RIR Estimation for Synthetic Data Acquisition

Kevin Venalainen, Philippe Moquin, Dinei Florencio

Microsoft

ABSTRACT - Automatic Speech Recognition (ASR) works best when the speech signal best matches the ones used for training. Training, however, may require thousands of hours of speech, and it is impractical to directly acquire them in a realistic scenario. Instead, we estimate Room Impulse Responses (RIRs), and convolve speech and noise signals with the estimated RIRs. This produces realistic signals, which can then be processed by the audio pipeline, and used for ASR training. In our research, a limited corpus of speech data as well as noise sources is recorded and the RIR at 27 positions is determined using a variety of methods (chirp, MLS, impulse, and noise). The convolved RIR with the “clean speech” is compared to the actual measurements.

Content

- Introduction
- Test signals
  - MLS
  - Swept Sine
  - White Noise
- Room and Test set-up
- Results

Introduction – Problem statement

- To validate Automatic Speech Recognition a large corpus of data in various acoustical environments is required.
- Testing is time consuming, difficult to replicate and subject to corruption by noise.
- Auralization has been proposed but the fidelity to real rooms is not accurate enough
- There are no commercial solutions which are capable of performing high resolution data synthesis for speech recognition.
- Our proposed solution is to measure room impulse response and convolve the test vectors to synthesise measured data to the fidelity required to accurately test ASR
- The other question we want to answer is what is the best way to measure the RIR.

Introduction – Proposed approach

Three different types of excitation signals:

- Maximum length sequences (MLS)
- Sine sweeps
- White noise.

Validation signals

- 8 sentences: 4 male and 4 female from ITU-T p.863
- (ITU) P.50 signal.

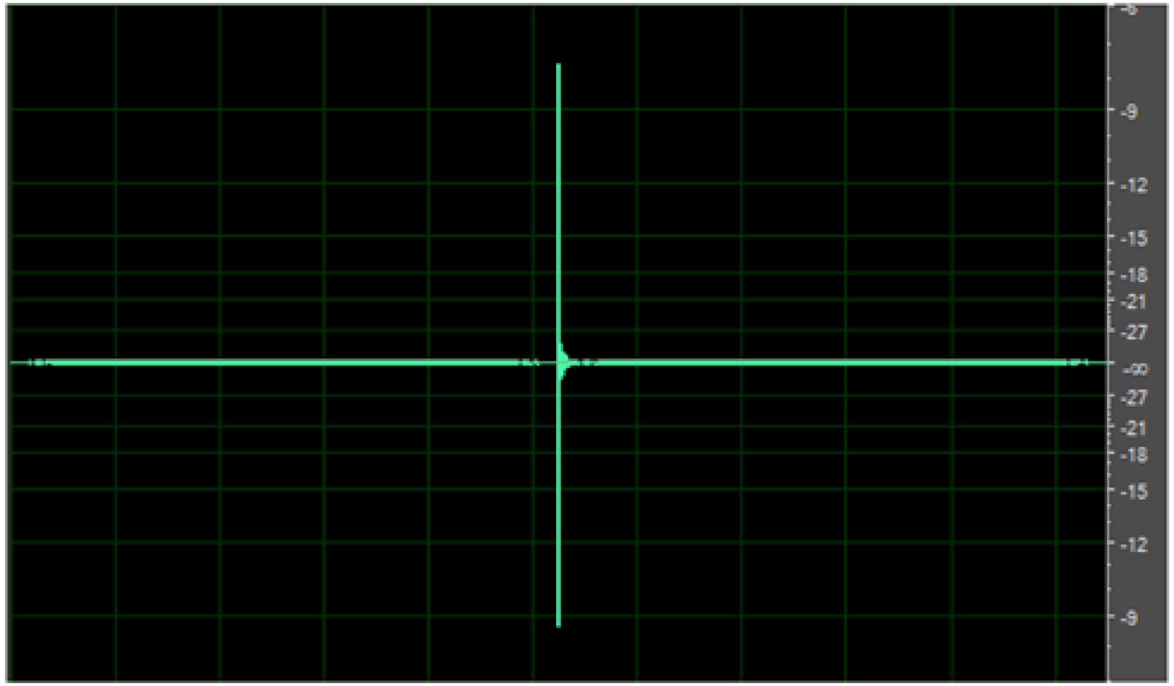
Test signals

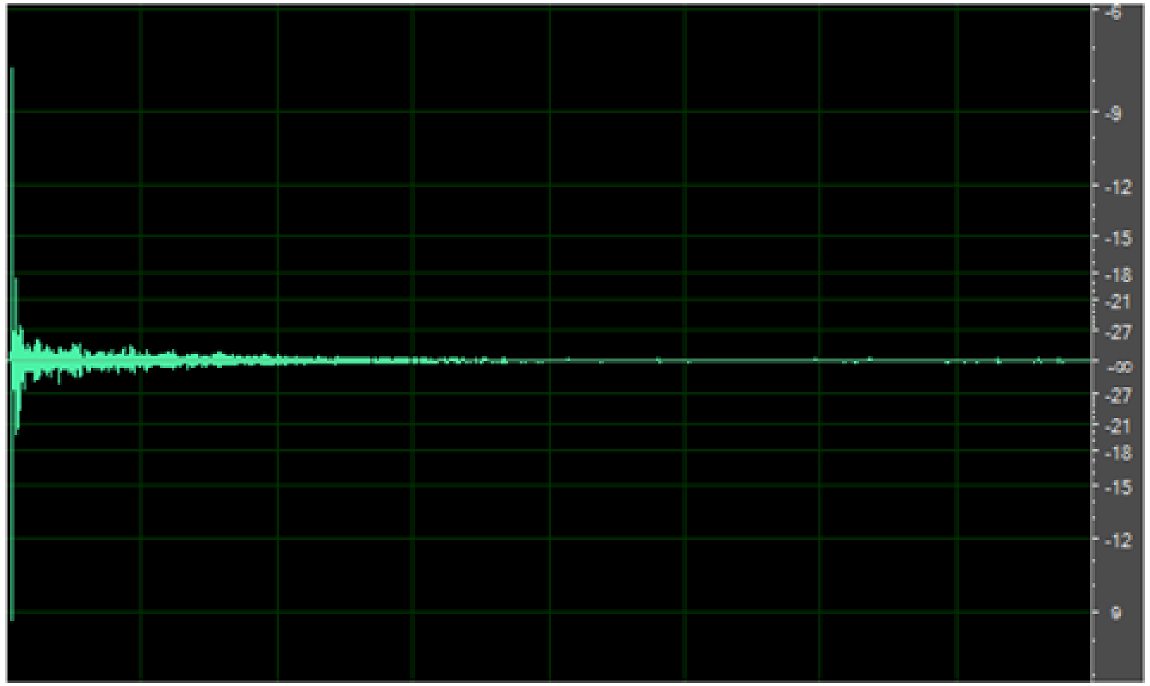
DEVICES & STUDIOS

Test signal – MLS

- Maximum length sequences (MLS) of length exactly 218 - 1 samples
- corresponds to approximately 5.5 seconds at sample rate of 48kHz.
- The MLS signal, after being recorded, is deconvolved into an impulse response via time-reversal convolution between the source MLS and the recording
- Microsoft’s audio lab RT60 ~ 300ms Therefore the IR is trimmed to contain 16000 samples (1/3 second).

Test signal – MLS – deconvolved signal



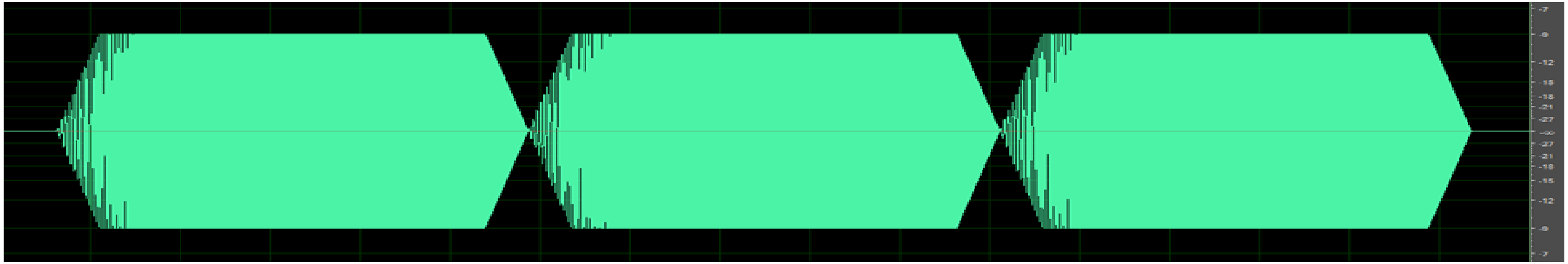


- complete deconvolved signal
- trimmed deconvolved signal

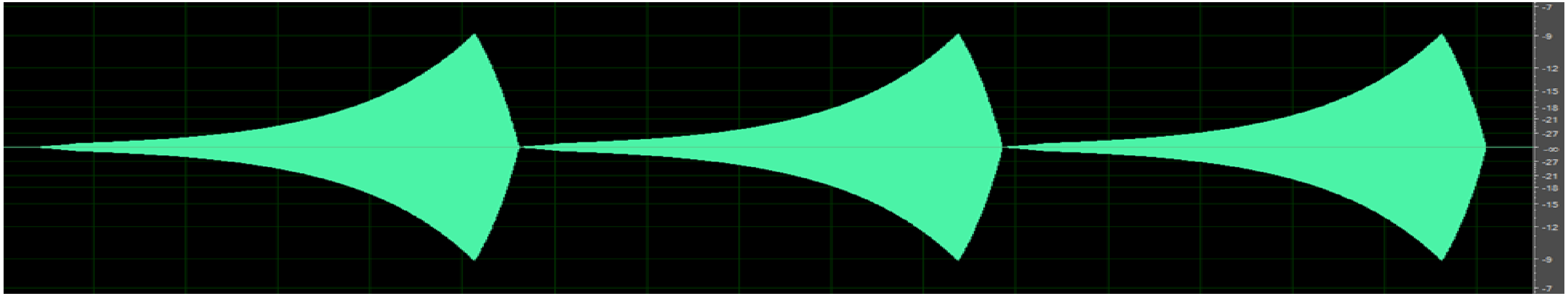
Test signal – Swept Sine

- sine signal of length of 218 samples.
- faded in and out at the end 0.5s in order to reduce distortion of the speakers due to startup transients.
- made continuous at the end points using a 1µHz bi-directional binary search pattern
- the exponential nature of the sweep is corrected by an exponentially growing amplification of the signal in the time domain before deconvolution
- immunity to room reflections and harmonic distortion of the components

Test signal – Swept sine with amplitude



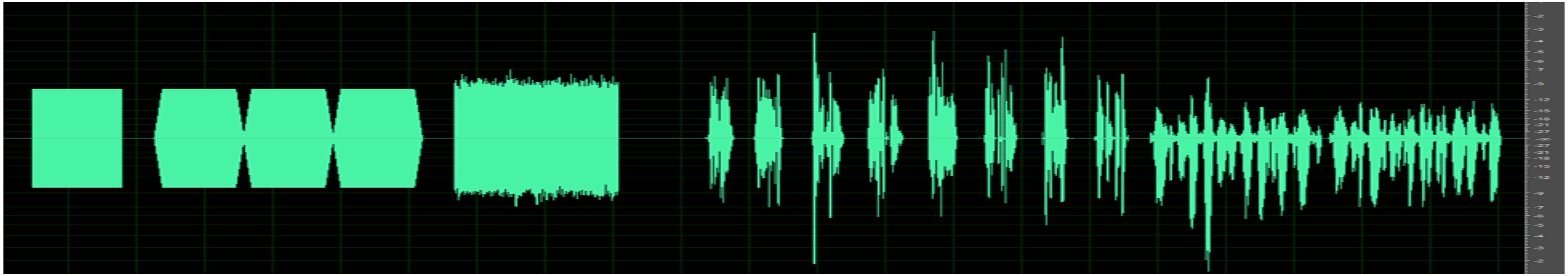
Swept sine signal for test



Amplitude corrected swept sine signal for deconvolution

Test signal – White noise & Speech

- White noise is generated for 1 second.
- The speech signals are inserted into the file following the white noise. The full excitation plus speech signal is shown below.
- MLS, swept sine, noise, real speech, finally p.50 signal

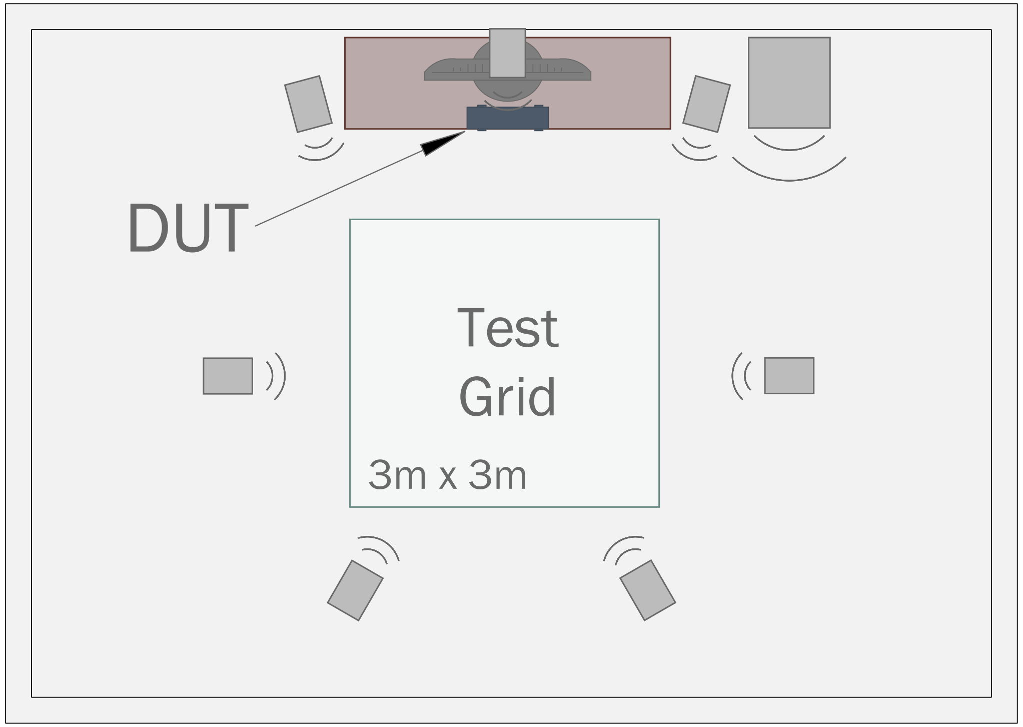


Room and Test setup

DEVICES & STUDIOS

Room Set-up

- 27 positions in a 3x3x3m grid ranging from 1m in front of the device to 3m back and 3m high.
- For each position each of the 7 JBL speakers also play the test signal as the RiR will change for each robot position



Data analysis

- Room Impulse Responses (RIR) are calculated by deconvolution
- RiRs are convolved with the speech section of the test signal, producing synthetic data for each position and configuration. 6510 RiRs were generated for analysis.
- A 8192 point FFT analysis compares the synthetic data to the directly measured data
- the mean-error is computed across 6 bands using the formula below.
  - Narrowband (300-3.4kHz),
  - Wideband (50-7kHz),
  - Super Wideband (50-12kHz),
  - Subwoofer Band (20-120Hz), and
  - Full Band (all points).

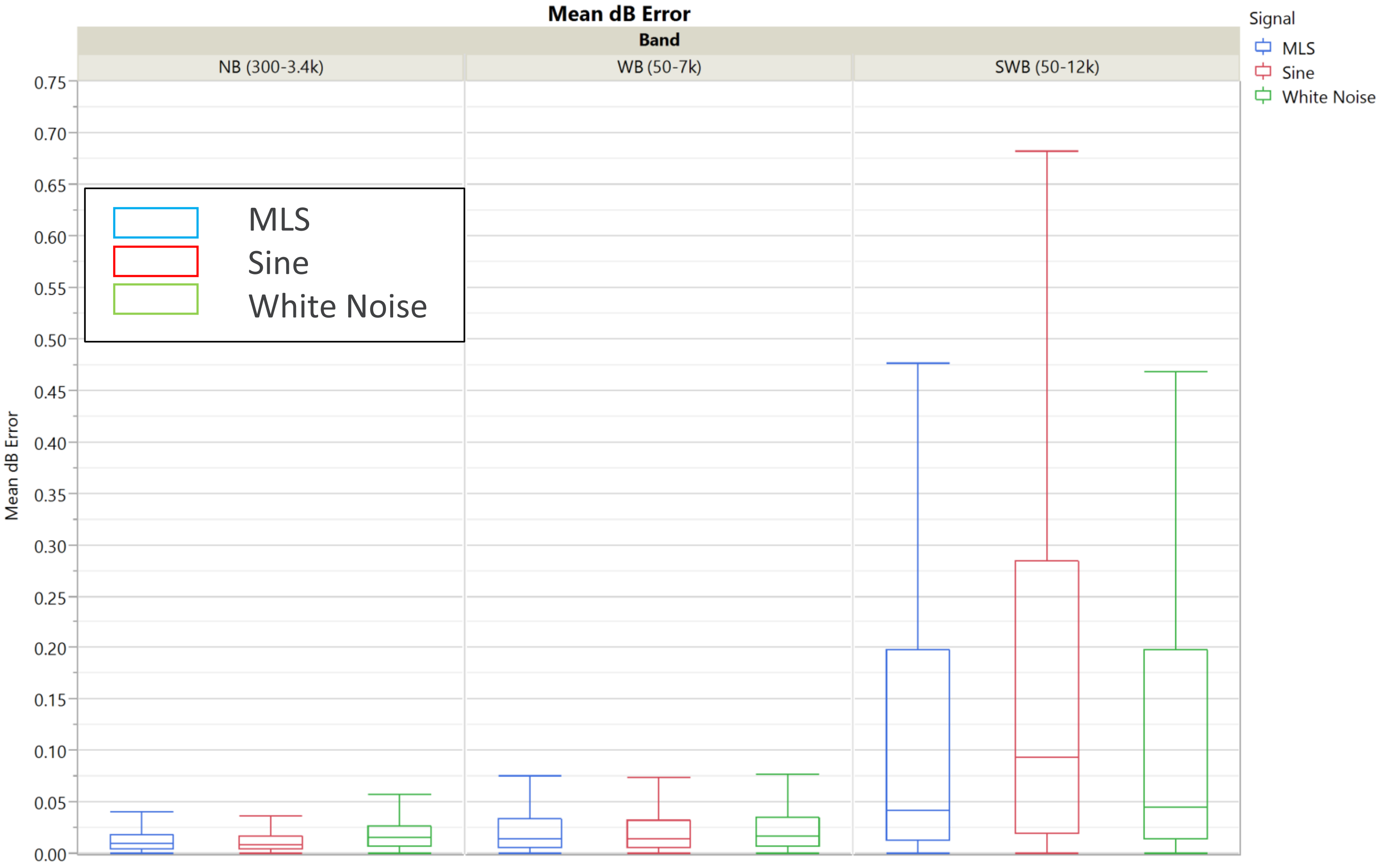
$$Error_{mean} = \sqrt{\frac{\sum_{i=1}^{N} E_i}{N}}$$

Results

DEVICES & STUDIOS

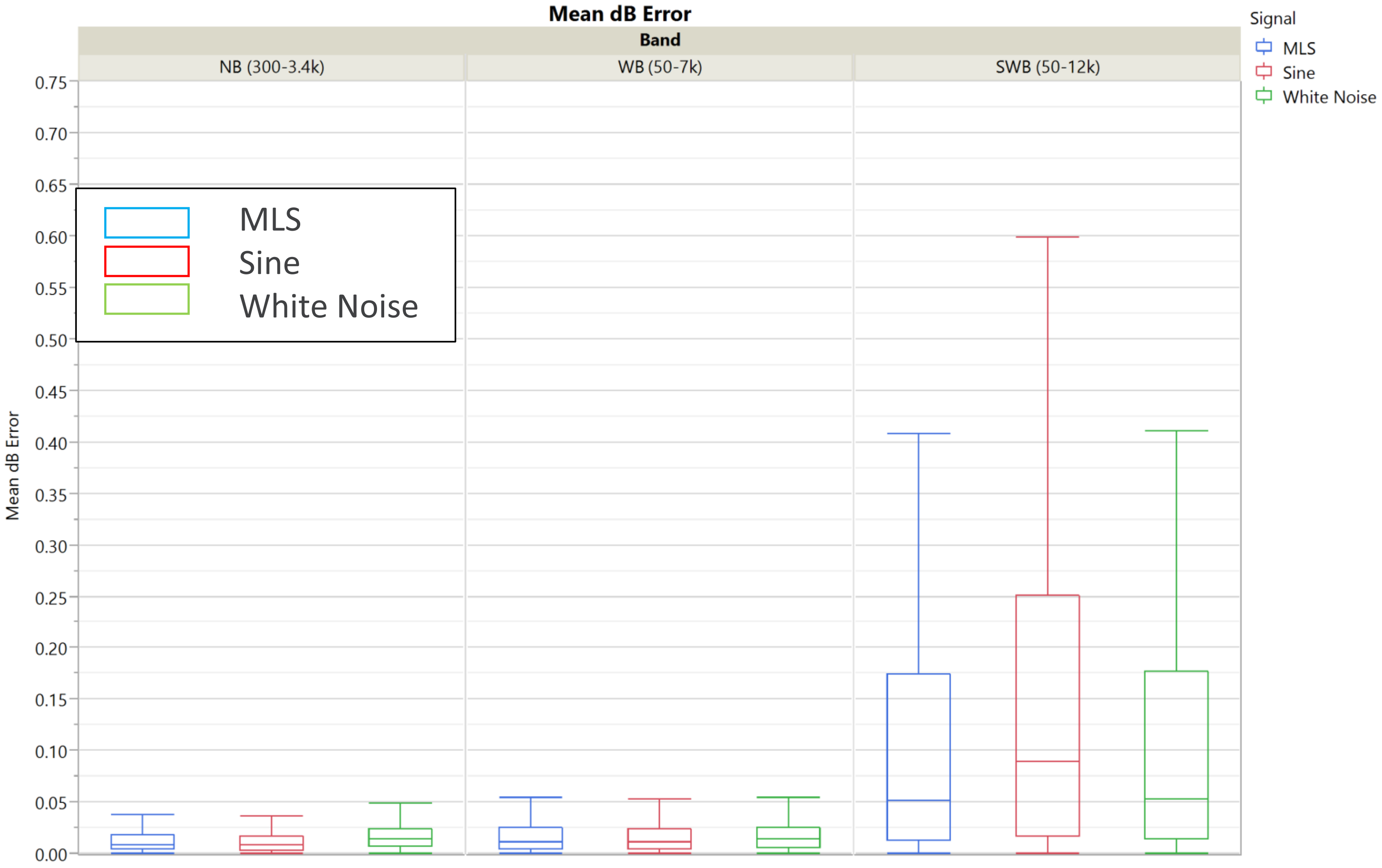
All Mics – All positions

- Overall summary
- Good in telephony bands, max error <0.7dB (outliers hidden)
- Error <0.5dB for MLS (excluding hidden outliers)
- MLS best method in Narrow Band, and is never worse than White Noise



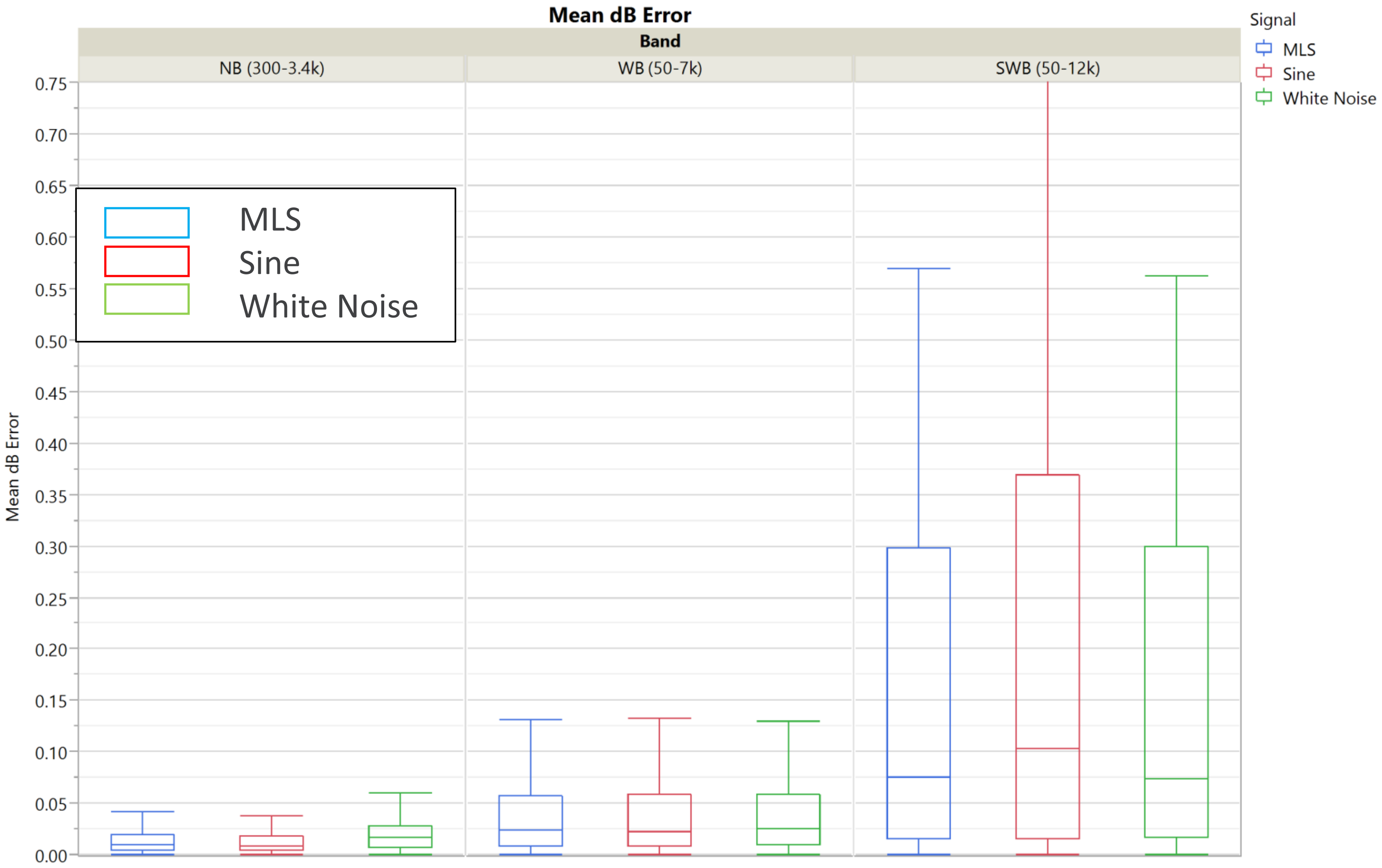
Production Array Mics – All positions

- Device mics only
- Excluding subwoofer
- MLS better in Narrow Band marginally
- Very good performance – average error of 0.05dB for MLS in Super-Wide Band



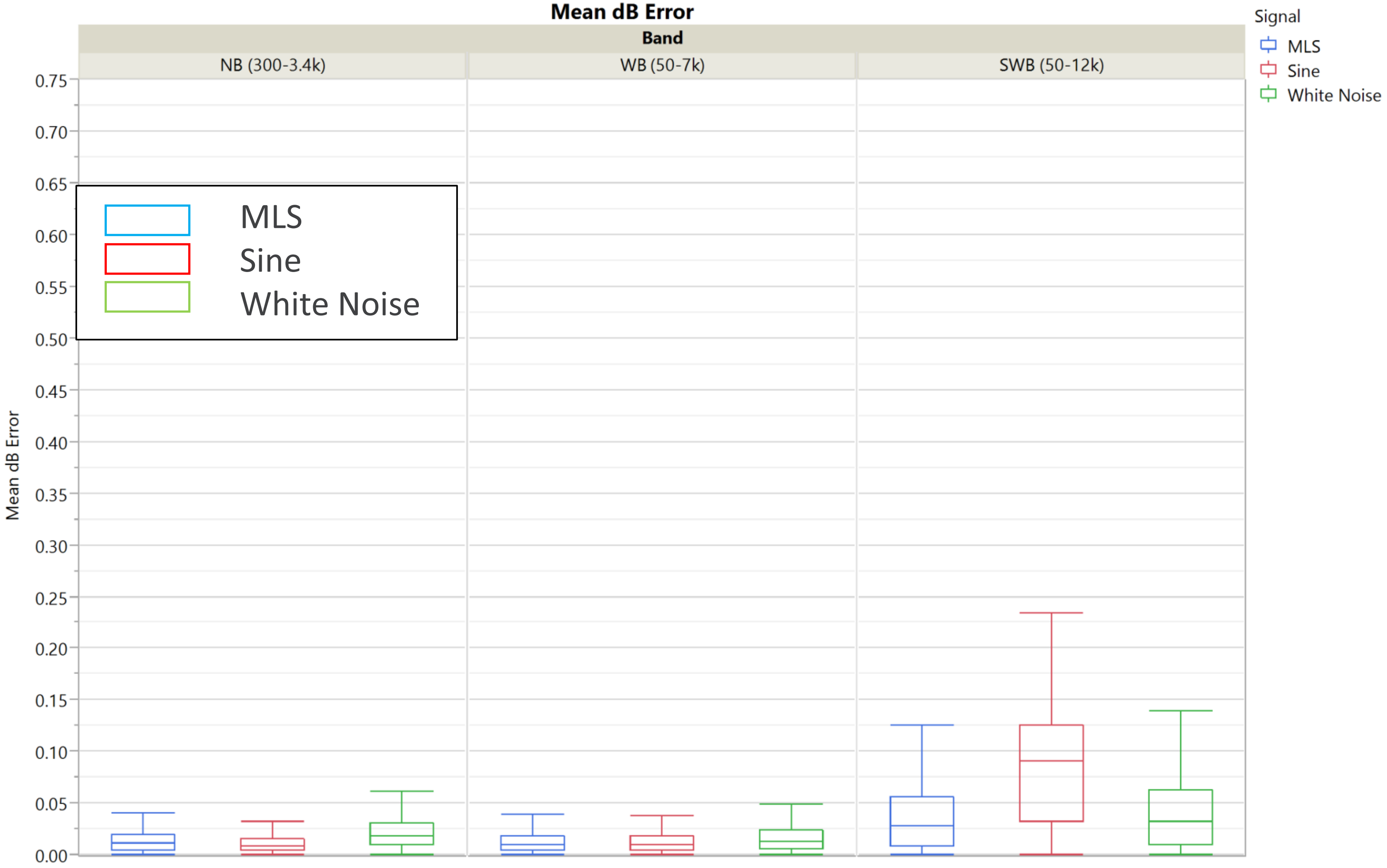
Prototype Mics– All positions

- Device mics only
- Excluding subwoofer
- Sine quite bad, exceeding plot range!



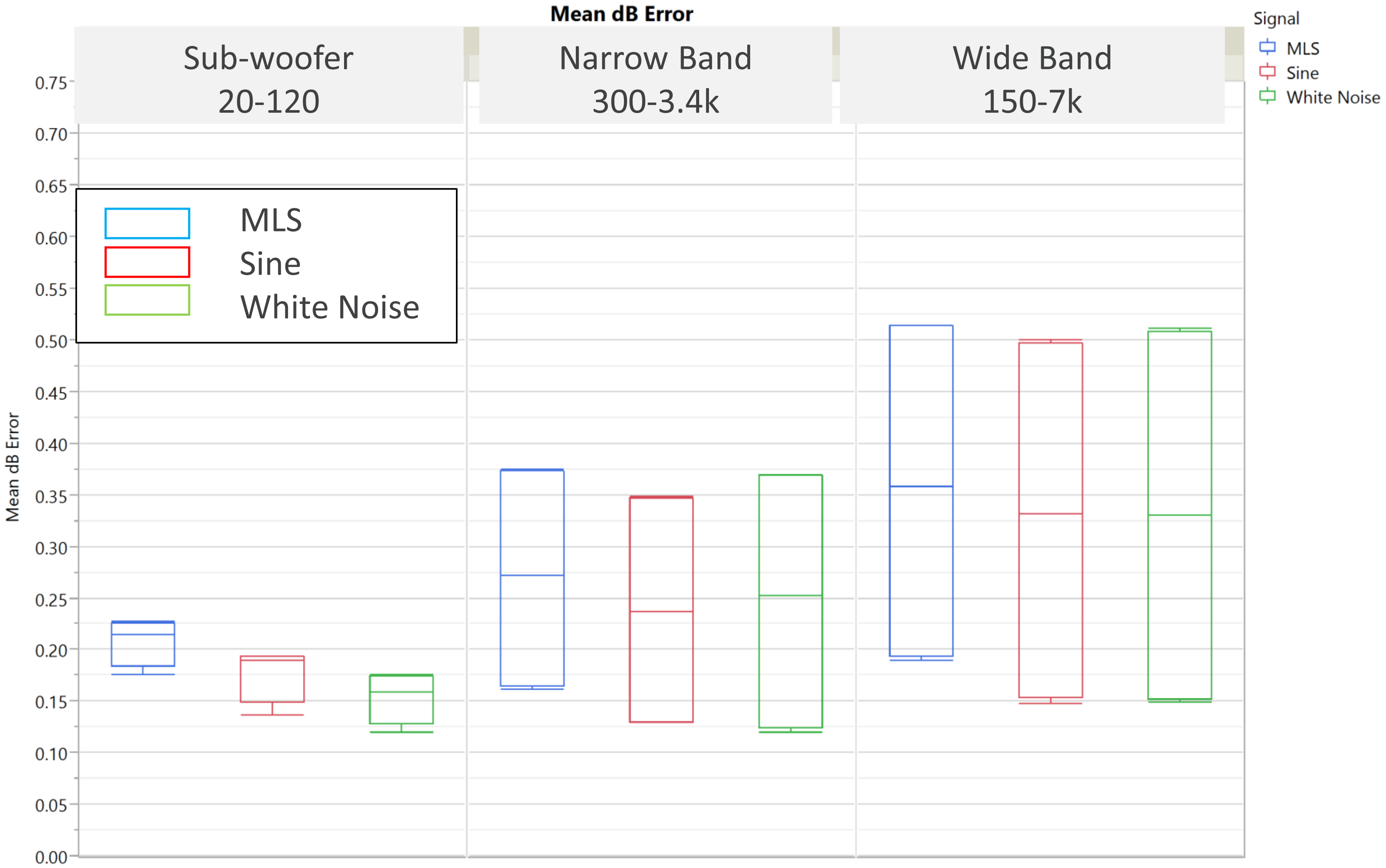
Reference Mics – All positions

- Excluding subwoofer
- MLS better in all categories

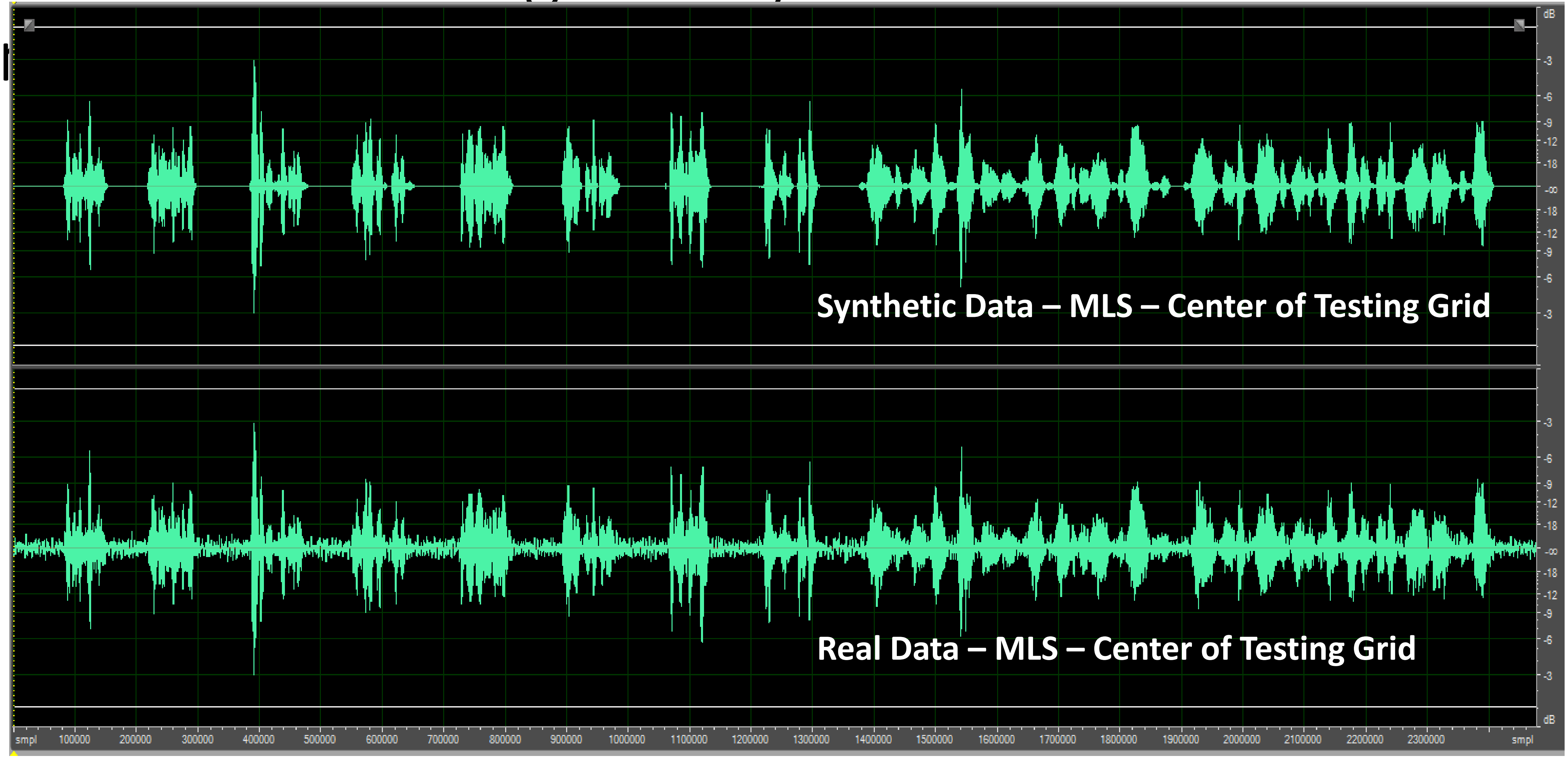


Reference Mics – Subwoofer

- Subwoofer only
- sine is better in NB/WB, noise is better in sub-woofer band
- Still good performance for all signals
- Possibly due to limited spatial averaging



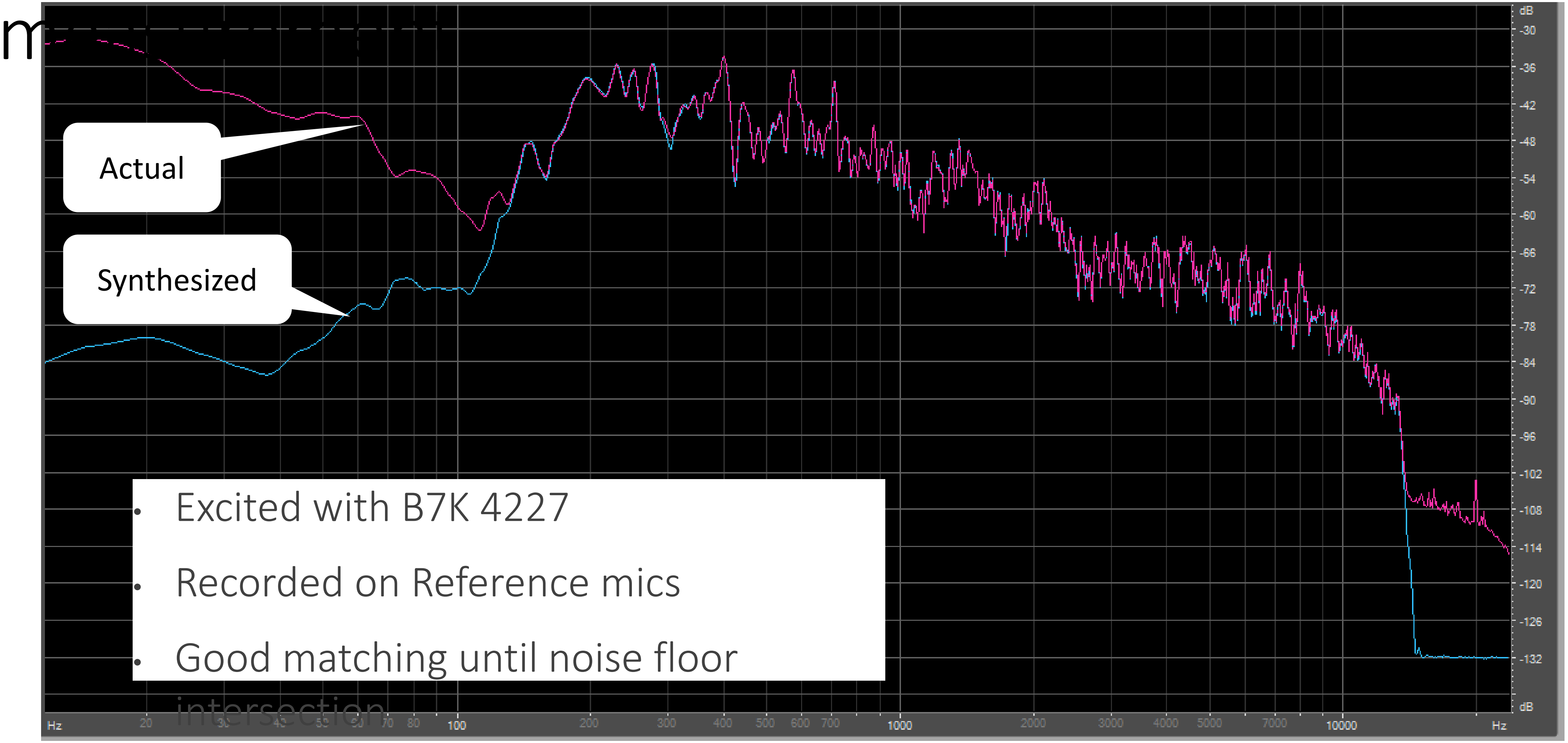
Results – Time signal – synthesized vs actual



Synthetic Data – MLS – Center of Testing Grid

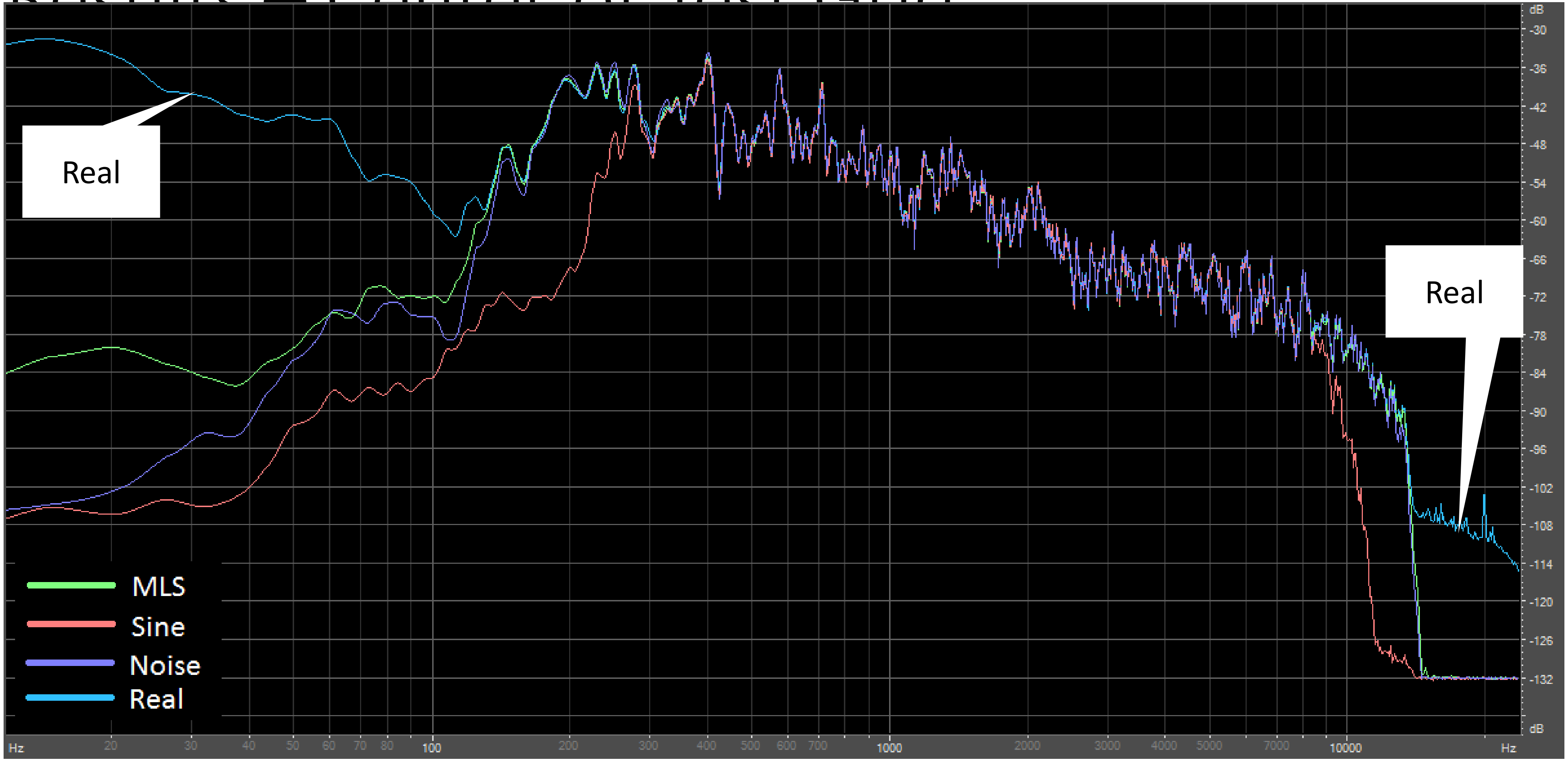
Real Data – MLS – Center of Testing Grid

Results – FFT synthesized vs actual measurement



- Excited with B7K 4227
- Recorded on Reference mics
- Good matching until noise floor intersection

Results – Center of Test Grid



Real

Real

Conclusions

- Of the three signals all provide reasonable performance
- The MLS test signal provides the best overall performance
- We now need to validate this approach on a full corpus run which is planned in the future.

Questions ?

THANK YOU!

DEVICES & STUDIOS

REFERENCES

[1] P. Moquin, K. Venalainen, and D. Florencio, “Determination of room impulse response for synthetic data acquisition,” The Journal of the Acoustical Society of America 136 (4), pp 2265.

[2] A Yellepeddi and D Florencio, “Sparse array-based room transfer function estimation for echo cancellation,” Signal Proc. Letters, IEEE, vol. 21, no. 2, pp. 230–234, 2014.

[3] D. Florencio and Z Zhang, “Maximum a posteriori estimation of room impulse responses,” in Proc. of ICASSP, 2015.

[4] M Song, C Zhang, D Florencio, and H Kang, “Personal 3D audio system with loudspeakers,” in Proc. of ICME, 2010.

[5] Y Huang, J Chen, and J. Benesty, “Immersive audio schemes,” Signal Processing Magazine, IEEE, vol. 28, no. 1, pp. 20–32, Jan 2011.

[6] J Patynen, S Tervo, and T Lokki, “Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses,” The Journal of the Acoustical Society of America, vol. 133, pp. 842, 2013.

[7] H Morgenstern and B Rafaely, “Analysis of acoustic mimo systems in enclosed sound fields,” in Proc. of ICASSP, 2012.

[8] S Goetze, E Albertin, M Kallinger, A Mertins, and K Kammeyer, “Quality assessment for listening-room compensation algorithms,” in Proc. of ICASSP, 2010.

[9] F Xiong, J Appell, and S Goetze, “System identification for listening-room compensation by means of acoustic echo cancellation and acoustic echo suppression filters,” in Proc. Of ICASSP, 2012.

[10] Y Rui, D Florencio, W Lam, and J Su, “Sound source localization for circular arrays of directional microphones,” in Proc. of ICASSP, 2005.

[11] S Tervo, J Patynen, and T Lokki, “Acoustic reflection localization from room impulse responses,” Acta Acustica, vol. 98, no. 3, pp. 418–440, 2012.

[12] S Tervo and T Tossavainen, “3D room geometry estimation from measured impulse responses,” in Proc. of ICASSP, 2012.

[13] F Ribeiro, D Florencio, D Ba, and C Zhang, “Geometrically constrained room modeling with compact microphone arrays,” Audio, Speech, and Language Processing, IEEE Trans. on, vol. 20, no. 5, pp. 1449–1460, 2012.

[14] D Ba, F Ribeiro, C Zhang, and D Florencio, “L1 regularized room modeling with compact microphone arrays,” in Proc. Of ICASSP, 2010.

[15] F Ribeiro, D Florencio, P Chou, and Z Zhang, “Auditory augmented reality: Object sonification for the visually impaired,” in Proc. of MMSP, 2012.

[16] J Klein, M Pollow, P Dietrich, and M Vorl¨ander, “Room impulse response measurements with arbitrary source directivity,” in 40th Italian (AIA) Annual Conference on Acoustics, 2013.

[17] R Mignot, L Daudet, and F Ollivier, “Room reverberation reconstruction: Interpolation of the early part using compressed sensing,” Audio, Speech, and Language Processing, IEEE Trans. on, vol. PP, no. 99, pp. 1–1, 2013.

[18] G Turin, “An introduction to matched filters,” Information Theory, IRE Trans. on, vol. 6, no. 3, pp. 311–329, 1960.

[19] J Vanderkooy, “Aspects of mls measuring systems,” Journal of the Audio Engineering Society, vol. 42, no. 4, pp. 219–231, 1994.

[20] M Vorlander and M Kob, “Practical aspects of mls measurements in building acoustics,” Applied Acoustics, vol. 52, no. 3, pp. 239–258, 1997.

[21] I Mateljan, “Signal selection for the room acoustics measurement,” in Proc. of WASPAA, 1999.

[22] S Schimmel, M Muller, and N Dillier, “A fast and accurate shoebox room acoustics simulator,” in Proc. of ICASSP, 2009.

[23] D Florencio and H Malvar, “Multichannel filtering for optimum noise reduction in microphone arrays,” in Proc. of ICASSP, 2001.

[24] F Ribeiro, D Florencio, C Zhang, andMSeltzer, “CrowdMOS: An approach for crowdsourcing mean opinion score studies,” in Proc. of ICASSP, 2011.

[25] F Ribeiro, D Florencio, and V Nascimento, “Crowdsourcing subjective image quality evaluation,” in Proc. of ICIP, 2011.

[26] F Ribeiro, C Zhang, D Florencio and D Ba, “Using reverberation to improve range and elevation discrimination for small array sound source localization,” Audio, Speech, and Language Processing, IEEE Transactions on, 18(7), pp 1781-1792, 2010.

[27] Y Rui, and D Florencio, “New direct approaches to robust sound source localization,” in Proc. of ICME, 2003.

[28] F. Ribeiro, D Florencio and V Nascimento, “Crowdsourcing subjective image quality evaluation,” in Proc. of ICIP, 2011.

[29] D. Florencio and R. Schafer, “Perfect reconstructing nonlinear filter banks,” in Proc. of ICASSP, 1996.

[30] A Conceicao, J Li and D Florencio, “Is IEEE 802.11 ready for VoIP?,” in Proc. of MMSP, 2006.

[31] F Ribeiro, D Ba, C Zhang and D Florencio, “Turning enemies into friends: using reflections to improve sound source localization,” in Proc. of ICME, 2010.

[32] M Song, C Zhang, D Florencio and H Kang, “An interactive 3-d audio system with loudspeakers,” Multimedia, IEEE Transactions on 13 (5), 844-855, 2011.