HEAD ORIENTATION COMPENSATION WITH VIDEO-INFORMED SINGLE CHANNEL SPEECH ENHANCEMENT

Soumitro Chakrabarty, Deepth Pilakeezhu, Emanuël Habets

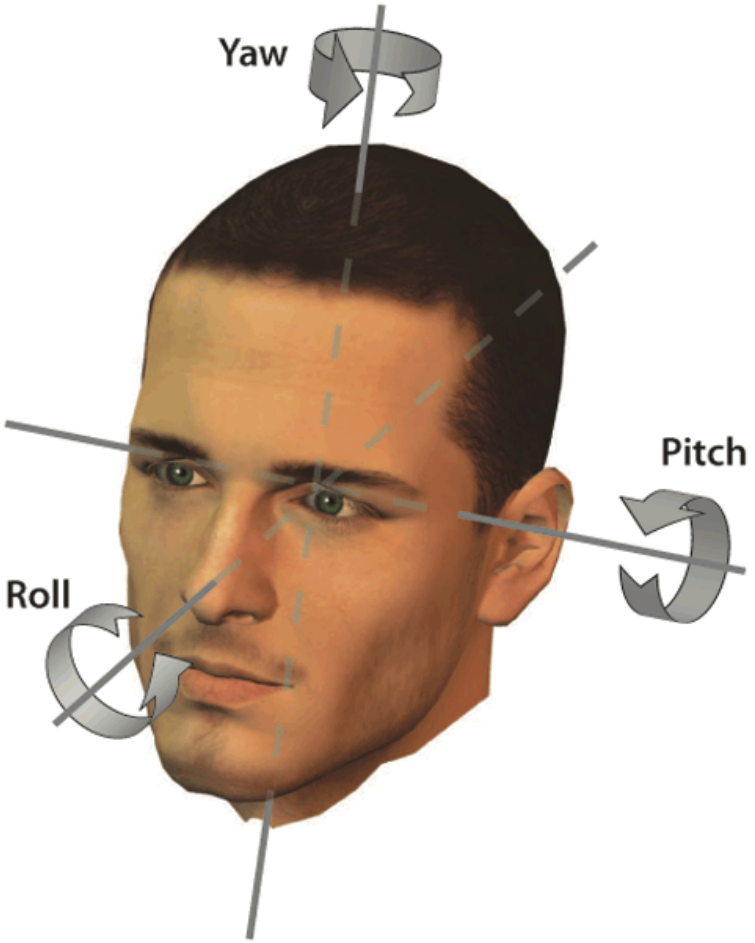
IWAENC 2016

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

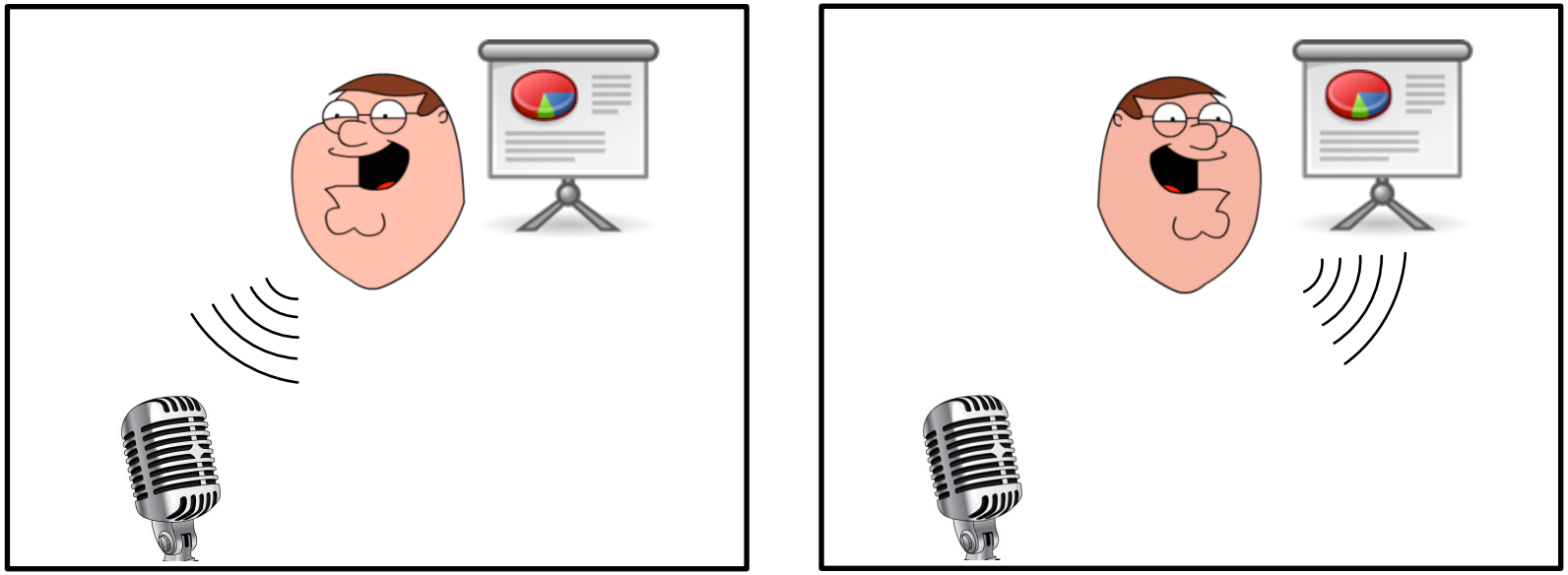
Fraunhofer IIS

Head Orientation

- Three degrees of freedom for head movement
- Propagation of sound waves in the horizontal plane
- Orientation of interest:
  YAW

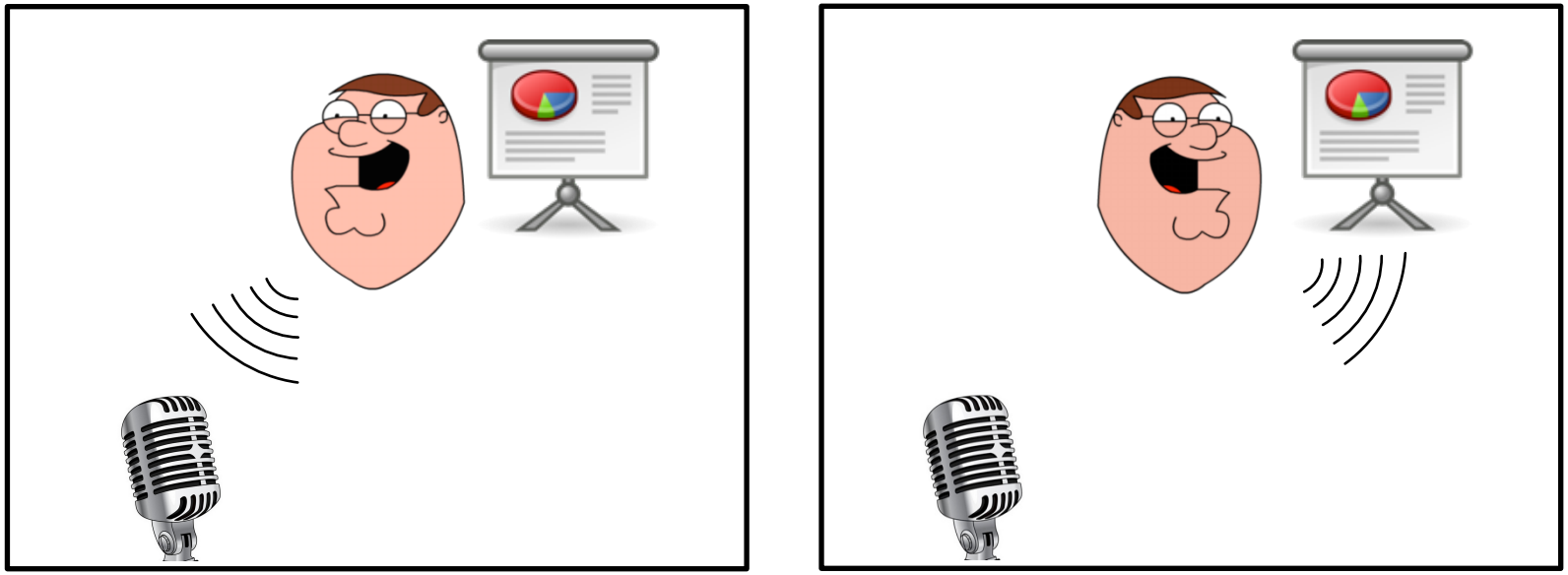


Motivation



- Similar problems can be witnessed in hands-free communication systems, speech based human machine interfaces etc.

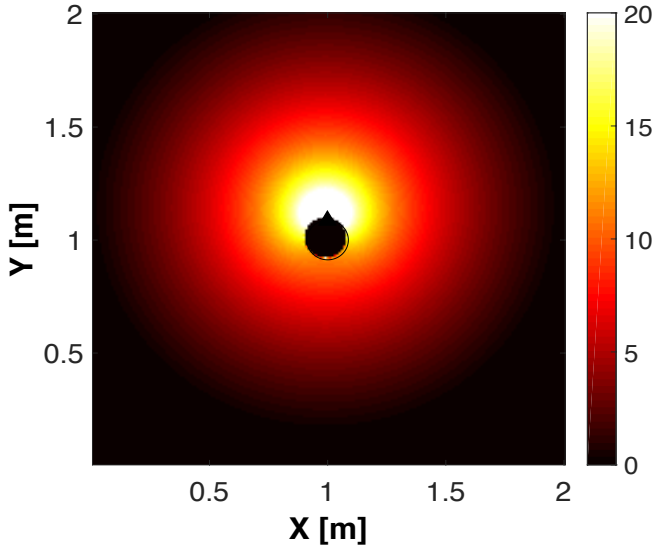
Motivation



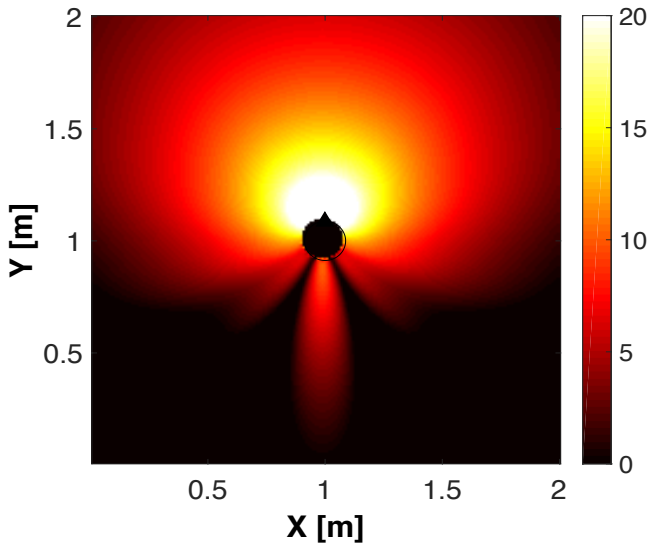
- Human speakers radiate sound primarily to the front [1]
- Also, the radiation pattern is frequency dependent [1]

[1] H. K. Dunn and D. W. Farnsworth, “Exploration of pressure field around the human head during speech,” Journal Acoust. Soc. of America, vol. 10, no. 1, pp. 83–83, 1938.

Sound Radiation Pattern
Frequency Dependency



(a) 100 Hz

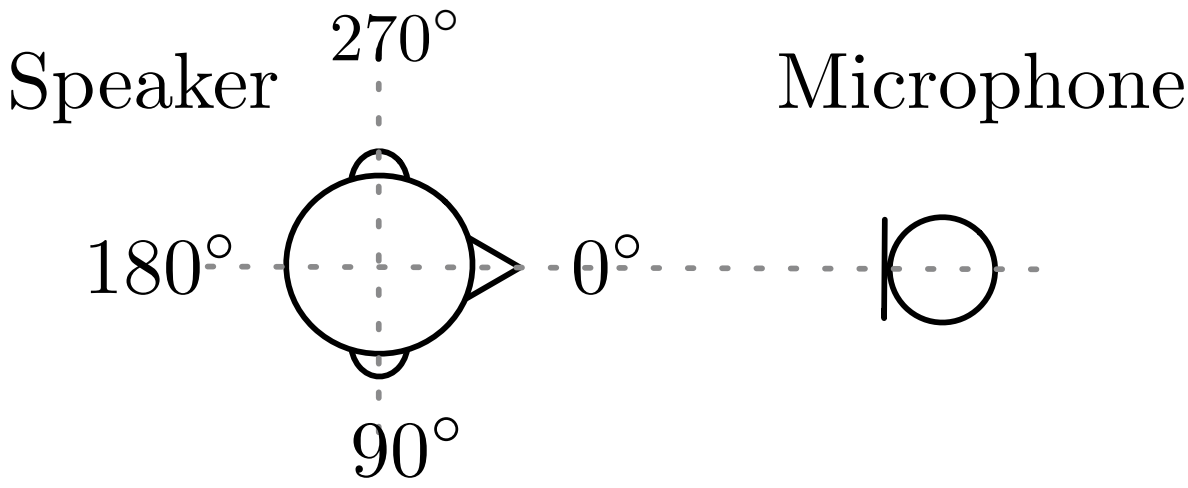


(b) 3 kHz

- Generated using spherical microphone impulse response generator (SMIRgen) [2]

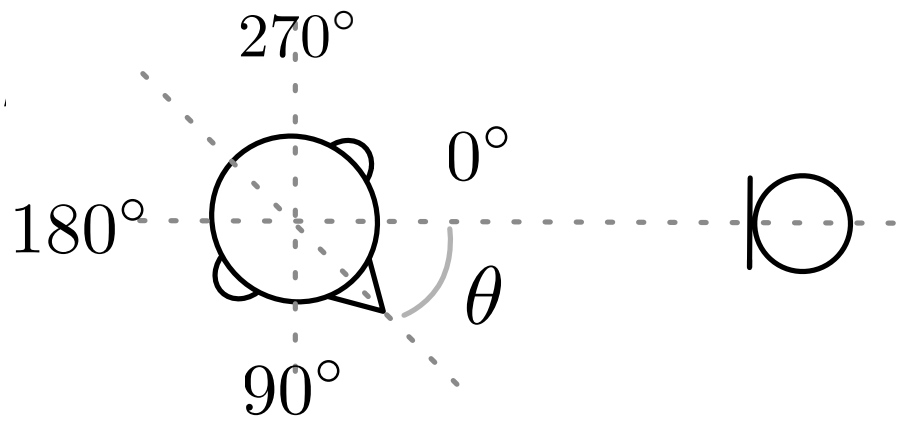
[2] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, “Rigid sphere room impulse response simulation: Algorithm and applications,” Journal Acoustical Society of America, vol. 132, pp. 1462, 2012.

Speaker-Microphone Setup



- Aim: Compensate for the reduction in sound energy due to the relative orientation of the speaker with respect to the microphone, while attenuating the noise.

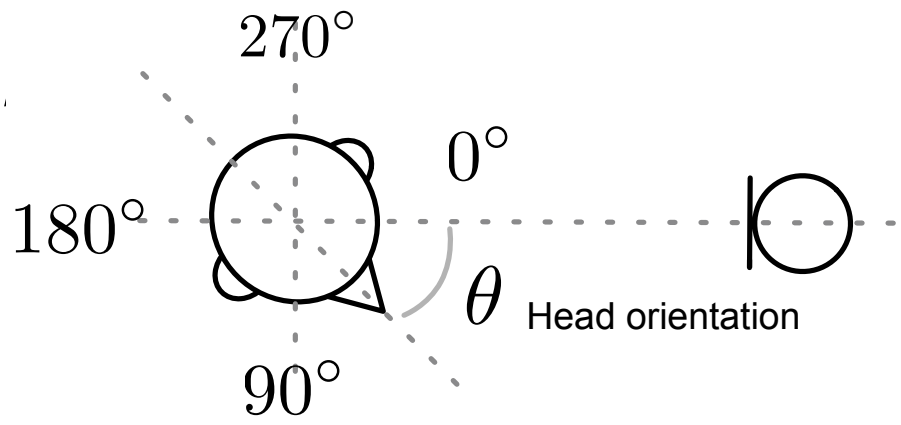
Problem Formulation



- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

Problem Formulation

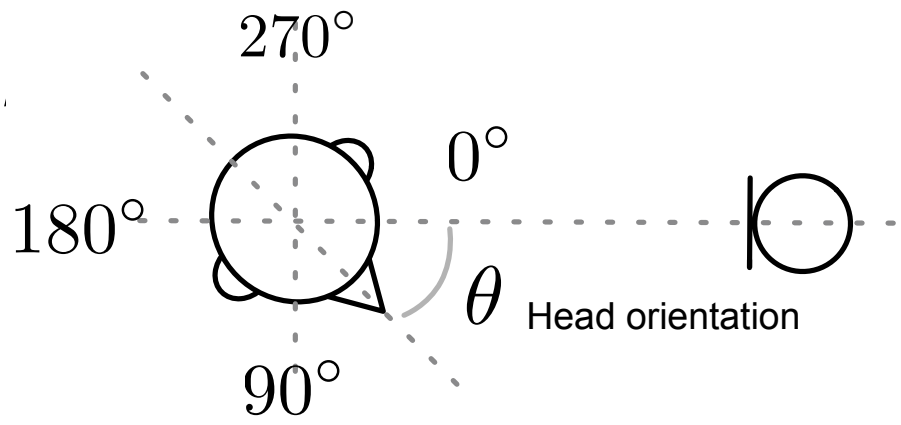


- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

Orientation-dependent ATF

Problem Formulation

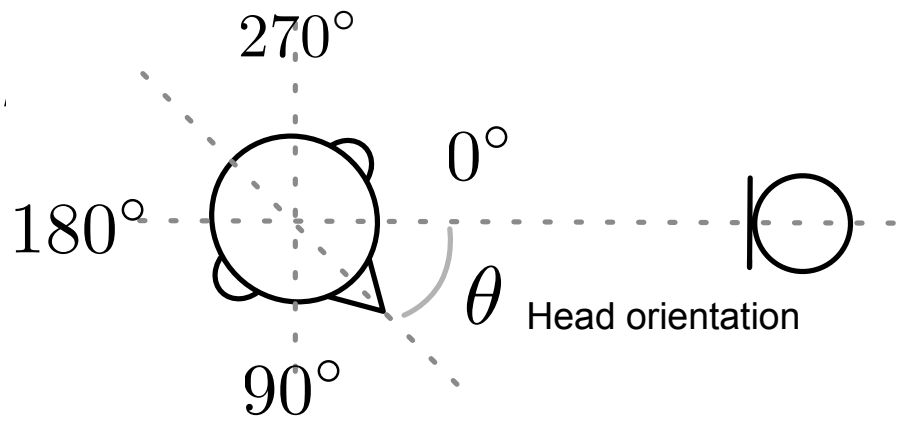


- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

Source signal

Problem Formulation

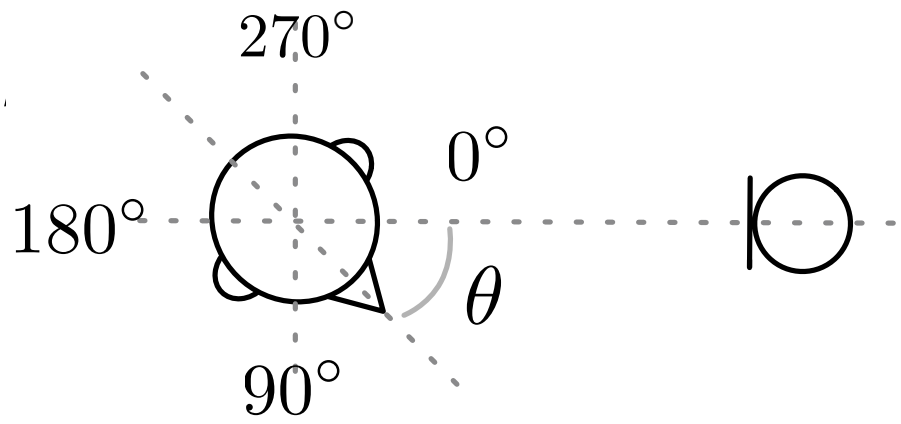


- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

Noise

Problem Formulation



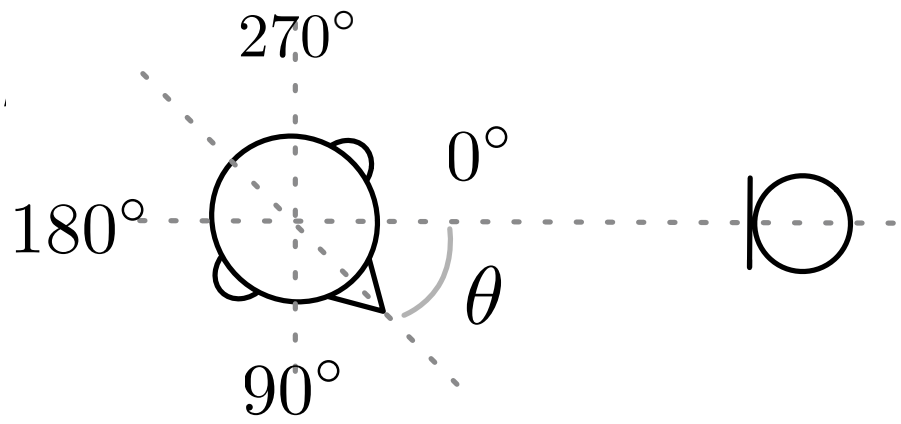
- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

- Can be formulated as

$$Y(n,k) = A(\theta,k)X(n,k) + V(n,k)$$

Problem Formulation



- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

- In terms of attenuation

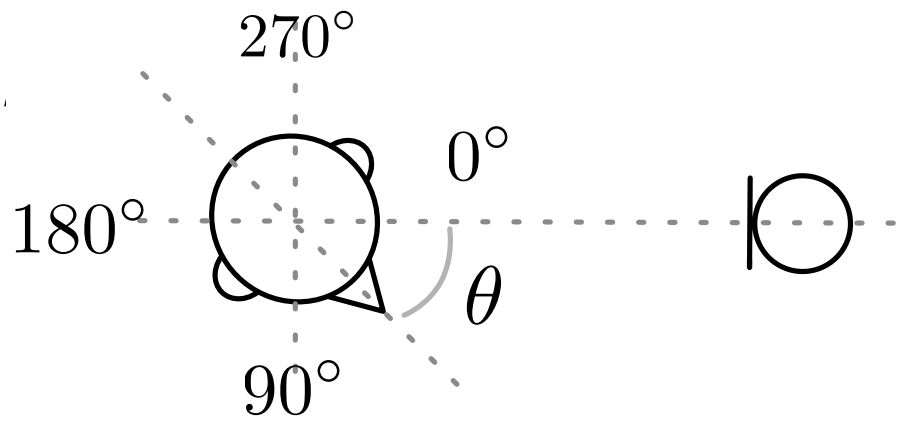
$$Y(n,k) = A(\theta,k)X(n,k) + V(n,k)$$

with

$$A(\theta,k) = \frac{H(\theta,k)}{H(0,k)} \quad \text{and} \quad X(n,k) = H(0,k)S(n,k)$$

Attenuation factor

Problem Formulation



- Microphone signal (STFT Domain)

$$Y(n,k) = H(\theta,k)S(n,k) + V(n,k)$$

- In terms of attenuation

$$Y(n,k) = A(\theta,k)X(n,k) + V(n,k)$$

with

$$A(\theta,k) = \frac{H(\theta,k)}{H(0,k)}$$ Attenuation factor

and

$$X(n,k) = H(0,k)S(n,k)$$ Desired Signal

Orientation Compensation Filter
Derivation

- Assuming all signal components to be independent

$$\phi_Y(n,k) = |A(\theta,k)|^2 \phi_X(n,k) + \phi_V(n,k)$$

Orientation Compensation Filter
Derivation

- Assuming all signal components to be independent

$$\phi_Y(n,k) = |A(\theta,k)|^2 \phi_X(n,k) + \phi_V(n,k)$$

Microphone signal PSD — Desired signal PSD — Noise PSD

Orientation Compensation Filter
Derivation

- Assuming all signal components to be independent

$$\phi_Y(n,k) = |A(\theta,k)|^2 \phi_X(n,k) + \phi_V(n,k)$$

- Estimate of desired signal

$$\hat{X}(n,k) = W(\theta,k) Y(n,k)$$

Orientation Compensation Filter
Derivation

- Assuming all signal components to be independent

$$\phi_Y(n,k) = |A(\theta,k)|^2 \phi_X(n,k) + \phi_V(n,k)$$

- Estimate of desired signal

$$\hat{X}(n,k) = W(\theta,k) Y(n,k)$$

- Using MMSE criterion

$$W(\theta,k) = \arg\min_W \ \mathrm{E}\{|WY(n,k) - X(n,k)|^2\}$$

- Solution:

$$W(\theta,k) = \frac{|A(\theta,k)|\phi_X}{|A(\theta,k)|^2\phi_X + \phi_V}$$

Orientation Compensation Filter
Orientation Dependent Gain

- Defining orientation dependent gain

$$G(\theta, k) = |A(\theta, k)|^{-1}$$

Orientation Compensation Filter
Orientation Dependent Gain

- Defining orientation dependent gain

$$G(\theta, k) = |A(\theta, k)|^{-1}$$

- Solution:

$$W(\theta, k) = G(\theta, k) \cdot \frac{\phi_X}{\phi_X + G^2(\theta, k)\phi_V}$$

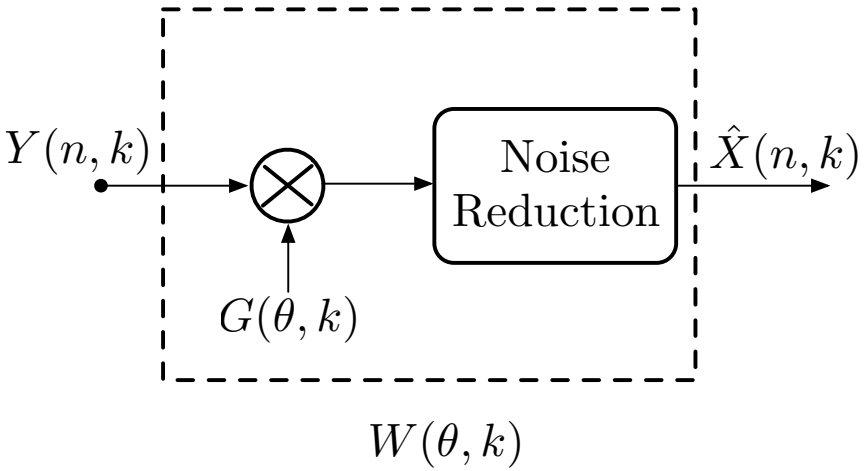
Orientation Compensation Filter
Orientation Dependent Gain

- Defining orientation dependent gain

$$G(\theta, k) = |A(\theta, k)|^{-1}$$

- Solution:

$$W(\theta, k) = G(\theta, k) \cdot \frac{\phi_X}{\phi_X + G^2(\theta, k)\phi_V}$$



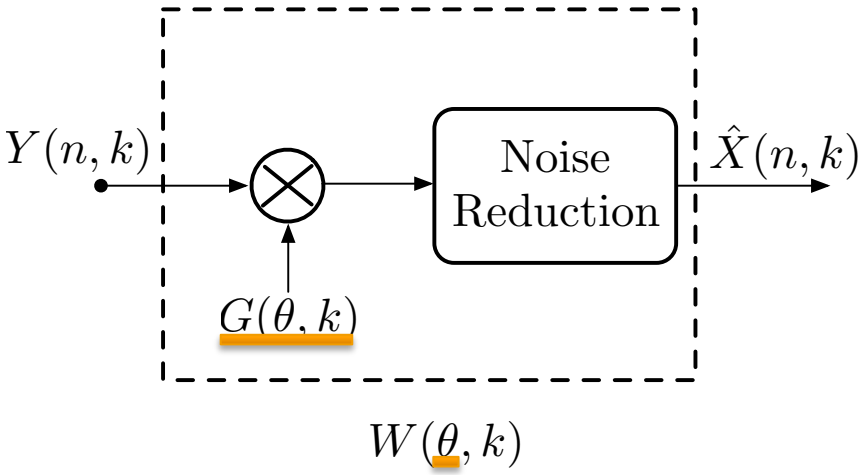
Orientation Compensation Filter
Orientation Dependent Gain

- Defining orientation dependent gain

$$G(\theta, k) = |A(\theta, k)|^{-1}$$

- Solution:

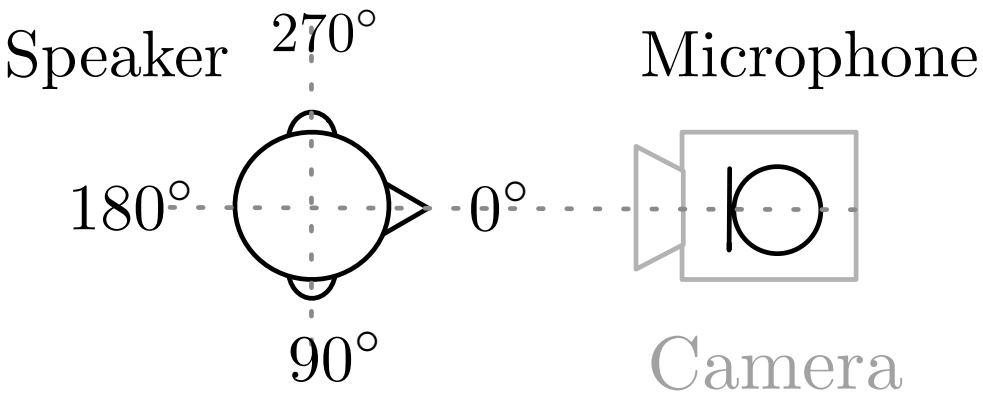
$$W(\theta, k) = G(\theta, k) \cdot \frac{\phi_X}{\phi_X + G^2(\theta, k)\phi_V}$$



Need to be estimated

Head Orientation Estimation

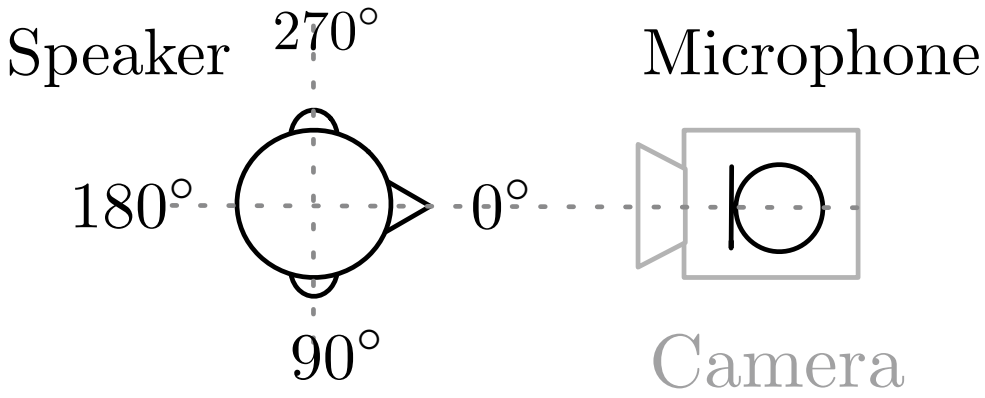
Video-based Method



- Camera is co-located with the microphone

Head Orientation Estimation

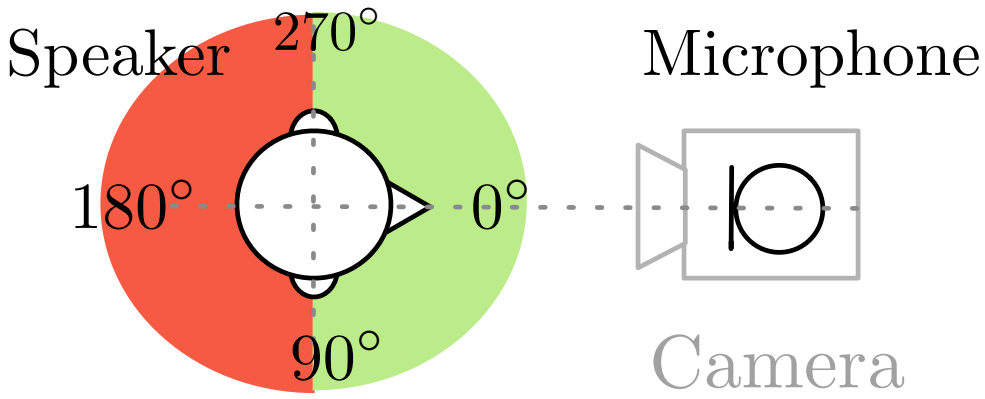
Video-based Method



- Camera is co-located with the microphone
- Proprietary software from Fraunhofer IIS, SHORE™, is used to obtain a single orientation estimate at each time frame

Head Orientation Estimation

Video-based Method



- Video-based orientation estimation
- Proprietary software from Fraunhofer IIS, SHORE™, is used to obtain a single orientation estimate at each time frame n

- Current limitation: We do not obtain estimates in the range of $[90°, 270°]$

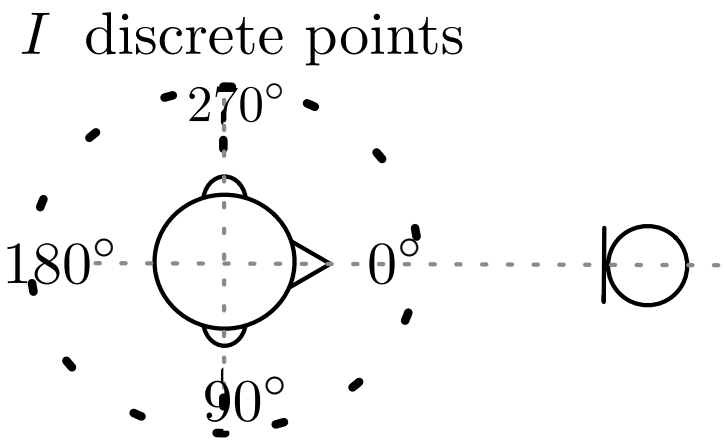
In this work, we assume the orientation to be known

Orientation Dependent Gain

- Use SMIRgen as a mouth simulator to compute a gain table
- Head is modeled as a rigid sphere
- Mouth is an omnidirectional point source placed on the sphere
- Orientation dependent gain is selected from the pre-computed gain table, at each time frame, based on the current estimate of the orientation

Gain Table
Computation

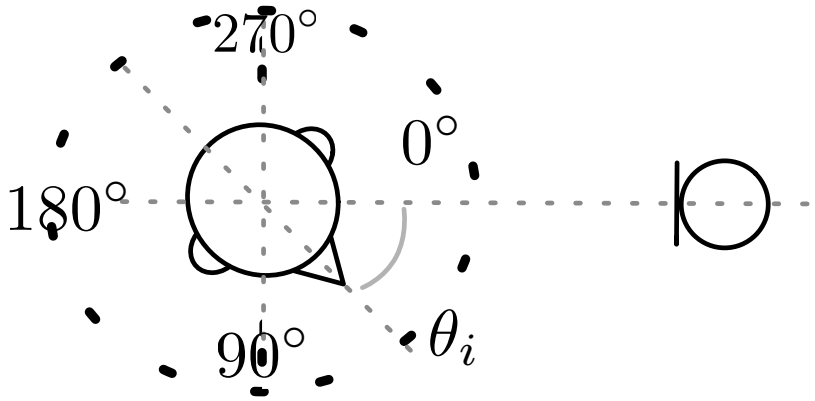
(1) Sample the orientation range at I points



Gain Table
Computation

(1) Sample the orientation range at I points

(2) Compute the ATF at each θ_i



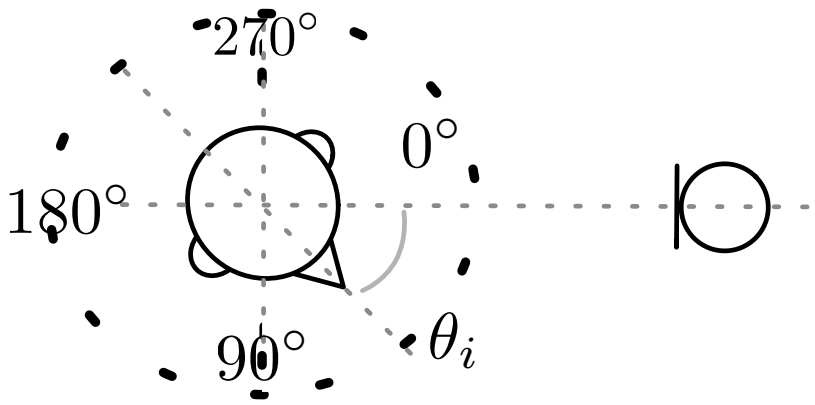
Gain Table
Computation

(1) Sample the orientation range at I points

(2) Compute the ATF at each θ_i

(3) Compute the corresponding gain at each point, for each bin, as

$$G(\theta_i, k) = \left| \frac{\hat{H}(\theta_i, k)}{\hat{H}(0, k)} \right|^{-1} \quad \text{follows from} \quad G(\theta, k) = |A(\theta, k)|^{-1}$$



Gain Table
Computation

(1) Sample the orientation range at I points

(2) Compute the ATF at each θ_i

(3) Compute the corresponding gain at each point, for each bin, as

$$G(\theta_i, k) = \left| \frac{\hat{H}(\theta_i, k)}{\hat{H}(0, k)} \right|^{-1} \quad \text{follows from} \quad G(\theta, k) = |A(\theta, k)|^{-1}$$

(4) The gain table is a matrix of size $I \times K$.

Gain Table

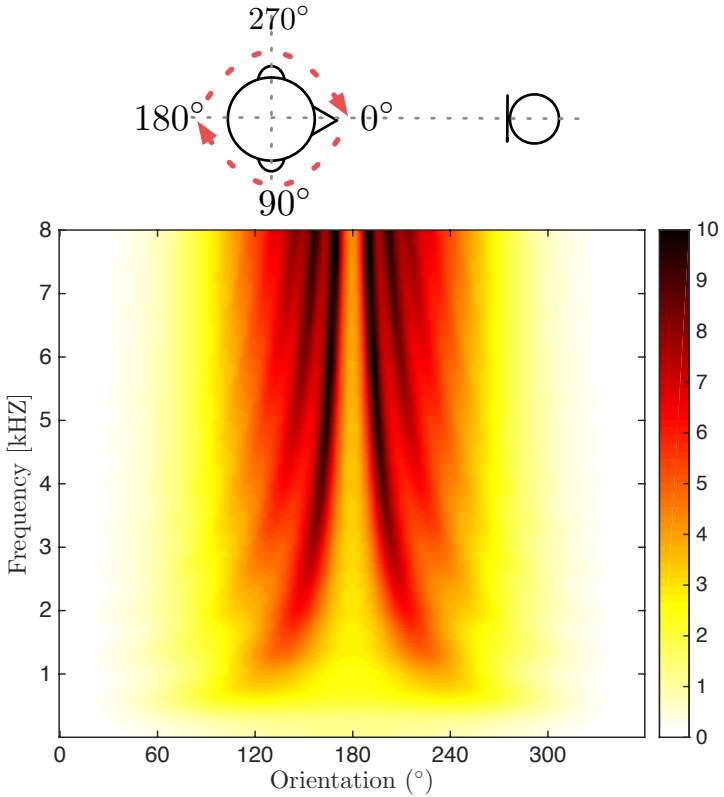


Figure: Gain table computed with SMIRgen for an anechoic environment

Gain Table

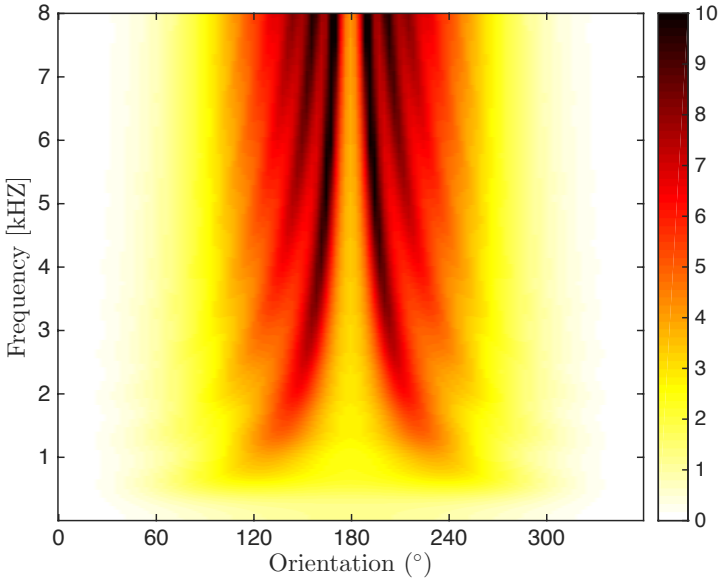
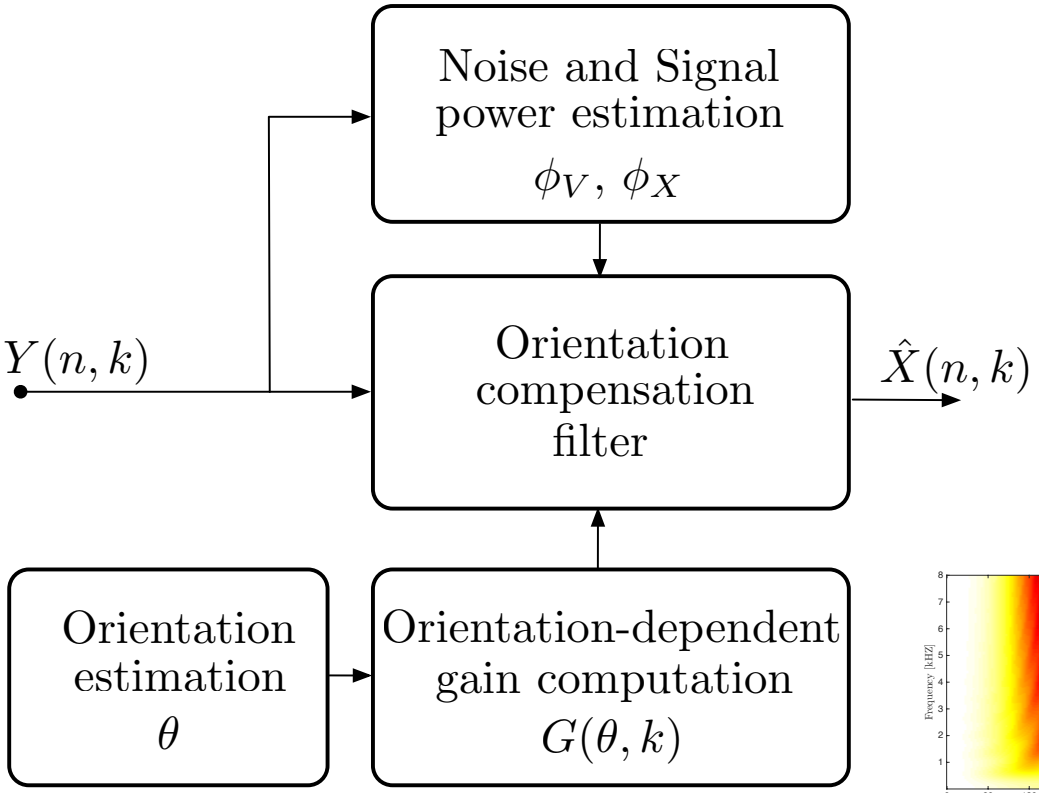
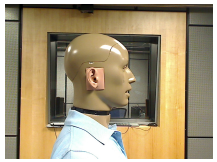


Figure: Gain table computed with SMIRgen for an anechoic environment

- Can be computed for reverberant environments using SMIRgen
- Can be computed using measured ATFs

System Overview

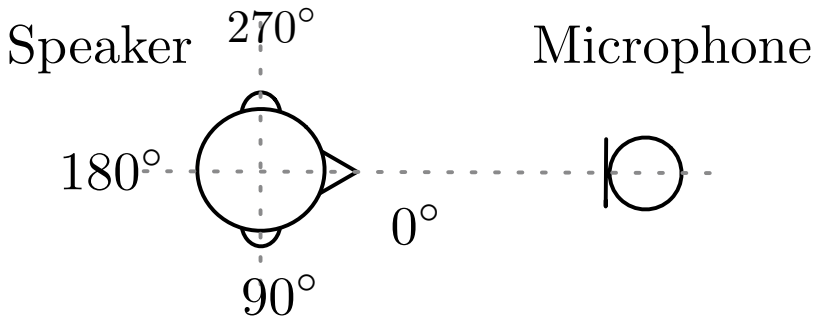




Experimental Results
Measured RIRs

- Measurement setup
  - Room size: $4.55\,\mathrm{m} \times 4.45\,\mathrm{m} \times 2.55\,\mathrm{m}$
  - Source-to-microphone distance: 1 m
  - $T_{60} = 0.17$s

KEMAR Dummy Head

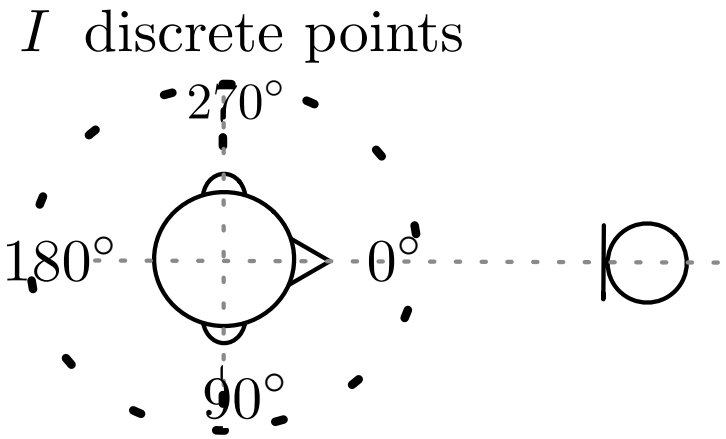


Experimental Results
Measured RIRs

- Measurement setup
  - Room size: $4.55\,\mathrm{m} \times 4.45\,\mathrm{m} \times 2.55\,\mathrm{m}$
  - Source-to-microphone distance: 1 m
  - $T_{60} = 0.17$s
- Stationary white noise with iSNR = 20 dB
- STFT parameters: 16 kHz sampling rate, frame length of 1024 samples with 50% overlap
- Noise PSD: estimated from silent frames
- Desired signal PSD: Decision directed approach [3]

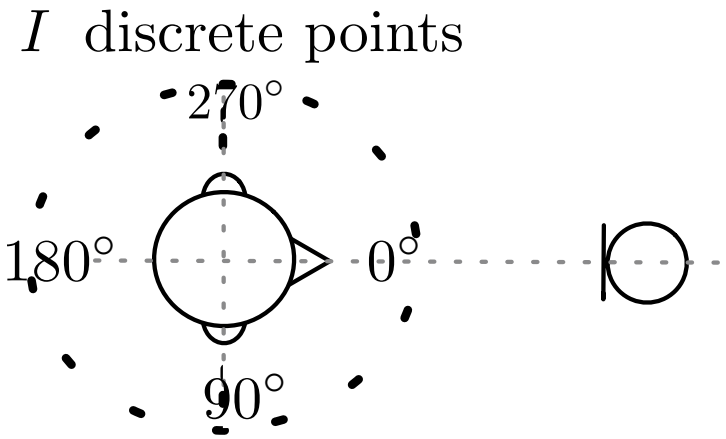
[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443–445, 1985.

Experimental Results
Measured RIRs



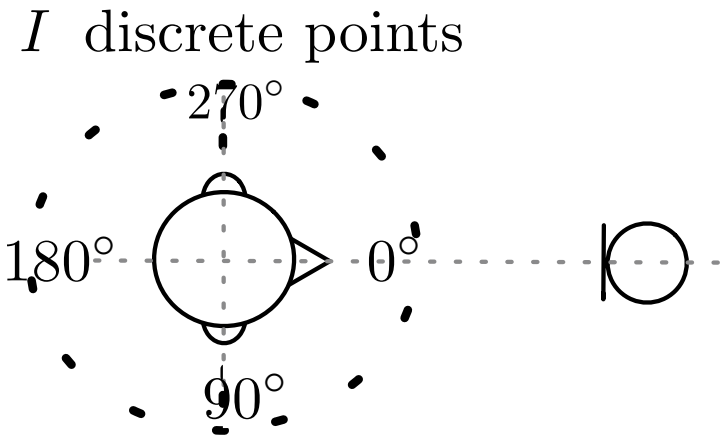
- Results presented for three different gain table computations with resolution of 30 degrees, plus for only noise reduction

Experimental Results
Measured RIRs



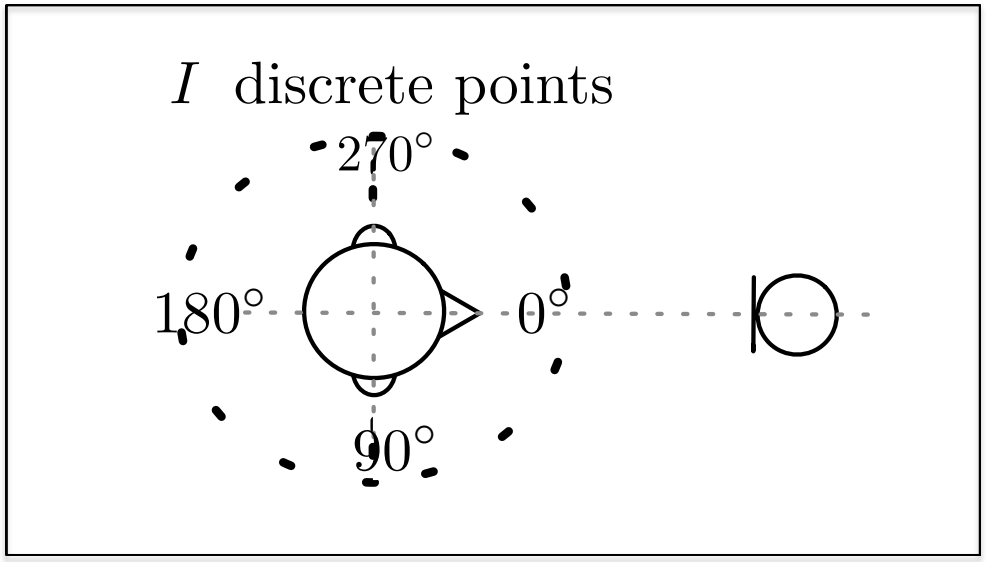
- Results presented for three different gain table computations with resolution of 30 degrees, plus for only noise reduction
  - NR: Only noise reduction, no application of orientation related gain

Experimental Results
Measured RIRs



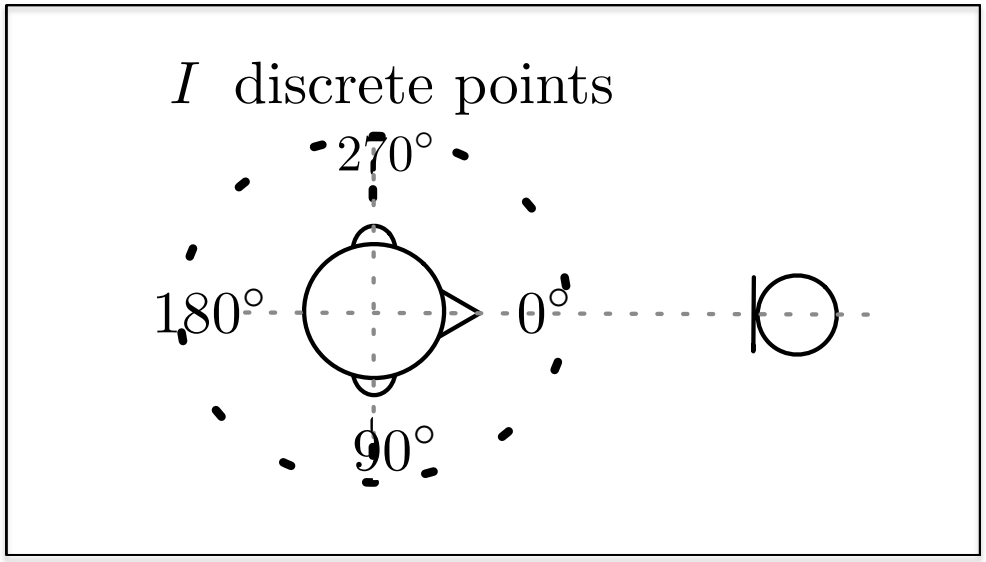
- Results presented for three different gain table computations with resolution of 30 degrees, plus for only noise reduction
  - NR: Only noise reduction, no application of orientation related gain
  - AG: Assuming anechoic environment (using SMIRgen)

Experimental Results
Measured RIRs



- Results presented for three different gain table computations with resolution of 30 degrees, plus for only noise reduction
  - NR: Only noise reduction, no application of orientation related gain
  - AG: Assuming anechoic environment (using SMIRgen)
  - RGSA: Spatially averaged reverberant gain (using SMIRgen)

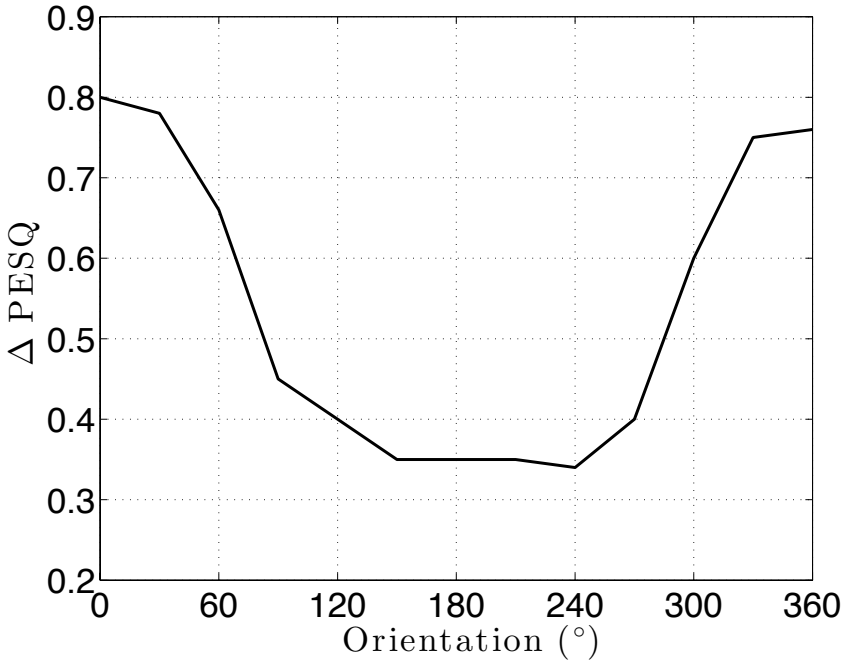
Experimental Results
Measured RIRs



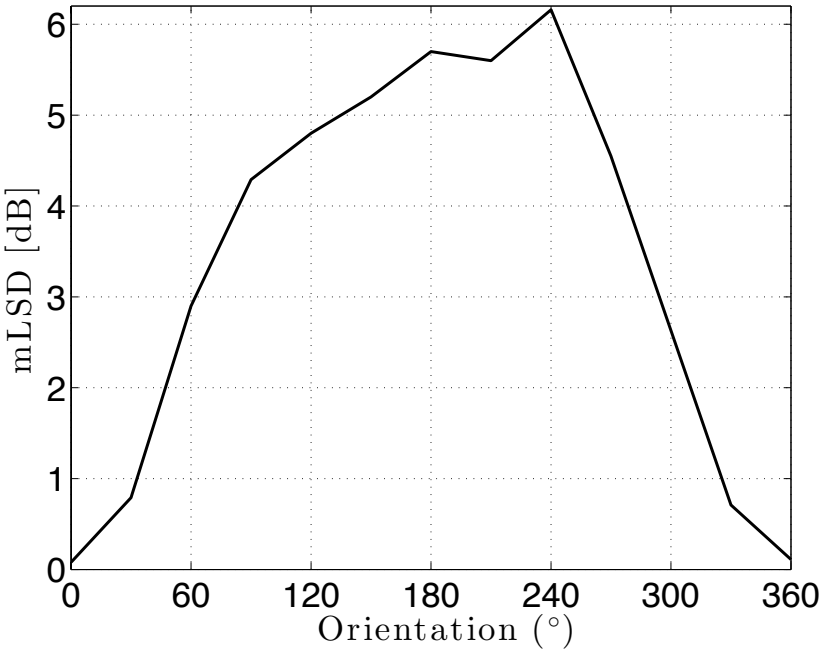
- Results presented for three different gain table computations with resolution of 30 degrees, plus for only noise reduction
  - NR: Only noise reduction, no application of orientation related gain
  - AG: Assuming anechoic environment (using SMIRgen)
  - RGSA: Spatially averaged reverberant gain (using SMIRgen)
  - RGMES: Using measured ATFs

Experimental Results
Measured RIRs

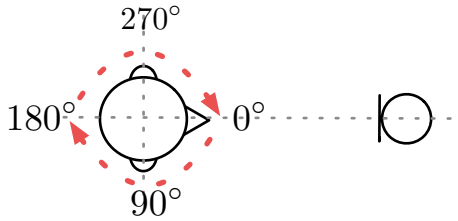
NR: Noise Reduction



(a) PESQ Improvement

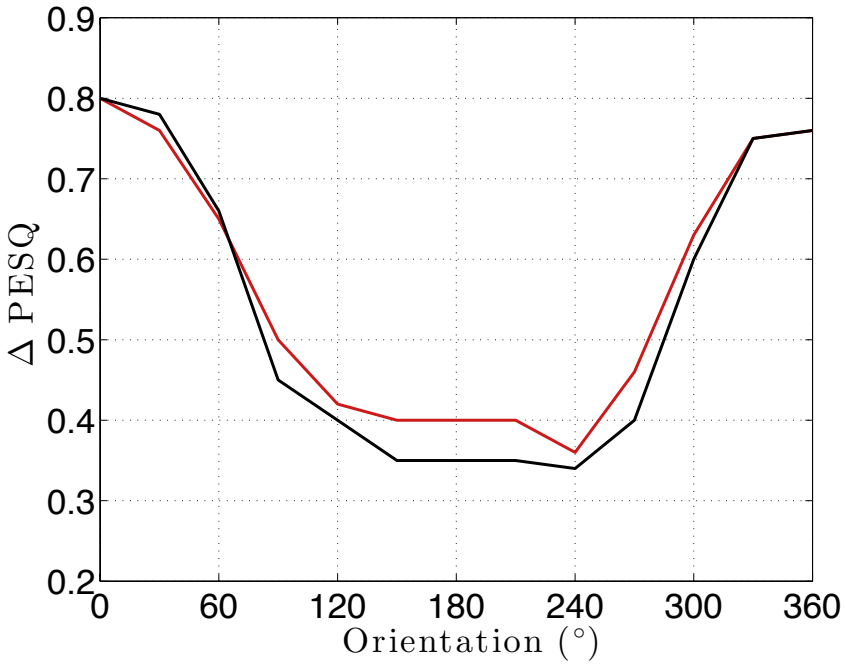


(b) meanLSD

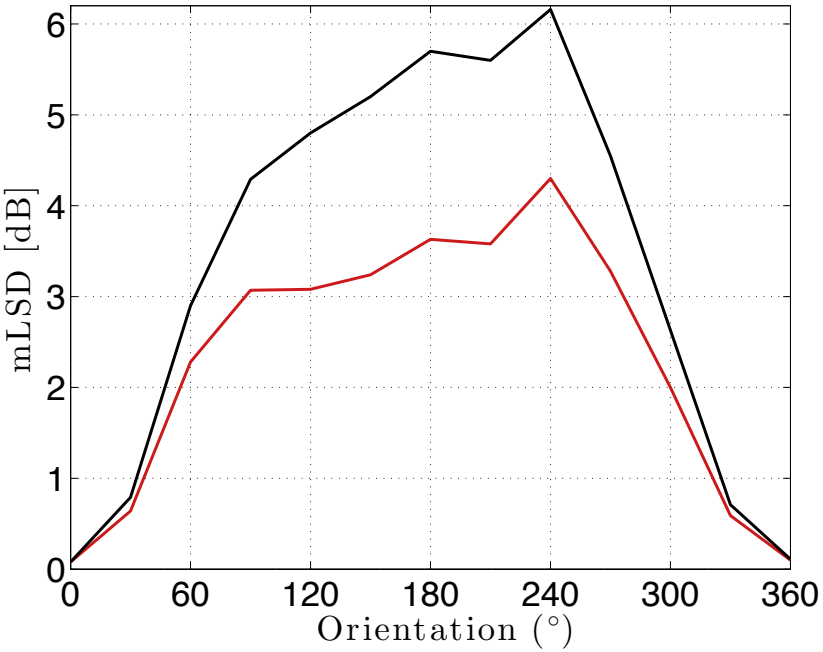


Experimental Results
Measured RIRs

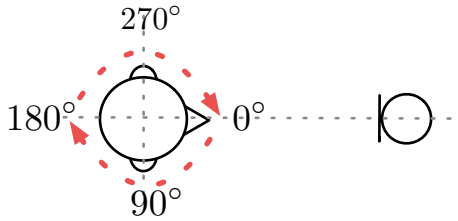
AG: Anechoic Gain (SMIRgen)



(a) PESQ Improvement



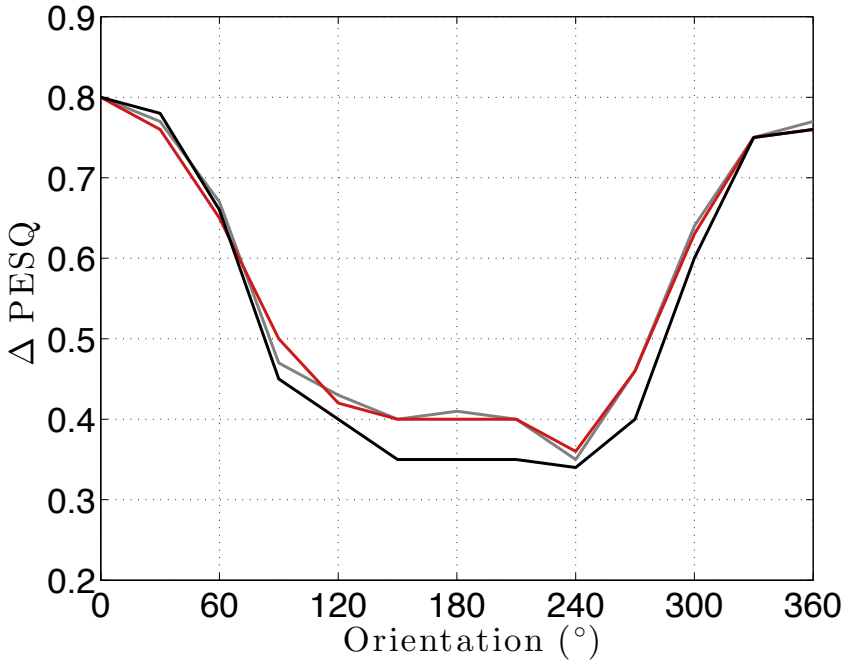
(b) meanLSD



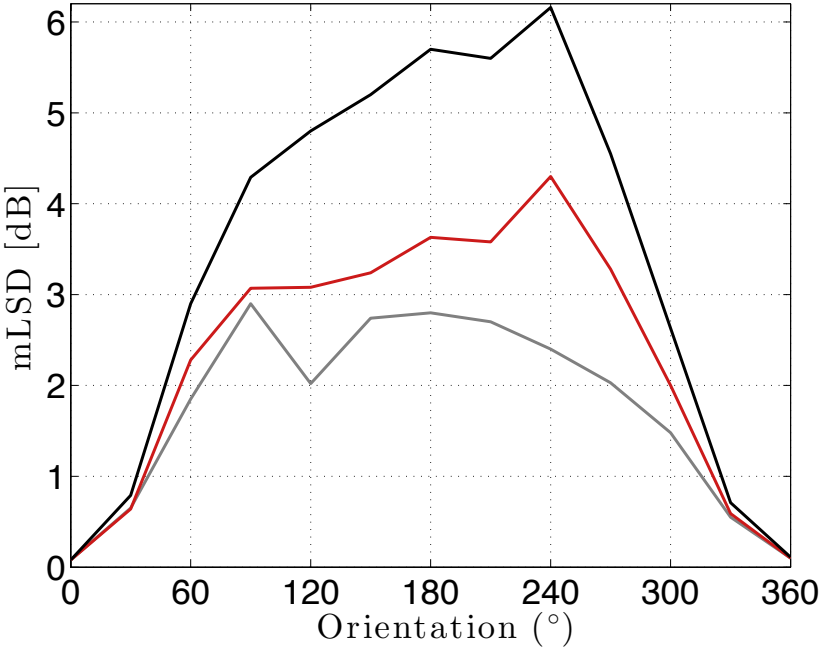
Experimental Results

Measured RIRs

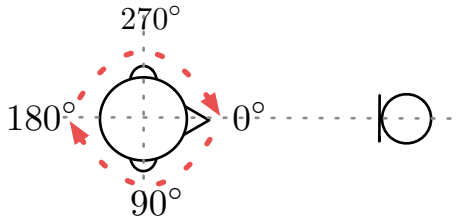
RGSA: Spatially Averaged Reverberatn Gain (SMIRgen)



(a) PESQ Improvement

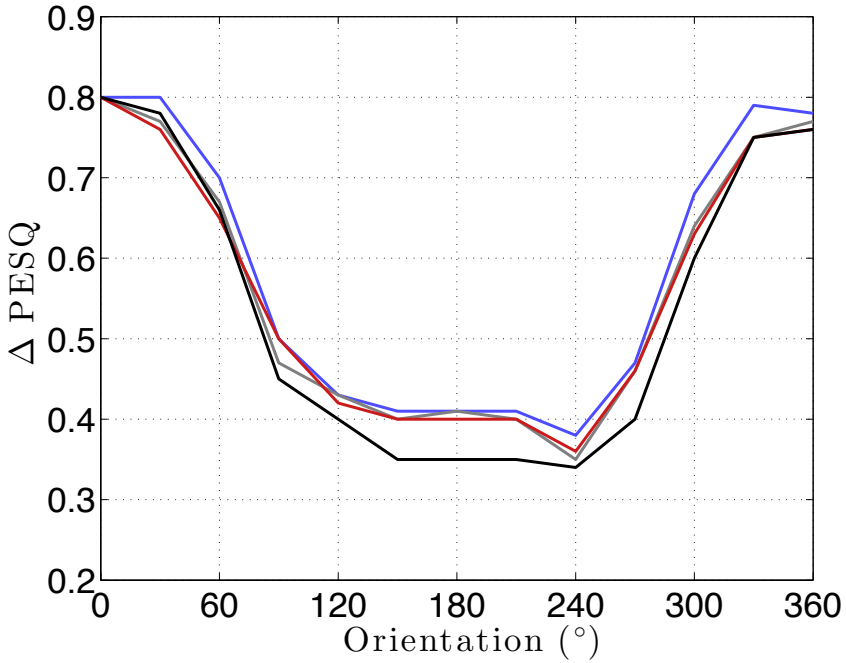


(b) meanLSD

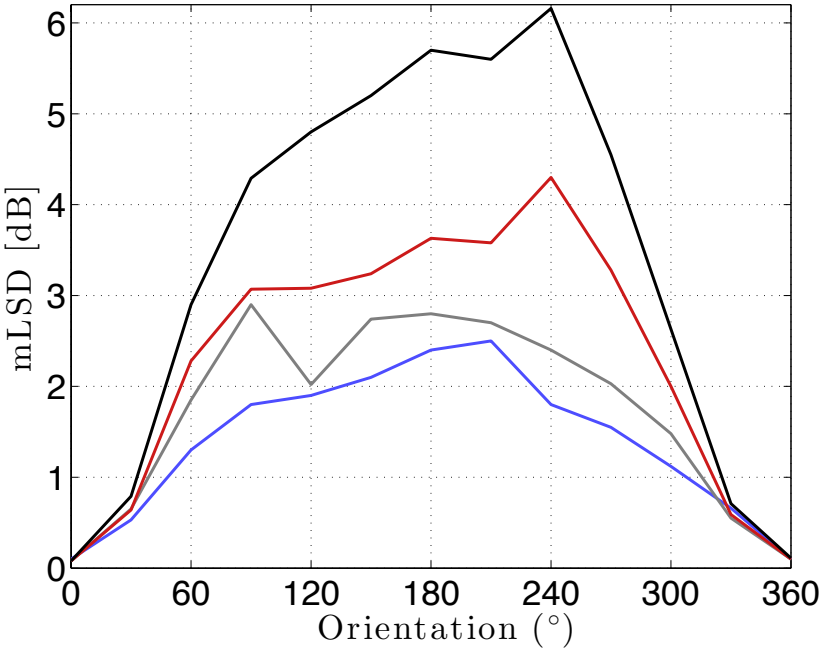


Experimental Results
Measured RIRs

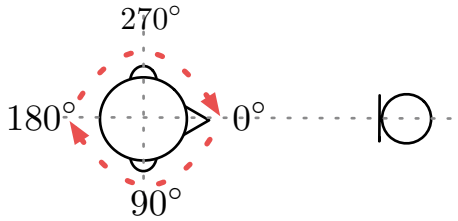
RGMES: Measured ATFs



(a) PESQ Improvement

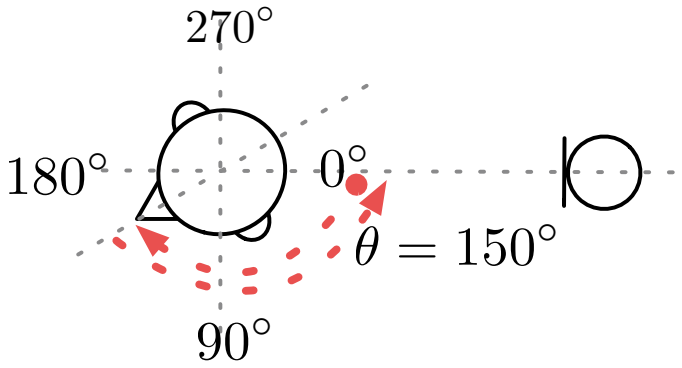


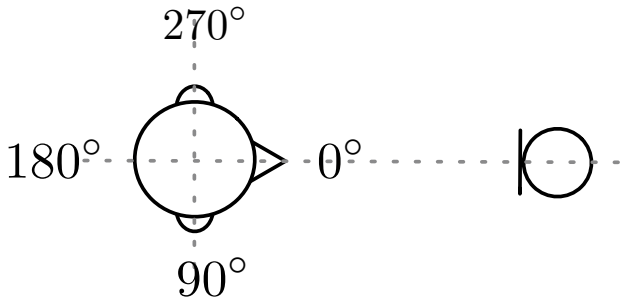
(b) meanLSD



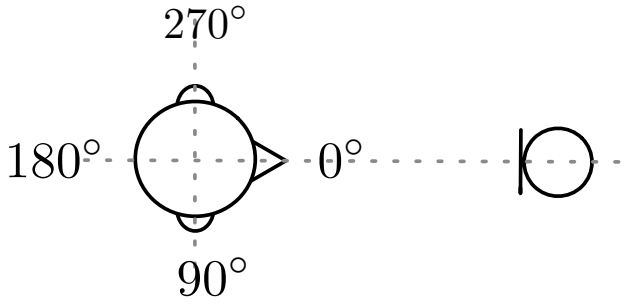
Audio Examples

Measured RIRs

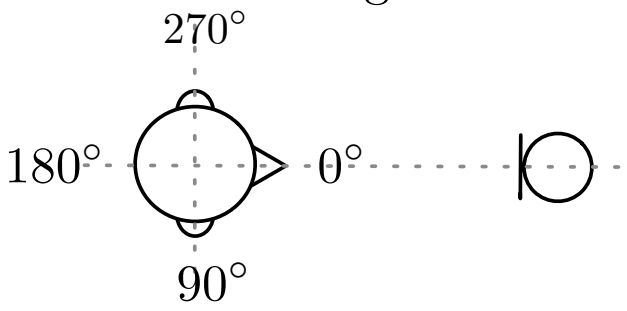




Noise Reduction Only



RGMES gain



Conclusions and Outlook

- A single channel speech enhancement framework, that incorporates head orientation information was presented
- Experimental results provided motivation for further exploring the significance head orientation information for speech enhancement
- Current research focuses on developing methods to learn the attenuation characteristics due to the orientation of the speaker to perform the compensation
- Future work would involve relaxing the constraints of the current system, and develop a method more suitable for a practical setting

Thank you for your attention

Questions?