PARAMETRIC HEAR THROUGH FOR AUGMENTED REALITY AUDIO

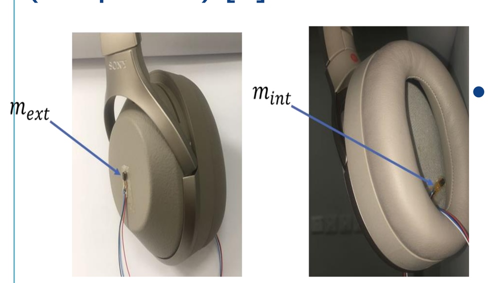
Rishabh Gupta[1], Rishabh Ranjan[1], Jianjun He[2], Woon-Seng Gan[1];

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore; [2]Maxim Integrated, San Jose, United States;

PAPER #3160

Motivation

- Augmented Reality (AR) audio needs real sounds unaltered and fusion of real and virtual sounds
- Hear-through (HT) used for real sounds and binaural rendering for virtual sounds
- Different playback devices have certain advantages and disadvantages for AR audio:

Playback devices	Advantages	Disadvantages
Open back headphones [1] 	• Passive hear-through • Hi-fidelity virtual playback • Open ear canal listening	• Comb filtering at high frequencies • Need to embed pinna cues in EQ
Closed in-ear headphones [2] 	• Individual Pinna cues preserved naturally • Hi-fidelity virtual playback	• Loose fitting -> comb filtering • Compensate for ear canal occlusion
Closed back headphones (Proposed) [3] 	• Open ear canal listening • Less/No comb filtering effects • Complete sound-field control	• Need to embed pinna cues in EQ • Compensate for headphone isolation
Open ear emitter 	• No need for HT EQ • Pinna cues are preserved • Open ear canal listening	• Poor isolation and bass • Leakage effects

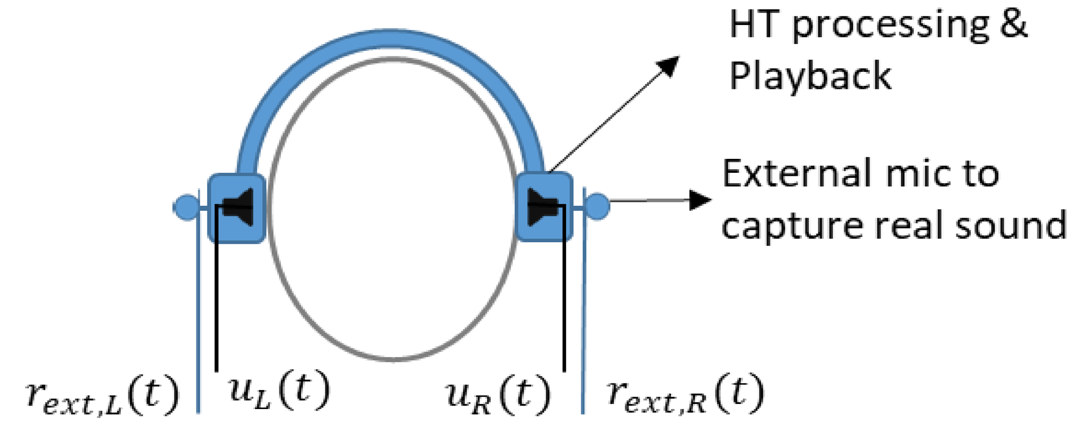
- Study by Gupta *et al*. [3] shows directional HT EQ filters outperform average HT and unequalized HT EQ for frequencies >2 kHz
- However, past studies have dealt with HT for single sound source
- This study uses parametric approach for multiple sound sources
- Benefits of parametric approach [4]:
  - Independent processing of spectral coefficients obtained for different sources
  - Differences in magnitude and phase in time-frequency domain similar to variation of human spectral cues

References

1. R. Ranjan and W. S. Gan, "Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1988-2002, 2015
2. A. Härmä, et al., "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, pp. 618–639, 2004
3. R. Gupta, R. Ranjan, J. He, and W. S. Gan "On the use of closed back headphones for active hear-through equalization in augmented reality applications" in *Proc. AES AVAR Conference*, Redmond, USA, Aug 2018
4. V. Pulkki, S. Delikaris-Manias, and A. Politis (Edited), *Parametric time-frequency domain spatial audio*, Wiley, 2018

Proposed system

AR audio headset prototype



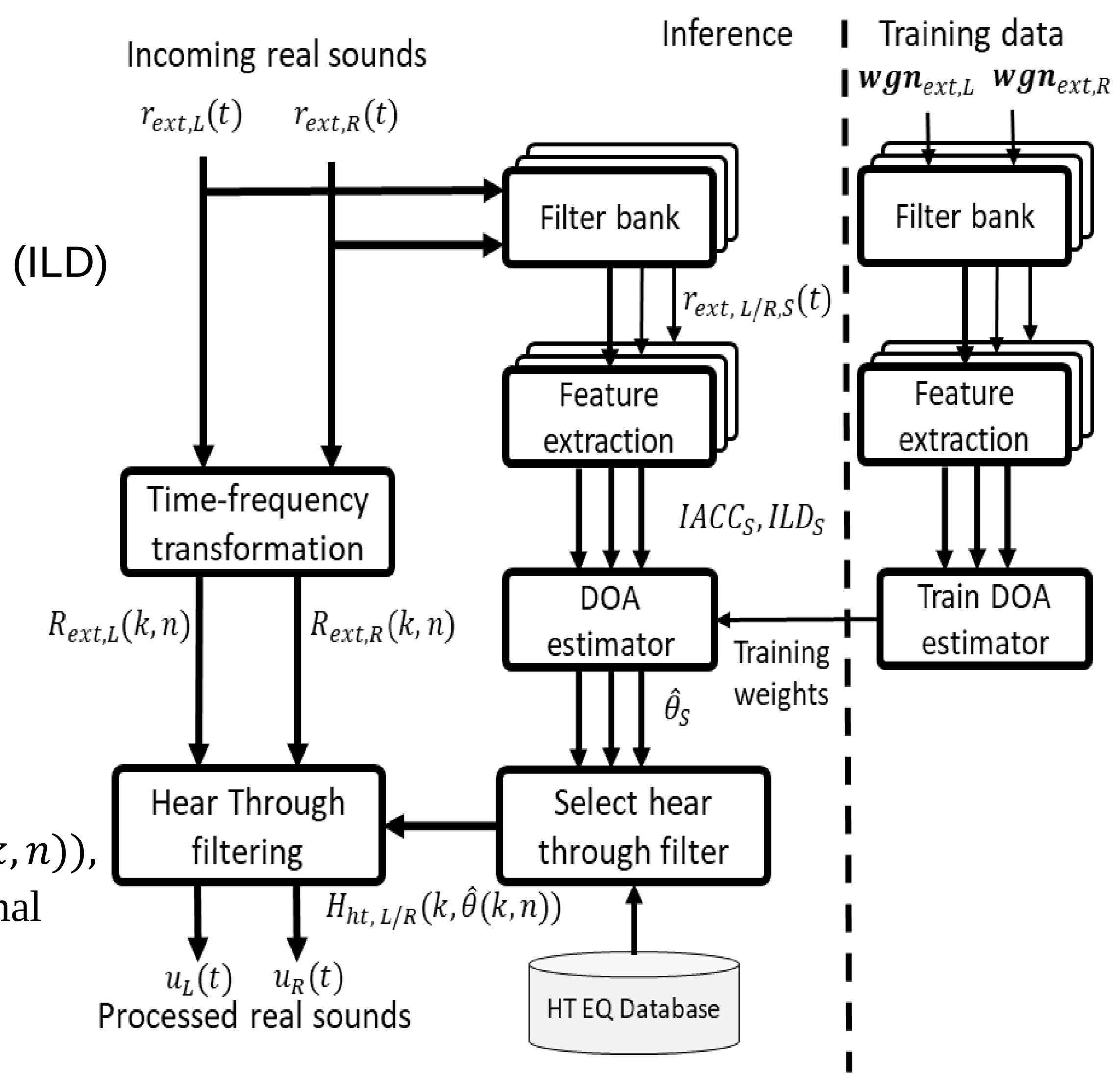
- AR audio headset prototype shown in previous study [1]
- External microphones denoted by $r_{ext,L/R}(t)$ and processed ear signals by $u_{L/R}(t)$
- Aim of the system: DoA estimation and directional HT filtering in time-frequency domain

Features: Interaural Cross Correlation (IACC) and Interaural Level Difference (ILD)

$$IACC_s(\Delta t) = \frac{\sum r_{ext,L,s}(t) r_{ext,R,s}(t-\Delta t)}{\sqrt{\sum r_{ext,L,s}^2(t) \sum r_{ext,R,s}^2(t-\Delta t)}}$$

$$ILD_s = 10 log_{10}(\frac{\sum r_{ext,L,s}^2(t)}{\sum r_{ext,R,s}^2(t)})$$

$U_{L/R}(k,n) = R_{ext,\,L/R}(k,n) H_{ht,\,L/R}(k, \hat{\theta}(k,n))$,
where $U_{L/R}(k,n)$ is the processed real signal



DoA estimation using neural network

- NN based model tested for frontal source directions (-90° to 90°)
- Model trained using 10 s white noise filtered with HRTFs measured at external microphone
- Simple network topology:

Input layer	Hidden layer	Output layer
128 nodes with 102 dimension input vector	Single hidden layer with 128 nodes	13 nodes
Training parameters: Optimizer: adam; Learning rate: 0.001; Batch size: 25 samples; maximum 100 epochs		

Parametric hear-through processing

- EQ filters pre computed to cover entire 360° at resolution of 15° called idealHT
- Zone based EQ filters for three zones (GroupedHT): frontal (-60° to 60°), lateral (60°-120°, -60° to -120°), and rear (120° to 180°, -120° to -180°)
- AvgHT: Average across all directional EQ filters
- Filtering: STFT of captured signal $R_{ext,L/R}(k,n)$ filtered by sub-band directional filter $H_{ht,L/R}(k, \hat{\theta}(k,n))$ chosen for each direction

Results and analysis

Signal synthesis

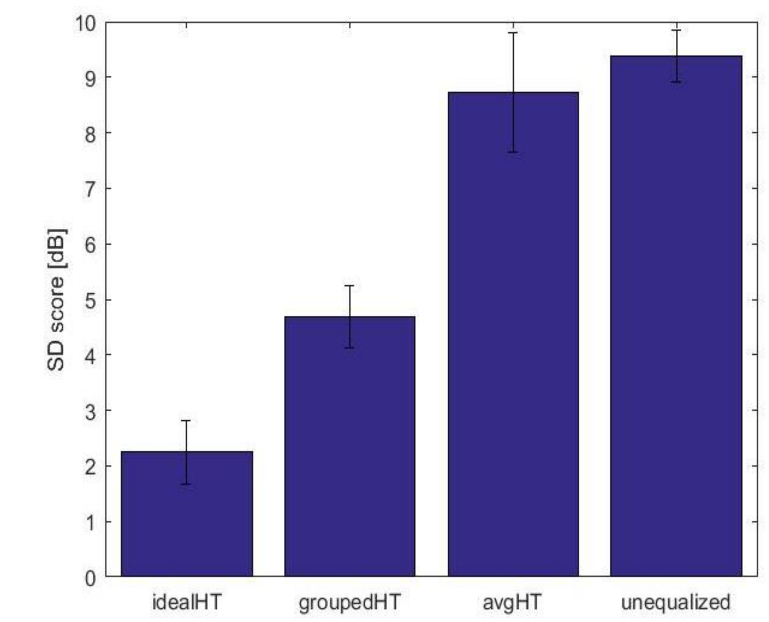
- 2 uncorrelated pink noise signals of 2s each filtered by 3 bandpass filters: 0.1-1 kHz (low), 1-5 kHz (middle), and 5-16 kHz (high)
- Obtained signals filtered with impulse response for two direction pairs: (0°, 30°) and (-15°, 75°)
- All combinations taken (total 12, 6 each for overlapping and non overlapping frequency bands)
- Real sound: broadband music and narrowband speech signal (4s each) convolved with 2 directions chosen randomly from set of 13 azimuthal positions (total 156 soundtracks)

Hear-through equalization results

$$SD_{combined} = \frac{SD_L P_L + SD_M P_M + SD_H P_H}{P_L + P_M + P_H}$$

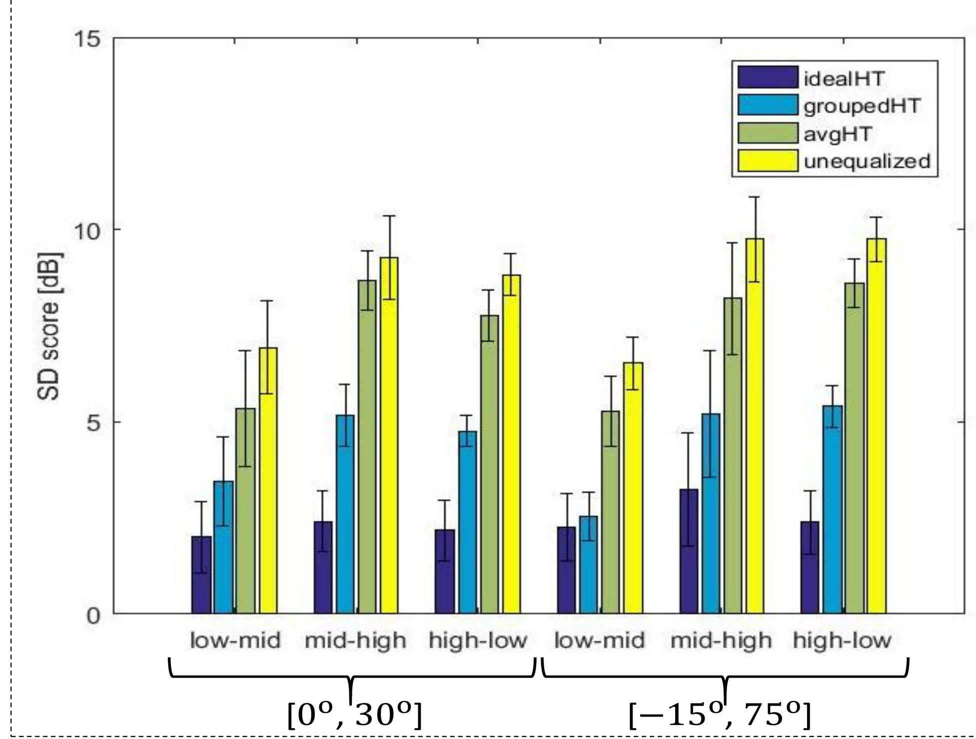
$$SD_{L/M/H} = \sqrt{\frac{1}{K_{L/M/H}} \sum_{K_{L/M/H}} \left| 10 log \frac{R_{ref,L}^2(k) + R_{ref,R}^2(k)}{\hat{R}_{ref,L}^2(k) + \hat{R}_{ref,R}^2(k)} \right|^2}$$

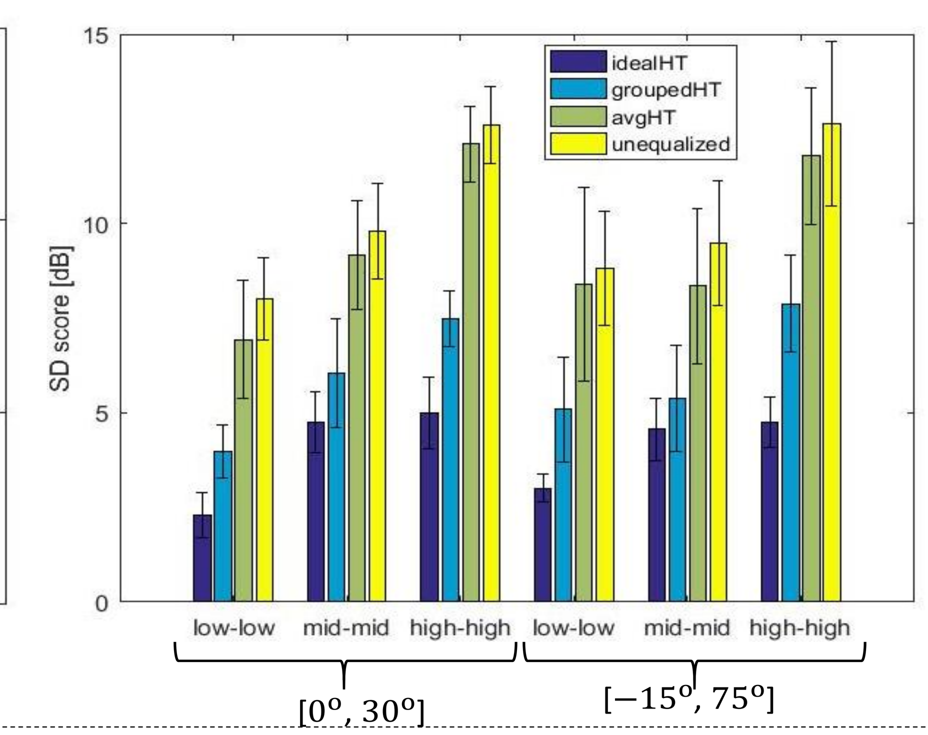
$$P_{L/M/H} = \sum_{K_{L/M/H}} \left(|R_{ext,L}(k)|^2 + |R_{ext,R}(k)|^2 \right)$$



$R_{ref,L/R}(k)$ is the frequency spectrum of open ear reference, $\hat{R}_{ref,L/R}(k)$ is the processed real sound recorded at the ear, and $K_{L/M/H}$ denotes the total number of frequency bins in each frequency band, respectively.

- All HT filters perform better in low frequencies (< 1 kHz)
- Performance of HT filters: IdealHT > groupedHT > AvgHT > UnequalizedHT
- Lowest SD values for idealHT (< 5 dB for all cases)





Conclusion and future work

Conclusions

- Directional EQ filters (IdealHT and groupedHT) show close match to reference for all cases, including real and overlapping sounds
- NN based DoA approach using IACC and ILD features shows good localization performance

Future work

- Real time system with NN based DoA estimation and parametric HT filtering
- Sound classification and HT for diffuse sounds

Acknowledgement

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060