General Audio Transfer
Requirements
CONFIDENTIAL

General Audio Transfer Requirements

Contents

1 Purpose of this document

This document should give a short overview about some building blocks present in an audio path and the temporal behavior of an audio device. It helps to understand the fundaments of audio data transfer in a better way and the requirements which result from this.
The actual requirements are not listed in this general document but are enumerated in separate requirements document related to the different Cerence products.

Please read carefully section 1.1 due to important definitions and terms which are necessary to completely understand following sections of this document.

1.1 Terms and Abbreviations

Term / Abbreviations	Description
audio sample	A numeric quantity expressing the audio value for a single channel at a point in time
bits per audio sample	Number of bits of information per audio sample. Usually this is 16. See also abbreviation *F* and F_{Byte}
bit depth	Identical to the term bits per audio sample
audio frame	A set of audio samples, one per channel, at a point in time. See also abbreviation *CH* and *N*.
audio buffer	Space to hold the *complete* audio data per one invocation of the audio device.
audio buffer size	Needed amount of samples to hold one audio buffer. See also abbreviation *F* and F_{Byte}
SR	Audio sample rate for which the audio device generates/demands audio samples. Occasionally the term audio frame rate is also used.
CH	Number of audio channels for which the audio device generates/demands audio samples
N	Number of audio sample **per** channel the audio device generates/demands per one invocation. Also, the term frame count is used here.
NBAS	Number of bytes per audio sample (1 byte corresponds to 8 bits). Usually this is 2 (NBAS = 2). See also abbreviation F_{Byte}.
F	Number of overall audio samples the audio device generates/demands per one invocation. $F = CH \cdot N$ Please note that the base unit for the term *F* is number of audio samples. See also the abbreviation “audio buffer size”.
F_{Byte}	The audio buffer size measured in quantities of bytes. Usually 1 Byte corresponds to 8 bits (Only on some architecture like e.g. some DSPs this might be different) $F_{Byte} = NBAS \cdot F = NBAS \cdot CH \cdot N$
T_S	Time interval between successive captured or played-back audio samples $T_S = \frac{1}{SR}$

T_P	Period time in seconds for one audio buffer of audio date $$T_P = \frac{N}{SR}$$
i	Discrete time index for succeeding audio buffers the audio device is generating or demanding
k	Discrete time index for succeeding audio samples within the same audio channel
m	Discrete time index for succeeding audio frames within an audio buffer
ADC	Analog to digital converter. A hardware module that converts an analog signal (continuous in both time and amplitude) to a digital signal (discrete in both time and amplitude).
DAC	Digital to analog converter. A hardware module that converts a digital signal (discrete in both time and amplitude) to an analog signal (continuous in both time and amplitude).
audio device	An audio device is an abstract entity which allows a software application to request or transfer audio data.
audio input device	Audio capture devices from application point of view
audio output device	Audio playback devices from application point of view
audio driver	An audio driver is a special application which on one side provides the audio application programming interface (audio API) to the application and on the other side can communicate to the ADC and DAC hardware. The audio driver is deeply embedded in the operating system and provided by the operating system itself or by third party vendors. Simplified, an audio driver can be understood as the software implementation of an audio device.
SRC	Sample rate converter in general. Could be either an SSRC or an ASRC.
SSRC	Synchronous sample rate converter
ASRC	Asynchronous sample rate converter
SB	Start-up burst of an audio output device (playback).
K	Number of needed start-up audio blocks to reach a temporal steady-state playback audio device behavior
dB(SPL)	Sound pressure level (SPL) is a logarithmic measure of the effective sound pressure of a sound relative to the reference value 20 µPa. 20 µPa is usually considered as the threshold of human hearing at 1 kHz.
NAD	In this document the Network Access Device (NAD) is the first device in the device chain of the telephony system which transfers audio data to the mobile network provider. E.g. for a mobile phone which is connected via Bluetooth to the telephony system the NAD is the Bluetooth baseband chip located on the telephony system and not the mobile phone itself.
T_{RefMic}	Relative signal latency between the reference signal (either external or internal) compared to the microphone. This latency can be easily measured by generating a short impulsive noise at the far-end side of the telephony system and recording microphone and reference audio signal in a synchronized way.

	The relative delay T_{RefMic} is then the time delay between the occurrences of the impulsive noise in these two audio signals.

1.2 History

Rev.	Section	Modification	Date	Name
1		Initial version of a document only for the general audio requirements based on the document *Audio_Transfer_Requirement_SimpleHandsfree.pdf*	26.01.2015	Martin Roessler
2	All	Change to Cerence document template.	2019-11-08	Martin Roessler

1.3 Referenced Documents

Ref. ID	Document name	Version	File	Author

2 General description

This section should give a short overview about the temporal behavior of audio devices and general building blocks present in the audio paths.

2.1 Audio sample organization within an audio buffer

2.1.1 Interleaved audio data

For an interleaved audio buffer audio samples from all channels of one point in time are located directly beside of each other constructing one audio frame for this time instance.

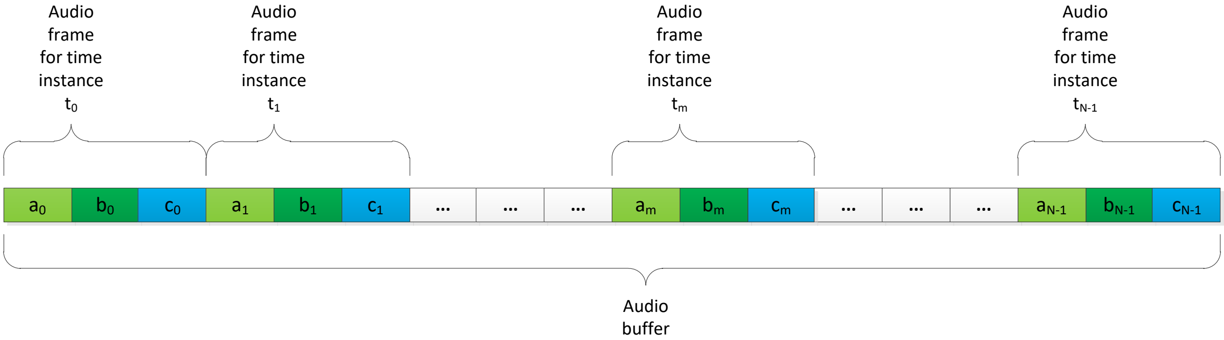
Audio frame for time instance t_0
Audio frame for time instance t_1
Audio frame for time instance t_m
Audio frame for time instance t_{N-1}
a_0 b_0 c_0 a_1 b_1 c_1 … … … a_m b_m c_m … … … a_{N-1} b_{N-1} c_{N-1}
Audio buffer

Figure 1 : Interleaved audio buffer for a 3 channel audio device (CH=3) with N audio samples per channel and the audio samples for channel one, channel two and channel three denoted as a_k, b_k and c_k.

2.1.2 Non-interleaved audio data

In a non-interleaved audio data buffer, all audio samples corresponding to one channel of the audio device are located beside of each other. Here the concept of an audio frame is not any more meaningful and should be avoided to prevent confusion.

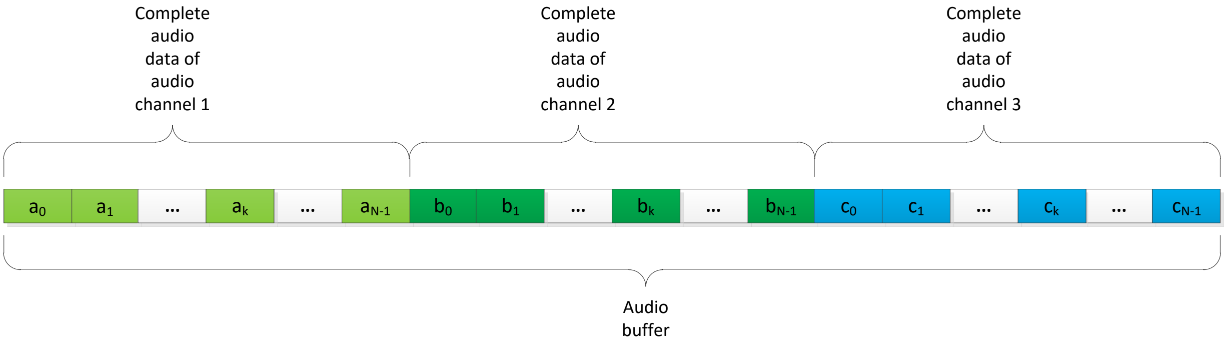
Complete audio data of audio channel 1
Complete audio data of audio channel 2
Complete audio data of audio channel 3
a_0 a_1 … a_k … a_{N-1} b_0 b_1 … b_k … b_{N-1} c_0 c_1 … c_k … c_{N-1}
Audio buffer

Figure 2 : Non-interleaved audio buffer for a 3 channel audio device (CH=3) with N audio samples per channel and the audio samples for channel one, channel two and channel three denoted as a_k, b_k and c_k.

2.1.3 Audio buffer with only one channel

If an audio buffer contains only one audio channels (CH=1) both the interleaved and the non-interleaved audio data organization converges into the identical audio sample grouping of the audio buffer.

2.2 Sample rate

2.2.1 Audio devices with identical sample rate

The sample rate of two audio devices SR_1 and SR_2 are only identical, if and only if both are sharing the same clock source or clock domain for the ADC or the DAC.

$$SR_2 = SR_1$$

2.2.2 Audio devices with the same nominal sample rate

If two audio devices are not sharing the same clock domain and both devices are opened with the same nominal sample rate SR, T_{P1} and T_{P2} are not identical. Following formula is describing the mismatch of the observable sample rates of the two audio devices

$$SR_2 = [1 + \beta(t)] \cdot SR_1$$

Usually the variation of β is rather slow and depends mainly on temperature. The value of β is usually in the per mill or ppm range.

2.2.3 Synchronous sample rate converter

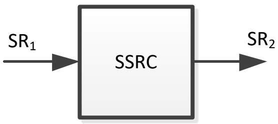

Figure 3 : SSRC block diagram

A synchronous sample rate converter (SSRC) is used to convert between audio data from one sample rate SR_1 to another sample rate SR_2 in the following way

$$SR_2 = Q \cdot SR_1$$

Please note that the factor Q is not time dependent in any way.

Dependent on the factor Q the sample rate conversion can be rather complex and so CPU-utilization costly for non-integer factors for Q (or the inverse of Q: 1/Q) like e.g. Q = 8000/44100 ≈ 0.18114. For an integer factor of Q (or the inverse of Q: 1/Q) the conversion could be implemented quite efficient like e.g. for Q = 16000/8000 = 2.

2.2.4 Asynchronous sample rate converter

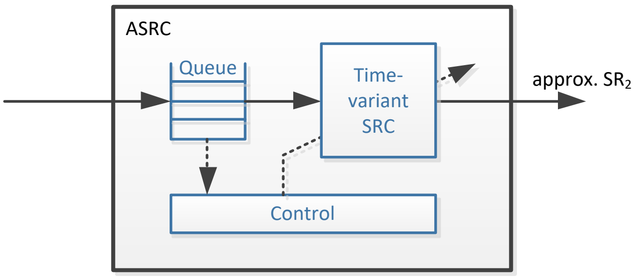

Figure 4 : Possible ASRC block diagram

An asynchronous sample rate converter (ASRC) is used to synchronize audio data coming from two audio devices not sharing the same clock domain. This is typically done by observing fill levels in the related audio queues and dependent on these fill levels a time-variable SRC is used within the ASRC module to temporal generate/demand more/less audio data to balance the mismatch $\beta(t)$ of the two clock domains (see also chapter 2.2.2).
It is very important to understand that a typical software based ASRC cannot exactly balance the mismatch between the two clock domains like e.g. a hardware phase-lock-loop circuit: It can only balance the clock mismatch of the two clock domains in a mean-sense.

2.3 Audio Sample Jitter

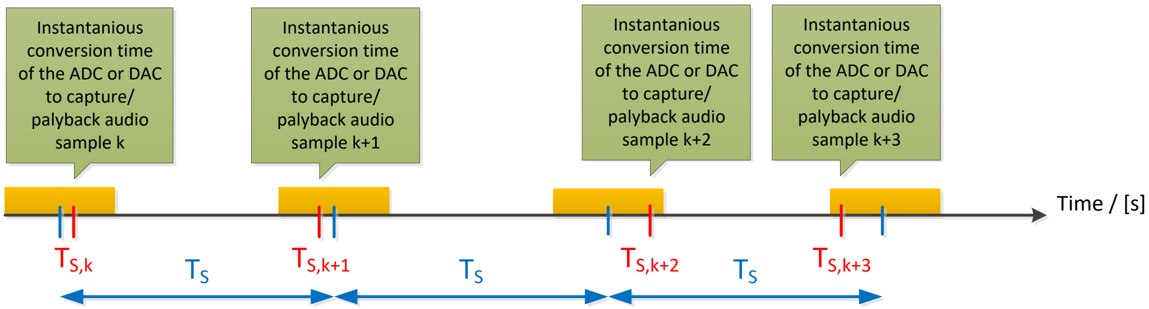
Instantanious conversion time of the ADC or DAC to capture/ palyback audio sample k
Instantanious conversion time of the ADC or DAC to capture/ palyback audio sample k+1
Instantanious conversion time of the ADC or DAC to capture/ palyback audio sample k+2
Instantanious conversion time of the ADC or DAC to capture/ palyback audio sample k+3
Time / [s]
TS,k
TS
TS,k+1
TS
TS,k+2
TS
TS,k+3

Figure 5 : Ideal (blue) and observable (red) sampling timing for ADC or DAC conversion

Ideally, each audio sample is captured/played-back in a time interval of T_S

$$T_S = \frac{1}{SR}$$

Jitter due to noise or interferences around the threshold region of the clock interface of an ADC can corrupt the ideal time interval T_S for which the analog audio signal is captured. Corresponding behavior holds also for the DAC.

A too high audio sample jitter results in a decreased signal-to-noise ratio (SNR) and increased spurious artifacts observable in the quantization noise spectra.

Please consult the clock reference design documents of your ADC and DAC codec vendor to minimize the audio sample jitter. As a rule of thumb for state-of-the-art audio ADC and DAC and properly designed clock distribution networks, audio sample jitter is usually not a common problem.

2.4 Audio Session

An audio session is defined as the time duration all involved audio devices are supplying/demanding audio data. This is actually the complete time between opening and closing all involved audio devices.
During an audio session the application is doing (but is not limited to) following steps

1. Capture audio data coming from the audio input devices
2. Process the audio input data: This means e.g. repackage and/or resample the audio data or apply algorithms of VoCon SSE to enhance the speech content of the audio data.
3. Play back the processed audio data to another audio output device or provide the audio data to the speech recognizer engine.

2.5 Capture audio devices

2.5.1 Ideal steady-state temporal behavior

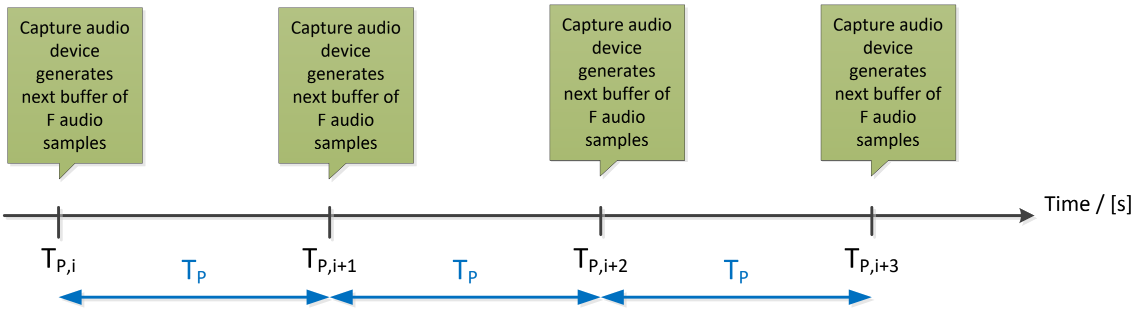
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Time / [s]
$T_{P,i}$
T_P
$T_{P,i+1}$
T_P
$T_{P,i+2}$
T_P
$T_{P,i+3}$

Figure 6 : Ideal temporal behavior of a capture audio device generating audio data

After opening a capture audio device, the device generates audio data in a well-defined temporal interval, the period time T_P. The period time can be calculated dependent on the sample rate and the number of audio frames within the audio buffer of the capture audio device it is operating on.

$$T_P = N \cdot T_S = \frac{N}{SR}$$

2.5.2 Ideal temporal startup behavior

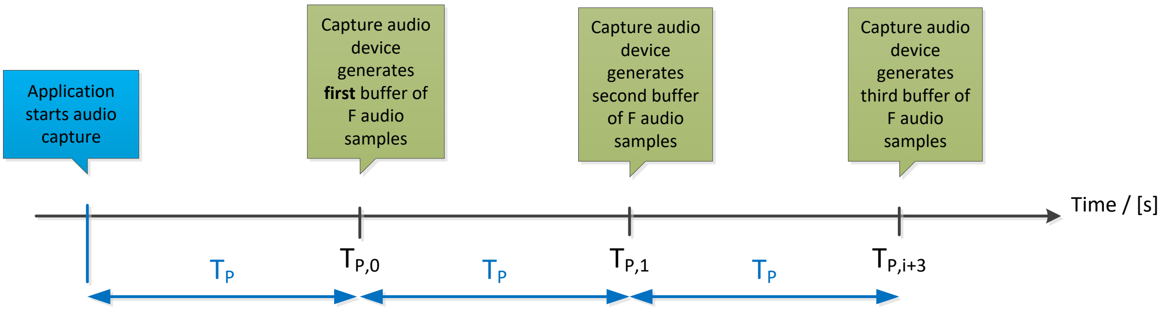
Application starts audio capture
Capture audio device generates **first** buffer of F audio samples
Capture audio device generates second buffer of F audio samples
Capture audio device generates third buffer of F audio samples
Time / [s]
T_P
$T_{P,0}$
T_P
$T_{P,1}$
T_P
$T_{P,i+3}$

Figure 7 : Ideal temporal startup behavior of a capture audio device

The ideal startup behavior of a capture audio device is identical to the steady-state behavior: The capture audio device generates the first audio block exactly after time T_P after starting the audio input device.

2.5.3 Observable steady-state temporal behavior

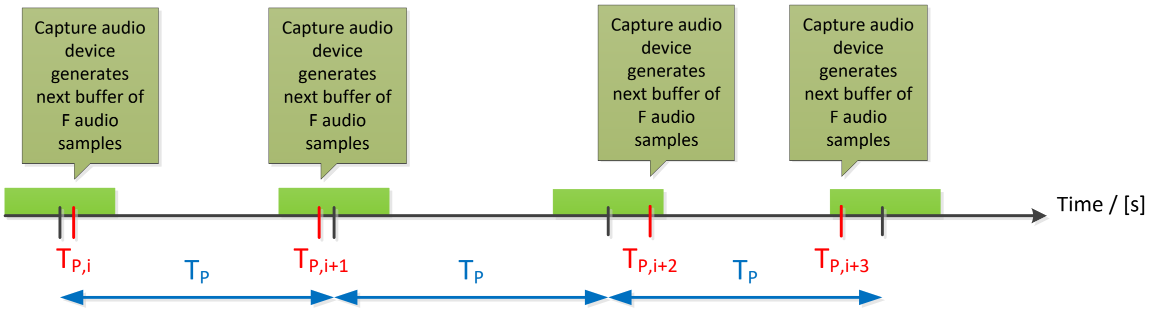
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Capture audio device generates next buffer of F audio samples
Time / [s]
$T_{P,i}$
T_P
$T_{P,i+1}$
T_P
$T_{P,i+2}$
T_P
$T_{P,i+3}$

Figure 8 : Observable temporal behavior of a capture audio device generating audio data

The difference of the observable steady-state temporal behavior compare to the ideal one is that the instant of time the audio device is generating an audio buffer (denoted as red lines on the time axis in Figure 8) is statistically jittering around the ideal time (denoted as black lines on the time axis in Figure 8).
This means that the time differences between successive audio buffers is not anymore exactly the period time T_P as defined in chapter 2.5.1. However, the mean of the time differences between generated successive audio buffers is still T_P.

2.5.4 Observable temporal startup behavior

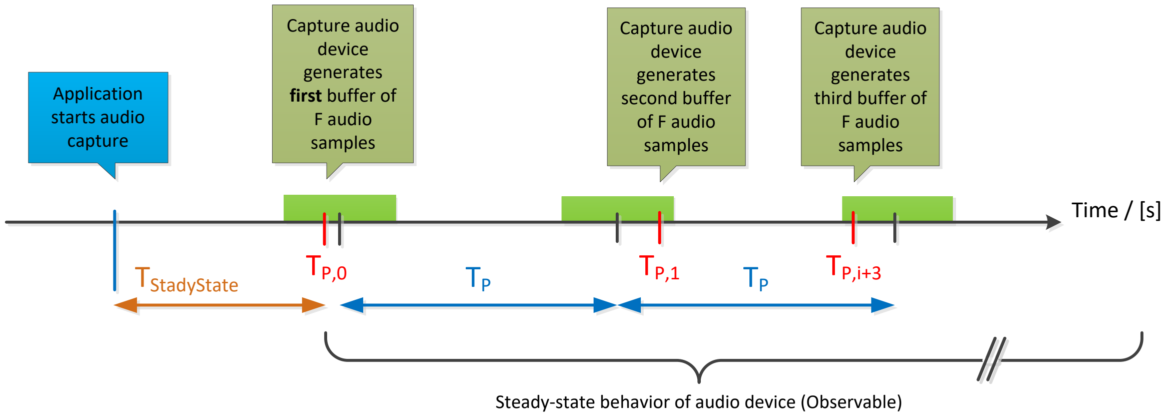
Application starts audio capture
Capture audio device generates **first** buffer of F audio samples
Capture audio device generates second buffer of F audio samples
Capture audio device generates third buffer of F audio samples
Time / [s]
$T_{StadyState}$
$T_{P,0}$
T_P
$T_{P,1}$
T_P
$T_{P,i+3}$
Steady-state behavior of audio device (Observable)

Figure 9 : Observable temporal behavior of a capture audio device generating audio data

The difference of the observable temporal startup behavior compared to the ideal one is that the instant of time the audio device is generating the first audio buffer $T_{SteadyState}$ is usually less than T_P.

2.6 Playback audio devices

2.6.1 Ideal steady-state temporal behavior

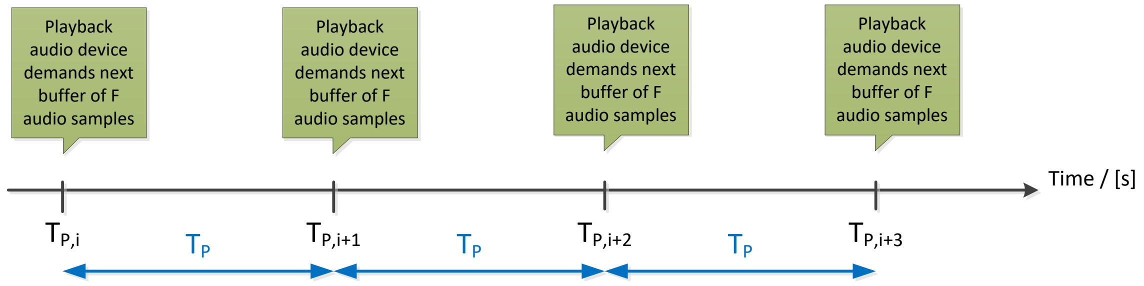
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Time / [s]
$T_{P,i}$
T_P
$T_{P,i+1}$
T_P
$T_{P,i+2}$
T_P
$T_{P,i+3}$

Figure 10 : Ideal temporal behavior of a playback audio device demanding audio data

After opening a playback audio device, the device demands audio data in a well-defined temporal interval, the period time T_P. The period size can be calculated dependent on the sample rate and the number of audio frames within the audio buffer of the playback audio device for which it was opened

$$T_P = \frac{N}{SR}$$

2.6.2 Ideal temporal startup behavior

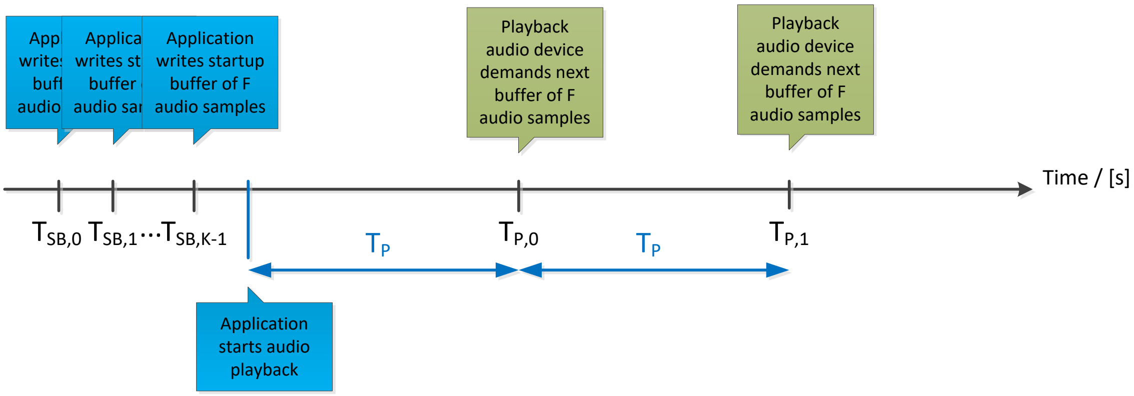
App writes buff audio
Applica writes st buffer audio sar
Application writes startup buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Time / [s]
T_SB,0 T_SB,1 … T_SB,K-1
T_P
T_P,0
T_P
T_P,1
Application starts audio playback

Figure 11: Startup behavior of a playback audio device with transition to steady-state behavior

Just after opening the playback audio device, it is needed to fill a well-defined number of audio buffers denoted by K straight away. After these K audio buffers have been written (referred to as start-up burst) and the audio playback was started, the playback audio device reaches instantaneously the steady-state temporal behavior like described in chapter 2.6.1.

2.6.3 Observable steady-state temporal behavior

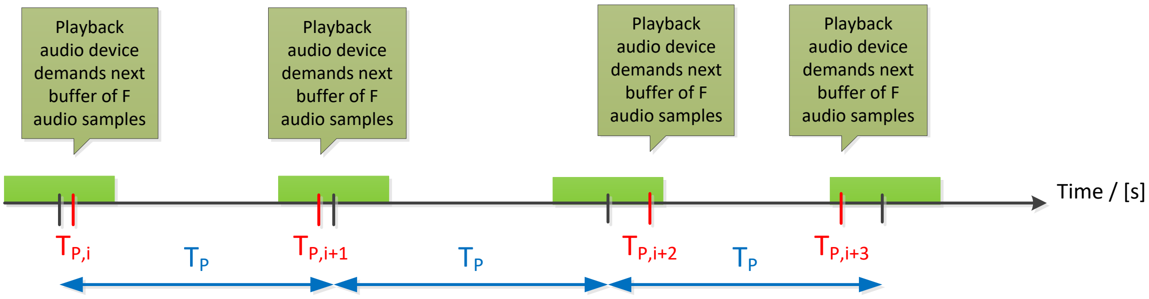
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Time / [s]
T_P,i
T_P
T_P,i+1
T_P
T_P,i+2
T_P
T_P,i+3

Figure 12 : Observable temporal behavior of a playback audio device demanding audio data

The difference of the observable steady-state temporal behavior compare to the ideal one is that the instant of time the audio device is demanding an audio block (denoted as red lines on the time axis in Figure 12) is statistically jittering around the ideal time (denoted as black lines on the time axis in Figure 12).

This means that the time differences between successive audio buffers is not anymore exactly the period time T_P as defined in chapter 2.6.1. However, the mean of the time differences between demanded successive audio blocks is still T_P.

2.6.4 Observable temporal startup behavior

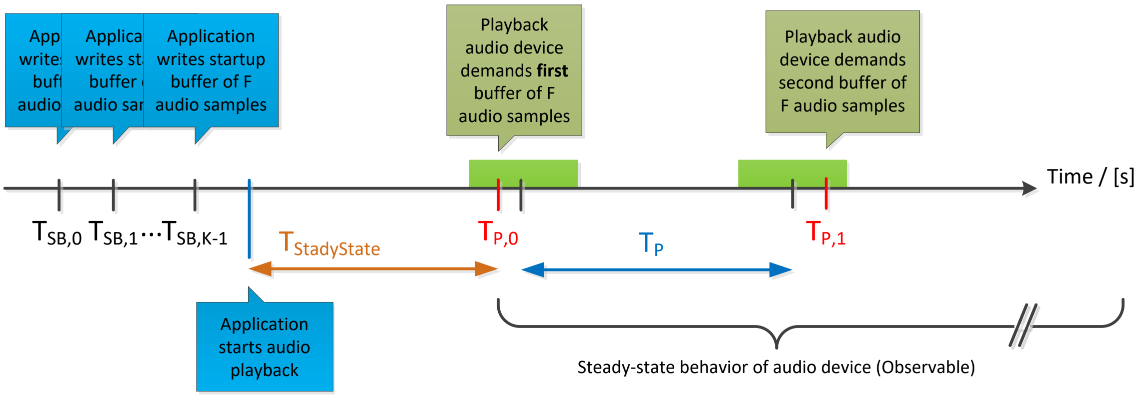
Application writes startup buffer of F audio samples
Playback audio device demands **first** buffer of F audio samples
Playback audio device demands second buffer of F audio samples
Time / [s]
$T_{SB,0}$ $T_{SB,1}$…$T_{SB,K-1}$
$T_{StadyState}$
$T_{P,0}$
T_P
$T_{P,1}$
Application starts audio playback
Steady-state behavior of audio device (Observable)

Figure 13 : Observable temporal startup behavior of a playback audio device

The difference of the observable temporal startup behavior compared to the ideal one is that the instant of time the audio device is demanding the first audio buffer $T_{SteadyState}$ is usually less than T_P.

2.6.5 Erroneous temporal startup behavior

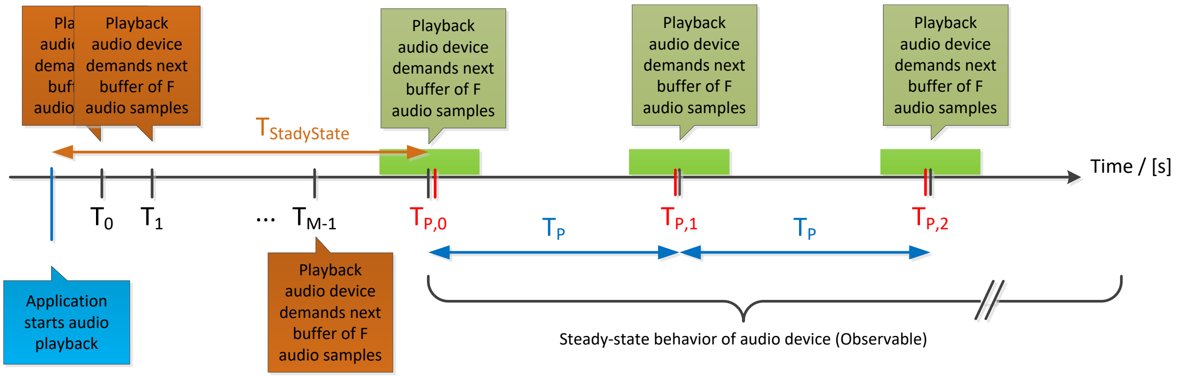
Playback audio device demands next buffer of F audio samples
$T_{StadyState}$
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Playback audio device demands next buffer of F audio samples
Time / [s]
T_0 T_1 … T_{M-1}
$T_{P,0}$
T_P
$T_{P,1}$
T_P
$T_{P,2}$
Application starts audio playback
Playback audio device demands next buffer of F audio samples
Steady-state behavior of audio device (Observable)

Figure 14 : Erroneous temporal startup behavior of a playback audio device

Unfortunately, there exists not only one erroneous temporal startup behavior because of the huge variety how an audio driver can be implemented. Instead, some general observations what could happen wrongly, are identified below:

1. The time to reach the steady-state behavior of the audio device $T_{StadyState}$ is not equal to T_P. This means that the first audio buffer the audio device is demanding from the application after starting the playback is significantly smaller or larger compared to the expected time period T_P.

2. Additional to point 1, M audio buffers are demanded by the audio device within the time the steady-state behavior of the audio device is reached. Often, the M audio buffers are requested within a few milliseconds after the application has started the playback.