Making Windows audio devices work great with Skype

Please Note!

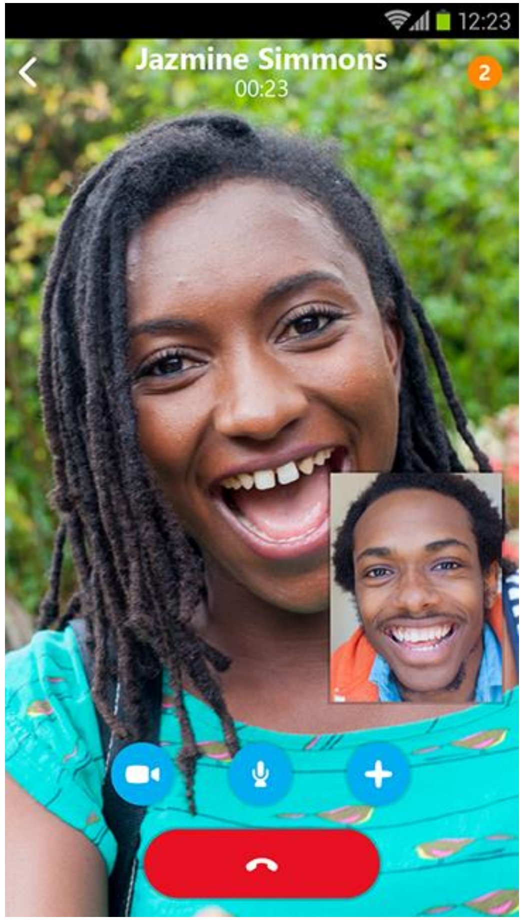
This document is provided “as-is”. Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal reference purposes.

Microsoft, Lync, and Skype are trademarks of the Microsoft group of companies. All other trademarks are property of their respective owners.

Why invest in making a device work great with Skype?

- Skype fast facts:
  - Skype users made 2 billion minutes of calls in one day
  - Skype continues to be an essential app, ranked in the top 10 most downloaded apps of all time on Windows Phone, iOS and Android At certain points in the day, there are more people using Skype from their phone or tablet than from a PC.
- Also improves non-Skype scenarios
  - Speech recognition
  - Image / video / audio capture
- Making video capture good is often just firmware tuning
- Making audio capture / render good needs minor improvements in design
- Making Skype great is typically < $2 extra COGS



Overview

Part 1: Audio test setup

A. Introduction: DUT classification
B. Environments
C. DUT test positions

Part 2: Audio specifications

A. Offloading vs. Non-offloading specification
B. Test methods

Part 3: Design guidance

A. Audio capture and renderer
B. Top failures
C. Advanced topics

References

Audio terminology

Term	Definition
dB	Decibel: A logarithmic unit that indicates the ratio of a physical quantity (usually power or intensity) relative to a specified or implied reference level (W_0) $L = 10\log_{10}\frac{W}{W_0}$ • 2X change in power is 3 dB • 4X change in power is 6 dB
dBA	Decibel A-weighted: The A-weighted sound level is the sound pressure level in dB SPL, weighted by use of metering characteristics and A-weighting specified in ANSI S1.4.
dBFS	Decibel full scale: The signal level of a digital signal relative to its overload or maximum level is given
SPL	Sound pressure level, a measure of loudness • Typical speech is 65 dBA SPL @ 0.5m • Quiet office is ~35 dBA SPL

Test setup: DUT classification

DUT classification based on acoustic UI

- **HEADSET**
  - Monaural headset
  - Binaural headset
- **HANDSET**
- **SPEAKERPHONE**
  - Handheld speakerphone – usage distance up to 0.3 m
  - Personal speakerphone* – usage distance up to 0.70 m
  - Group speakerphone – usage distance up to 1.5 m
  - Long range speakerphone – usage distance up to 5.0 m

DUT classification based on audio processing

Requirement to be tested against audio offloading specification:

DUT has acoustic echo canceler (AEC) and noise suppression (NS).

Requirement to be tested against non-offloading specification

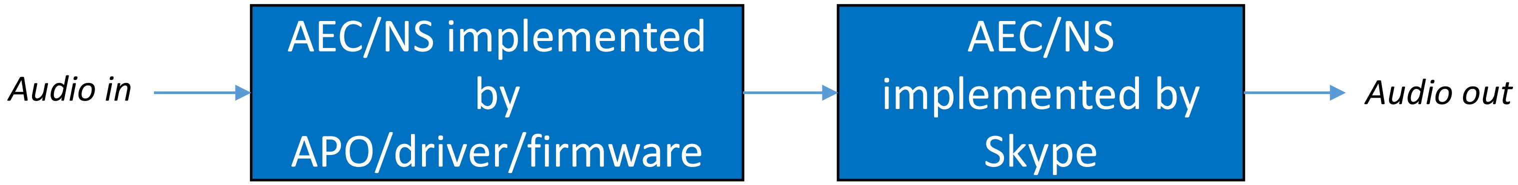
DUT provides raw capture stream that does not include any non-linear processing.

Device must not have any EFX conflicting with Skype processing.

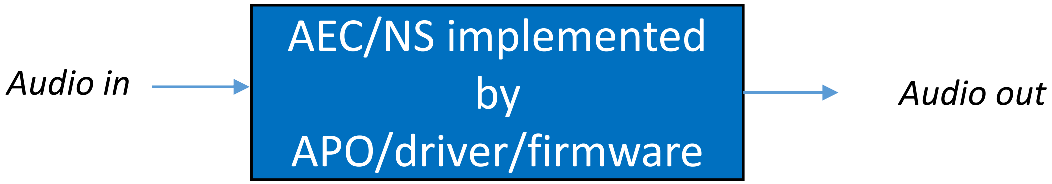
Some complex solutions enable both modes. In such cases, it is up to vendor chooses which requirements to test against.

However, if neither condition is fulfilled, then the device is not within the logo program scope.

Different processing modes



Windows 8 and earlier scenario: cascaded processing



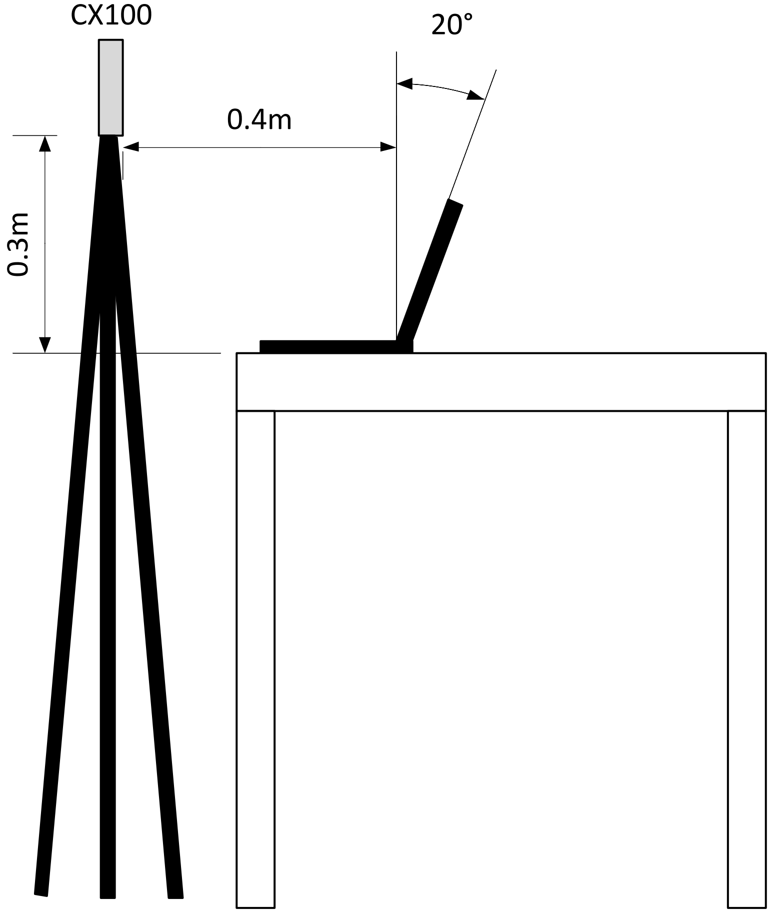
Current scenario for listed devices on Windows 8.1 Skype, only driver processing.
Future classic clients also will exhibit this behavior.

Current scenario for non-listed devices on Windows 8.1 Skype, only Skype processing. This is the default behavior for non-listed devices.
Future classic clients will also exhibit this behavior.

Test setup: anechoic and reverberant room

HCK audio capture / render test setup

- Equipment
  - Polycom CX100
  - Tripod
  - Ruler
- Test environment
  - Noise: ≤ 35 dBA SPL



Logo program and HCK audio requirements comparison

Logo test specification	Some coverage in HCK
4.1.1 Send - total quality loss	No
4.1.2 Send - end to end latency	No
4.1.3 Send - signal level with loud speech	No
4.1.4 Send - signal level with normal speech	Yes
4.1.5 Send - signal level with quiet speech	No
4.1.6 Send - idle channel SNR	No
4.1.7 Send - active channel SNR	Yes
4.1.8 Send - single frequency interference	No
4.1.9 Send - distortion and noise	No
4.1.10 Send - frequency response	No
4.1.11 Send - directivity	No

Logo test specification	Some coverage in HCK
4.2.1 Receive - output level	Yes
4.2.2 Receive - total quality loss	No
4.2.3 Receive - end to end latency	No
4.2.4 Receive - idle channel noise	No
4.2.5 Receive - single frequency interference	No
4.2.6 Receive - distortion and noise	No
4.2.7 Receive - frequency response	No
4.2.8 Receive - maximum output level	No
4.3. Echo path tests	No
(4.4 Acoustic echo canceler performance tests)	No
(4.4 3QUEST tests)	No

There are very few tests directly mapping between HCK and Logo program. The HCK cover more device level items and it is entry criteria for Logo Program audio tests.

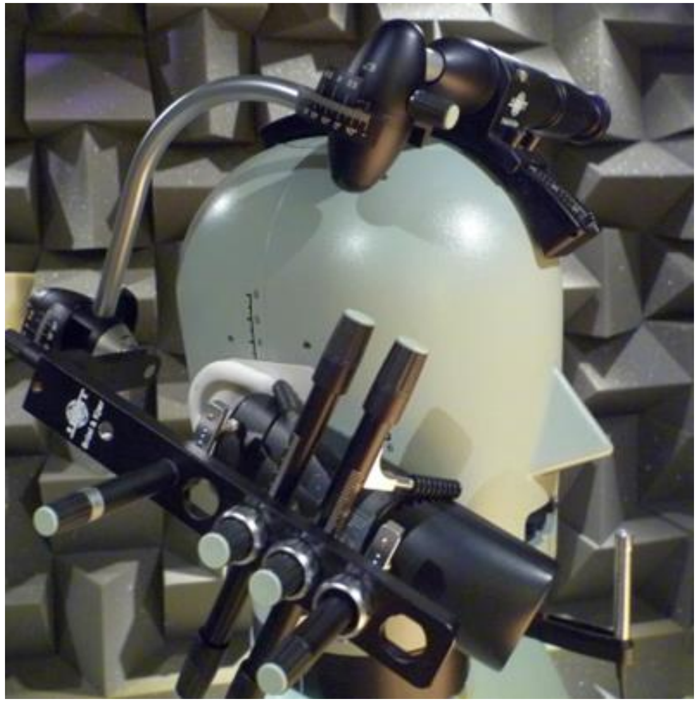
Windows HCK audio capture / render requirements

Test	Criteria	Impact if failure
Speech-to-noise ratio	≥ 30 dB	Speech is less intelligible
Send level with 65 dBA speech source at 0.5m	[-50,-4] dBFS	Near end will be hard to hear
Coupling clipping (10 ms frames)	≤ 2	Can cause echo and distortion
Receive level at 0.5m with -24 dBFS speech	≥ 65 dBA	Far end will be hard to hear
Max echo leak using Windows AEC	≤ 4%	Echo will be very noticeable
Unreported timestamp latency	[0,40] ms	Can cause echo and distortion
Microphone and loopback timestamp error	<0.5 ms	Can cause echo and distortion
Microphone and loopback timestamp glitches	<3 per minute	Can cause echo and distortion
Microphone and loopback timestamp drift	<0.08%	Can cause echo and distortion
Mouth to ear latency	≤ 100 ms	Makes conversations more difficult

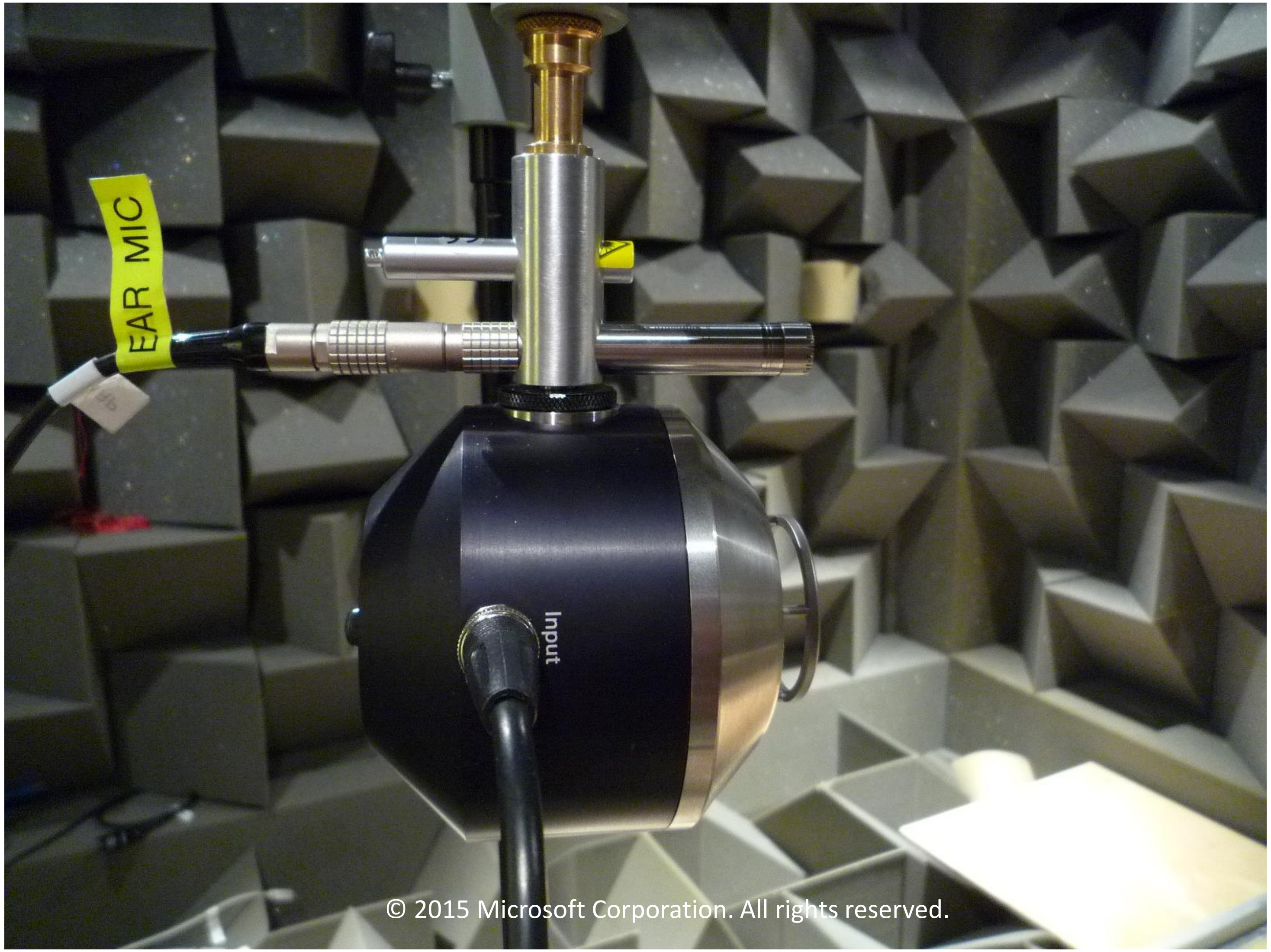
Anechoic room

The anechoic chamber or semi-anechoic chamber used should fulfill the acoustical requirements for high-quality wideband VoIP and telecom product testing. Please refer to ITU-T P.341 for the recommended parameters for the test room.

Headsets and handsets are tested with HATS in an anechoic room



All speakerphones are tested with an artificial mouth and free a field microphone in the anechoic room

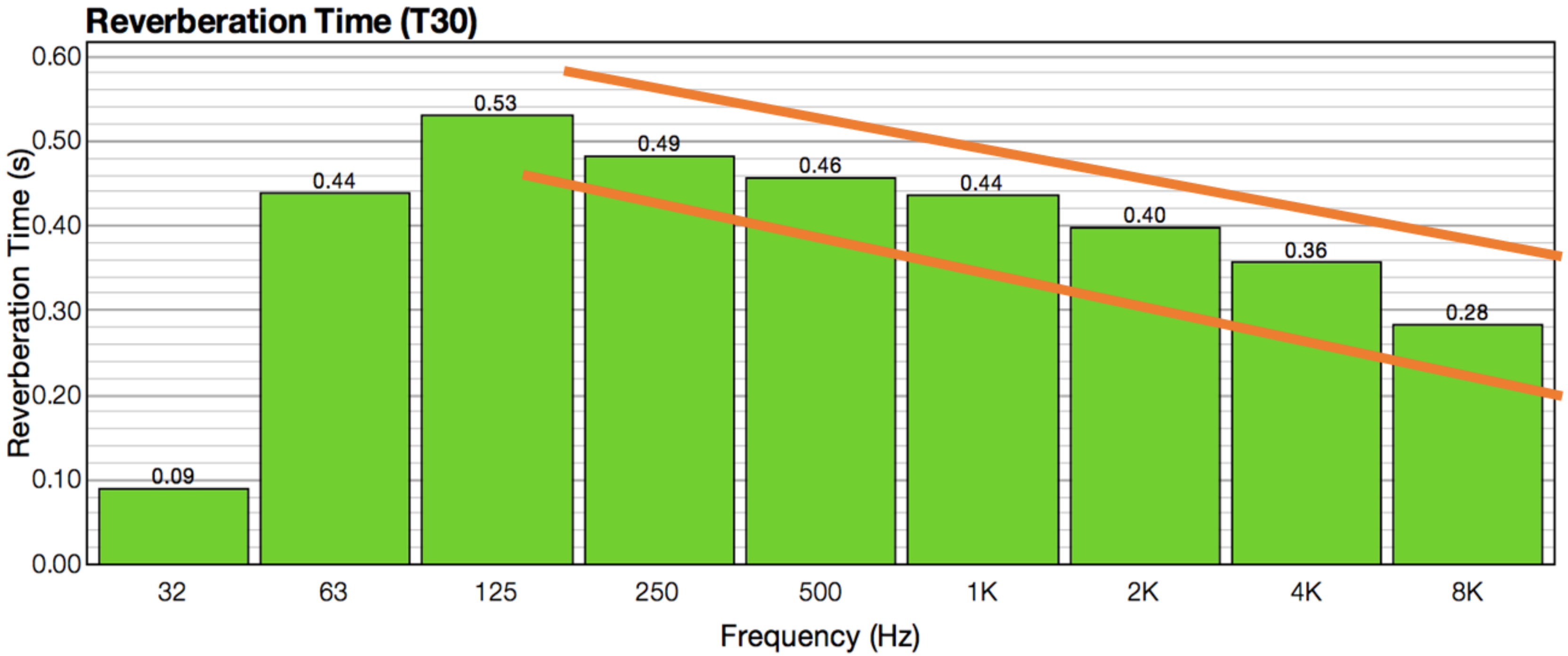


ETSI room

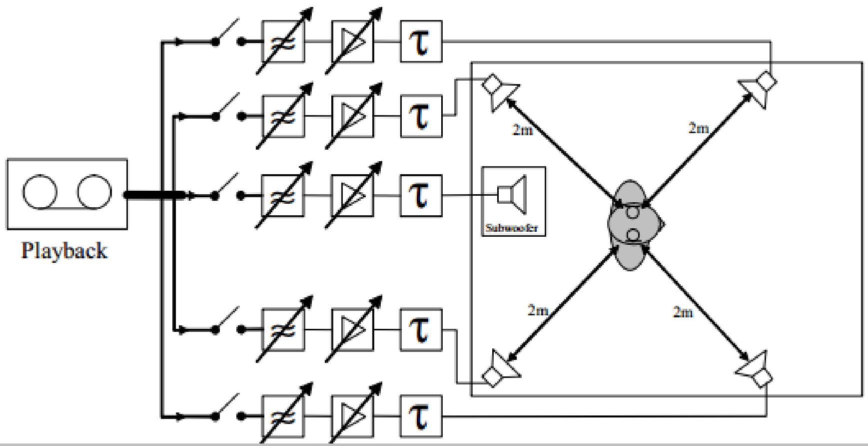
Room size – the room size should be in range between 2.7 m x 3.7 m and 3.5 m x 4.4 m. Room height 2.2 m to 3.25 m.

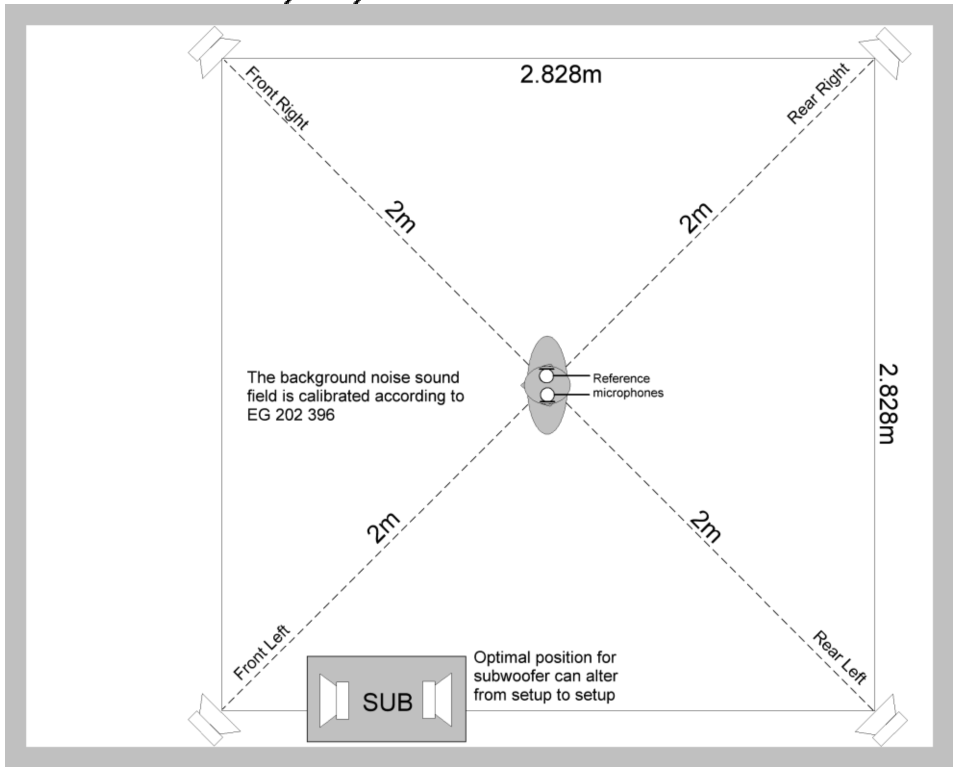
Treatment of the room – the reverberation time of the room should be less than 0.7 sec, but higher than 0.4 sec in frequency range between 100 Hz and 8 kHz. The reverb time should be declining toward high frequencies, but should not have dips or peaks in some octave bands that deviate more than 0.2 sec compared to the adjacent octave bands on either side. Such a declining trend of reverb time versus frequency represents a common meeting room or living room acoustics.

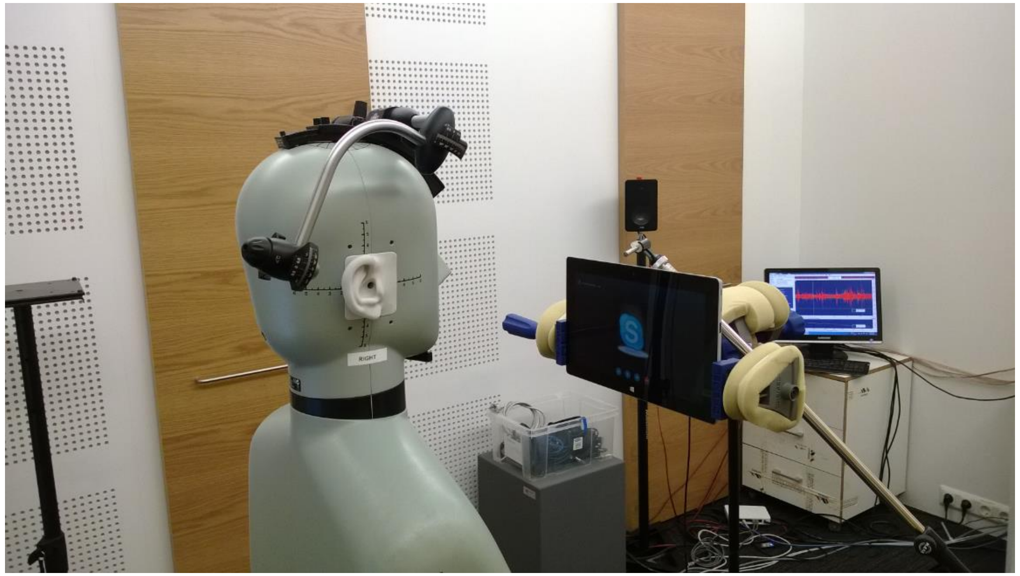
Noise floor – to avoid room noise influencing the test results the average noise floor in room should be 28 dBSPL(A) or below.



ETSI background noise test room







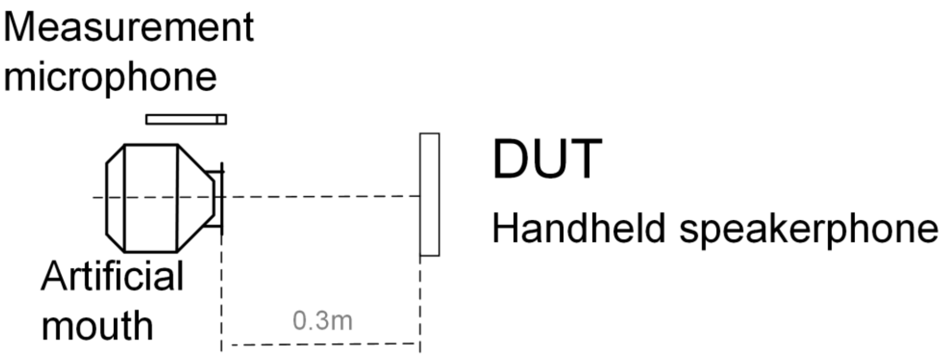
An originally binaural noise recording is played back with four loudspeakers (and possibly a subwoofer). Time delay is added to make the noise sources not phase aligned.

All speakers are equalized at HATS position for headset/handset or at DUT position for speakerphones / tablets / laptops, etc.

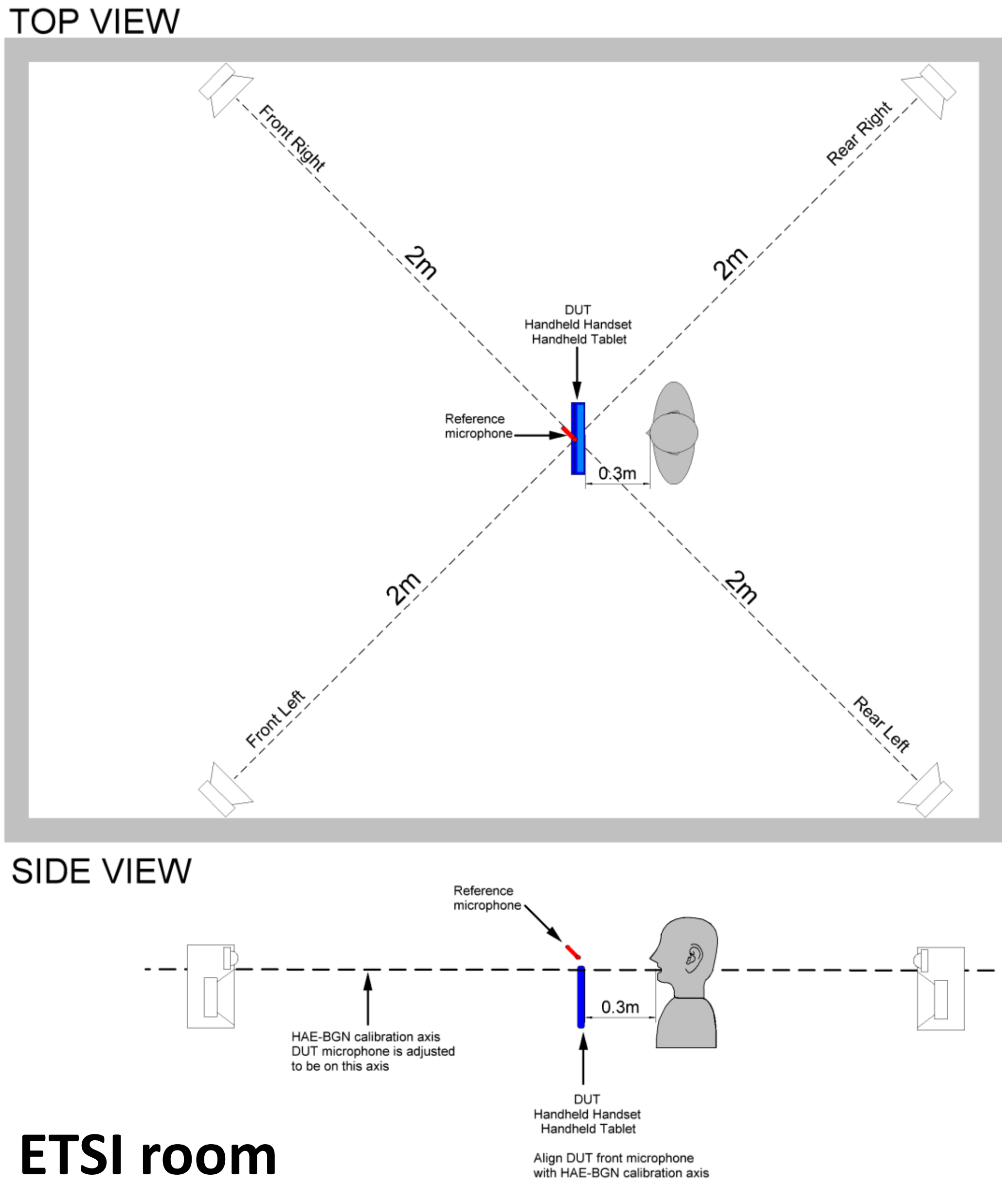
ACQUA system can start HAE-BGN background noise playback automatically during a test sequence.

Test setup: positioning speakerphones for testing

Handheld speakerphone

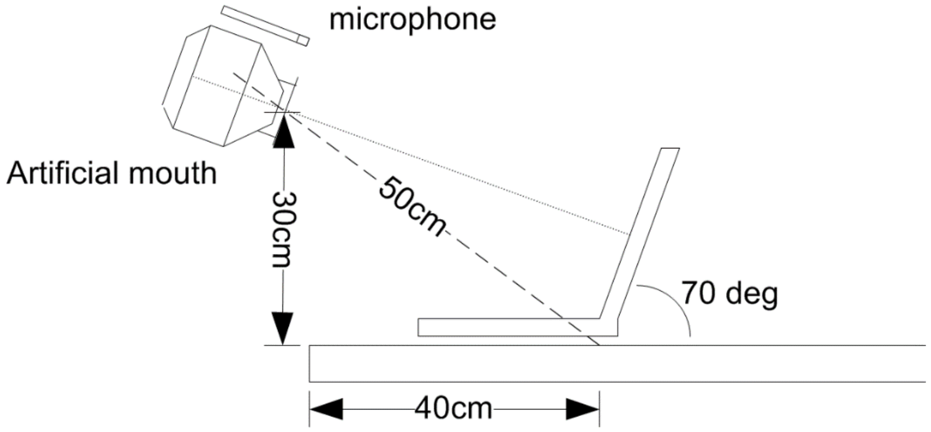


Anechoic room
(both specifications)

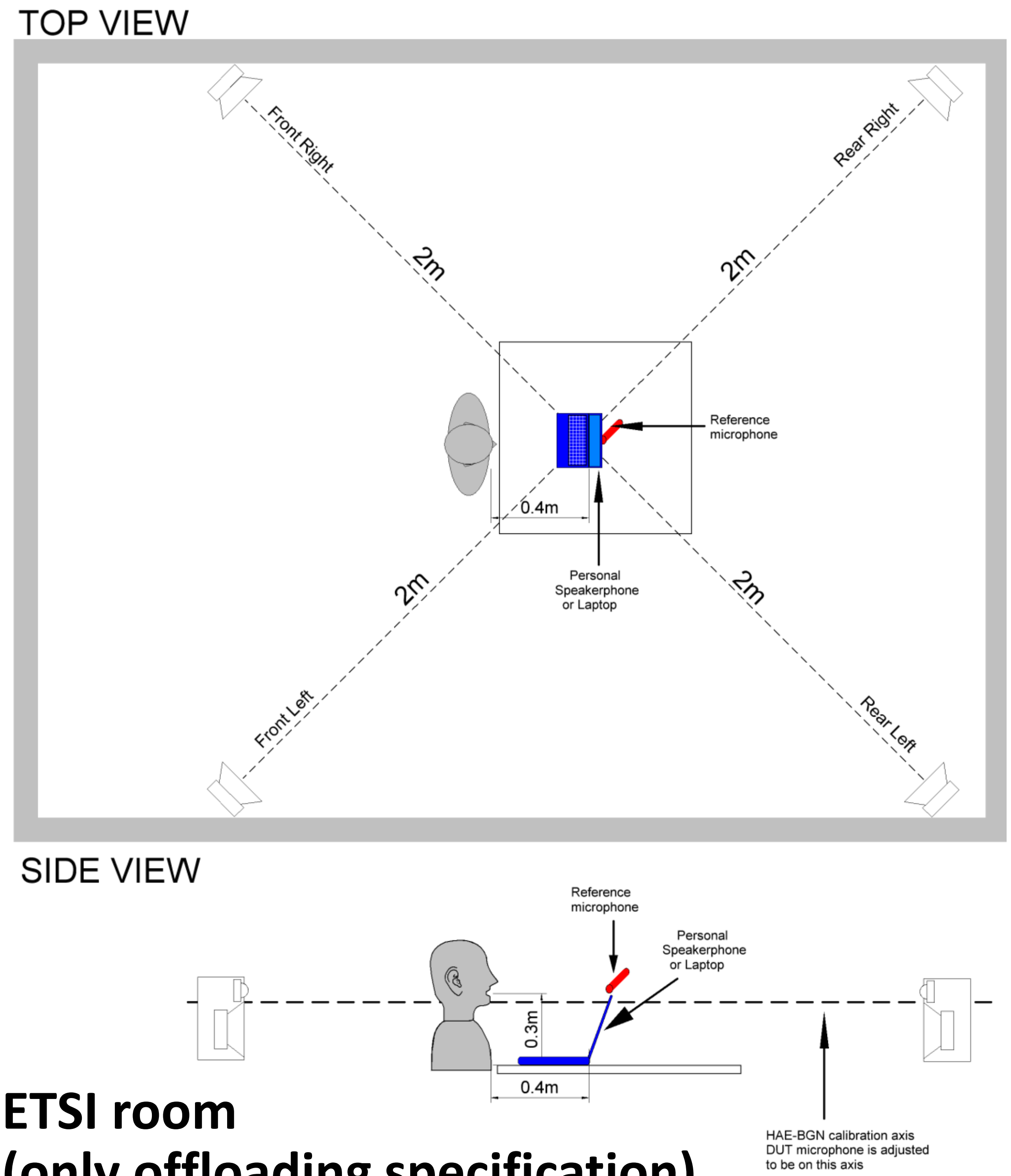


ETSI room
(only offloading specification)

Personal speakerphone

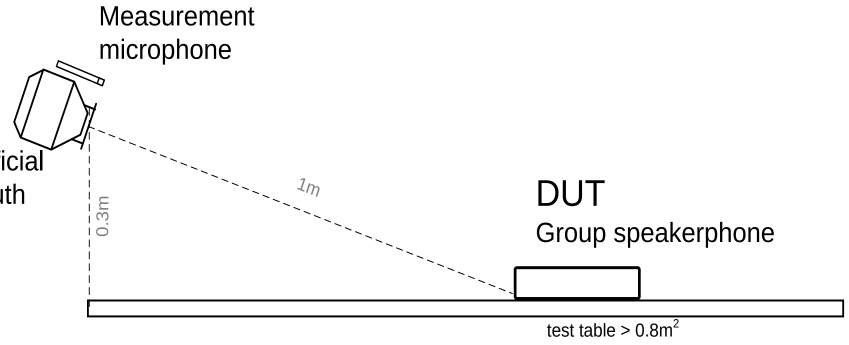


Anechoic room
(both specifications)

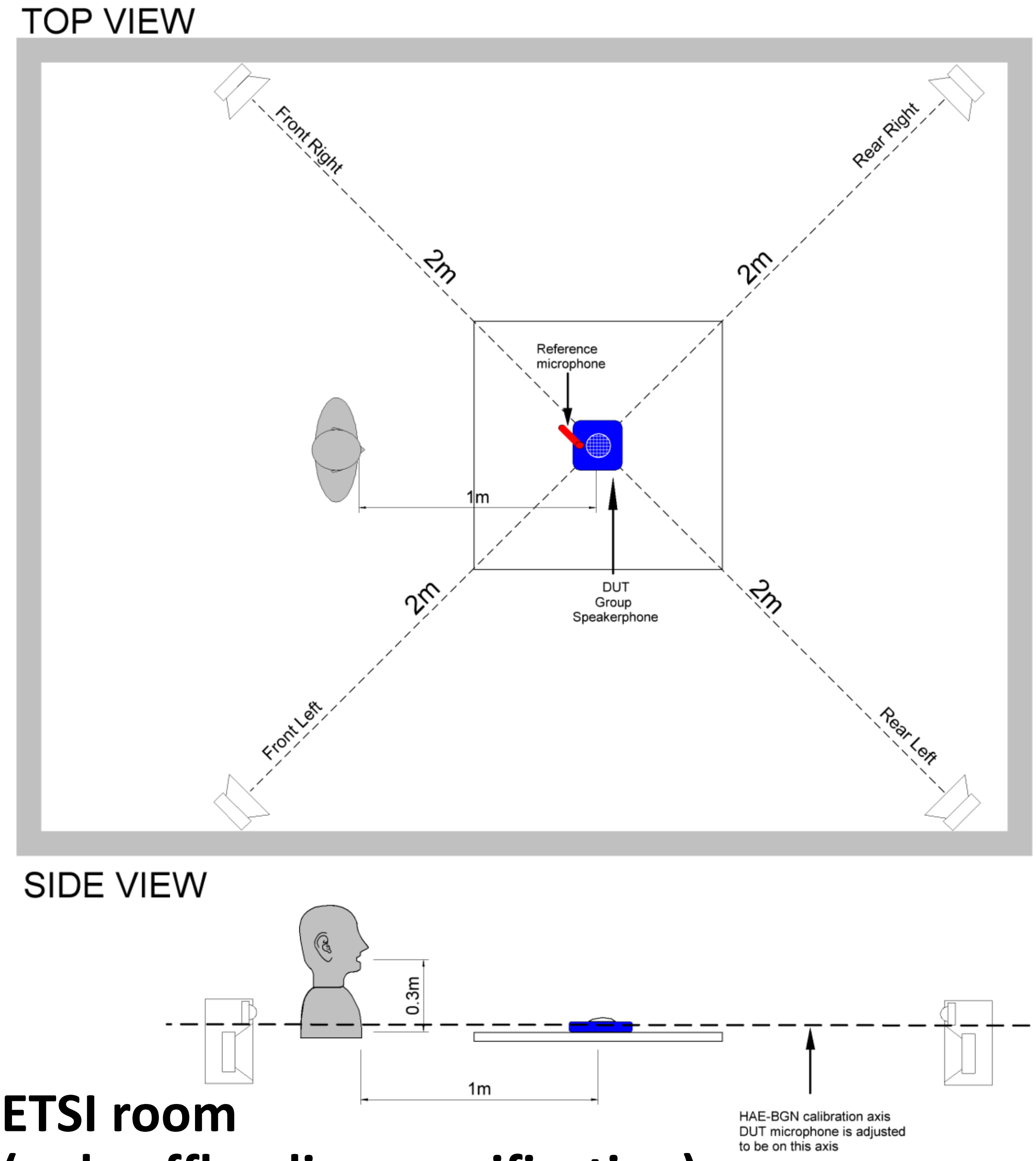


ETSI room
(only offloading specification)

Group speakerphone



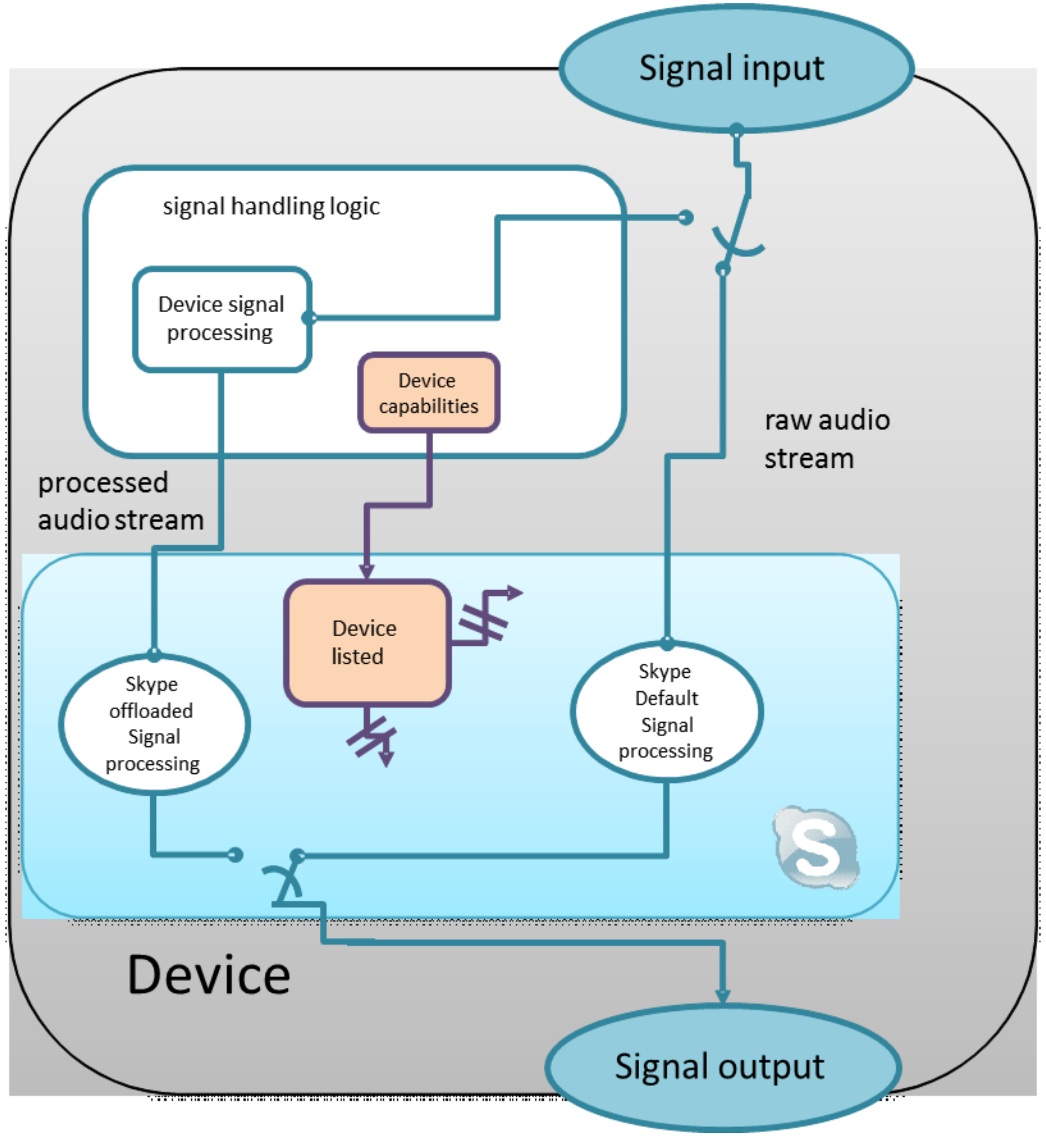
Anechoic room
(both specifications)

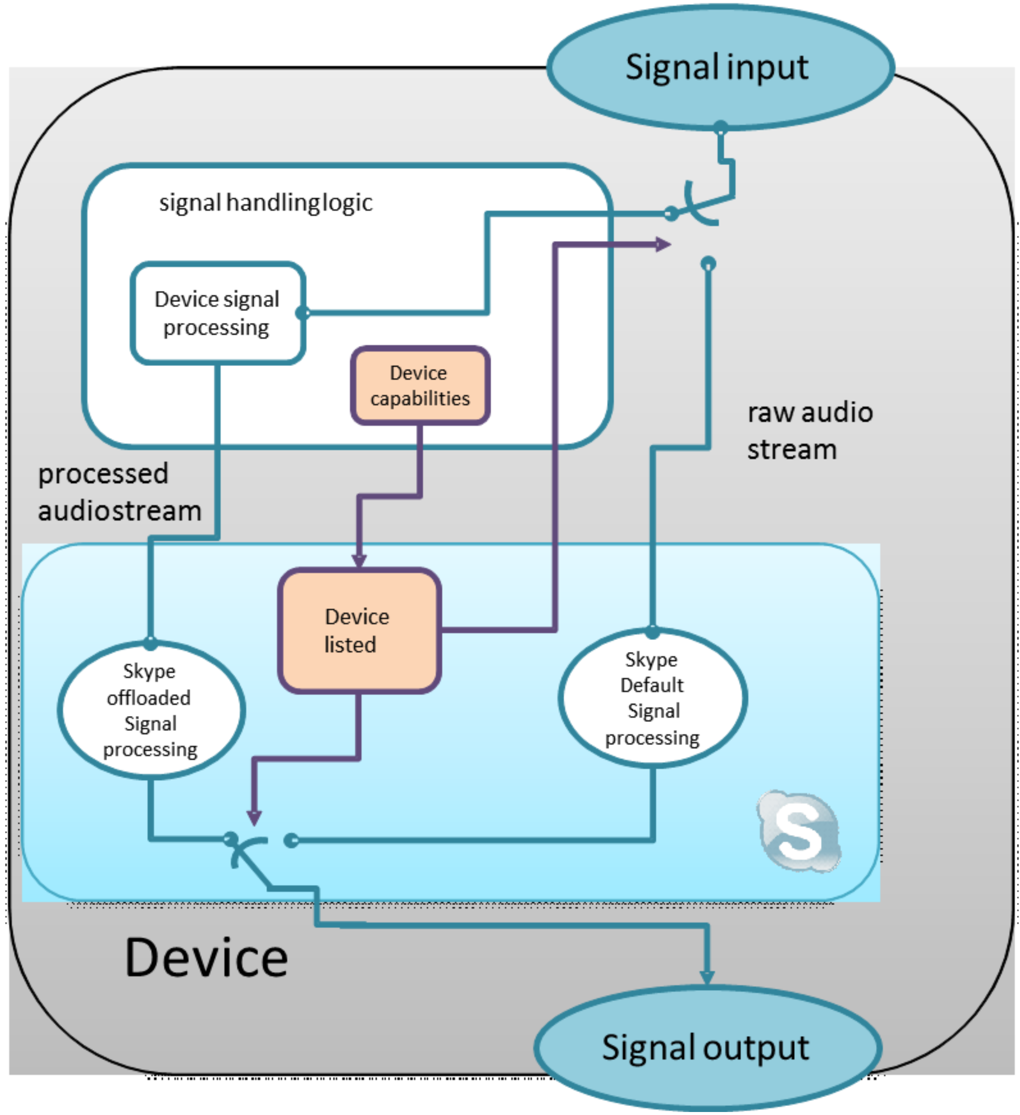


ETSI room
(only offloading specification)

Test specifications: offloading vs. non-offloading

Non-offloading vs. Offloading





Purpose: to test hardware performance

Offloading specification

- Skype pre-processors components are not duplicating the custom ones

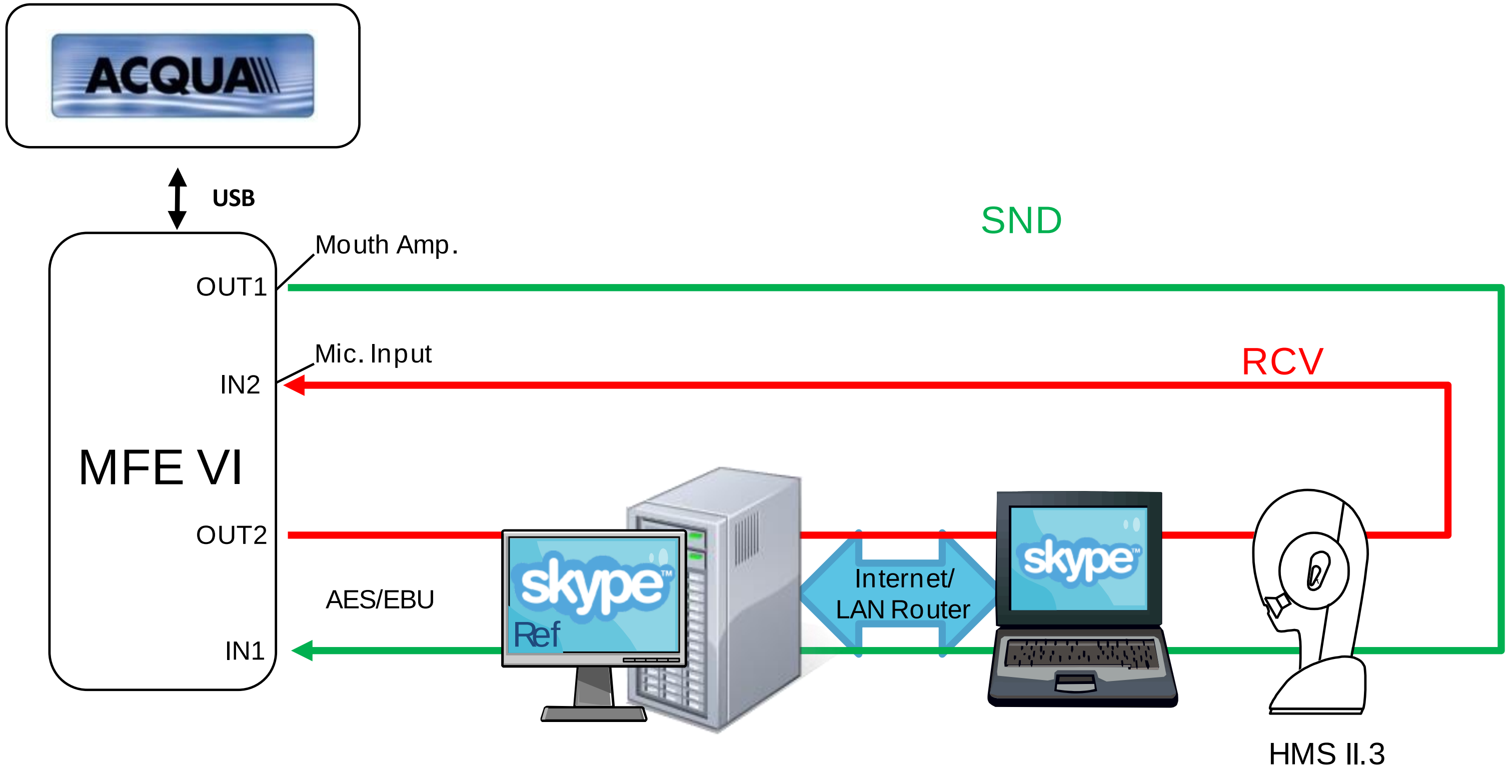
Example: if DUT has only AEC and NS, then Skype takes care of AGC.

Non-offloading specification

- Skype pre-processor is controlled to have minimal impact for the measurements

Example: AEC is disabled when measuring TCLw, but enabled when measuring EQUEST.

Testing over Skype call



Skype software settings for test

DUT Editor V.4

Enable DUT Client mode	<ForceAlgorithmControlString />
	<DisableABTesting> true</DisableABTesting>
	<DisableNS>true</DisableNS>
	<DisableCNG>true</DisableCNG>
	<DisableDigitalFarEndAGC>true</DisableDigitalFarEndAGC>
	<DisableDigitalNearEndAGC>true</DisableDigitalNearEndAGC>
Disable AGC	<DisableAGC>true</DisableAGC>
Disable AEC	<DisableAEC>true</DisableAEC>

REF Editor V.4

Enable Reference Client mode	<ForceAlgorithmControlString />
	<DisableWholeVQE>true</DisableWholeVQE>
	<DisableABTesting> true</DisableABTesting>
Capture Right Ch. (MFE VI)	<ForceInputChannel>1</ForceInputChannel>

Offloading spec:
Skype NS and AEC are OFF, AGC on/off depends on specific DUT.

Non-offloading spec:

AGC in Skype audio test specification context means the recording device input level slider.
AGC is turned on or turned off based on testcase nature.

Skype AEC is generally disabled throughout testing. The performance with Skype AEC enabled should still be verified unless the device is a full audio offloading device.

Skype DigitalAGC, noise suppression, comfort noise, and playback volume/compressor are always off during the certification testing on the DUT side.

Analog AGC – on/off logic in DEFAULT spec

4.1	**Send path tests**
4.1.1	Send path - total quality loss (Skype end to end test)
4.1.2	Send path - latency
4.1.3	Send path - signal level with loud speech
4.1.4	Send path - signal level with normal speech
4.1.5	Send path - signal level with quiet speech
4.1.6	Send path - idle channel SNR
4.1.7	Send path - active channel SpNR
4.1.8	Send path - single frequency interference
4.1.9	Send path - distortion and noise
4.1.10	Send path - frequency response
4.1.11	Send path - directivity
4.2	**Receive path**
4.2.1	Receive path - output level
4.2.2	Receive path - total quality loss (Skype end to end test)
4.2.3	Receive path - latency
4.2.4	Receive path - idle channel noise
4.2.5	Receive path - single frequency interference
4.2.6	Receive path - distortion and noise
4.2.7	Receive path - frequency response
4.2.8	Receive path – maximum output level
4.3	**Echo path**
4.3.1	Echo path - terminal coupling loss weighted (TCLw)
4.3.2	Echo path – EQUEST MOS at nominal playback volume
4.3.3	Echo path – echo control characteristics
4.3.4	Echo path – EQUEST MOS at max playback volume
4.3.5	Sidetone Masking Rating
4.3.6	Sidetone Latency

Skype adjusting AGC allowed – preceded with preparation testcase

Skype adjusting AGC allowed – preceded with preparation testcase

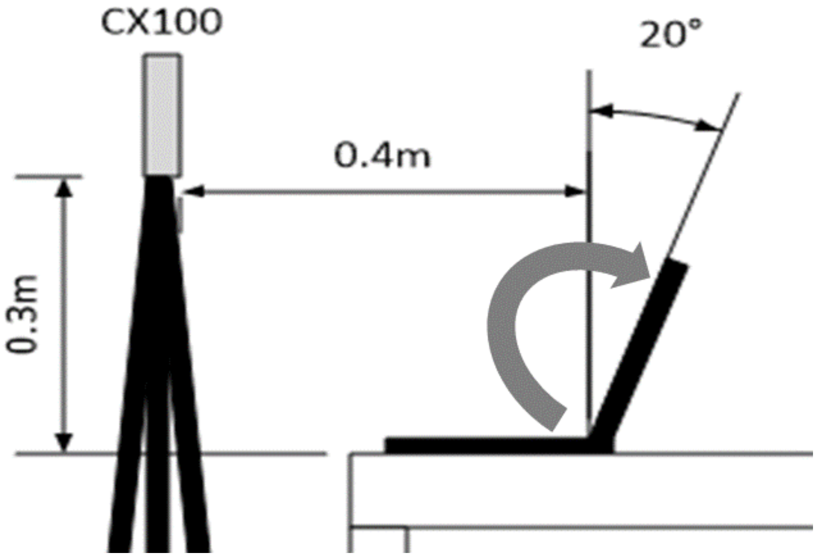
This helps to make sure AGC has enough steps to lower the input sensitivity so it would not clip with loud speech or for webcams for expected speaker echo signal

Skype adjusting AGC disabled – same AGC setting used as for normal speech.
This helps make sure these results are all in reference to signal levels with normal speech just as all the standards generally have.

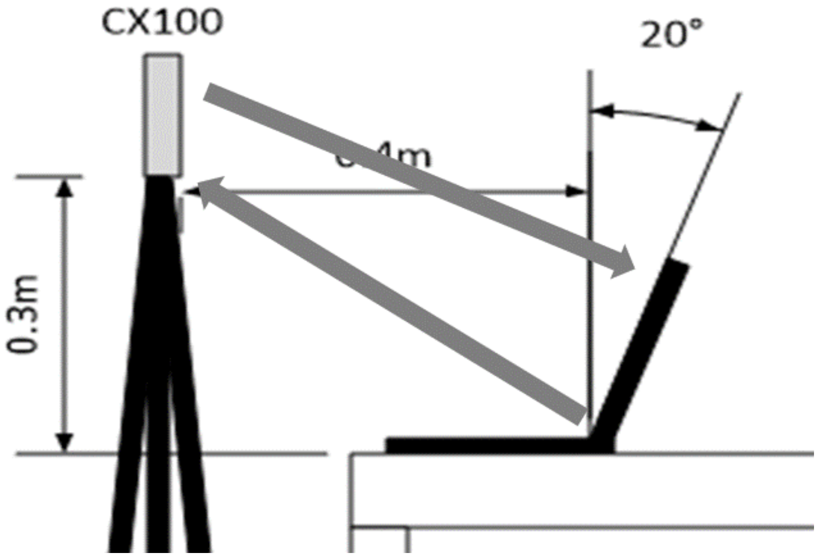
Skype adjusting AGC allowed during preparation test case, this makes sure the AGC setting is set to accommodate the loud acoustic echo signal
For the actual test cases the AGC adjustment is disabled and the fix setting used is determined during the preparation sequence

Test specifications: Test methods

Latency measurement in NON-OFFLOAD vs. OFFLOAD using Windows HCK setup



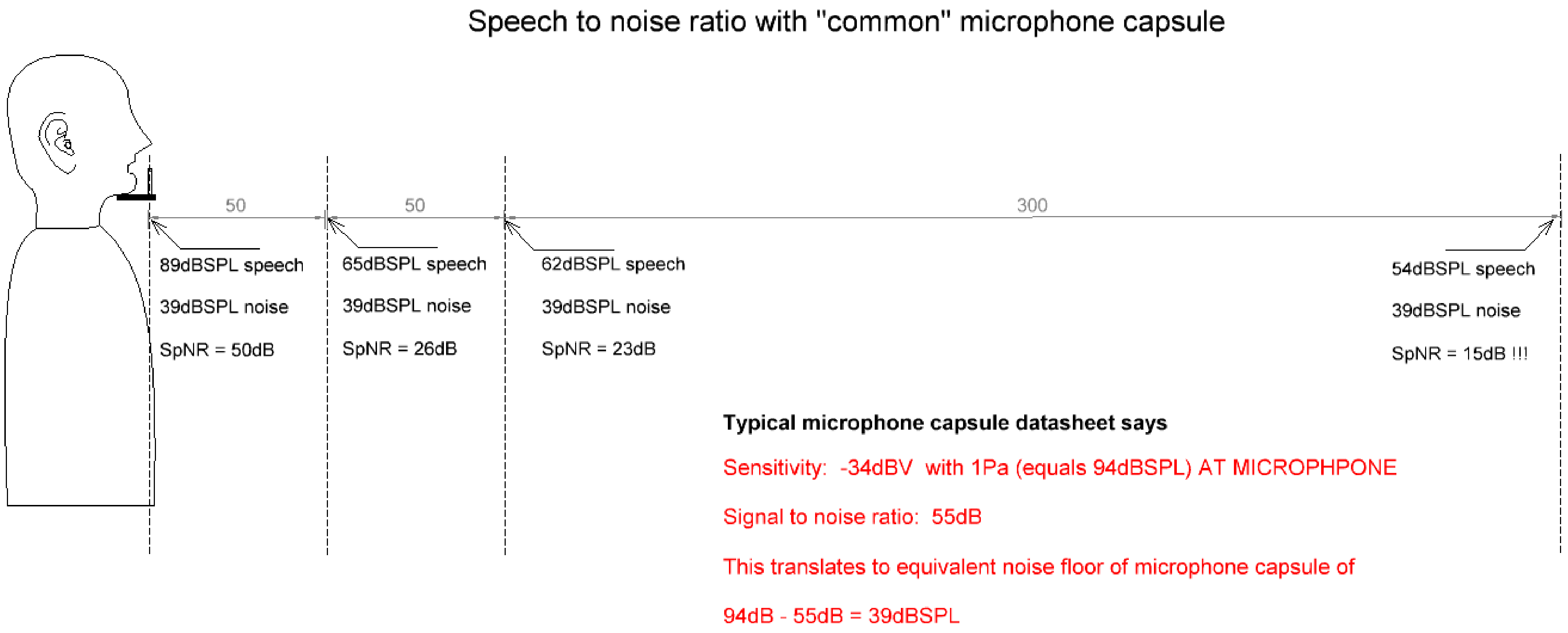
RAW – as there is no AEC, the loopback latency can be measured from DUT speaker back to the microphone



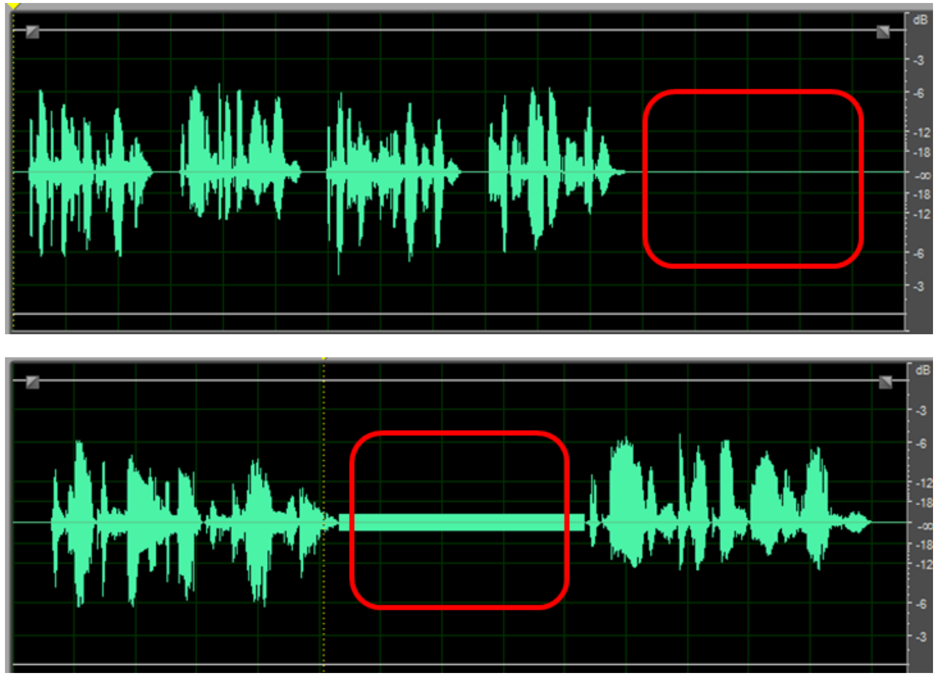
Offload – as there is an AEC, the loopback method would not work, so a path from

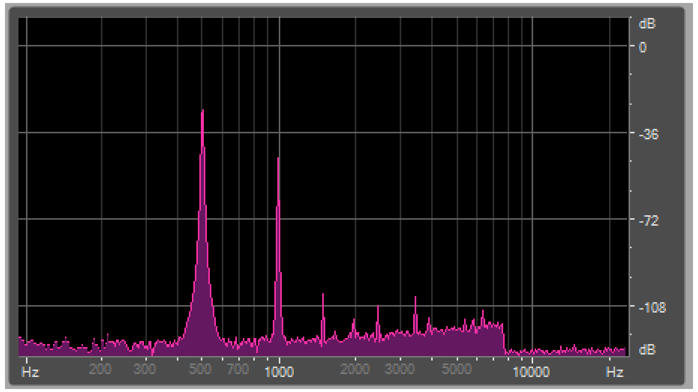
- DUT speaker to CX100 microphone is measured
- CX100 speaker to DUT microphone is measured

SNR sample with same mic, but different category

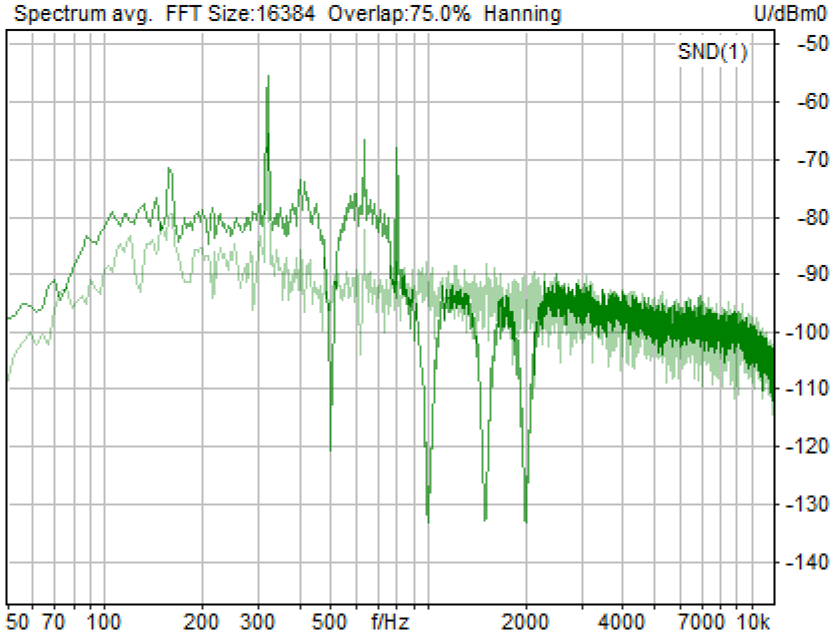


SNR and SpNR





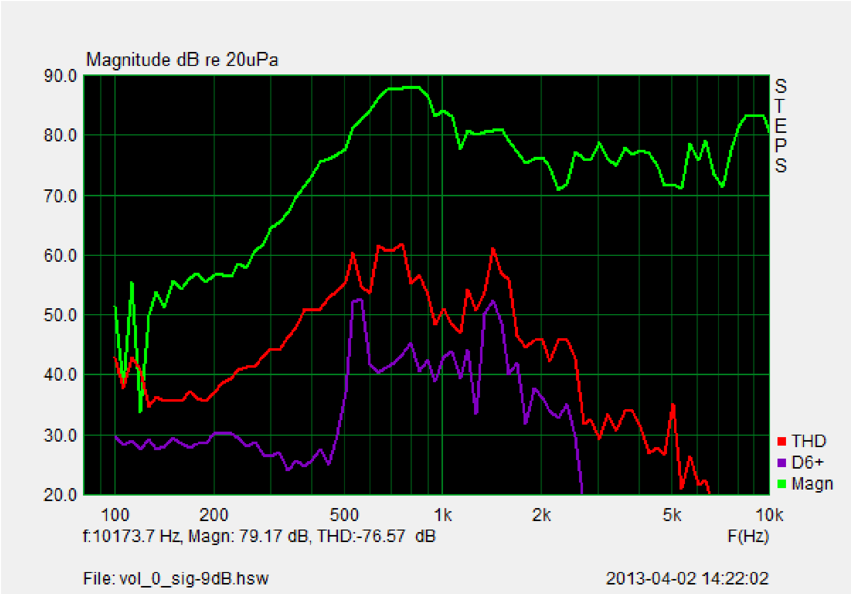
The added sines are short enough so NS won’t yet learn it as a noise



Here is a sample where
SNR = 46 dB
SpNR = 41 dB

From the resultant diagrams, it is evident that the noise between 100 Hz and 800 Hz is suppressed more during silence than it is during active channel

Distortion testing (1)



Ideally, we would like to run a sine sweep test and get a plot of not only the frequency response and overall THD, but ability to see individual 2nd, 3rd harmonic and then 6th and higher harmonics.

Those 6th and higher only occur due to case resonances, rubb and buzz, and rattle. They never occur in a speaker unit itself unless very severely overdriven.

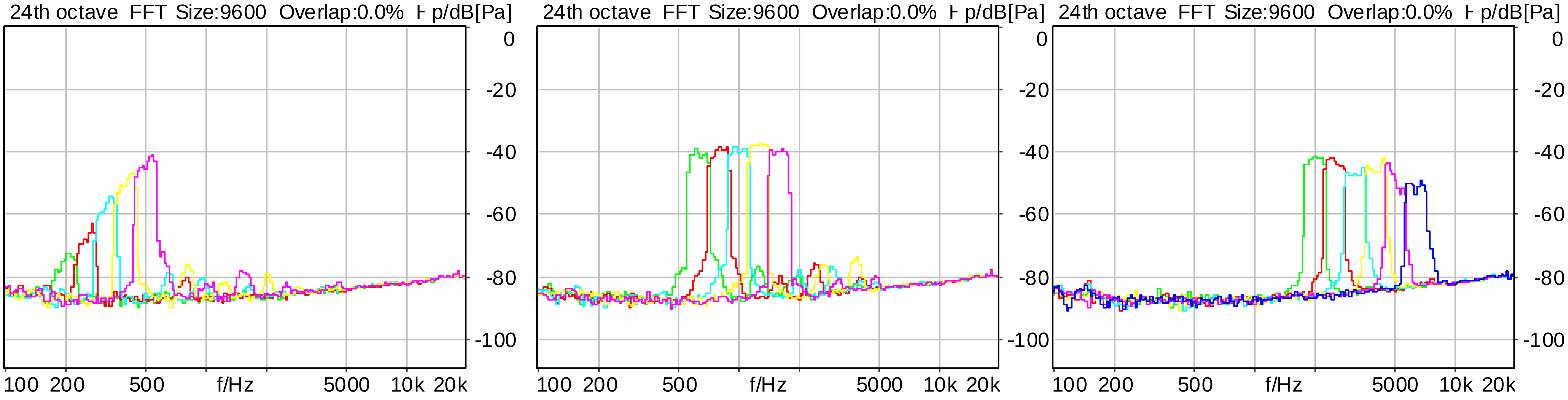
Unfortunately, today's ACQUA system does not allow this.

			SDNR (pulsed noise signal-to-distortion-and-noise ratio)			
			Standard (dB)		Premium (dB)	
		Frequency band	Level: -22dBFS	Level: -16dBFS	Level: -22dBFS	Level: -16dBFS
Handheld speakerphone	Laptop	178-224	NA	NA	NA	NA
		224-282	NA	NA	NA	NA
		282-355	NA	NA	≥24	≥24
		355-447	NA	NA	≥24	≥24
		447-562	NA	NA	≥24	≥24
		562-708	≥20	≥20	≥26	≥26
		708-891	≥22	≥22	≥26	≥26
		891-1122	≥24	≥24	≥28	≥26
		1122-1413	≥24	≥24	≥28	≥26
		1413-1778	≥26	≥24	≥28	≥26
		1778-2239	≥26	≥24	≥28	≥26
		2239-2818	≥26	≥24	≥28	≥26
		2818-3548	≥26	≥24	≥28	≥26
		3548-4467	≥26	≥24	≥26	≥26
		4467-5623	≥24	≥24	≥26	≥26
		5623-7079	≥24	≥24	≥26	≥26

The second best choice is the SDNR test based on IEEE-269, IEEE-1329. There a 3rd octave wide pulsed noise is used. This way each frequency run covers one octave wide band, and with about 13 runs the spectrum we most care about gets covered. With sine signals getting similar coverage with single tones is just not possible and the reality is that speakers in laptops and tablets often have issues in quite narrow frequency bands due to complex case design, output ports, etc.

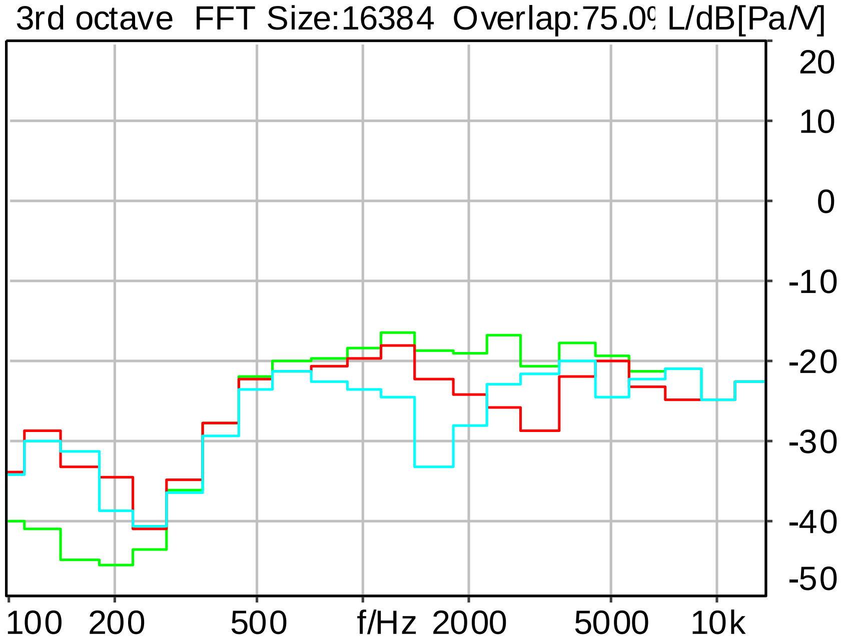
These narrow band issues are often the cause of AEC leaks.

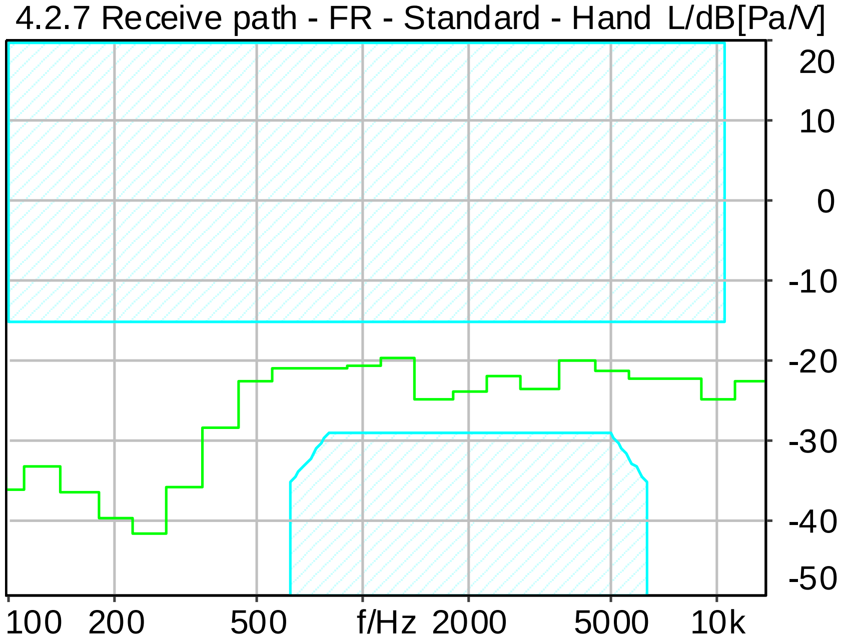
Distortion testing (2)

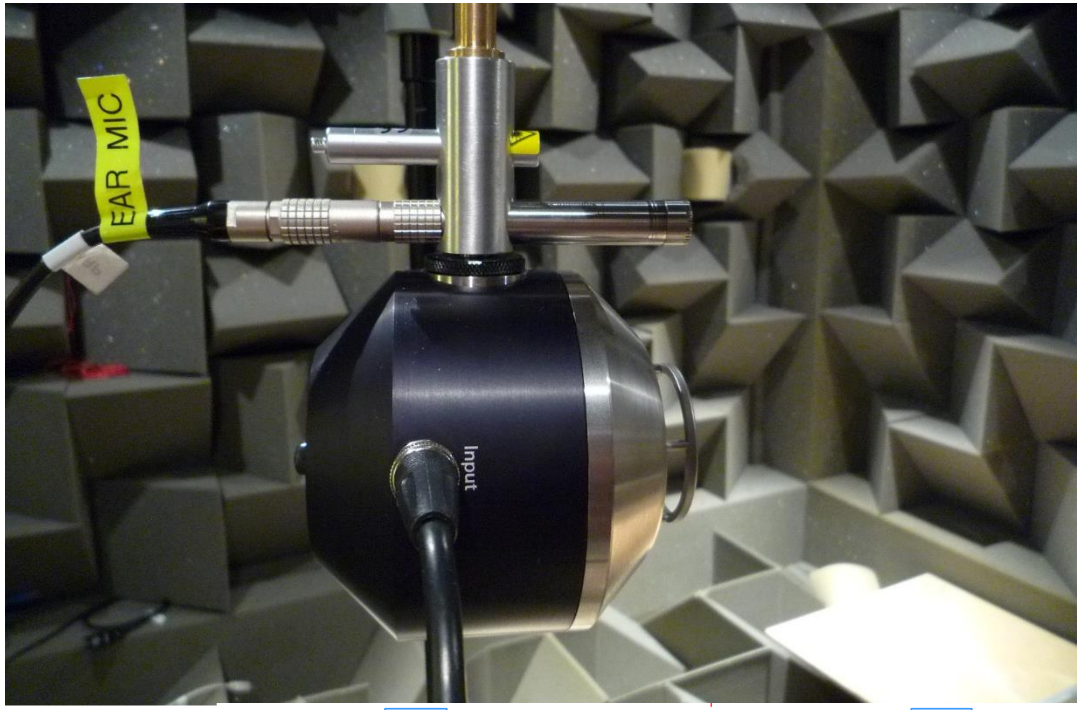


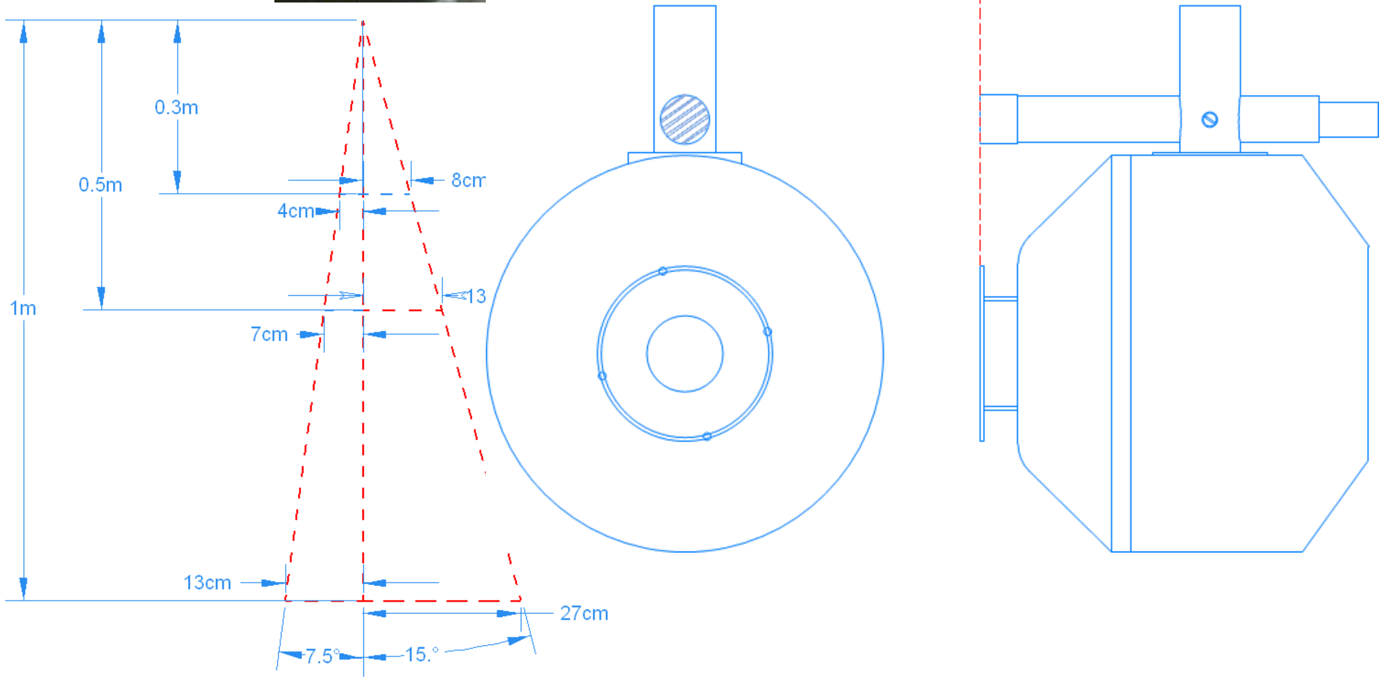
Example of the reported RCV distortion report.

Frequency response testing

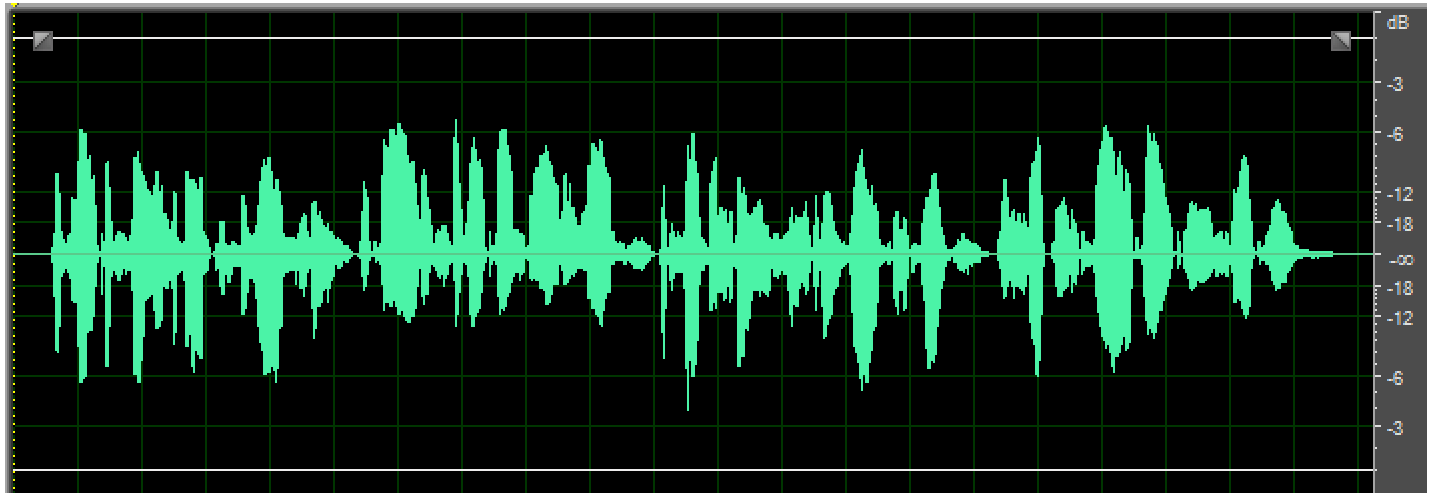


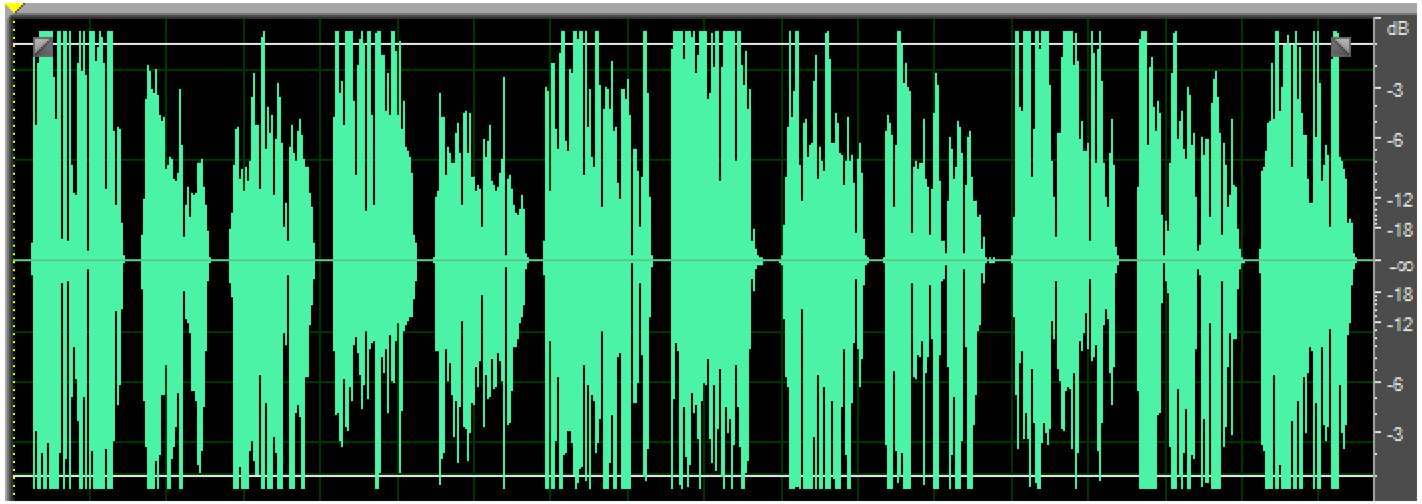






TCLw





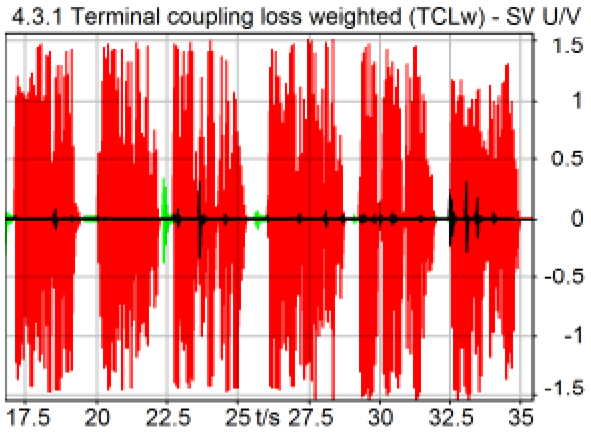
The industry standard method is still used, but the test signal has changed from “normal male speech” to a test signal recommended in the latest 3GPP standards and specified in ITU-T P.501, the latest amendment.

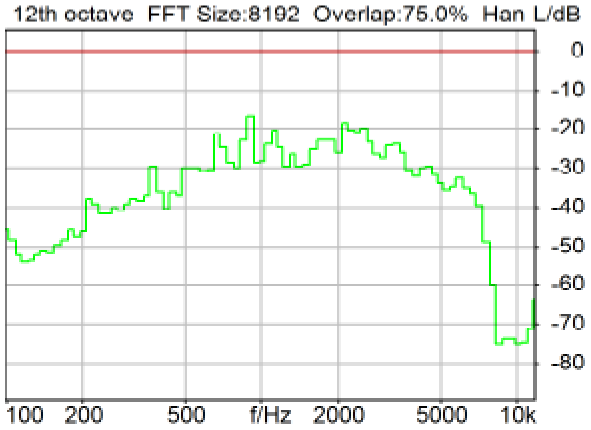
This new signal uses extra compression in different frequency bands, thus giving a more uniform spectrum that improves the measurement accuracy, especially in high frequency.

An added benefit is that the signal goes closer to 0 dBFS in playback, which is what the Skype client does in real calls.

EQUEST– echo performance -> MOS score

4.3.1 Terminal coupling loss weighted (TCLw) - SWB





Correction

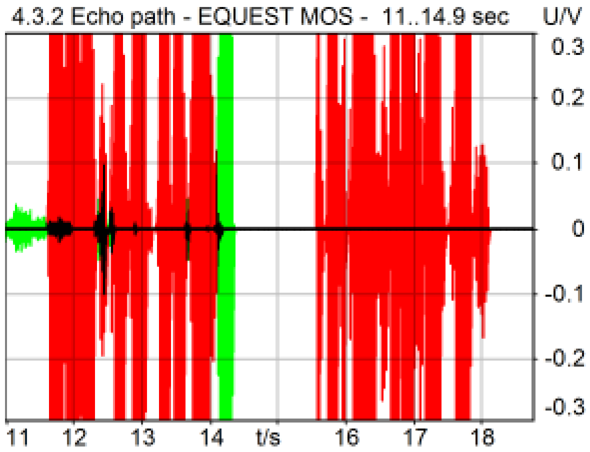
Echo Loss: 27.10 dB
Corrected Echo Loss: 39.73 dB

Fail

The conventional TCLw test analyzes:

- Level of echo residuals
- Applies frequency weighting (per G.122)

4.3.2 Echo path - EQUEST MOS - 11..14.9 sec



Time range: 11.0 .. 14.9 s	
MOS	1.6
Delay	363.0 ms
Echo Level	-20.58 dB
Avg. Delta Rel.App.	2.12 cPa
Max. Correlation	14.62 %
Time range: 14.9 .. 18.8 s	
MOS	4.6
Delay	363.0 ms
Echo Level	-82.42 dB
Avg. Delta Rel.App.	0.00 cPa
Max. Correlation	4.26 %

MOS (Average): 3.1 Fail

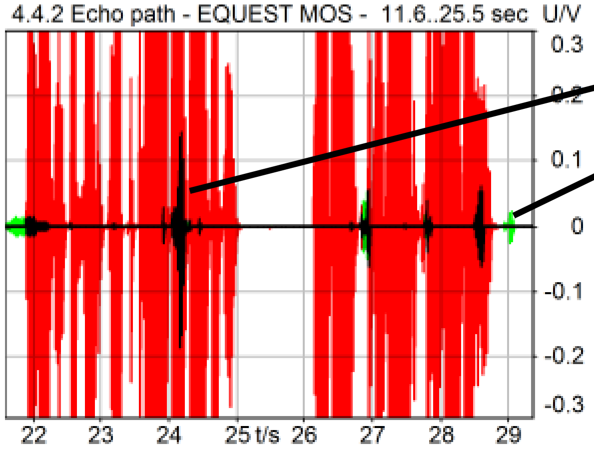
Fail

The new EQUEST test analyzes:

- Level of echo residuals
- Applies frequency weighting (dependent of Round Trip Delay)
- Considers Round Trip Delay (higher delay -> more disturbing echo)
- Considers time when echo leaks happen
- Calculates MOS score

EQUEST– measured sample

4.4.2 Echo path - EQUEST MOS - 11.6..25.5 sec - 2



First echo leak is rated less objectionable as it happens during talking while the second echo comes after talking and, thus, is very audible due to lack of masking effect

Time range: 21.6 .. 25.5 s	
MOS	2.8
Delay	354.0 ms
Echo Level	-42.24 dB
Avg. Delta Rel.App.	0.40 cPa
Max. Correlation	5.58 %
Time range: 25.5 .. 29.4 s	
MOS	2.1
Delay	354.0 ms
Echo Level	-48.11 dB
Avg. Delta Rel.App.	0.41 cPa
Max. Correlation	0.61 %

MOS (Average): 2.5 Fail

Fail

Different criteria is used for the same echo leak depending on the length of Round Trip Delay

Echo performance Mean Opinion Score (MOS) is calculated

www.iwaenc.org/proceedings/2008/contents/papers/9067.pdf

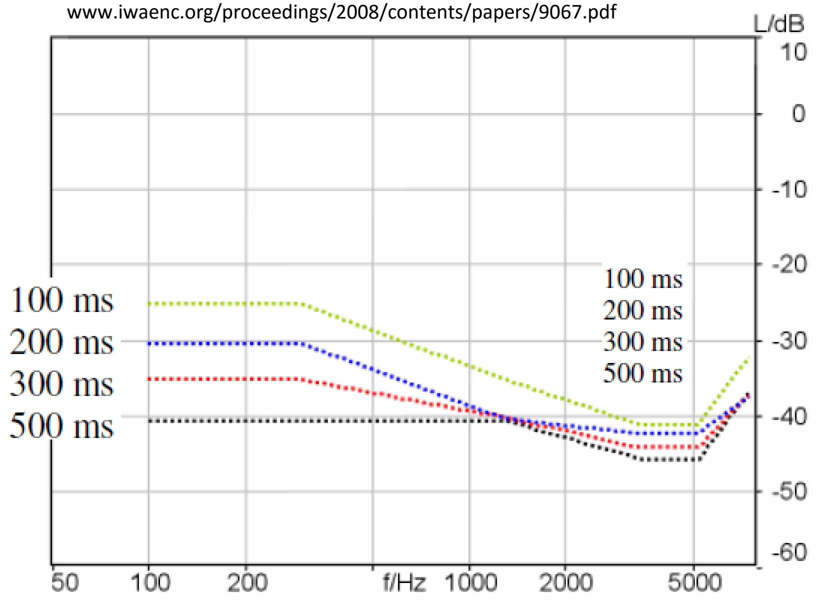
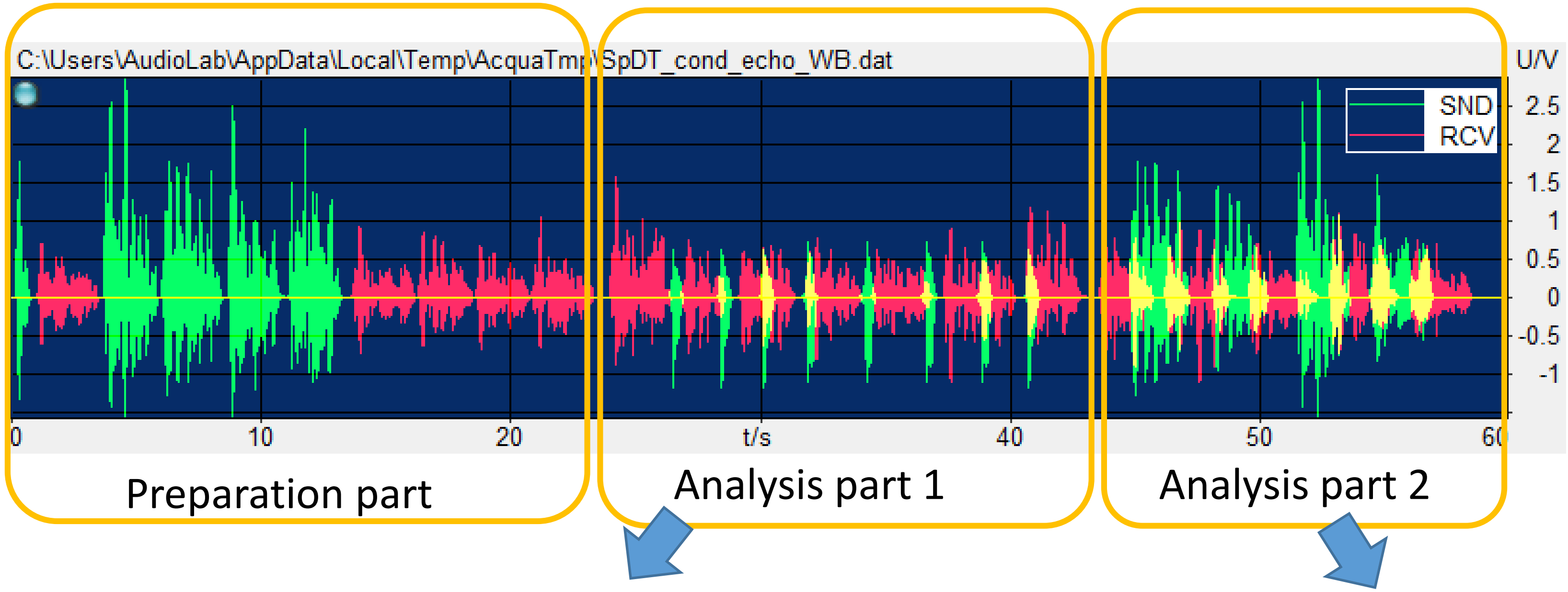
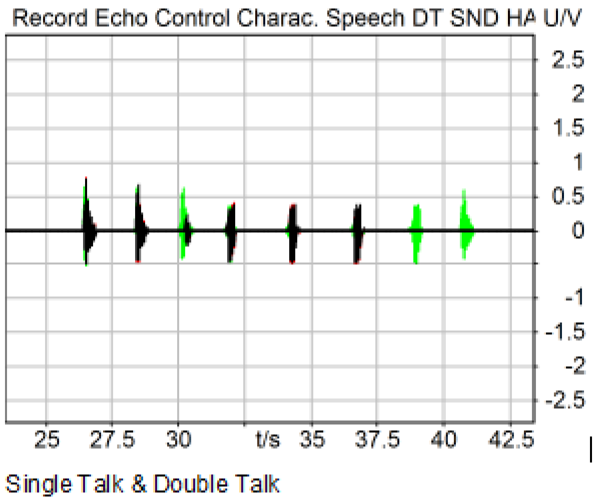


Figure 4.1: Suggestion for spectral tolerance schemes for different round trip delays

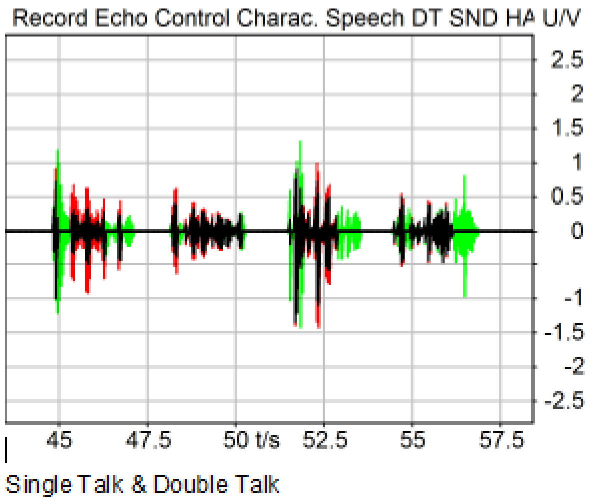
Echo Control Characteristics – ECC



4.3.3. Echo Control Charac. Speech DT SND 1o2 HAWB

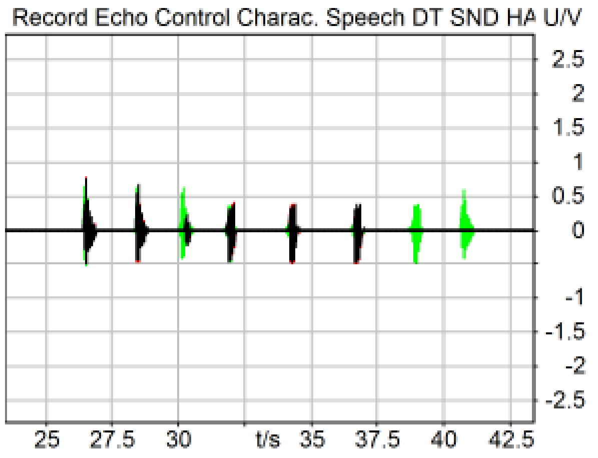


4.3.3. Echo Control Charac. Speech DT SND 2o2 HAWB

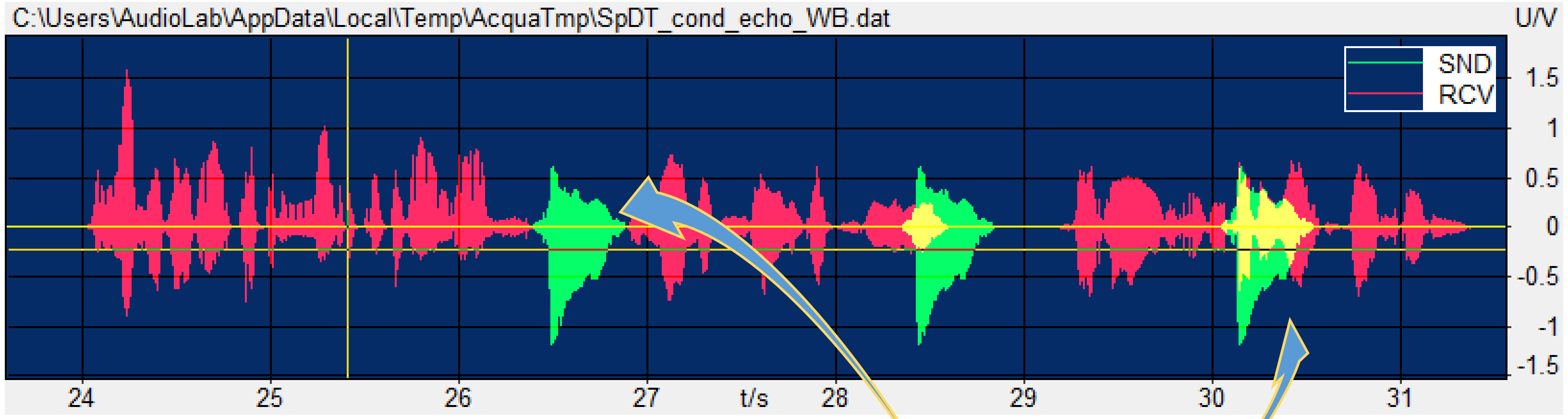


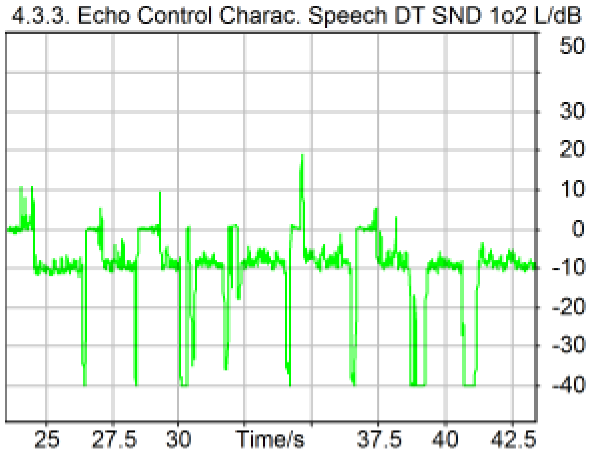
ECC – Single Talk / Double Talk / Echo

4.3.3. Echo Control Charac. Speech DT SND 1o2 HAWB



Single Talk & Double Talk



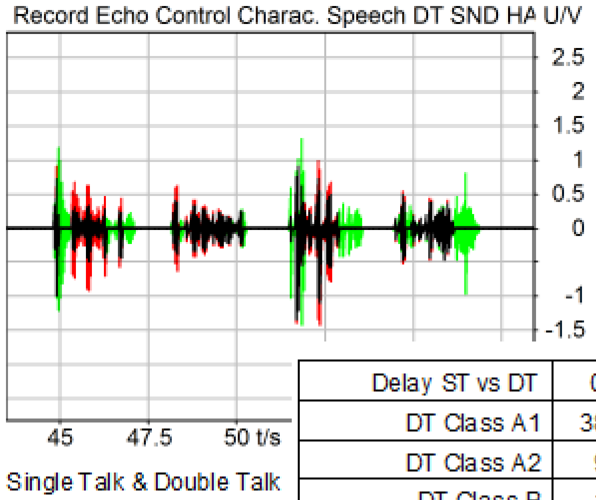


Level vs. Time (Double Talk) - Level vs. Time (Single Talk)

Delay ST vs DT	0.001 s	Delay SND vs Source	-0.122 s
DT Class A1	38.98 %	ST Class A1	7.49 %
DT Class A2	9.42 %	ST Class A2	91.27 %
DT Class B	1.04 %	ST Class B	0.67 %
DT Class C	15.21 %	ST Class C	0.58 %
DT Class D	26.45 %	ST Class D	0.00 %
DT Class E	0.61 %	ST Class E	0.00 %
DT Class F	3.89 %	ST Class F	0.00 %
DT Class G	4.41 %	ST Class G	0.00 %
Double Talk Activity	39.14 %	Single Talk Activity	35.25 %

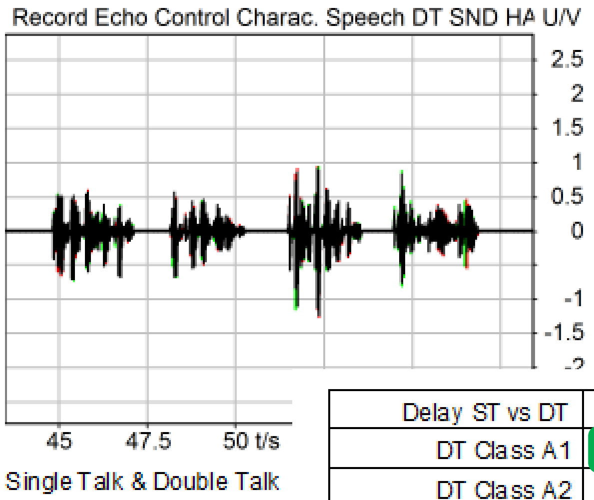
Note: The ratio of Single Talk and Double Talk is different for part 1 and part 2

Echo Contol Characteristics – result samples



Skype to Skype call on Handheld speakerphone (smartphone)

Delay ST vs DT	0.001 s	Delay SND vs Source	-0.122 s
DT Class A1	38.98 %	ST Class A1	7.49 %
DT Class A2	9.42 %	ST Class A2	91.27 %
DT Class B	1.04 %	ST Class B	0.67 %
DT Class C	15.21 %	ST Class C	0.58 %
DT Class D	26.45 %	ST Class D	0.00 %
DT Class E	0.61 %	ST Class E	0.00 %
DT Class F	3.89 %	ST Class F	0.00 %
DT Class G	4.41 %	ST Class G	0.00 %
Double Talk Activity	39.14 %	Single Talk Activity	35.25 %

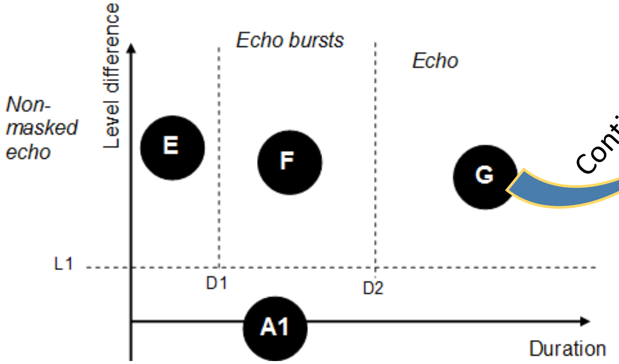


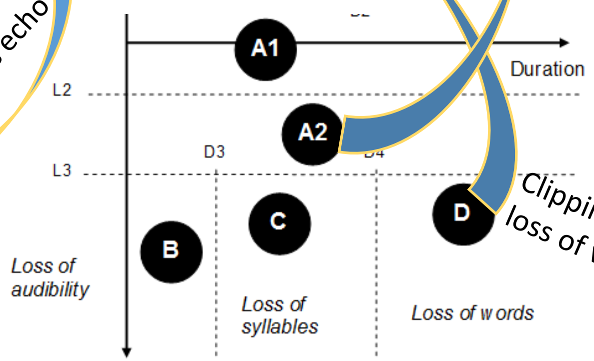
Skype to Skype call on Handset (smartphone)

Delay ST vs DT	0.001 s	Delay SND vs Source	-0.121 s
DT Class A1	100.00 %	ST Class A1	100.00 %
DT Class A2	0.00 %	ST Class A2	0.00 %
DT Class B	0.00 %	ST Class B	0.00 %
DT Class C	0.00 %	ST Class C	0.00 %
DT Class D	0.00 %	ST Class D	0.00 %
DT Class E	0.00 %	ST Class E	0.00 %
DT Class F	0.00 %	ST Class F	0.00 %
DT Class G	0.00 %	ST Class G	0.00 %
Double Talk Activity	37.96 %	Single Talk Activity	35.52 %

Full-duplex with level loss in Tx

Continuous echo





Clipping resulting in loss of words

Methods in unified specification version 2

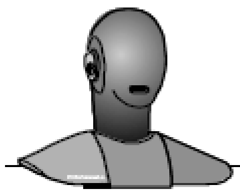
For specification version 2, there are no plans to start using additional methods and the currently described remains valid.

The updates in the next version are mainly focusing on fine-tuning the required values for different use cases.

3QUEST – speech signal and noise suppression quality

Noises used
(same as 3GPP / HDVoice)

1) Pub Noise
2) Outside Traffic Road
3) Outside Traffic Crossroads 4) Train Station
5) Fullsize Car1 130 kmh
6) Cafeteria Noise
7) Mensa
8) Work Noise Office Callcenter



Application Note 3QUEST

Determination of subjective speech MOS (S-MOS)	Determination of subjective noise MOS (N-NOS)	Determination of subjective global MOS (G-MOS)
5 – NOT DISTORTED 4 – SLIGHTLY DISTORTED 3 – SOMEWHAT DISTORTED 2 – FAIRLY DISTORTED 1 – VERY DISTORTED	5 – NOT NOTICEABLE 4 – SLIGHTLY NOTICEABLE 3 – NOTICEABLE BUT NOT INTRUSIVE 2 – SOMEWHAT INTRUSIVE 1 – VERY INTRUSIVE	5 – EXCELLENT 4 – GOOD 3 – FAIR 2 – POOR 1 – BAD

Table 1: Instructions and scales acc. to ITU-T P.835

4.4.4.2 Requirement

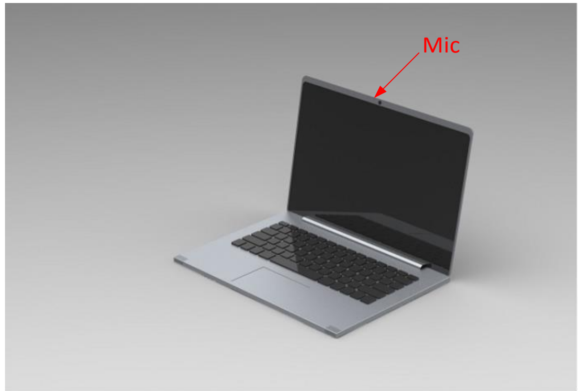
	3Quest results (MOS LQOw)	
	Standard	Premium
S-MOS (average score of tested noise cases)	≥ 3.3	≥ 3.5
N-MOS (average score of tested noise cases)	≥ 2.3	≥ 3.0
G-MOS (average score of tested noise cases)	-	-
S-MOS min (lowest score of tested noise cases)	≥ 3.0	≥ 3.3
N-MOS min (lowest score of tested noise cases)	≥ 2.0	≥ 2.3
G-MOS min (lowest score of tested noise cases)	-	-

Table 38: Speech quality requirements in presence of background noise requirements for Handset / Headset

Design guidance: audio capture and render

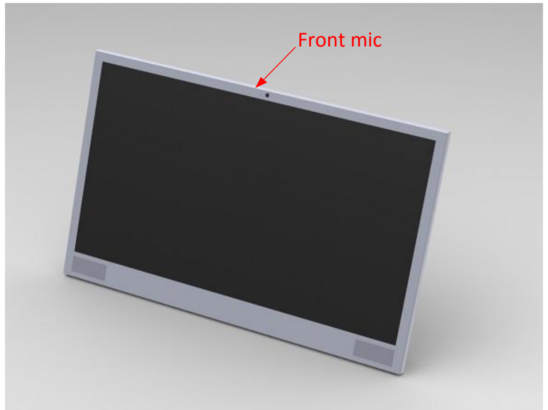
Audio capture guidelines

- Use microphone locations shown here
- 2-element omni left/right mic array will give 1-2 dB SNR improvement over single mic if done right; most mic arrays are worse than a good single mic!
  - Windows mic array spec requires cardioid microphones
  - It is better / cheaper to use a single 61 dB SNR mic than two 56 dB mics
- Cardioid mic will give ~4.5 dB better SNR than omnidirection mic; cardioid needs back-port
- 2-element omni front/back mic array can give 10-30 dB SNR improvement over single mic
- Microphone needs to be well sealed and have sufficient open grill area
  - Well sealed: maintains 1 PSI (~7000 Pascals) for 1 minute
- Mount microphone(s) in isolated rubber boot to minimize mechanical coupling
  - Don't mount the microphone on the PCB
- The microphone locations must be discoverable by DEVPKEY_Device_PhysicalDeviceLocation (same data as ACPI _PLD)
- Don't include a mic boost for all microphone types
- Don't include any gain for MEMS microphones in audio driver; OS will provide digital gain



Notebook

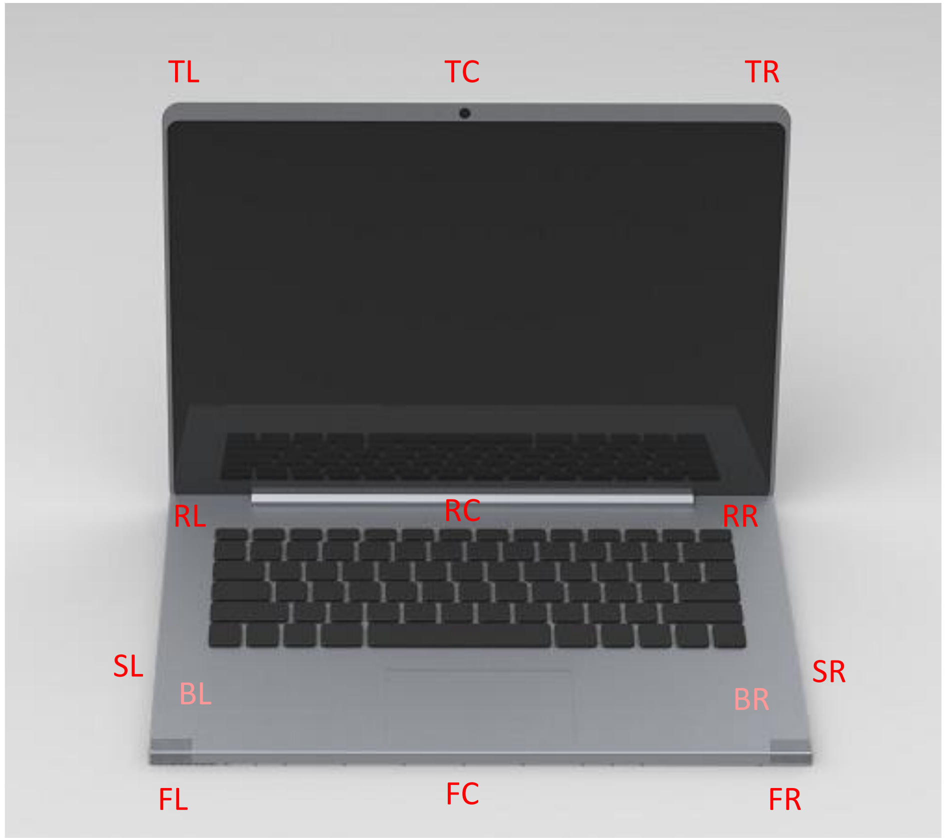
Tablet



All-in-one

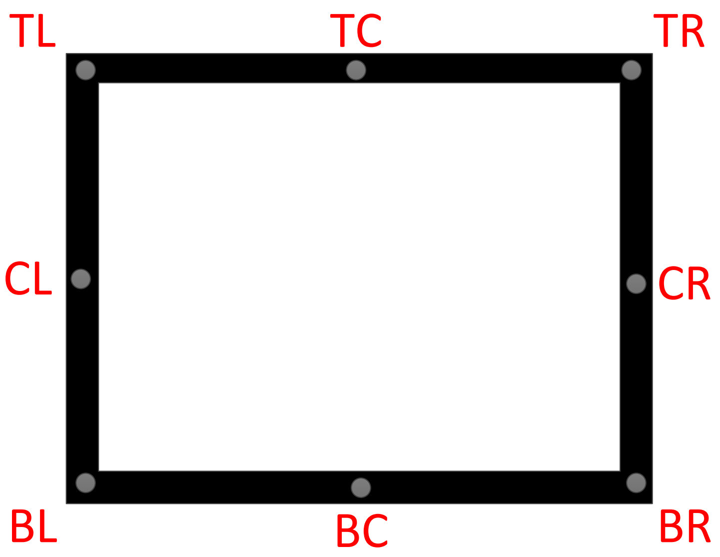
Microphone position analysis: Notebooks

Microphone location	Comment
RL, RR	Picks up keyboard, fan noise High, asymmetric coupling with speaker SL/SR
RC	Picks up keyboard, fan
SL, SR	Picks up keyboard, fan noise High, asymmetric coupling with speaker FL/RL, FR/RR
TL, TR	Moderate, asymmetric coupling with speaker RL/RR
FC	High coupling with speaker FL, FR
FL, FR	High, asymmetric coupling with speaker FL, FR
TC	Lowest symmetric coupling, lowest keyboard, fan noise



Microphone position analysis: Tablets

Microphone location	Comment
CL, CR	Can get occluded by hands in typical location (see Figure 2)
BL, BC, BR	Will get occluded by hands, lap, bed covers
TL, TR	Close to loudspeaker at TL, TR
TC	Good



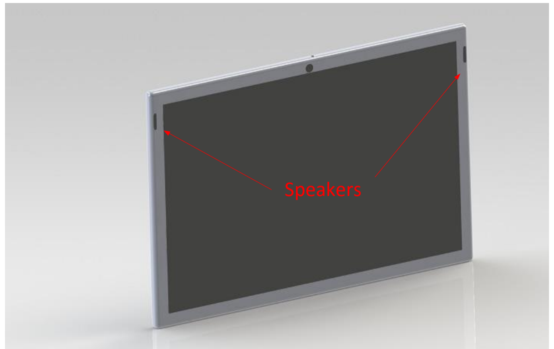
Audio capture component guidelines

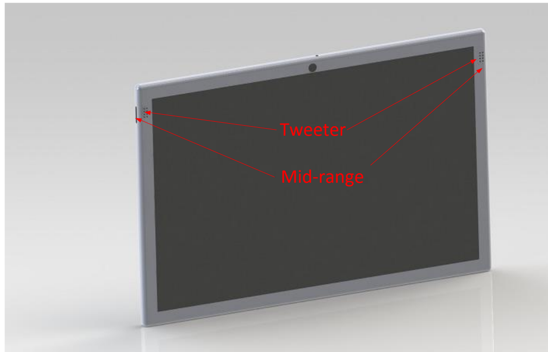
Area	Guideline
Microphone type	Standard: Omnidirectional microphone Premium: • Cardioid microphone • Front/Back omni mic
Microphone SNR	Standard: > 60 dBA @ 94 dBA SPL @ 0.5m 100-7000 Hz Premium: > 64 dBA @ 94 dBA SPL @ 0.5m 100-12000 Hz
Microphone sensitivity	Digital microphone: -26 +/- 2 dBFS (94 dB SPL @ 1kHz)
Microphone frequency response	Standard: [100,7000] Hz +/- 4 dB Premium: [100,12000] Hz +/- 4 dB
Microphone high-pass filter cut-off	100Hz at -3 dB for analog microphone 60Hz at -3 dB for digital microphone
Microphone high-pass filter slope	Better than 18 dB/oct
ADC/DAC resolution	≥16bits
Clipping	No mic clipping when speaker at max volume with full scale sine wave at all frequencies

Audio render guidelines

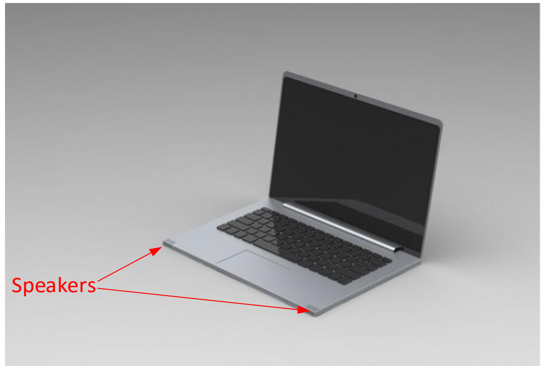
- Use locations recommended here
- Speakers should point toward user
  - Pointing to sides or back attenuates amplitude
  - To minimize front speaker area on tablets tweeters can be used with side mid-range
  - High frequencies are most directional
  - Worst speaker location: facing back
    - Attenuates high frequencies > 14 dB making speech narrowband
- Isolate speakers and microphone(s) to minimize coupling
- Remove loose parts in case to reduce rattles
  - Keyboard, cables, etc.
- Seal speakers to grill: maintain 2 psi (~14000 Pascals) for 1 min

Tablet





Notebook



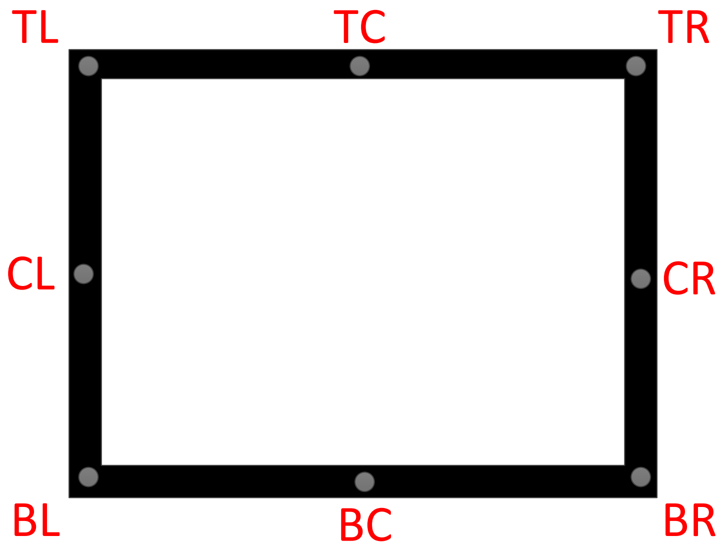
All-in-one

Speakers position: Notebooks

Speaker location	Comment
BL, BR	Can get occluded lap, covers Poor high frequency response
SL, SR	Lower loudness, poor high frequency response
RL, RR	High coupling with mic at RC/TL/TC/TR Lower loudness
FL, FR	Highest loudness, best frequency response

Speakers position: Tablets

Speaker location	Comment
TL, TR	Good
CL, CR	Will get occluded by hands
BL, BR	Will get occluded by hands, lap, bed covers



Audio render component guidelines

Area	Guideline
Frequency response	Standard: [300,7000] Hz +/- 5 dB Premium: [150,12000] Hz +/- 5 dB
Total harmonic distortion (THD)	Standard: <3% for [300,7000] Hz measured with volume set at 80 dB SPL at 0.5m at 1kHz Premium: <3% for [150,12000] Hz measured with volume set at 86 dB SPL at 0.5m at 1kHz
Grill	Open area > 50% speaker cone area
Driver type	Standard: Single driver (2X for stereo) Premium: Tweeter+woofer (2X for stereo)

Considering use cases for speech pick-up in case of audio offloading

Scenarios for DUT with camera(s)	Desired behavior
Front-facing camera in use, Handheld speakerphone audio UI	Front +/- 45^{o}, suppresses sounds from rear
Rear-facing camera video call, Handheld speakerphone audio UI	Omni-directional
Personal speakerphone audio UI	Omnidirectional or Front +/- 90^{o}
Group speakerphone audio UI	Omnidirectional or Front +/- 90^{o}
Headset audio UI	Switches to headset processing.
Handheld speakerphone audio UI (audio playback through headphones)	Switches to appropriate speakerphone audio UI mode.
Handset audio UI	Switches to handset processing

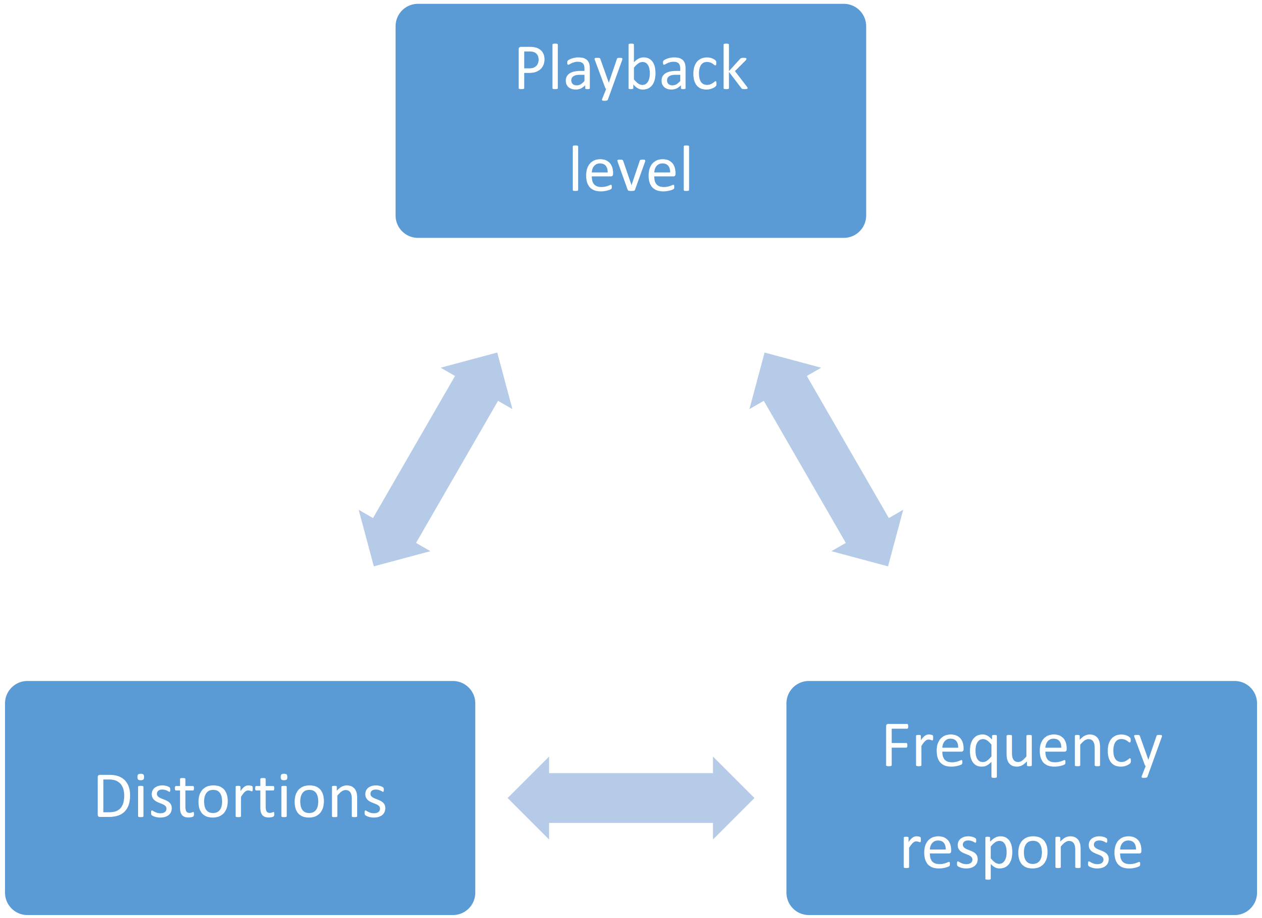
Design guidance: Top failures

Top audio capture / render issues in Windows 8 devices

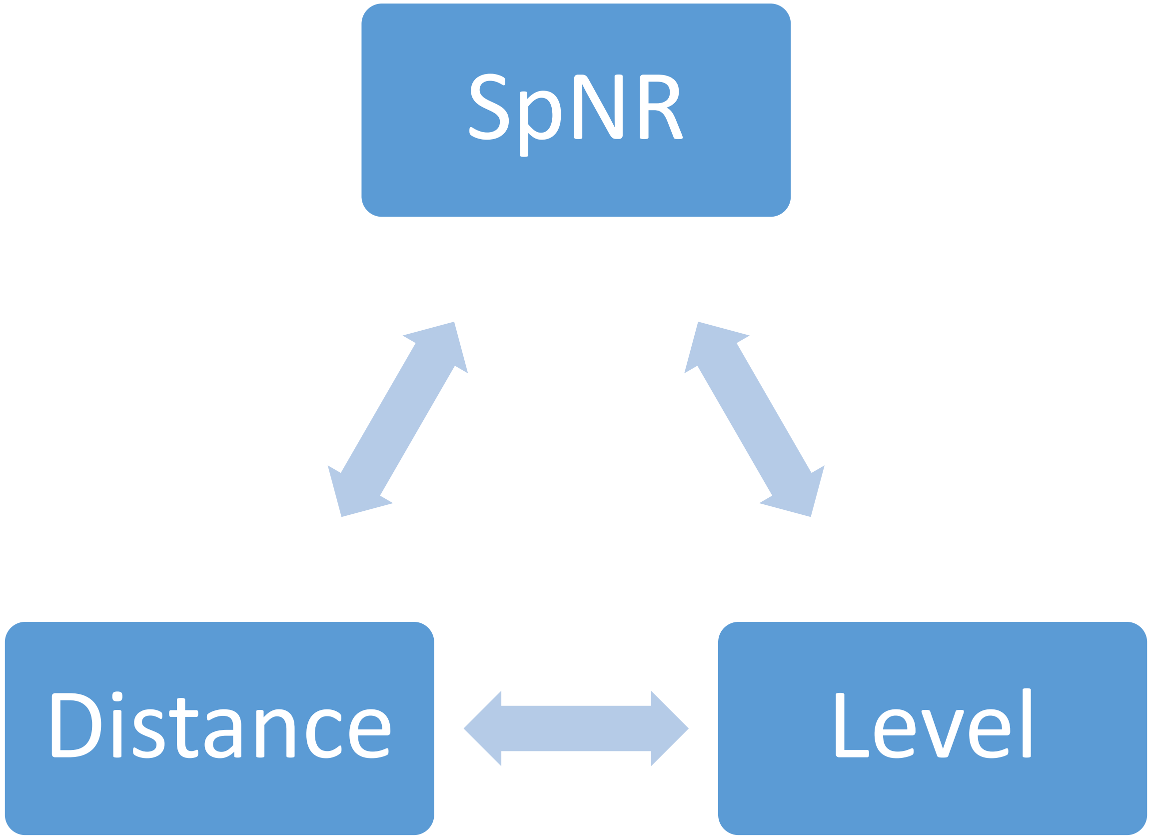
Issue	How common	Solutions
Echo leaks and distorted double-talk	30% Windows 8 devices	Reduce nonlinear coupling (rattles, etc.) Disable onboard nonlinear processing Fix microphone clipping Reduce coupling (TCLw)
High send noise	11% Windows 8 devices	Use mic with Signal-to-Noise Ratio > 60 dB Position mic away from fan
Low receive volume	80% tablets	Orient speakers toward user Increase speaker size/power

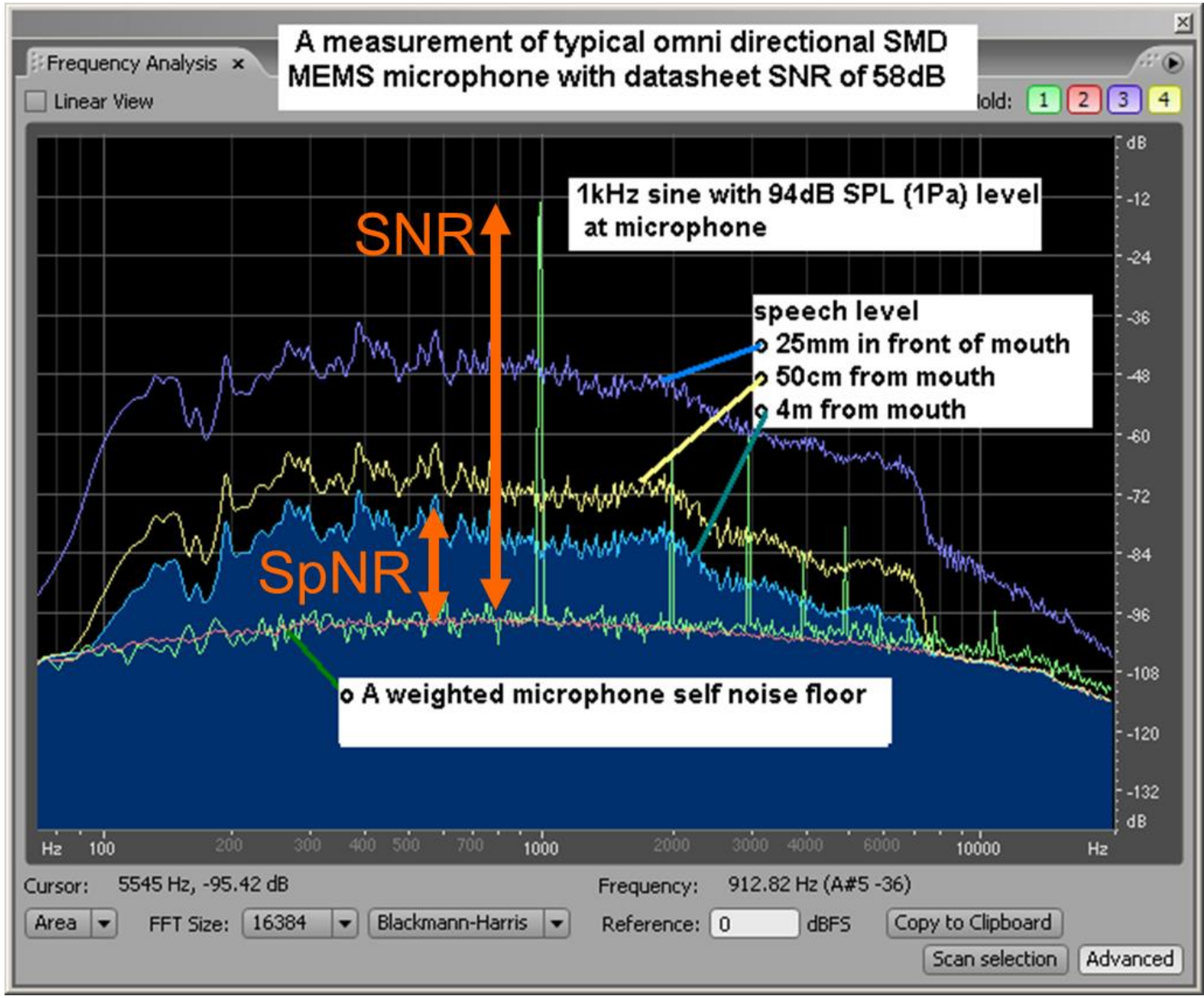
- **43% Windows 8 devices failed audio subjective tests** (MOS < 3 in 1 to 5 scale)
- N=71 Windows 8 devices tested (notebooks, tablets, all-in-ones)

Receive path challenge

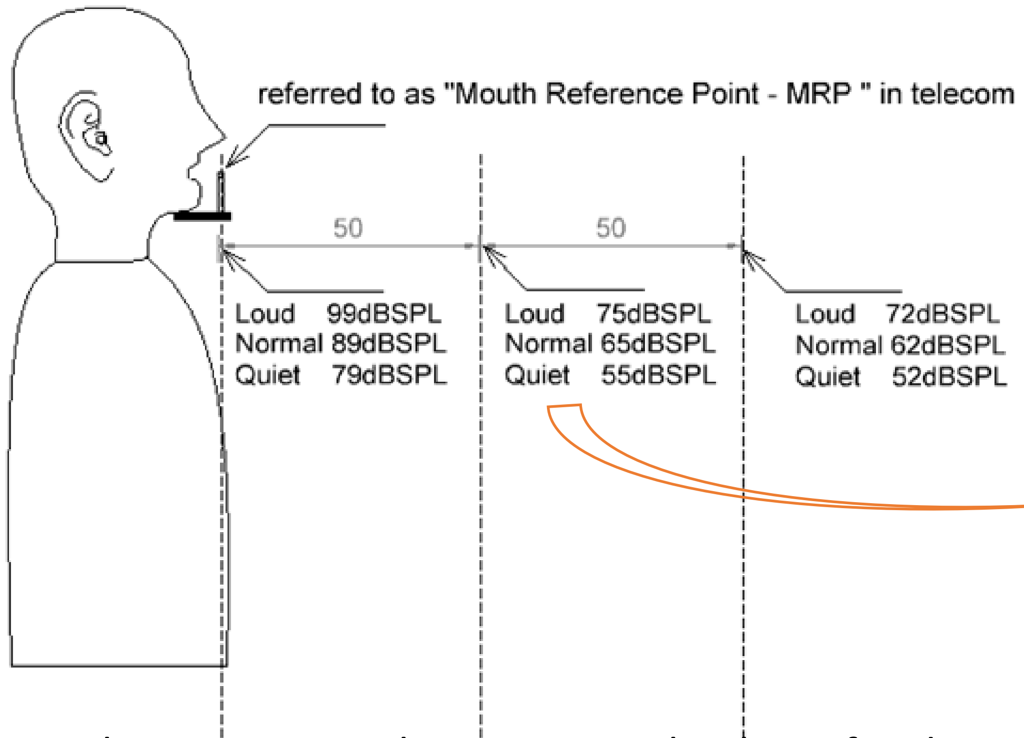


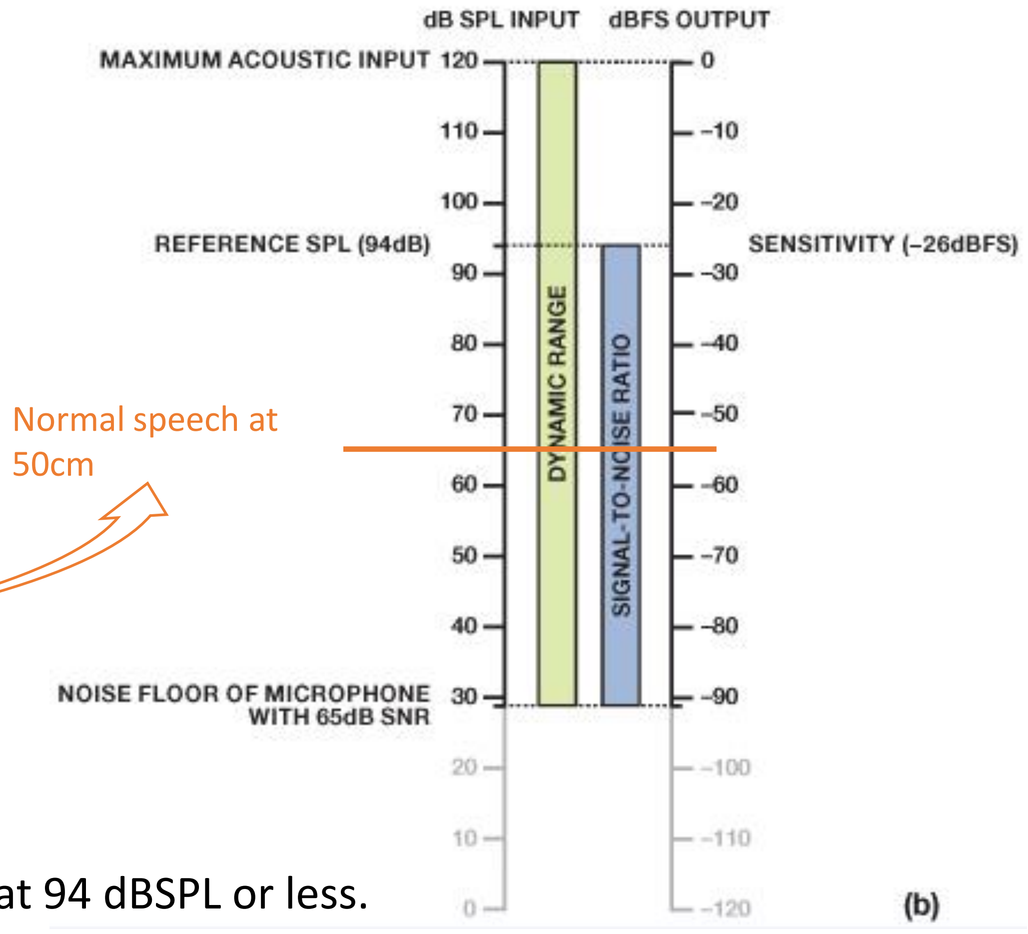
Send path challenge





SpNR



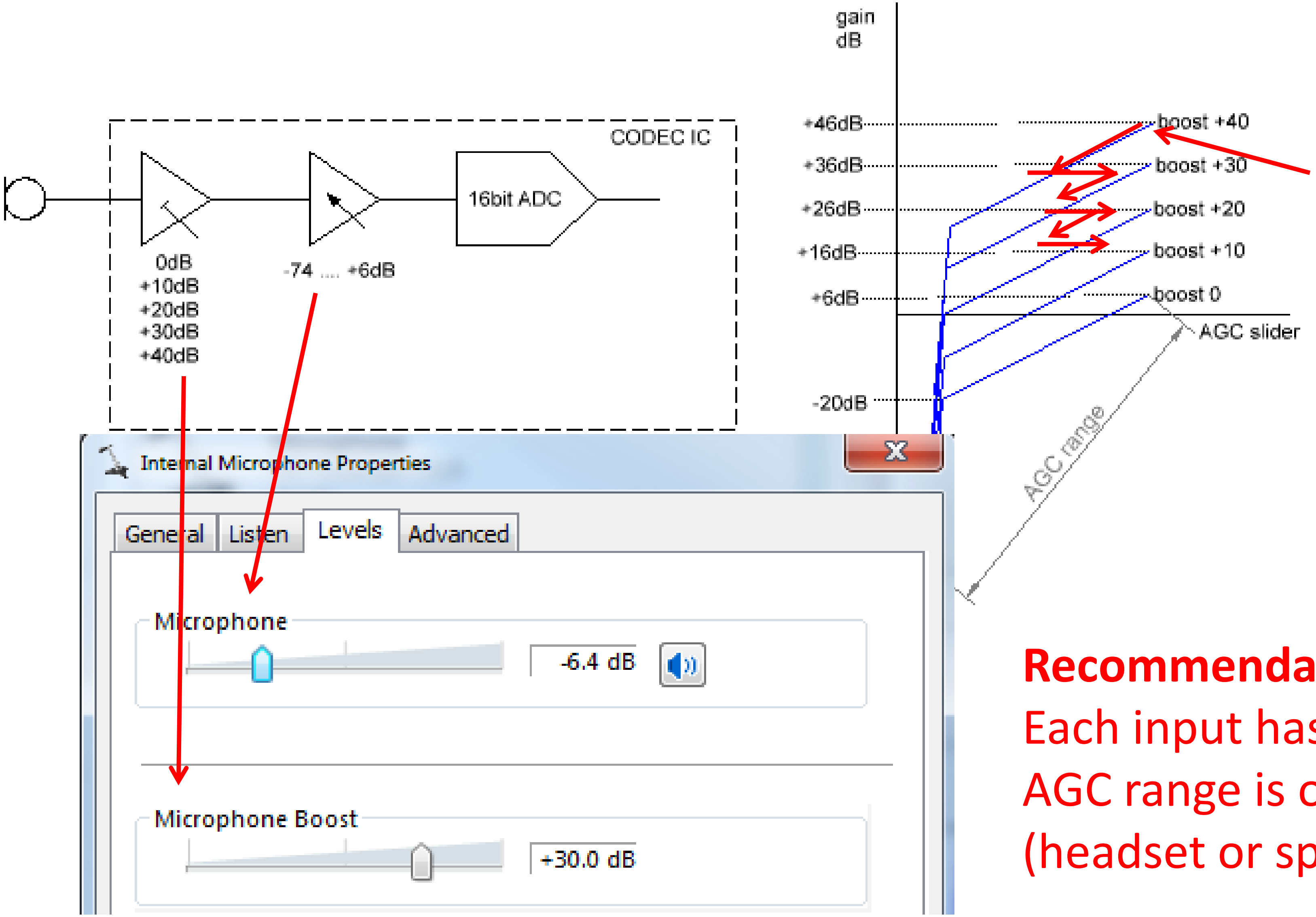


Digital MEMS microphones on market have fixed sensitivity of -26 dBFS at 94 dBSPL or less.

If mic sensitivity is -26 dBFS, then acoustic speech pickup level when user mouth is at 50cm is -55 dBFS rms. This is lower than the Skype requirement of -32 dBFS minimum – so it needs digital gain to meet the criteria. The signal on any other voice recording app also would be too quite without extra digital gain.

Best SNR of these microphones is approximately 62 dB -> making the equivalent self noise floor 94-62=32dBSPL -> thus Speech to Noise ratio will be 55 – 32 = 23 dB without noise suppression algorithms (Skype minimum requirement is 25 dB).

AGC problem based on Lenovo T510 / T420S



With the sample Lenovo T420S case – if a call starts at "boost=+40" setting and AGC slider is at 50% and input is overloading, Skype would need to go through many boost settings and AGC settings to reach the condition of no overload.

As each API call takes time, this might end up with a long condition of poor call quality and leaking AEC.

Recommendation:

Each input has only AGC gain slider (no boost!)

AGC range is optimized for each input (headset or speakerphone mode, etc.)

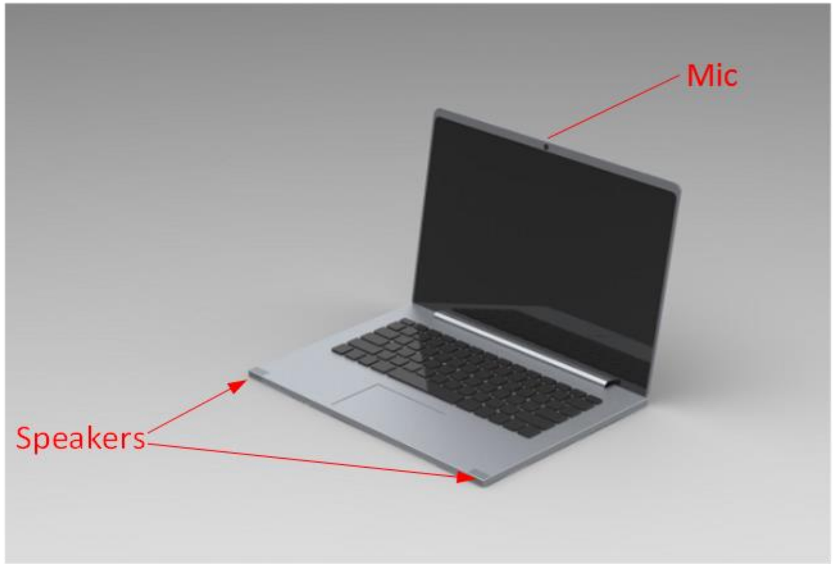
Coupling between microphone and loudspeakers

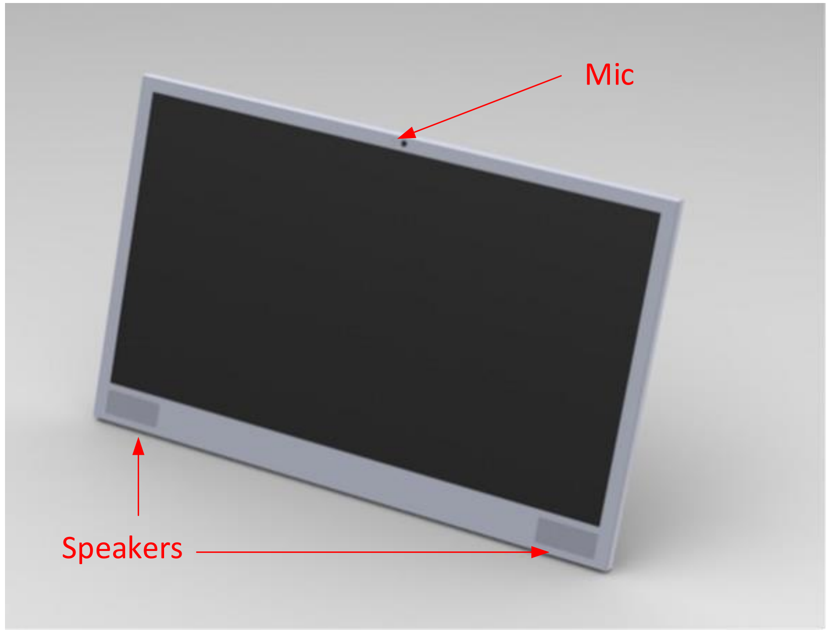
Distance

- Maximize the distance
- Do not forget usage scenarios

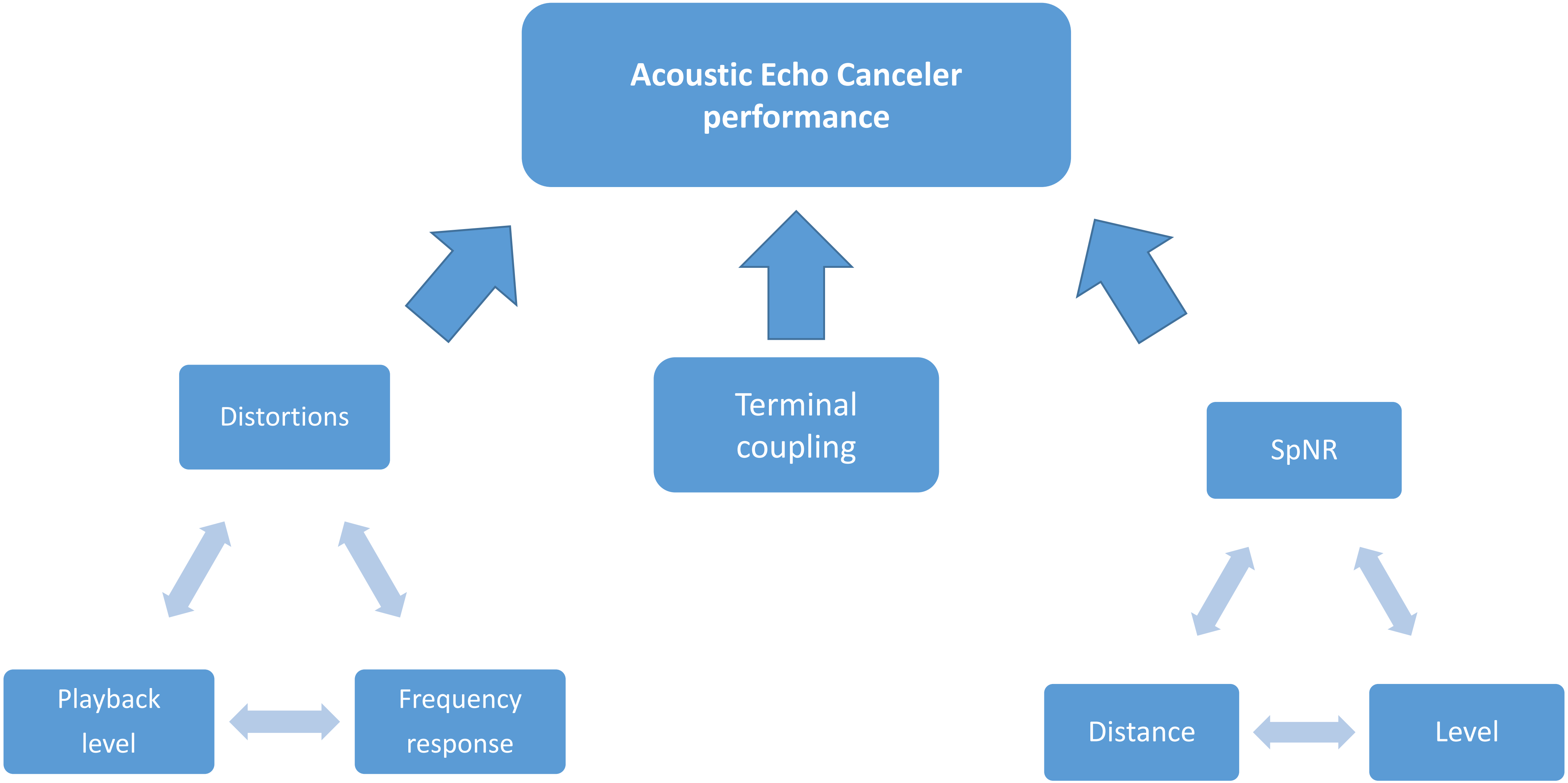
Acoustic sealing

- Foam boot between the microphone and port hole





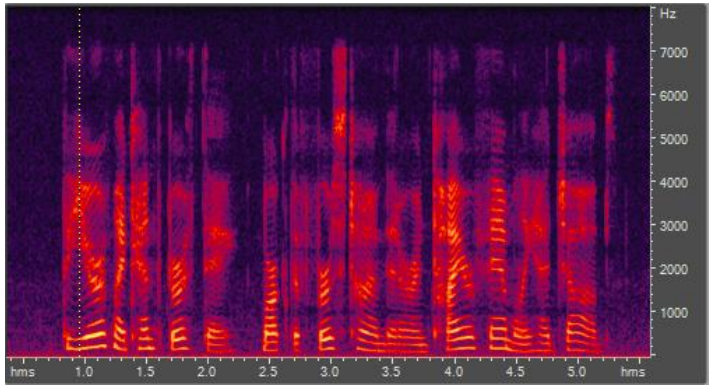
Echo canceler constraints

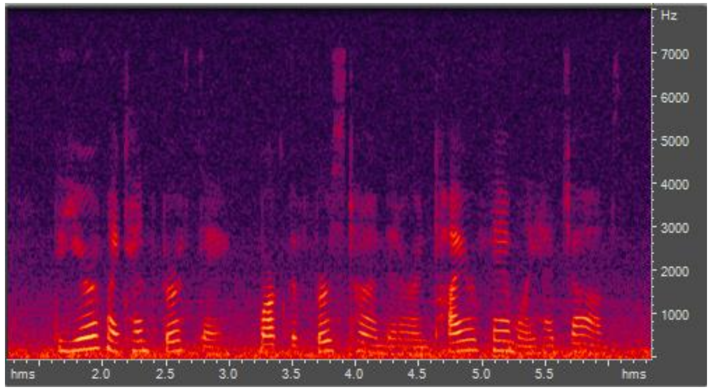


Design guidance: Examples of failures

Audio capture comparison: more details

- Two high-end notebooks
  - Captured in an anechoic chamber, background noise of 25 dBA SPL
  - Speech 70 dBA SPL @ 0.5 m, 0% CPU usage
  - Best case scenario
  - Two very different capture qualities (SNR, send noise, MOS)





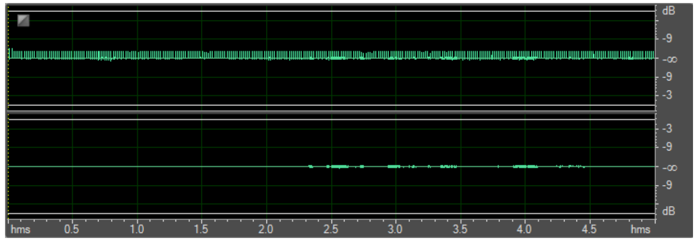
Raw SNR: 11.5 dB

Processed SNR: 26.8 dB

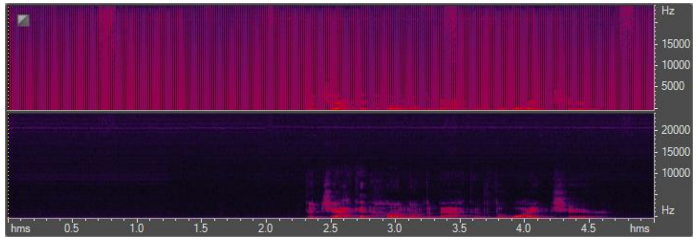
Raw SNR: 6.4 dB

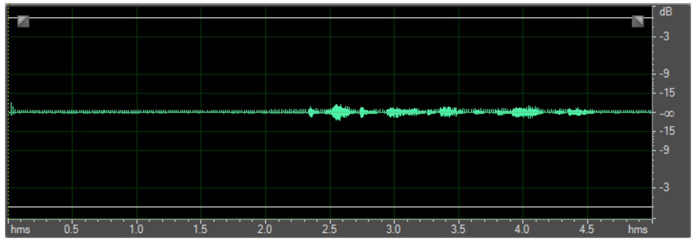
Processed SNR: 15.0 dB

Audio example from Windows 8 tests



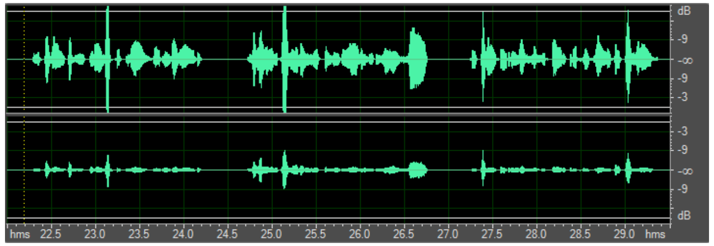
Left channel contaminated by noise



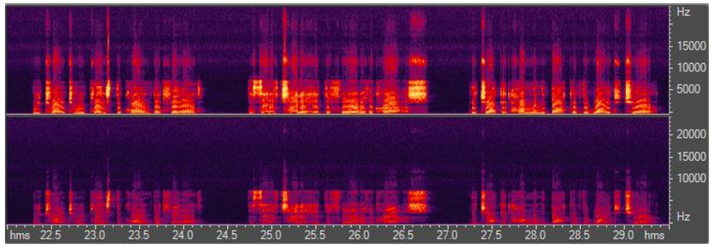


Output is unusable – PC doesn’t work with Skype!

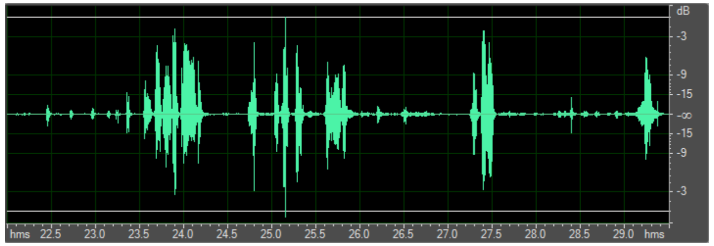
Audio example from Windows 8 tests (2)



Left channel speaker/mic coupling much higher



Asymmetric clipping and THDN

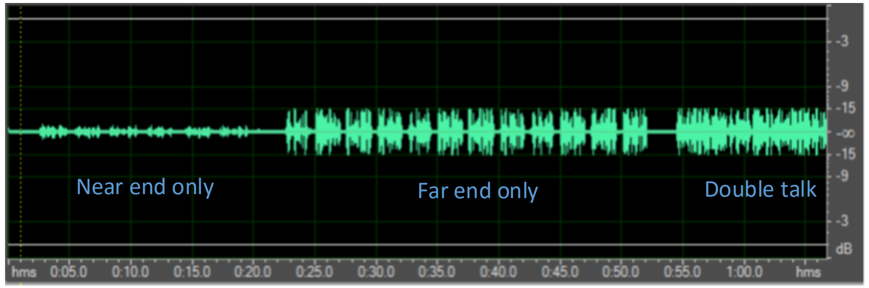


Unusable full echo for Skype

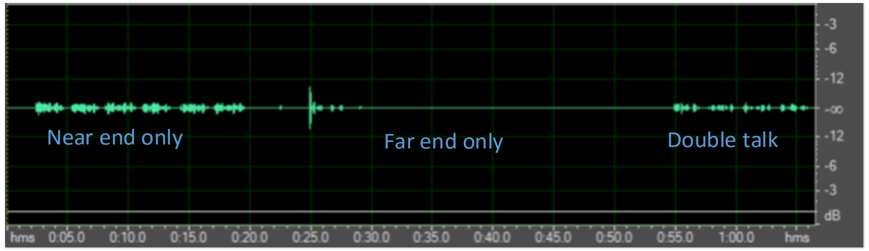
Design guidance: Advanced topics

Diagnosing audio issues using HCK test signal

- HCK results are super useful – always run HCK and make modifications to pass before submitting to Logo program
- Listen to the Mic and MicOut wav files to better understand issues (subjective quality)
  - Far end MicOut should be silent
  - Double talk should be undistorted speech



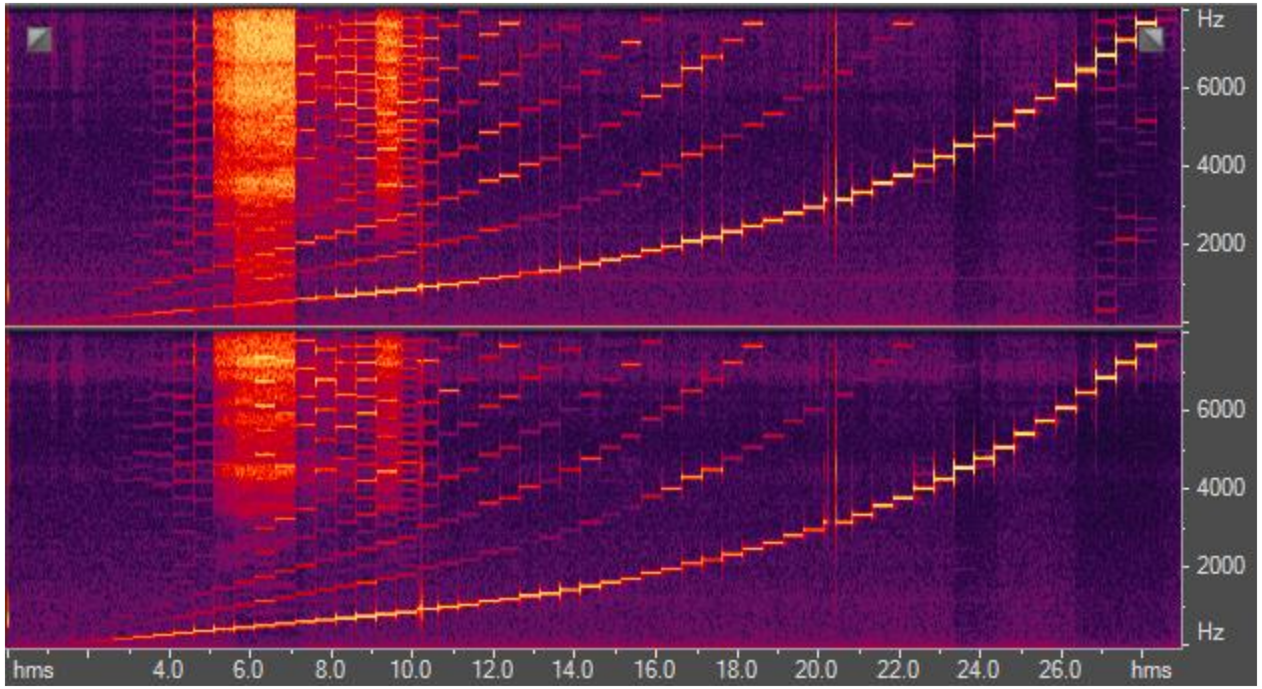
AECTest_DUTMic.wav (no AEC and NS)



AECTest_MicOut.wav (AEC and NS)

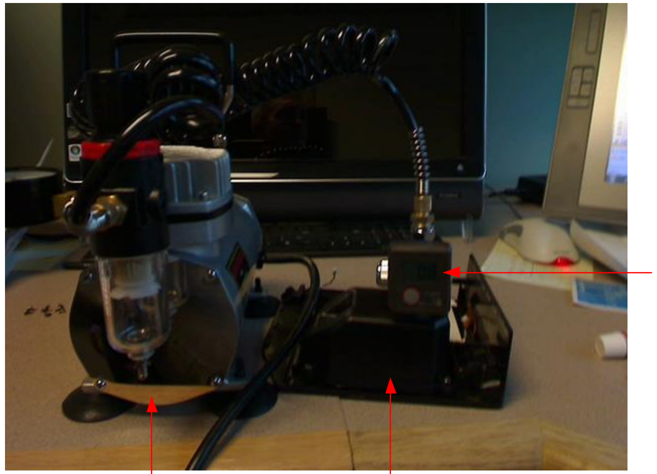
Diagnosing audio issues using sweep signals

- Stepped sweep signal very useful to detecting clipping at specific frequencies
- Also used to compute THDN and listen for distortion
- Good for glitches, too



Sealing microphone and speakers

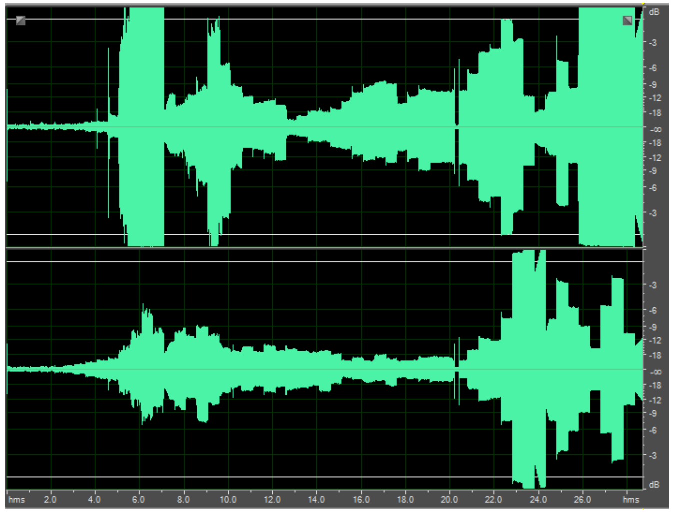
- Sealing microphones and speakers to front grill is the easiest, least expensive improvement
- Nearly all Windows 8 devices tested did not do this!
- Seal should hold 1-2 psi for 1 minute to be safe
- Use pressure-sensitive paper to help detect leaks

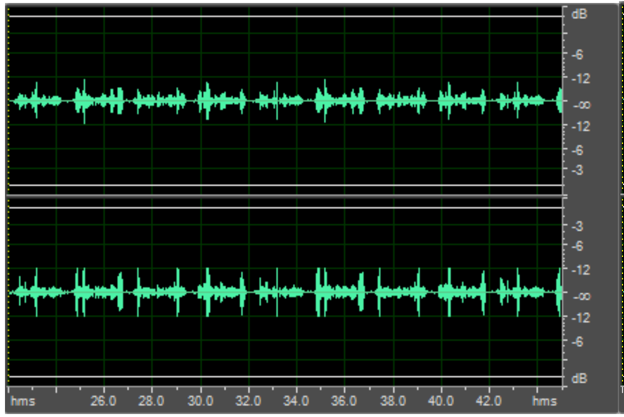


Pressure-sensitive paper

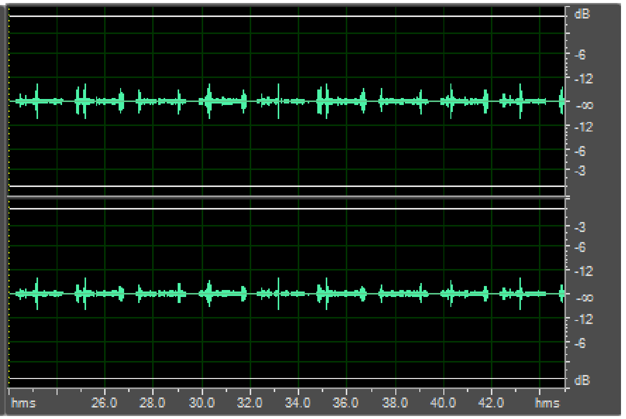
Windows 8 tablet case study

- Microphones and speakers unsealed
- Asymmetric coupling
- Unit reworked to seal microphones and speakers



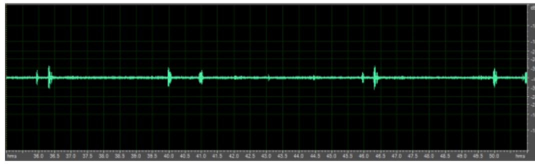


Before

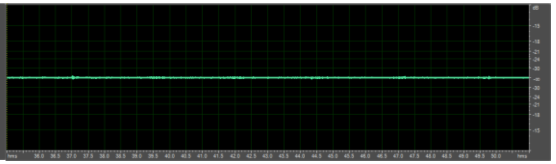


After

	Before	After
PostAEC Echo	2.8%	0.3%



Before



After

Microphone boot design

- Don’t mount microphone to PCB as it increases speaker coupling
- Mount microphone in supple rubber boot that is mechanically isolated from system
- Include an “sound barrier” that further seals the microphone from the internal sounds (reduces coupling)
- Seal the microphone boot to the microphone port hole (e.g., press fit)

References to materials used

- Skype test specifications for USB peripherals, PCs, and room systems http://technet.microsoft.com/en-us/lync/gg278181.aspx
- Windows HCK Communications Audio Fidelity Test – http://msdn.microsoft.com/en-us/library/windows/hardware/dn390880.aspx
- ODM Academy 400: Making Windows devices work great with Skype and Lync – ODM Academy 400 part 2
- Analog Devices: Understanding Microphone Sensitivity: http://www.analog.com/library/analogDialogue/archives/46-05/understanding_microphone_sensitivity.html

References to materials used (2)

- HEAD acoustics support: Application Notes
  http://www.head-acoustics.de/eng/telecom_application_notes.htm
- ETSI EG 202 396
- Windows Engineering Guidance