are used. Asynchronous sample rate converters (ASRC) are not allowed in this path because typically they don't have a constant latency and therefore are not time invariant (see also section 3.15.2).

If a SSRC is needed, the anti-aliasing filter must suppress aliasing components by at least 50 dB (see section 3.15.3) to avoid non-linear behavior.

3.4.3 Delay of Microphone Path

Because the delay in the microphone path is part of the echo path, it is very important to keep it constant during an active use-case (on sub-audio-sample base; Performance Class 1 for "Limits for the clock drift between the interfaces" in ITU-T P.1130(06/2015) [4]). The delay

of the microphone path is relevant for the relative delay between microphone channel and reference channel and therefore a critical characteristic for AEC (see also section 2.2).

It is recommended to keep this delay as short as possible to fulfill the delay requirements of ITU-T P.1100 / P.1110 [2][3] and Apple CarPlay [5].

3.4.4 Multi-Microphone Configurations

For multi-microphone configurations, the microphone paths of all involved microphones shall have the same **internal delay** (on sub-audio-sample base). Neither variable nor static latency are permitted between the microphone signals.

The **sensitivity** and the **frequency response** of each microphone path may not differ by more than 1dB for all microphone paths.

3.5 Reference Channel Requirements

This section describes reference channel generation as it is typically used in automotive applications. Nevertheless, the basic facts are valid for other applications like IoT or home entertainment as well and the requirements hold for these applications as well.

In a system which uses AEC, the reference channel is a fundamental signal. The reference channel allows the identification of the parts of the microphone signal that stem from the loudspeaker (and should be cancelled) and which part is background noise or local speech (and should not be distorted).

Each independent audio source requires a separate reference channel. For sources like telephony downlink or TTS prompts, a single reference channel is suitable, a stereo signal requires 2 reference channels (2 independent signals “left” + “right”) and for surround sound at least 5 independent reference channels are needed.

NOTE:
Special consideration must be taken for volume control. Since the volume control is time-variant, it must be applied before reference channel generation.

3.5.1 Signal Level and Frequency Response of the Reference Channel

The signal level in the reference channel must be proportional to the signal level at the according loudspeakers, and the reference channel must operate linearly up to the highest

required signal level (corresponding to SPLMAX). This means that there must be a constant gain between reference channel and loudspeaker signal. At SPLMAX, the maximum signal level in the reference channel shall be as high as possible but the signal must not clip.

The recommended level for a reference signal is:

- If only a single audio source (e.g., telephone downlink or TTS prompts) is included in the reference channel: At SPLMAX, an input signal to the audio sub-system with 0 dBFSpk shall result in a reference signal level between 0 dBFSpk and -6 dBFSpk.
- If two audio sources (e.g., telephone downlink + navigation prompts) are included in the reference channel: At SPLMAX, a single input signal of the audio sub-system with 0 dBFSpk shall result in a reference signal level between -6 dBFSpk and -12 dBFSpk.
- If more than two audio sources (which are not permanently active) are included in the reference channel: Under the assumption that not all sources will be active in parallel, at SPLMAX a single input signal of the audio sub-system with 0 dBFSpk shall result in a reference signal level between -9 dBFSpk and -12 dBFSpk.

The shape of the frequency response of the reference channel shall be the same (within +/- 5dB) as the frequency response measured at the position of the driver's right ear for left-hand drive vehicles.

3.5.2 Reference Channel Generation from Source Signals

For this method, the reference channel is picked up directly after the signal specific volume has been applied to the source signal (alternative: the reference channel is picked up before volume setting and the volume is applied to the reference channel and the loudspeaker signal at the same time and in the same way). Typical signals which can be handled here are the telephone downlink or TTS / navigation prompts. Stereo signals for media could be handled this way as well, if no upmixing (e.g., stereo → 5.1) is done afterwards.

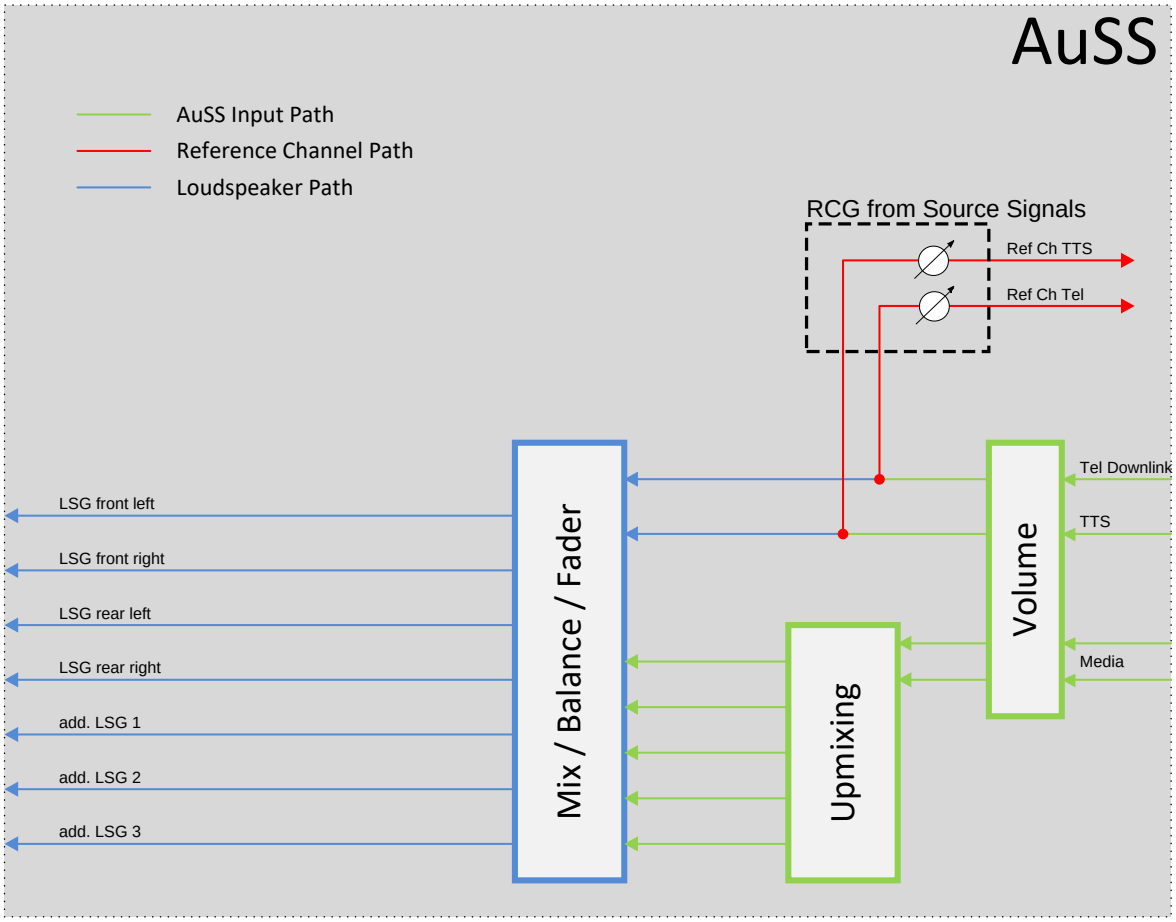

Figure 7: Generic structure of AuSS for RCG from source signals; loudspeaker path (blue) and reference path (red) are relevant for AEC und must be linear and time-invariant

The advantage of the mechanism shown in **Figure 7** is that for each independent audio source a dedicated reference channel will be generated. There are no implicit dependencies between the dedicated reference channels, and every signal is handled independently from each other. On the other hand, it must be considered that the RCG precedes the input to the AuSS. This limits the further processing in the AuSS (and subsequent "loudspeaker path"!) to linear and time-invariant modules. Linear processing like equalizing or balance/fader settings are estimated in the AEC filter and need not be considered in the reference channel.

3.5.3 Reference Channel Generation from Loudspeaker Signals

The other general method for RCG is the generation from the loudspeaker signals. Each reference channel is a weighted sum of loudspeaker signals or loudspeaker group signals. I.e.

the reference channels no longer represent individual audio sources. The volume setting is implicitly considered.

To avoid comb-filter effects in the RCG, it is not allowed to add signals from loudspeakers or loudspeaker groups which have a relative delay between them.

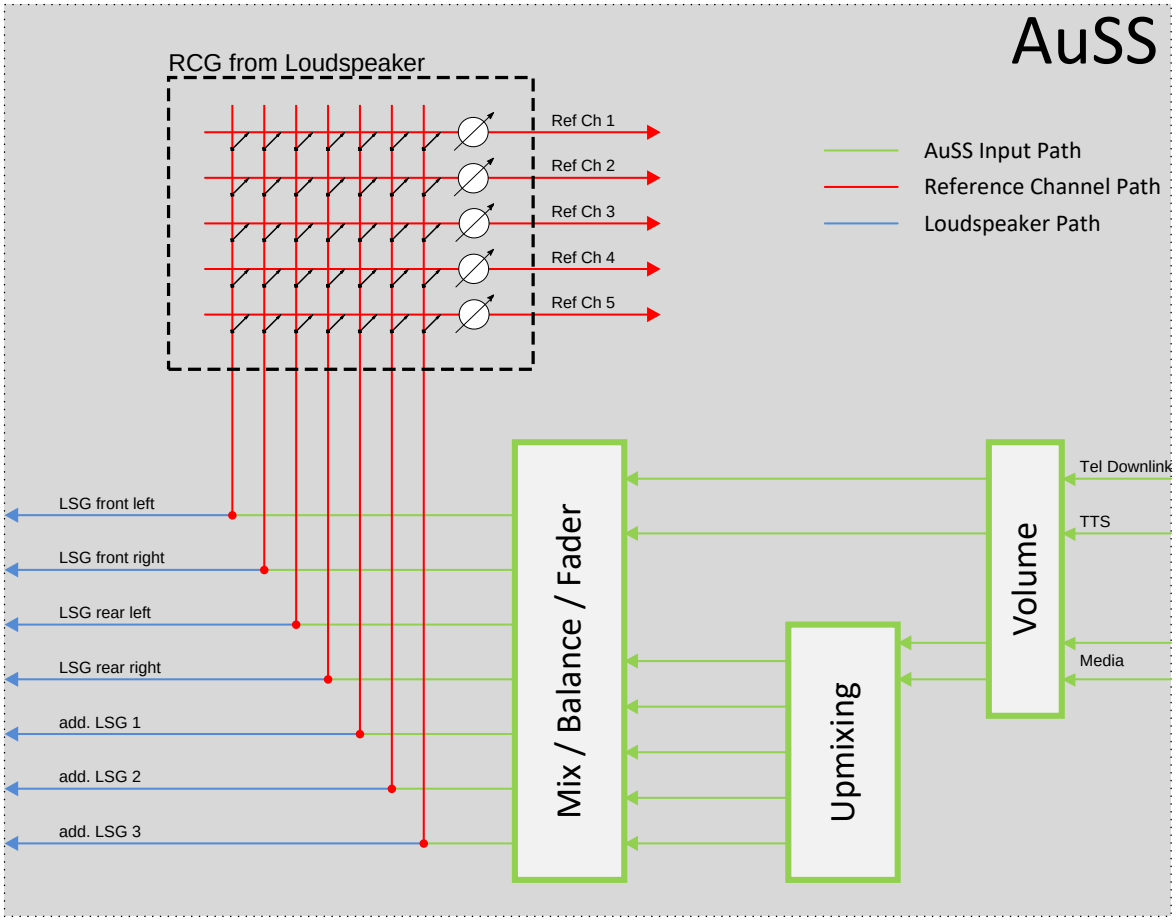

Figure 8: Generic structure of AuSS for RCG from loudspeaker signals; loudspeaker path (blue) and reference path (red) are relevant for AEC und must be linear and time-invariant

In real vehicles, loudspeakers can be split into groups where each member of a loudspeaker group (LSG) gets essentially the same signal (e.g., woofer / mid-range / tweeter for front left, separated by a cross-over network).

If there are no non-linear operations in the loudspeaker group processing (see **Figure 9**), it is possible to use the input signal of an LSG directly as shown in **Figure 8**.

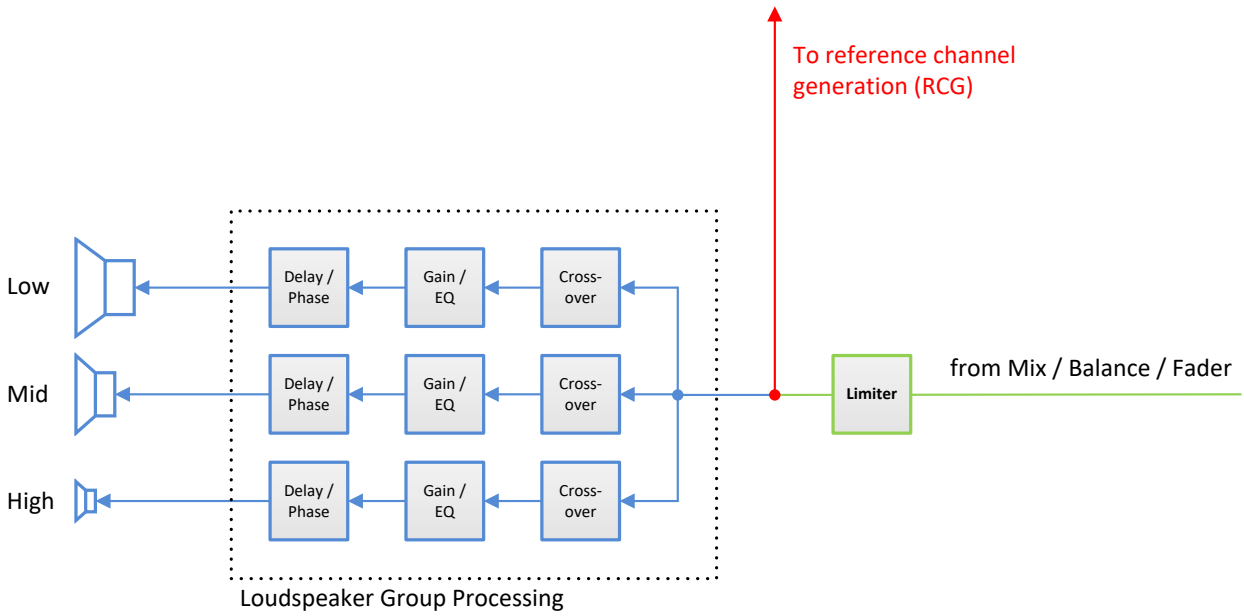

Figure 9: Reference Channel Generation for LSG processing after limiter. Here, the reference channel can be generated from the LSG input signal

In terms of non-linear processing, the critical component is typically the **limiter**. In **Figure 9** the limiter is located before the RCG. This means that the limiter will affect the signal of all loudspeakers within this LSG in the same way.

If the limiter is part of the LSG processing, the reference channel cannot be generated from the LSG processing input. In this case, the reference channel must be generated from the individual loudspeaker signals:

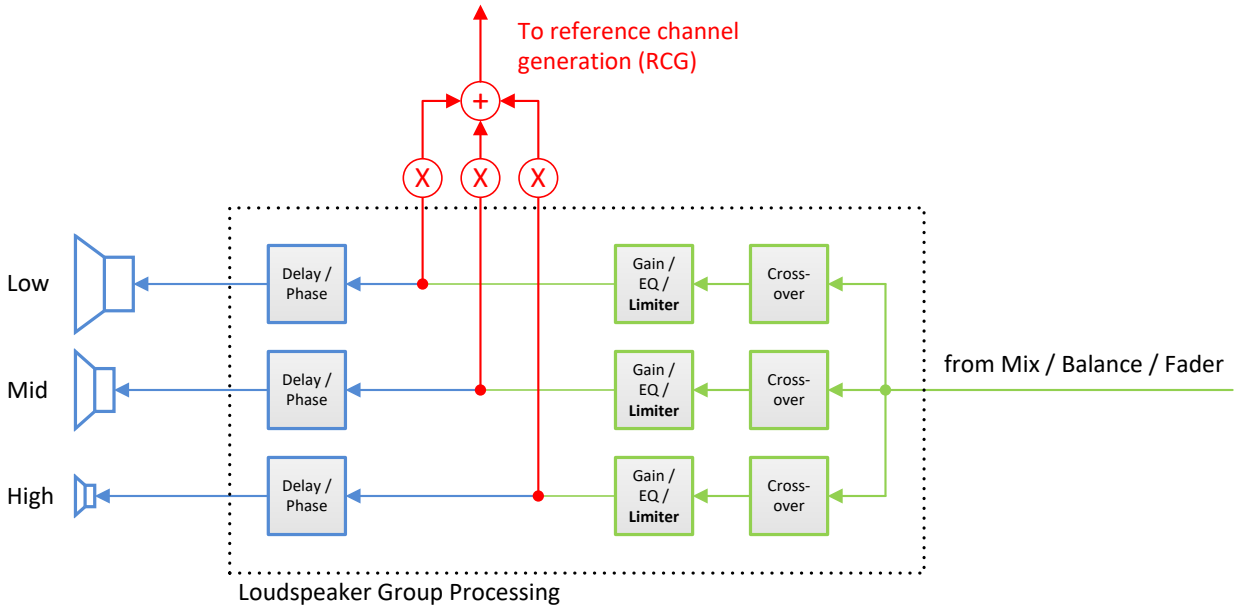

Figure 10: Reference Channel Generation for LSG processing with integrated limiter. Here, the reference channel must be generated from the weighted individual loudspeaker signals but before delay processing.

Figure 10 shows a generic implementation of an LSG with the limiter as non-linear module inside the LSG processing. In this case, it is necessary to generate the reference channel as a weighted sum of the individual loudspeaker signals of the LSG.

The **overlap of the frequency intervals** of the loudspeakers shall be very small to avoid comb-filter effects due to summation in the reference signal generation.

If **individual delays** are applied to the loudspeakers, it is important that these delays are **not** considered in the reference channel because they would lead to interference effects (comb-filtering) and reverberation in the reference signal which has negative effects on the AEC performance.

Recommended reference channel setup

In an early phase of a project, an initial set of reference channels will be determined based on the loudspeaker channels and the upmixing algorithm. Cerence expects that a suitable data base will be created in cooperation with the vendor of the sound system. The final set of reference channels must be determined during the tuning process and depends both on the upmixing algorithm and on the positions of loudspeakers and microphone(s) in the target environment. Therefore, it is necessary that the mixing and weighting of the loudspeaker signals into the reference channels can be modified by Cerence during the SSE tuning sessions.

The initial configuration of RCG should be as follows:

- 2 reference channel configuration (Stereo only): “left” / “right”
- 5 reference channel configuration (Stereo / Surround): “front left” / “front right” / “Center” / “Surround left” / “Surround right”. The “Surround” channels should be seen as place holders for the most dominant remaining channels, so depending on the loudspeaker configuration in the vehicle they might be replaced by other channels (e.g. “rear” or “3D”) in the tuning.

For each LSG, at least the mid loudspeaker must be taken into the reference channel. Depending on the amplifier and the cross-over networks it might be beneficial to add further loudspeaker signals (e.g., from bass speakers) in order to better cover other frequency ranges.

For a robust ASR performance, especially in the case of One-Shot wake-up-word activation, it is recommended to use a reference channel set composed of M = 5 reference channels if the system supports surround sound or applies upmixing algorithms for stereo signals.

3.5.4 Reference Channel Generation Depending on Use-Cases and Audio Configuration

The optimal method for reference channel generation depends on the specific use-cases:

3.5.4.1 Telephony or Barge-In Without Interfering Audio Sources

In this application, only a single audio source is active. The reference channel can be generated either from the source signal or from the loudspeaker signals (e.g., (“front left” + “front right”)/2). The recommended mechanism is to generate it from the source signal because then there are no restrictions regarding balance and fader settings. The disadvantage of limited processing afterwards is not so relevant here because the audio source is an announcement signal which doesn’t benefit from sophisticated, non-linear processing afterwards.

3.5.4.2 Telephony or Barge-In with Interfering Mono Audio Sources (same Center of Sound)

If more than one mono source is played back in the vehicle at the same time (e.g., telephony downlink + navigation prompts) **and** if they have the same center of sound (balance and fader), this is equivalent to playing back the sum of the signals over the loudspeakers. Effectively the sum can be considered a single mono signal. So, the reference channel can be generated as the sum of the signals as well. The relative gains between the signals at the loudspeaker must, however, also be considered in the reference channel.

3.5.4.3 Telephony or Barge-In with Interfering Mono Audio Sources (Different Center of Sound)

If two (or more) signals are played back with different balance and fader settings (e.g., telephone downlink at “front left” and “front right” and navigation prompts at “front left” only), then the transmission paths for both sources are different and cannot be modelled by a single AEC filter. In this case a single reference channel is not sufficient because for each independent signal a reference channel is needed. The recommended setup would be to generate the reference channels from the source signals, so that each source gets its “own” reference channel. The advantage of this solution is that each signal can be handled independently in Cerence SSE and does not interfere with the others.

If it is not possible to generate the reference signals from the audio sources, there are two alternatives:

1. The audio management takes care that all active signals have the same center of sound (see section 3.5.4.2).
2. The reference channel is generated from the loudspeaker signals (e.g., “front left” and “front right”). However, this has a certain risk of echoes: As long as only a single source is active, the AEC filters cannot decide which part of the microphone signal comes from which loudspeaker – only the sum of both parts will be treated. Doing this, the AEC will make an implicit decision which might be wrong because the reference channels are correlated (same source signal). As soon as a second source (with different balance setting) becomes active, the situation becomes unambiguous: The reference channels are no longer correlated because the weighting of the two sources is different. Typically, in this moment both AEC filters are not correctly adapted: Before this the adaptation was only valid for the sum but not for the individual sources. This means that now both AEC filters need to re-adapt which can cause echoes or can degrade double-talk performance for both signals during the adaptation time.
3. It is even possible that the second source cannot be treated correctly at all: Assume that the two reference channels are generated as “front left” and “front right”. If the first signal is played back over these two loudspeakers, the echo can be cancelled correctly. If the second source is played back over further loudspeakers like “rear left” and “rear right”, the reference channels will not change, but the transfer function for the second signal is completely different. This means there exists no single transfer function which can model the microphone signal correctly based on the sum of both signals (the reference signals). And it is not possible to benefit from

the second reference channel because it is correlated to the first one and has no additional information to distinguish between the audio sources. As a result, no correct AEC is possible here at all as soon as both sources become active.

3.5.4.4 Barge-In with Interfering Media Sources

The idea for this configuration would be a barge-in system in which the entertainment sources are not muted during the barge-in session.

To achieve a robust barge-in behavior, it is not recommended to keep entertainment active during a barge-in session at all. There is always a risk of residual echoes which can lead to self-barge-in:

1. **2 reference channels ("left" and "right"), same fader setting for entertainment and TTS:** Depending on the signal content (media, TTS) and balance setting for TTS, the two reference channels can be highly correlated. This means that AEC will tend to adapt to the (correlated part of the) sum of both reference channels. However, the correlation between both reference channels changes when TTS is active or not – so there is a high risk of residual echoes in these situations which can lead to self-barge-in.
2. **2 reference channels ("left" and "right"), different fader setting for entertainment and TTS:** For each reference channel the impulse response is significantly different for entertainment and TTS because the fader setting is implicitly modelled in the impulse response. Therefore, the AEC has to re-adapt at each beginning and end of a TTS prompt. This will cause echoes which can lead to self-barge-in.
3. **5 reference channels:** The basic situation is the same as described in item 1 above. Depending on the entertainment signal, some reference channels can be highly correlated (e.g., "front left" and "rear left") so that the AEC adapts to their sum. As soon as a TTS prompt is played back, the correlation will be reduced, and the AEC has to re-adapt to the (now less correlated) channels themselves. This will also cause residual echoes which can lead to self-barge-in. In following dialog phases when only entertainment is active, AEC might re-adapt to the (now again correlated) signals again, so that the next TTS prompt requires a further re-adaptation again.

NOTE:
The effects which are described in this section can also occur during wake-up-word. However, for barge-in the situation is much more critical because residual echoes can cause self-barge-in which unexpectedly stops the play-back of a prompt. In the wake-up-word use-case the situation is much more robust because the ASR engine will

trigger only if the correct wake-up-word is recognized which is very unlikely for residual echoes.

3.5.4.5 Wake-up-word (Media in Stereo)

If the media playback only consists of stereo signals (without upmixing!), both methods for reference channel generation are possible. Cerence recommends generating the reference channel from the loudspeaker signals, because this gives more flexibility for the signal processing in the AuSS (potential non-linear or time variant processing).

If a navigation prompt (or another mono source) is played in parallel, it must be played back with the same fader settings as the media. If this is not possible, the AEC must estimate more than 2 impulse responses and therefore needs more than two reference channels:

- For RCG from source signals, the navigation prompt is a third audio source and needs a separate reference channel (3 reference channels in total).
- For RCG from loudspeaker signals, 4 reference channels are required (“front left”, “front right”, rear left”, rear right”) because both the “left side” and the “right side” have to handle the navigation prompting separately from media playback.

3.5.4.6 Wake-up-word (Media in Surround Sound)

For wake-up-word during active media sources with more than two independent audio channels (e.g., 5.1 or other upmixing from stereo), it is necessary to generate the reference channel from the loudspeakers. If more than 5 independent audio signals are available, Cerence recommends using 5 reference channels as a compromise between AEC performance and computational complexity. Tests have shown a significantly higher performance than using 2 reference channels for this use-case. Typically, the signals for “front left”, “front right”, “rear left”, “rear right” and the most dominant channel of the remaining (e.g., “3D” channel) are used.

3.5.5 Wake-up-word in combination with In-Car-Communication (ICC)

In systems which are applying wake-up-word and ICC at the same time, SSE and ICC need to access the loudspeaker signals at the same time: SSE gets reference channels for its AEC for Wake-up-Word while the speech signal reinforced by ICC is mixed to the loudspeaker signals. Within the signal path of the amplifier, the reference signals for SSE need to be tapped before the ICC output signal is added to the loudspeaker signals. Otherwise, due to the low-latency characteristic of ICC and the periodic signal in voiced speech, an AEC module

could interpret local speech as an echo because it's "at the same time" in the microphone signal and in the reference channel.

During hands-free calls or barge-in sessions ICC shall be muted.

3.5.6 Engine Sound Simulation and Active Noise Cancellation

Acoustic effects like engine sound simulation (ESS) or active noise cancellation (ANC) shall not be included into the reference channels.

Within the signal path of the amplifier, the reference signals for SSE need to be tapped before the ESS or ANC output signals are added to the loudspeaker signals.

3.6 Reference Channel Path

The reference channel path is the audio signal path from reference channel generation (RCG) to the reference input of Cerence SSE (marked in red in **Figure 1**). It is relevant for AEC and therefore it must be linear and time invariant.

The complete reference channel path between reference channel generation (RCG) and Cerence SSE must fulfill the following requirements:

- No additional gain between RCG and Cerence SSE.
- THD+N < 1% for all relevant signal levels.
- Frequency response of SRC between RCG and Cerence SSE as required for passband (see section 3.15.3).

3.6.1 Signal Quality

The reference channel path must not add any signal components to the reference signal which are not played back over the AuSS.

The typical implementation of the reference channel path will be a pure digital path which doesn't affect the reference signal. However, if the reference channel path contains an analog part (e.g., combination of DAC and ADC), no electrical noise must be generated there.

3.6.2 Sample Rate Conversion

If a sample rate converter is needed in the reference channel, it is important that only synchronous sample rate converters (SSRC) are used. Asynchronous sample rate converters (ASRC) are not allowed in this path because typically they don't have a constant latency and therefore are not time invariant (see also section 3.15.2).

If an SSRC is used, the anti-aliasing filter must suppress aliasing components (see section 3.15.3) to avoid non-linear behavior.

3.6.3 Delay of Reference Channel Path

During an active use-case, the delay of the reference channel path must be constant (even if the delay is just a fraction of a sample) for all use-cases to guarantee a constant relative delay between microphone signal and reference channel. This is vital for high quality AEC.

For telephone applications, it is strongly recommended that this delay is shorter than the sum of the delays in loudspeaker and microphone paths: AEC requires that a signal arrives from the reference channel before the corresponding signal from the microphone path (otherwise AEC cannot estimate which signal components in the microphone signal are echo components). If the reference channel arrives after the microphone signal, it is necessary to delay the microphone signal internally in Cerence SSE which will increase the delay in sending direction and might cause problems with the roundtrip requirement from ITU-T [2][3] and Apple CarPlay [5].

3.7 Audio Sub-System Requirements

The audio sub-system (AuSS) can be split into two different areas: the processing that is done before reference channel generation ("RCG") for the Cerence SSE and the processing that is done after RCG.

3.7.1 Processing before Reference Channel Generation

If non-linear and time-variant processing (NLTV) is performed within the AuSS, this must be done before the RCG. Typical modules that perform non-linear or time-variant processing within an AuSS are:

- Automatic/dynamic volume control
- Compressor/limiter/expander

- Bandwidth extension
- Fader
- Modulation effects (chorus etc.)
- Surround sound processing
- Noise suppression

Since non-linear processing cannot be modeled by the AEC and time-variant processing requires permanent re-adaptation of AEC, this type of processing must be represented within the reference signal.

3.7.2 Processing after Reference Channel Generation

In contrast to the previously discussed processing before RCG, the processing after RCG must be linear and time-invariant (LTI). Typical modules that perform linear and time-invariant processing are:

- Fixed Equalizer (doesn't change during operation)
- Reverberation/convolution with a finite (shorter than the AEC filter length) impulse response
- Fixed gain

It is important to note that surround sound processing is time-variant and sometimes non-linear as well and therefore not allowed to be applied after reference channel generation.

All linear and time-invariant processing that is done after RCG can be modeled by the AEC since it can be interpreted as a finite impulse response. This is treated as a part of the room impulse response. If it prolongs the room impulse response that is seen by the AEC, the filter length of the AEC in the Cerence SSE must be chosen appropriately.

3.7.3 Maximum Sound Pressure Level

The maximum SPL must be defined at the beginning of a project (see 3.1). SPLMAX must consider the fundamental requirement, that the complete system remains linear (and time-invariant). This means that the whole audio path must stay linear. For automotive applications, typically a maximum level corresponding to RLR = -18 dB is recommended for telephony and barge-in, which fits quite well to the microphone specification as

recommended in ITU-T [2][3]. The maximum amplification in the AuSS must be chosen according to the maximum needed volume in the car. In automotive applications, an electric output level of the AuSS of +17 dBV or higher is required for this.

For this maximum SPL (and all lower values) the AuSS used in the system shall not clip and the distortions shall be THD < 1%.

The electrical SNR (relative to max. output level) shall be > 70 dB.

In wake-up-word applications, a higher SPLMAX might be required (SPL is defined by media volume setting). However, as soon as the system becomes non-linear, the wake-up-word performance will be degraded.

3.7.4 Frequency Response

The frequency response of the AuSS shall fulfill the requirements for the anti-imaging filtering for sample rate converters (see section 3.15.3) in up-sampling mode.

NOTE:
If a vehicle specific sound tuning is applied to the AuSS, the frequency response will not necessarily be flat. In a tuned system, the overall acoustic frequency response of AuSS and loudspeakers will be optimized instead of the electrical frequency response of the AuSS only.

3.7.5 Volume Control

Special consideration must be taken for volume control. Since the volume control is time variant, it must be applied before reference channel generation. Only then can the AEC still operate in an optimal manner when volume changes occur.

If the audio architecture of the telephony application does not allow the implementation of the volume control before the RCG, this is acceptable **for telephony use-cases only** and the customer must be aware that residual echoes might be audible for the remote communication partner for a short duration after the volume is changed. For voice recognition use-cases in combination with barge-in and wake-up-word detection, volume control shall be applied before the RCG to prevent false detections of the wake-up-word for a duration after the volume is changed.

It is important to keep in mind that in some automotive sound systems the volume is changed frequently and automatically depending on the speed of the vehicle or depending on the background noise in the cabin.

3.7.6 Loudspeaker Group Processing

This section describes sound processing as it is typically used in automotive applications. Nevertheless, the basic facts are valid for other applications like IoT or home entertainment as well.

Most high quality automotive sound systems use different loudspeakers to transmit different frequency ranges of a single audio signal. The “front left” signal might be transmitted by a woofer (for low frequencies), a mid-range loudspeaker and a tweeter (for high frequencies). Such configurations are called “loudspeaker groups”. The separation of the individual loudspeaker signals from the signal of the loudspeaker group is done by a cross-over network, which can be built either electrically (with resistors, coils and capacitors) or digitally as part of the signal processing. **Figure 11** shows a generic structure for this:

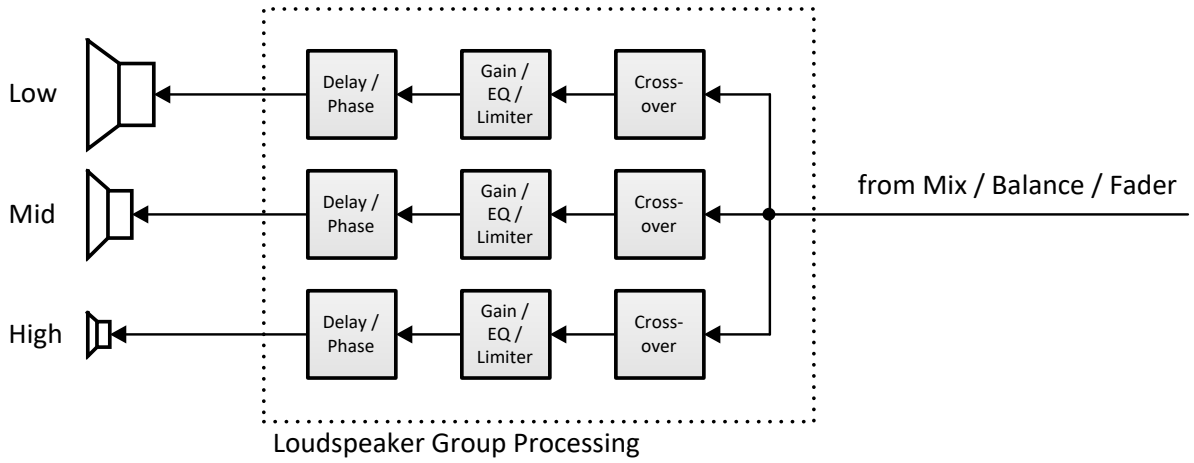

Figure 11: Loudspeaker group processing with three individual loudspeakers

The overlap of the frequency intervals of the loudspeakers shall be very small to avoid comb-filter effects due to summation in the reference signal if the reference channel is generated from individual loudspeaker signals.

If loudspeaker groups are used in a vehicle and the reference channel is generated according to **Figure 10** (section 3.5.3), the individual delay for each loudspeaker must be less than 5 ms.

For LSG processing the limiter can either be placed in the individual loudspeaker signals of an LSG (as shown in **Figure 11**) or in front of the whole LSG on the input signal of the LSG. This has an impact on the reference channel generation from the loudspeaker signals (see section 3.5.3)

3.8 Audio Sub-System Input Paths

The AuSS input paths (marked in green in **Figure 1**) are all paths from audio sources like telephony downlink, prompter, or media to the AuSS. In the context of this document, they can be considered as part of the audio sources, so no specific requirements apply.

However, the AuSS input paths affect the audio quality of played signals. Hence, it is recommended to avoid jitter, clipping or other distortions in these paths as well to avoid problems when certifying the system according ITU-T P.1100/1110 [2][3] or Apple CarPlay [5]. The audio bandwidths of the input paths shall be large enough to avoid signal degradations.

3.9 Loudspeaker Requirements

The loudspeakers used in the system shall not clip or produce a THD>3% for all required volume settings in the relevant use-cases.

It is important that the physical installation of the loudspeakers prevents resonances which can be caused by mechanical coupling to other components like speaker grills or covers.

3.10 Loudspeaker Path

The loudspeaker path is the audio signal path from RCG to the loudspeaker itself (marked in blue in **Figure 1**). It is relevant for the operation of AEC and therefore it must be linear and time invariant.

NOTE:
Typically, there will be some further signal processing in this path like the whole loudspeaker group processing (EQ, cross-over, loudspeaker specific gain, …). All these processing modules must be linear and time-invariant (see Figure 9 and Figure 10)!

The frequency response of the loudspeaker path depends on the sound tuning. A general and minimum requirement is an audio bandwidth which supports all signals required for the relevant audio scenarios.

3.10.1 Protection Limiter before DAC

It is important to understand that a limiter is a highly non-linear module. In many amplifiers "protection limiters" are used directly before the DAC – for an ideal system these limiters should be considered in the reference channel as well. Depending on the method for RCG this might be difficult to implement – in this situation it is recommended to reduce the output signal to levels which don't activate these limiters.

3.10.2 Sample Rate Conversion

Asynchronous sample rate converters (ASRC) are not allowed in this path because typically ASRC don't have a constant latency and therefore are not time invariant (see also section 3.15.2).

3.10.3 Delay of Loudspeaker Path

It is very important to keep the delay of the loudspeaker path constant because it is part of the echo path. The delay of the loudspeaker path is also relevant for the relative delay between microphone channel and reference channel and, therefore, is a critical characteristic for AEC (see also section 2.2).

It is recommended to keep this delay as short as possible to fulfill the delay requirements of ITU-T P.1100 / P.1110 [2][3] and Apple CarPlay [5]: At AEC input, the reference channel must arrive before the microphone signal. If the delay of the reference channel is too high, it will be necessary to add a delay on the microphone signal inside Cerence SSE. This will increase the delay from microphone to NAD accordingly.

3.11 Network Access Devices (NAD)

The term "network access device" (NAD) covers several kinds of devices:

- BT interface to mobile phone (typically via HFP)
- Integrated modem for network access (GSM, UMTS, CDMA, LTE, …)
- Smartphone connected via USB or Wi-Fi (e.g., Apple CarPlay, Android Auto)

Depending on use-case and system, the appropriate definition must be chosen.

3.11.1 Sensitivity

An input signal of 0 dBFSpk at the input of an NAD shall result in an output signal in the interval of [-6 dBFSpk; 0 dBFSpk] for both sending and receiving direction.

The CVSD codec which is used for BT narrow-band transmission, has an intrinsic slew-rate limitation. Therefore, this requirement cannot be fulfilled for sine signals with a frequency higher than 315 Hz and such signals are not appropriate for measurement and testing.

3.11.2 Frequency Response

In both directions the NAD shall have a flat frequency response (for frequencies > 50 Hz) according to the "passband" of sample rate conversion (see section 3.15.3). The frequency response is defined at moderate signal levels only to avoid non-linear effects of the involved speech codecs.

3.11.3 Distortions

The distortions shall be kept at a minimum to ensure good speech quality and fulfill certification requirements (e.g., Apple CarPlay [5]).

NOTE:
Some smartphones have internal signal processing which cannot be deactivated. If the NAD testing involves a smartphone, it is recommended to use a smartphone listed on the ITU-T whitelist (http://www.itu.int/en/ITU-T/C-I/Pages/HFT-mobile-tests/HFT_testing.aspx).

3.12 NAD Audio Paths

The NAD paths (marked in purple in **Figure 1**) affect the overall signal quality. Therefore, these paths shall be linear without clipping or other distortions for all signals up to 0 dBFSpk. As all audio paths relevant for AEC must be linear and time-invariant, in telephony use-cases the NAD audio paths are the only audio paths in which asynchronous sample rate converters (ASRC) are allowed. If the NAD has a separate audio clock generator than the other audio components, an ASRC must be placed between Cerence SSE and the NAD (even if the nominal sample rates of Cerence SSE and the NAD are the same!).

3.12.1 Uplink Direction

The NAD uplink path shall transport processed audio signals from Cerence SSE to the NAD.

If sample rate conversion is needed, the frequency response must fulfill the tolerance scheme for sample rate conversion (see section 3.15.3).

3.12.2 Downlink Direction

The NAD downlink path shall transport audio signals from the NAD to Cerence SSE without any changes. If no sample rate conversion is needed, the audio signals shall not be modified at all.

If sample rate conversion is needed, the frequency response must fulfill the tolerance scheme for sample rate conversion (see section 3.15.3).

3.13 Audio Path to ASR Engine

The audio path to the ASR engine (marked in purple in **Figure 1**) is the signal path between Cerence SSE and the speech recognition engine. It must be linear without clipping or other distortions for all signals up to 0 dBFSpk.

3.13.1 Cerence SSE Meta Information

For barge-in and wake-up-word applications in combination with Cerence ASR engine, Cerence SSE can provide additional information (“meta information”) encoded into the “Mic Processed” audio signal. This encoded information helps to reduce false alarms and false triggers in combination with the VoCon speech recognizer. If this highly recommended feature is used in an application, the output signal of Cerence SSE must be sent bit-exact to the speech recognizer. This means that no sample rate conversion, filtering, gain, or attenuation must be applied between Cerence SSE and the speech recognizer – if any sample rate conversion is needed, it is mandatory to use the sample rate converter inside the VoCon speech recognizer.

3.14 Internal System Delays Relevant for AEC

3.14.1 Delay between RCG and Loudspeaker

This delay must be constant (time invariant) if the tap of the reference channel is before the amplifier input. If the reference channel is generated within the amplifier, the delay of the signal processing chain after that tap in the amplifier ("loudspeaker path") must be constant. It is recommended to keep this delay as short as possible to fulfill the delay requirements of ITU-T P.1100 / P.1110 [2][3].

3.14.2 Relative Delay between Reference Channel and Microphone Signal

This relative delay T_{MicRef} is measured at Cerence SSE API between *Mic In* and *Reference In*. It is determined by the difference of the latency of the echo path and the latency of the reference path.

The relative delay T_{MicRef} needs to be constant ($T_{MicRef} = const.$) with a value smaller than 500 ms measured on a sub-audio-sample time base over the whole processing time period of Cerence SSE. This corresponds to Performance Class 1 for "Limits for the clock drift between the interfaces" in ITU-T P.1130(06/2015)[4]. If the Cerence SSE processing is restarted e.g., due to a new telephone call, a new voice recognition interaction or an audio stack close/open operation, a constant relative time delay of $T_{MicRef,N} = T_{MicRef} \pm 5ms = const.$ for the Cerence SSE lifecycle $N=\{1,2,3\ldots\}$ must again be present between the microphone and reference channel.

If this delay is time variant, the AEC cannot adapt, and this would permanently cause echoes or poor double-talk performance.

3.15 Sample Rate Conversion (SRC)

It has to be noted that all inputs and outputs of the Cerence SSE must operate at the same sampling frequency. All sample rate conversion must be done in the audio paths outside of Cerence SSE.

In general, two different types of sample rate conversions exist: Synchronous and asynchronous sample rate converters.

3.15.1 Synchronous SRC

Synchronous sample rate converters (SSRC) operate within the same clock domain (same clock generator) and perform either an up-sampling or a down-sampling. In some cases, this is needed since different modules operate on different sampling rates. Assuming proper anti-imaging and anti-aliasing filtering (section 3.15.3), synchronous sample rate converters are non-critical. Synchronous sample rate conversion must be implemented in a manner such that no time-varying delay occurs. It is recommended to use simple up-sampling or down-sampling factors which allow simple decimation or interpolation.

3.15.2 Asynchronous SRC

Asynchronous sample rate converters (ASRC) typically operate between two different clock domains. These clock domains may even have the same nominal clock rate, but their clock rates are derived from independent clock generators. Asynchronous sample rate converters may also be applied within the same clock domain if the ratio of the two sampling rates makes it difficult to perform a simple decimation or interpolation.

Asynchronous sample rate converters buffer samples and interpolate or extrapolate between samples. This results in a variable delay in the path where asynchronous sample rate converters are used. A variable delay in the echo path or in the reference path will result in decreased performance of AEC (the delay must be compensated by repeated adaptation) and double-talk. Therefore, the complete echo path and the reference path must operate in the same clock domain as the Cerence SSE and no ASRCs are allowed there.

If ASRCs are used in the NAD paths, the variation of their latency should be kept as small as possible to fulfill the low-latency requirements for ITU-T [2][3] or CarPlay [5].

3.15.3 Anti-Aliasing and Anti-Imaging Filtering

To avoid artifacts due to sample rate conversion, low-pass filtering is necessary to limit the audio bandwidth in the high-frequency domain. In **Figure 12** a tolerance scheme for a generic filter is shown.

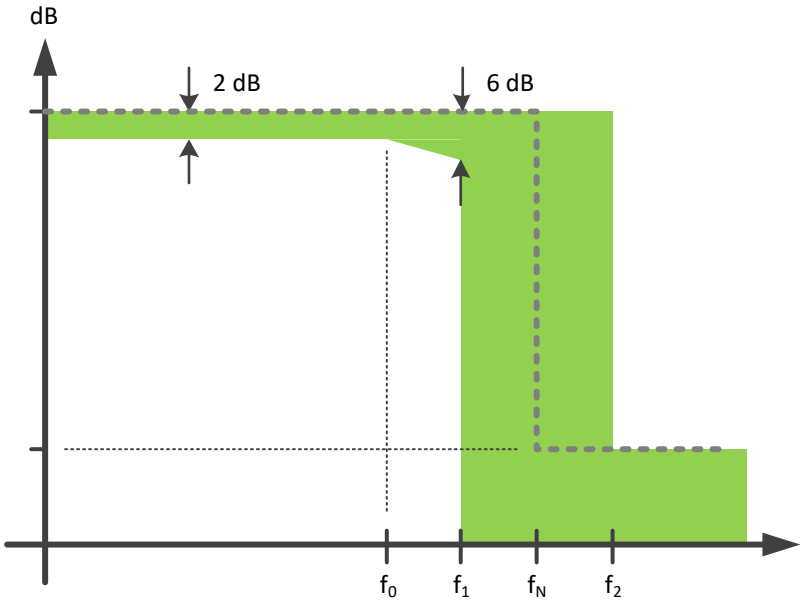

Figure 12: Generic frequency response for an anti-aliasing or anti-imaging filter

Depending on the sample rate f_s of the signal processing in Cerence SSE, the following values shall be chosen:

SSE Processing Sample Rate f_s	SSE Audio Bandwidth	f_0	f_1	$f_N = f_s / 2$	f_2
8,000 Hz	3,400 Hz	2,800 Hz	3,400 Hz	4,000 Hz	4,600 Hz
16,000 Hz	7,000 Hz	6,000 Hz	7,000 Hz	8,000 Hz	9,000 Hz
24,000 Hz	10,500 Hz	9,000 Hz	10,500 Hz	12,000 Hz	13,500 Hz

The frequency range [0 Hz; f_1] is called “pass band”, the frequency range [f_1; f_2] is the “transition band” and the frequencies >f_2 belong to the “stop band”.

The attenuation (relative to 1 kHz) in the stop band shall be:

- -50 dB for down-sampling (anti-aliasing filter) for proper AEC performance
- -70 dB for up-sampling (anti-imaging filter) to avoid audible artefacts

Besides this, the filter shall have a smooth frequency response without ripples (see **Figure 13**):

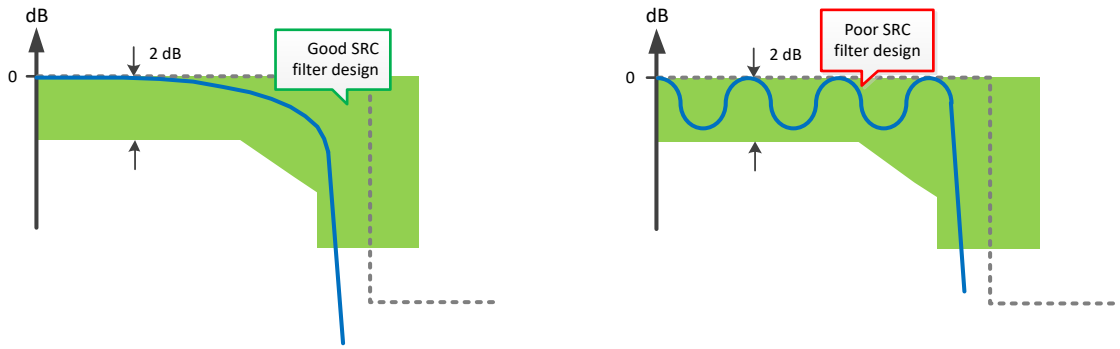

Figure 13: The filter shall have a smooth frequency response without ripples

4 Use-Case Specific Requirements

4.1 Telephony according ITU-T P.1100 / ITU-T P.1110

For telephony applications, it is recommended that the complete system fulfills ITU-T specifications P.1100 [2] (narrow-band) and P.1110 [3] (wide-band). These specifications describe a complete hands-free system and make implicit requirements for the whole system. These requirements are described in more detail in ITU-T P.1130(06/2015) [4], so this specification should be considered in the design of the complete audio architecture as well.

It is important that the audio paths are designed for “low latency” requirements, because the ITU-T specifications [2][3] have strong requirements on the overall latency in both sending and receiving direction.

General guidance on this can be found in ITU-T P.1130(06/2015)_[4], sections 9.1-9.4, 10.2, 10.3.

4.2 Apple CarPlay

The audio specification for the use-cases Telephony, FaceTime and Siri is based on ITU-T P.1100 [2] and P.1110 [3]. Therefore, these specifications (and corresponding ITU-T P.1130(06/2015)_[4]) shall be considered for the audio architecture here as well.

It is important that the audio paths are designed for “low latency” requirements, because the CarPlay specifications have strong requirements on the overall latency in both sending and receiving direction.

According Apple’s “Accessory Interface Specification” [[5]], section “26.2.8.2.2 Main Audio - Telephony”, the signal processing module (here: Cerence SSE) has to remove the DTMF tones, which are played through the cabin speakers, from the uplink signal. This requires that the DTMF tones must be included in the reference channel and that they are played with the same center of sound as the telephony downlink (see also section 3.5.4.2). Please check whether the DTMF tones are part of the telephony downlink or whether they are played back over a separate audio channel. Besides all other specifications regarding CarPlay, especially the “Accessory Interface Specification” [[5]] describes requirements for the audio system. Please be aware that the Accessory Interface Specification is updated frequently by Apple.

4.3 Android Auto

Regarding hands-free audio, Android Auto refers to the ITU-T specifications P.1100_[2] and P.1110 [3]. Therefore, these specifications (and corresponding ITU-T P.1130(06/2015 [4])) shall be considered for the audio architecture here as well.

Please check the relevant specifications from Google for further requirements (especially regarding delays).

4.4 eCall according GOST 33468

According GOST 33468, two methods of volume control in receiving direction are allowed:

- Manual volume control: The user can control the volume in a certain range manually.
- Automatic volume control: The systems controls the volume in receiving direction automatically based on the background noise inside the car, the user has no impact on the volume setting.

It is strongly recommended by Cerence to design the eCall device for automatic volume control and to not provide manual volume control for this use-case.

Please check the GOST 3368 specification for further requirements (e.g. delays and audio levels). **All** requirements of GOST 33468 must be fulfilled to get the authorization to sell the car in Russia.