Imperial College
London

Processing pipelines for efficient, physically-accurate simulation of microphone array signals in dynamic sound scenes

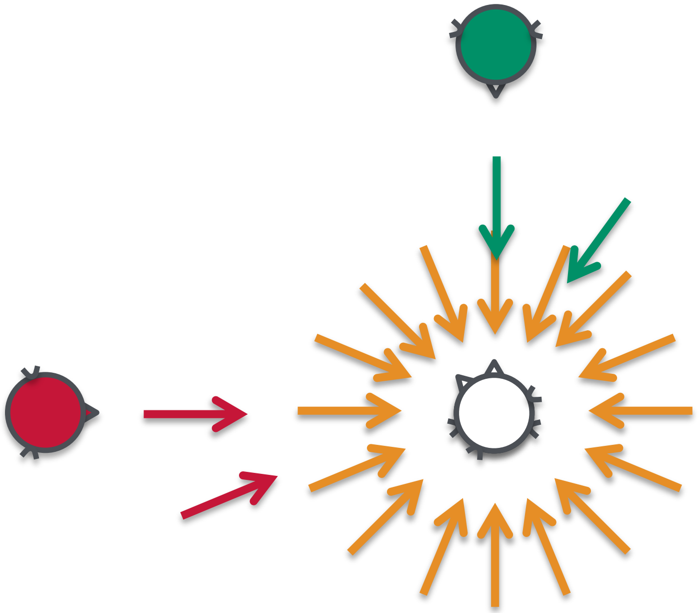
Alastair H. Moore, Rebecca R. Vos,
Patrick A. Naylor and Mike Brookes

ICASSP 2021

Motivation

“Listener-in-the-loop” perceptual experiments

Motivation



Motivation

Processing pipelines for efficient, physically-accurate simulation of microphone array signals in dynamic sound scenes

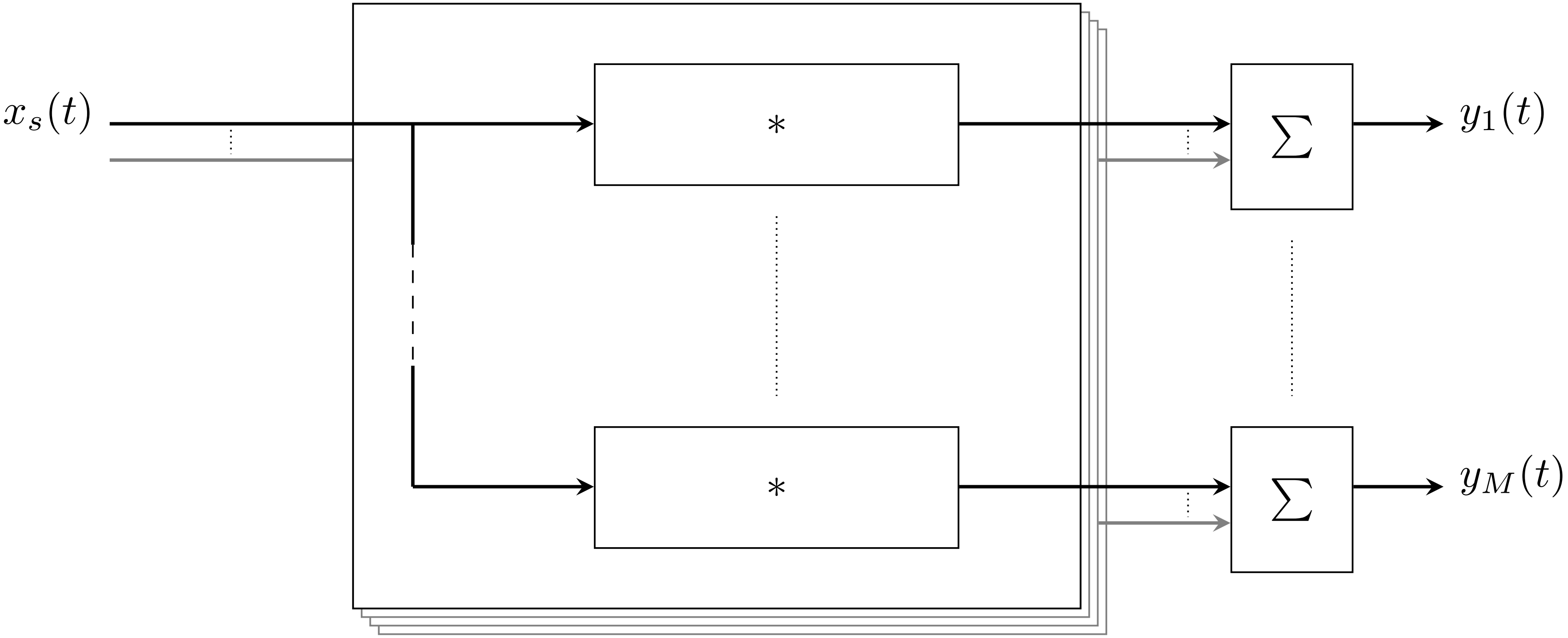
Outline

- Task and assumptions
- Pipelines
- Evaluation of accuracy
- Efficiency comparison

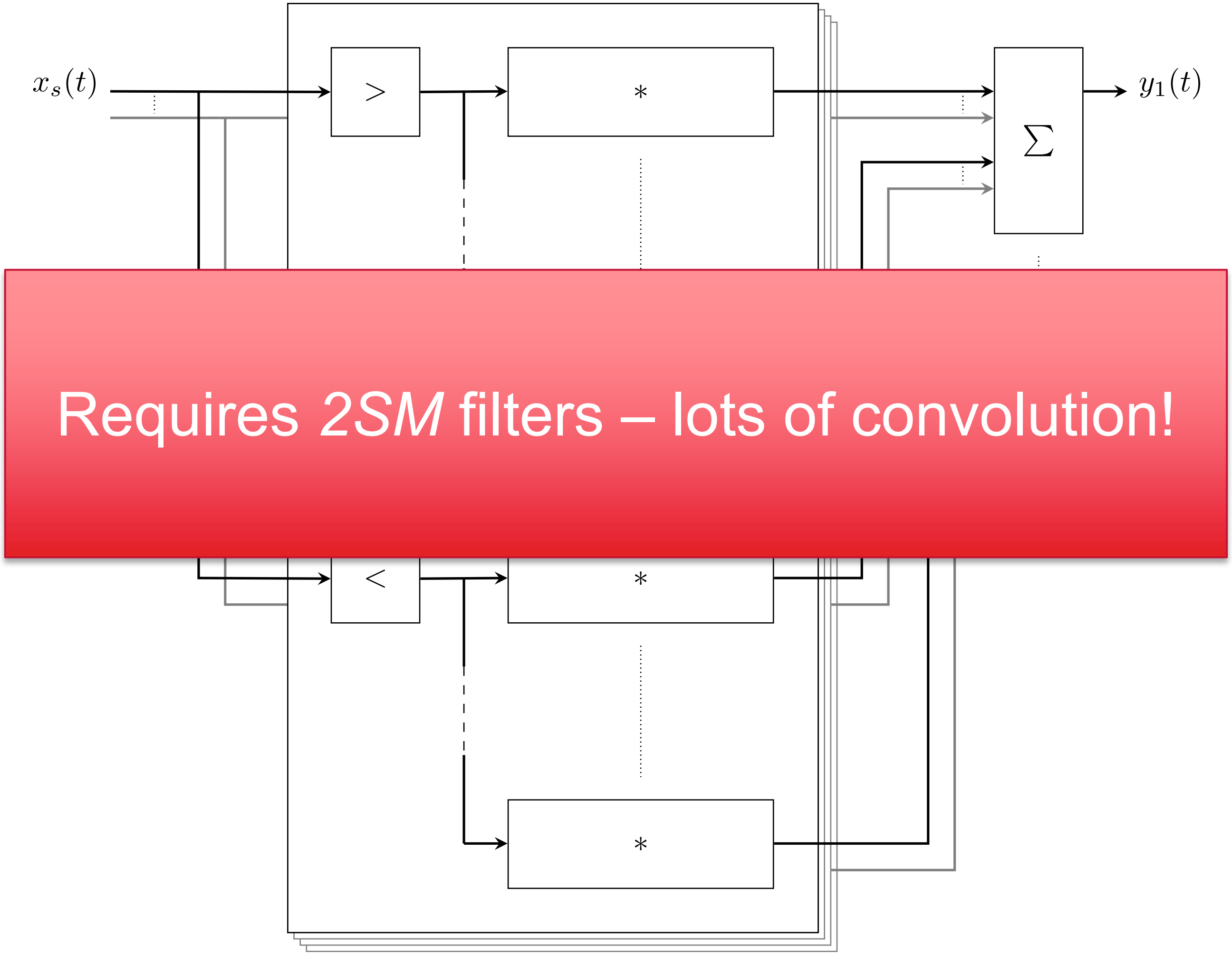
Plane wave spatialisation

- Acoustic room simulation calculates propagation from all sources in a scene to a single point
  - E.g. centre of head/array
- Contribution of each incident sound wave to to the microphone signals calculated according to its direction of arrival (DOA)
- Source and/or array movement causes DOA to change over time

S wavefronts, M microphones

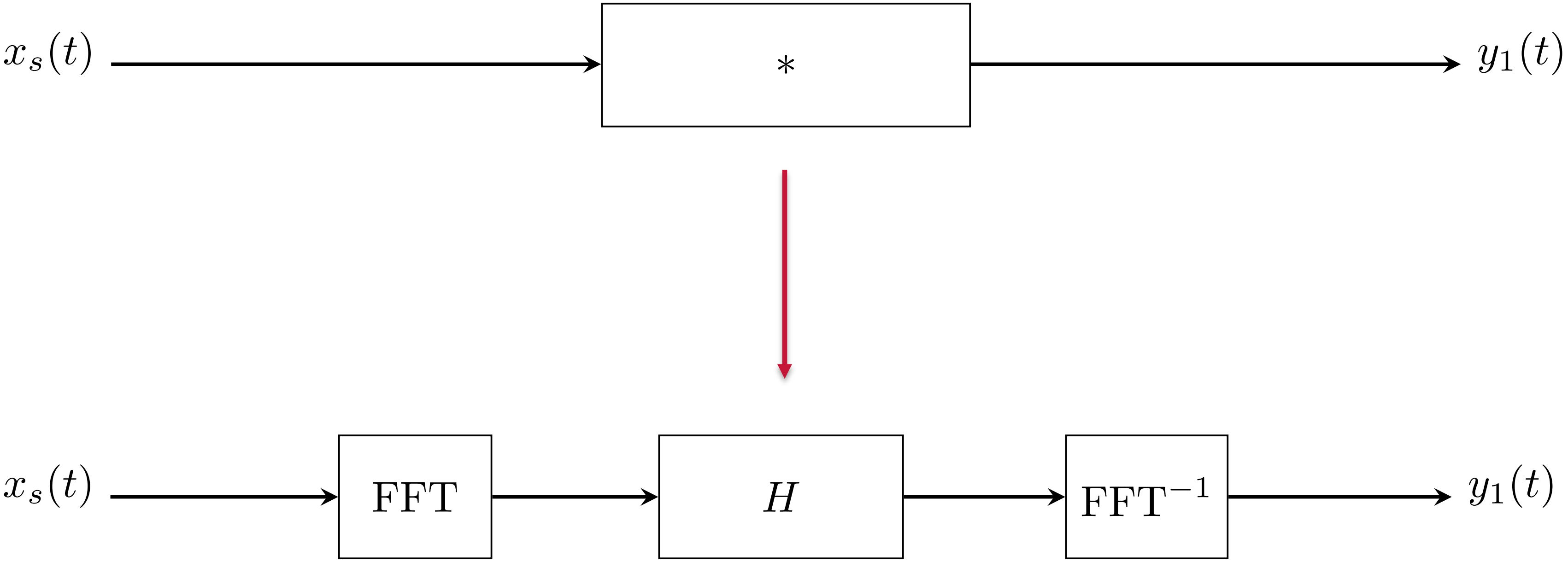


S wavefronts, M microphones - dynamic

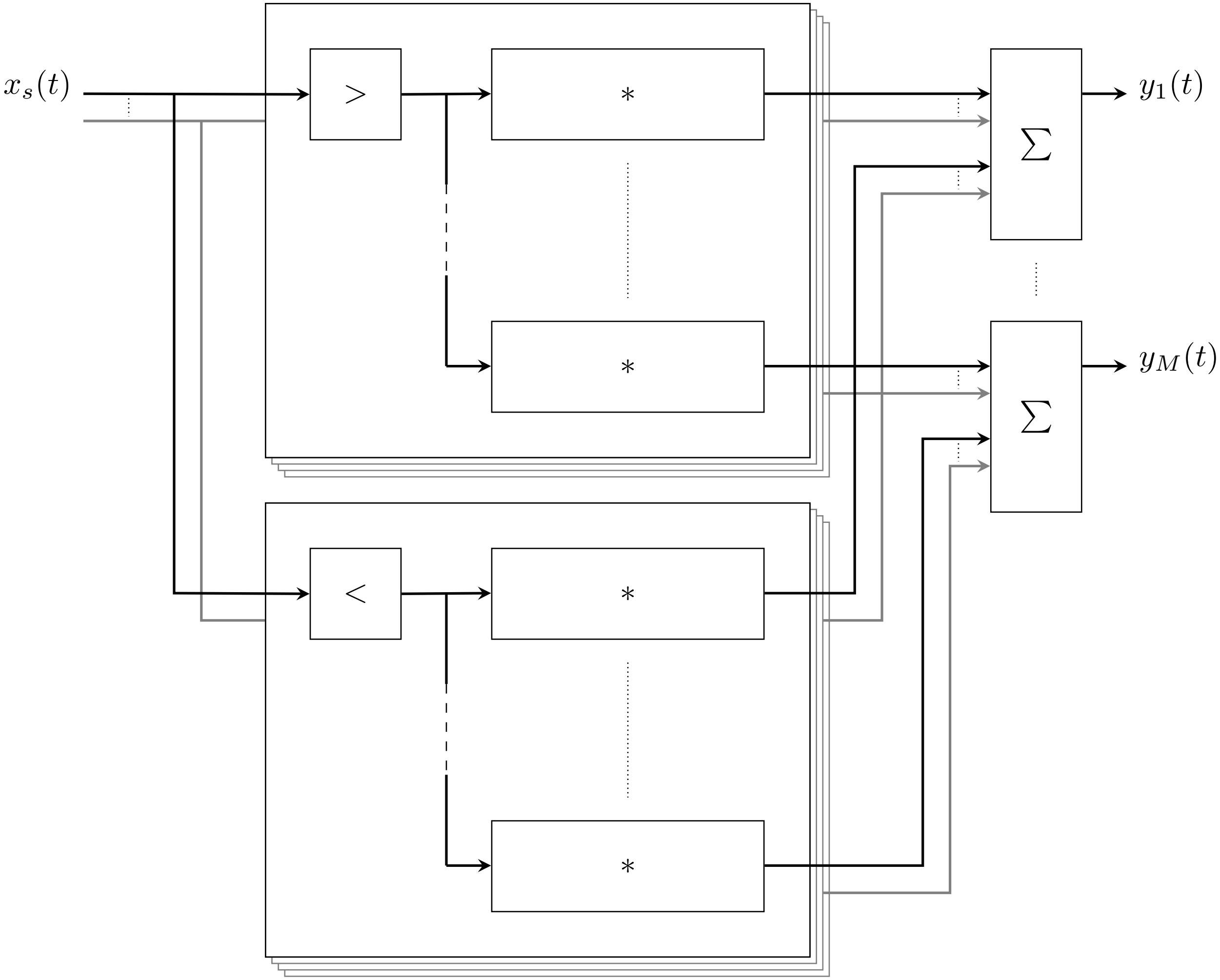


Requires *2SM* filters – lots of convolution!

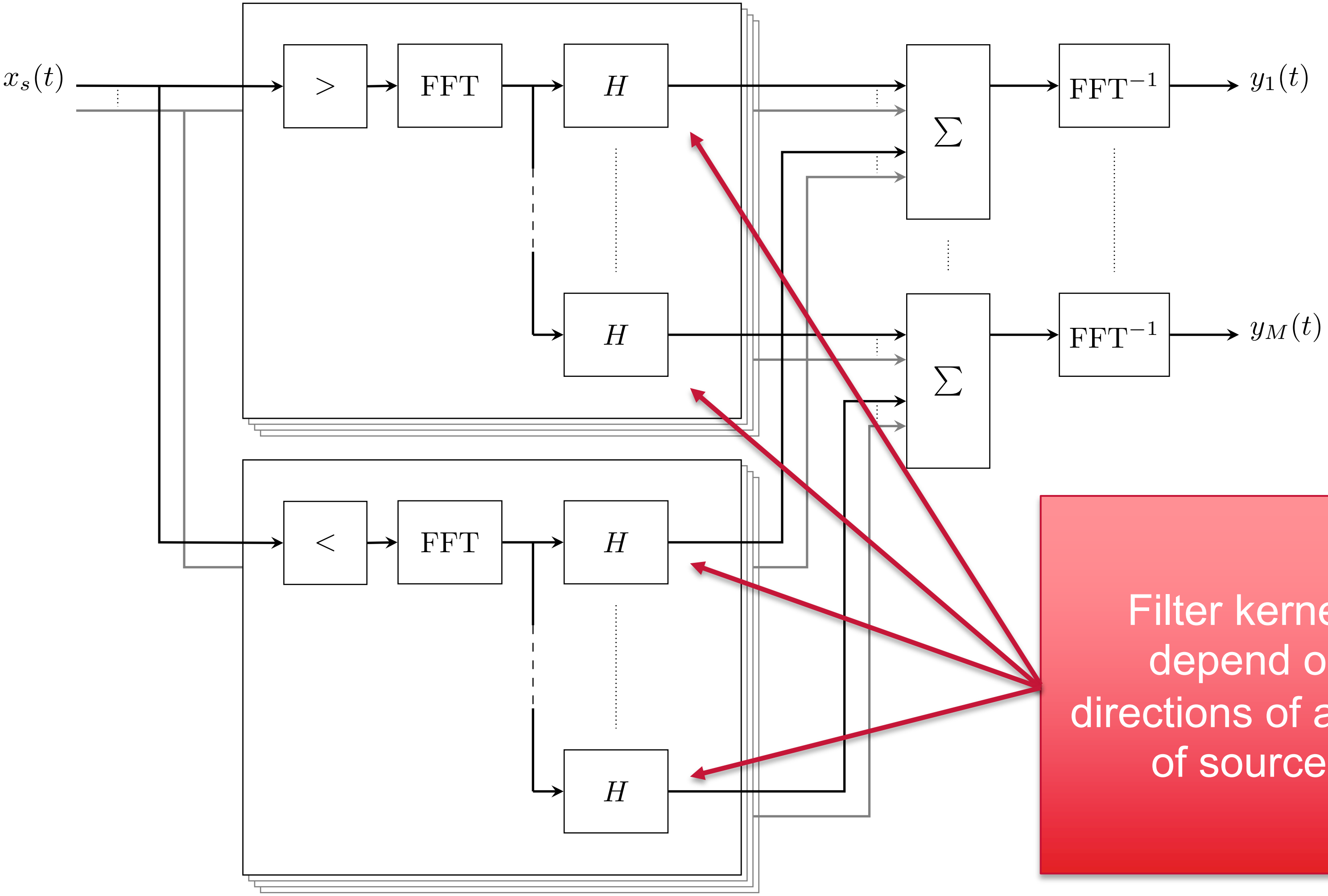
Fast convolution



S wavefronts, M microphones - dynamic



Direct synthesis (baseline)

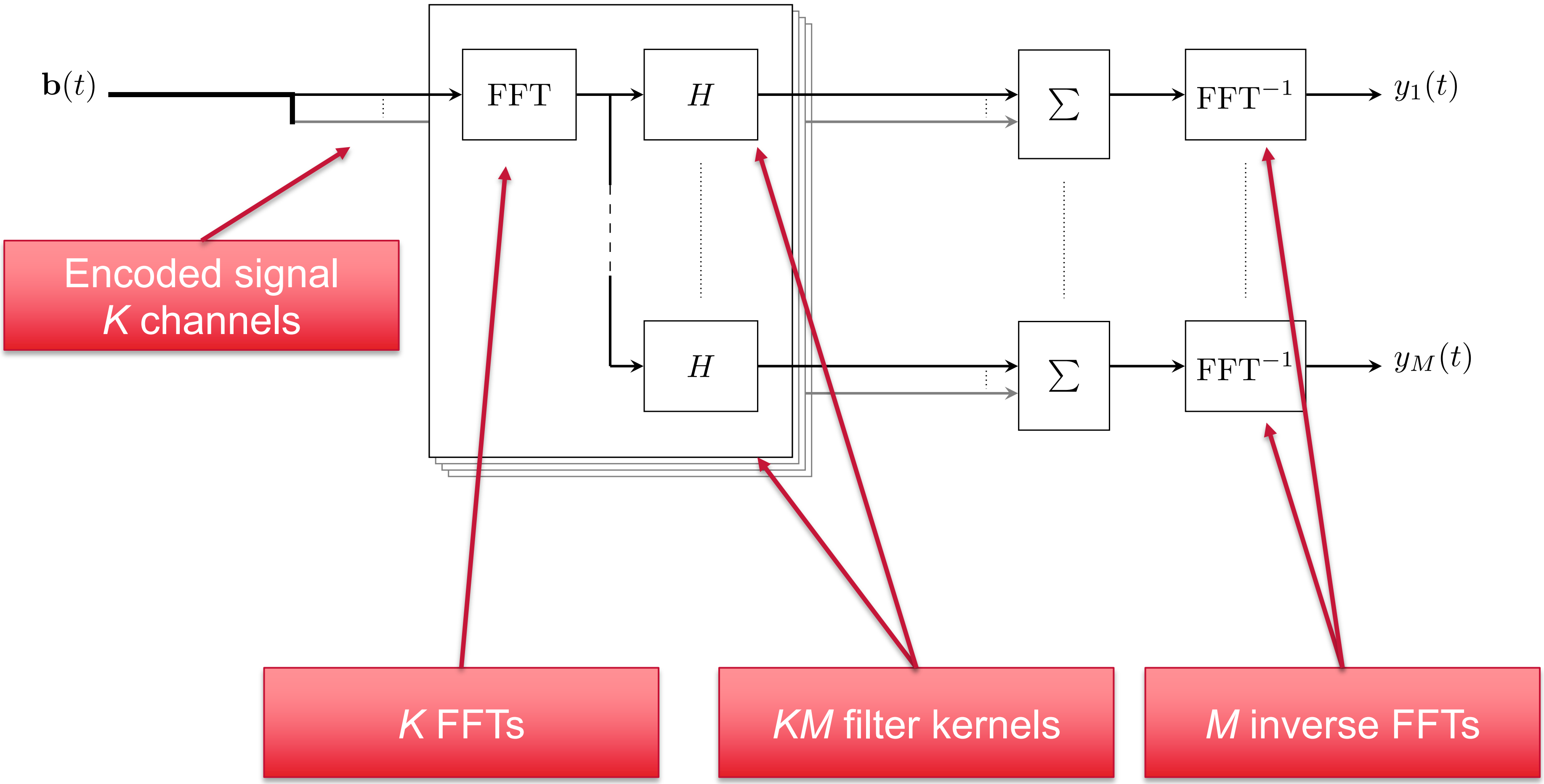


Filter kernels depend on directions of arrival of sources

Shared kernels

- Always evaluate a fixed set of filter kernels
- For each source
  - Find weights required to approximate the required impulse response using a combination of available kernels
  - Apply weights to the input signals
  - Add scaled signals to bus

Microphone independent encoding



Encoded signal
K channels

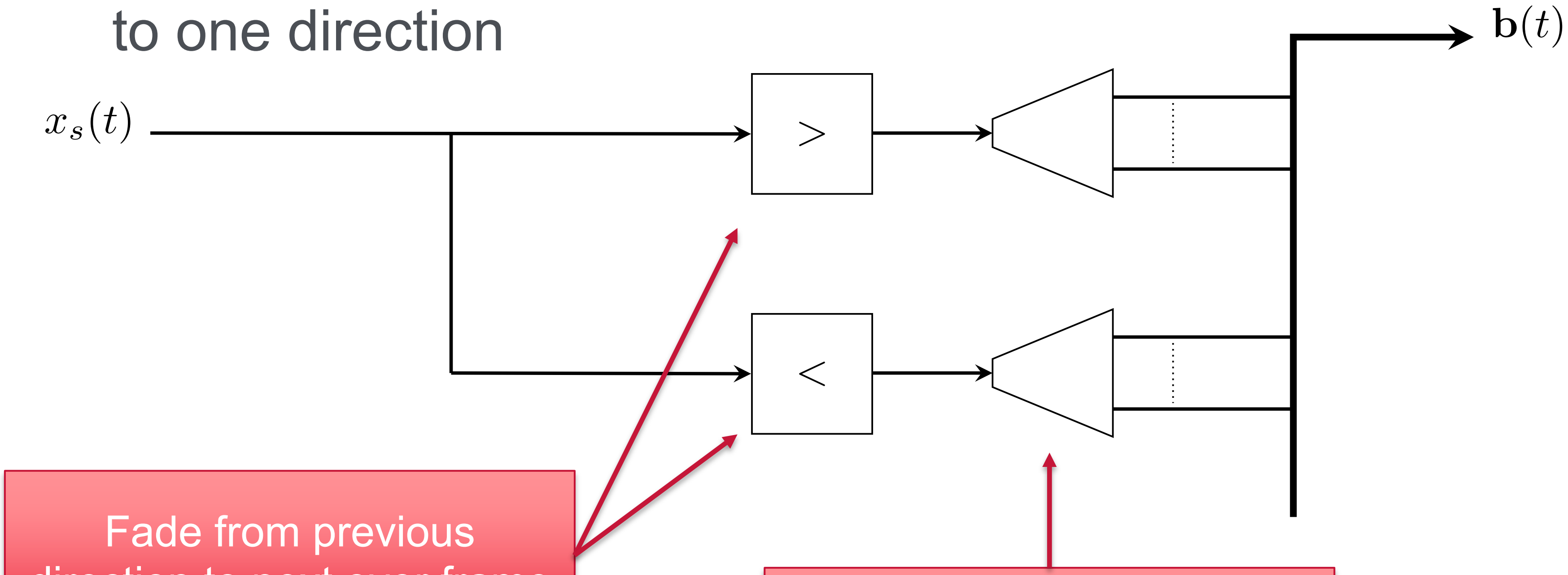
K FFTs

KM filter kernels

M inverse FFTs

Virtual speaker encoding (1)

- Kernels correspond to fixed directions of arrival
- More directions → Increases spatial resolution
- Nearest speaker encoder (NSPK) assigns each source to one direction



Fade from previous direction to next over frame

Assign signal to 1 of *K* directions

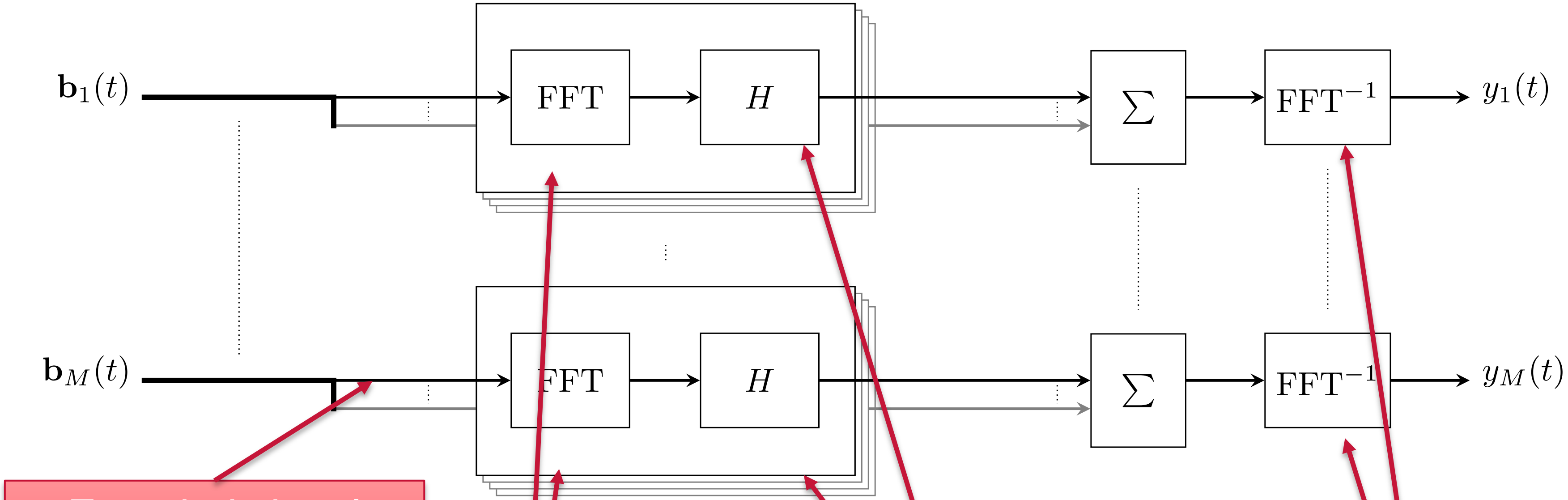
Virtual speaker encoding (2)

- Kernels correspond to fixed directions of arrival
- More directions → Increases spatial resolution
- Vector base amplitude panning (VBAP) assigns a portion of signal to multiple (*J*) virtual speakers
- Weights depend on direction of arrival

Spherical harmonic encoding

- Kernels correspond to spherical harmonic transform of the array manifold
  - Different coefficients for each microphone
  - Increasing order → Increases spatial resolution
- Source weights depend on direction of arrival
  - Obtained directly from spherical harmonic basis functions
  - Independent of microphone
- Fade weights between directions at start and end of frame

Microphone dependent encoding



Encoded signal
KM channels

KM FFTs

KM filter kernels

M inverse FFTs

Principal component analysis

- Kernels correspond to principal components of the array manifold
  - Different basis functions for each microphone
  - Increasing order → Increases spatial resolution
- Source weights depend on direction of arrival
  - Obtained from PCA
  - **Dependent on microphone**
- Fade weights between directions at start and end of frame

Pipelines

- Microphone-independent encoders
  - Nearest speaker
  - VBAP
  - SH
- Microphone-dependent encoders
  - PCA

Can we use fewer kernels by time-aligning impulse responses?
Does it reduce the overall computational cost?

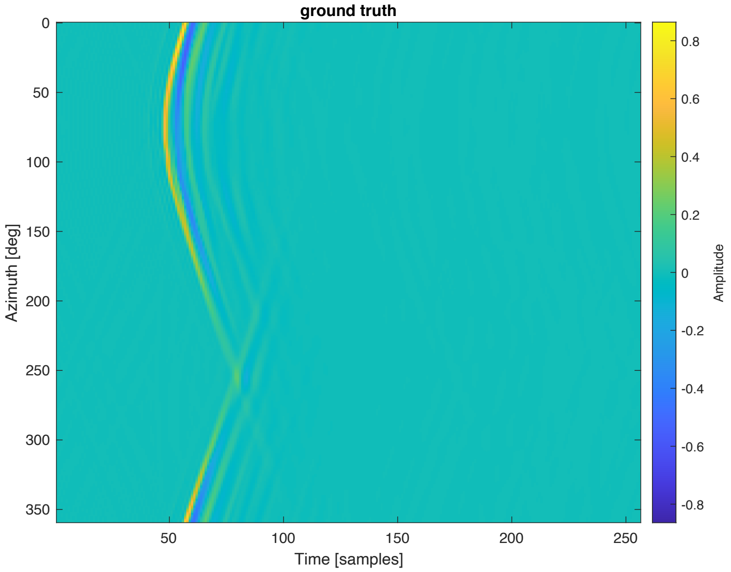
Time aligned kernels

- Remove direction-dependent delay from filters
  - Estimated using group delay
- Group delay aligned (GDA) impulse responses are more consistent
  - Better interpolation?
  - Lower order approximation?
- Direction-dependent delay must be added to each incident signal **before encoding**
- Delay is different for each microphone
- Sinc interpolation using *D* coefficients from precomputed

Time-aligned PCA spatialization is novel approach

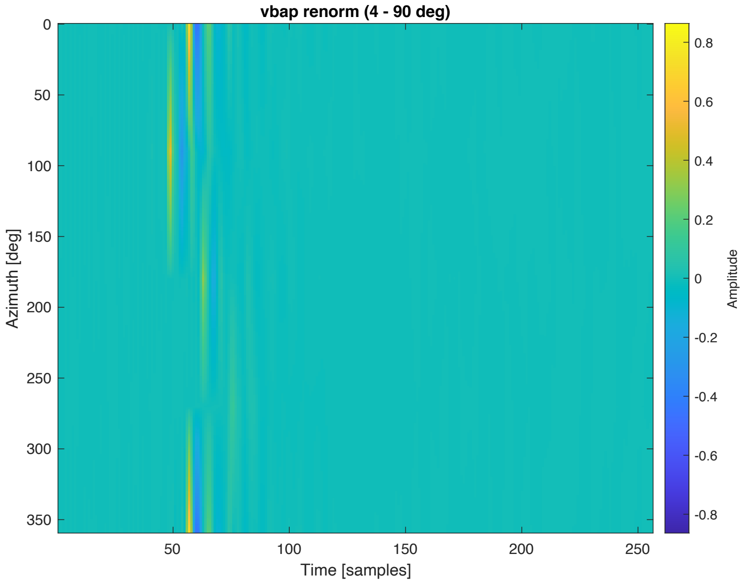
Time alignment example

- Front left channel of hearing aid array

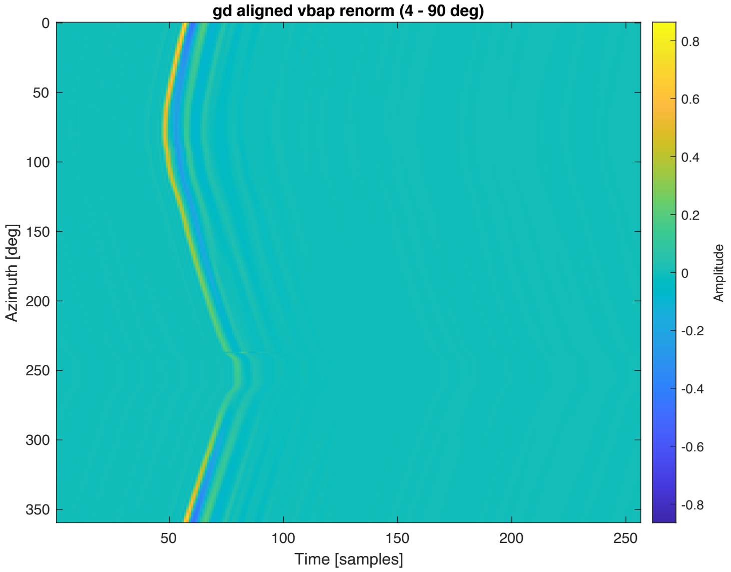


Ground truth

VBAP – 4 kernels

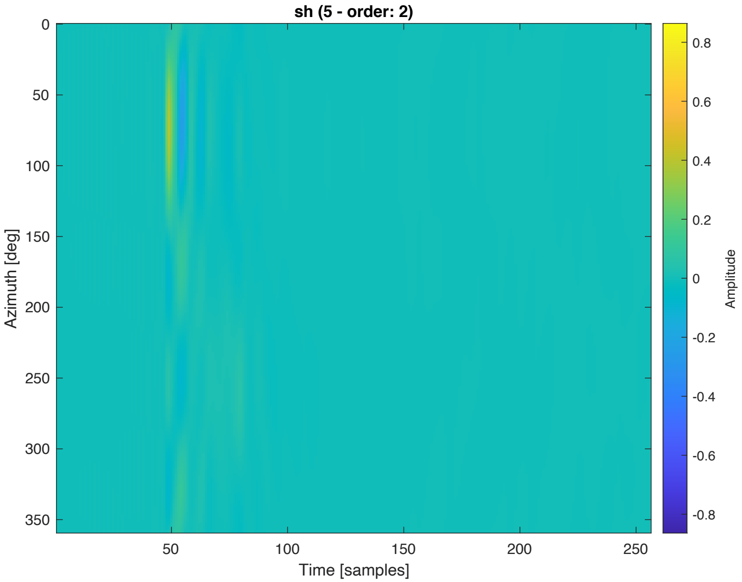


Original

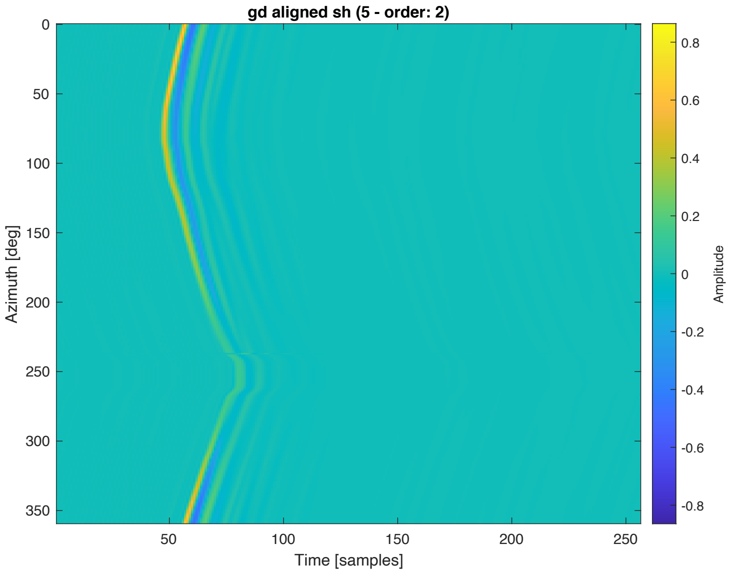


Aligned

SH – 5 kernels

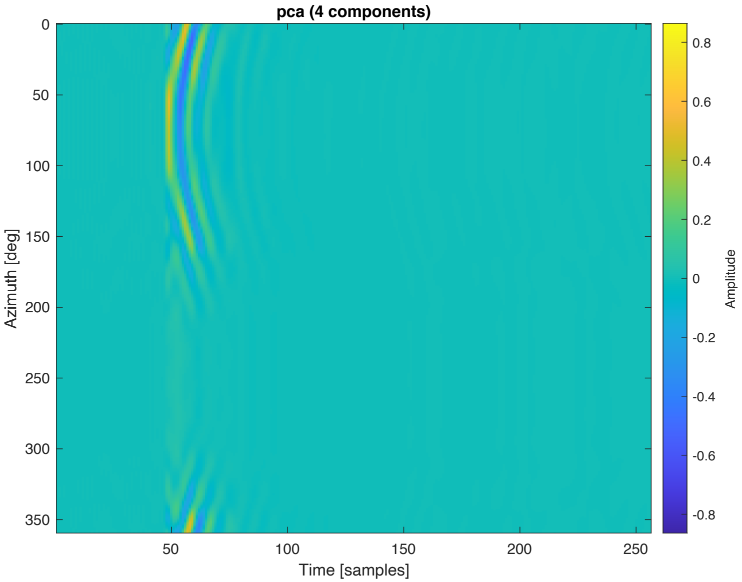


Original

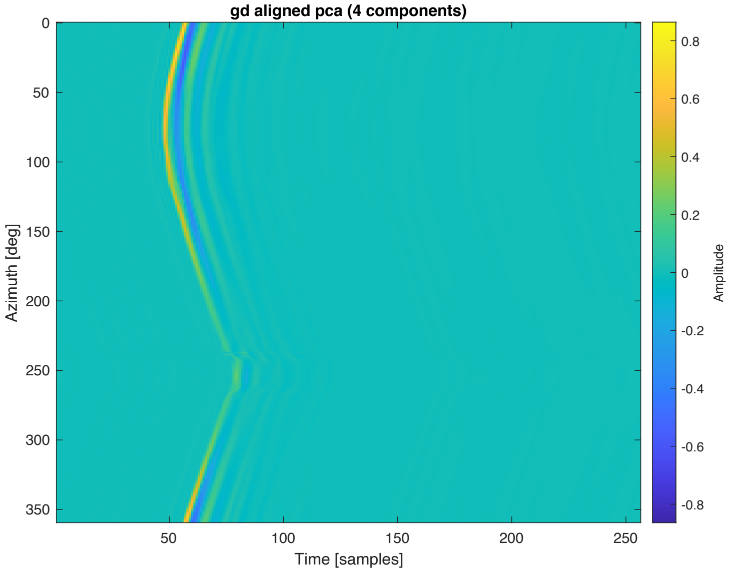


Aligned

PCA – 4 kernels



Original

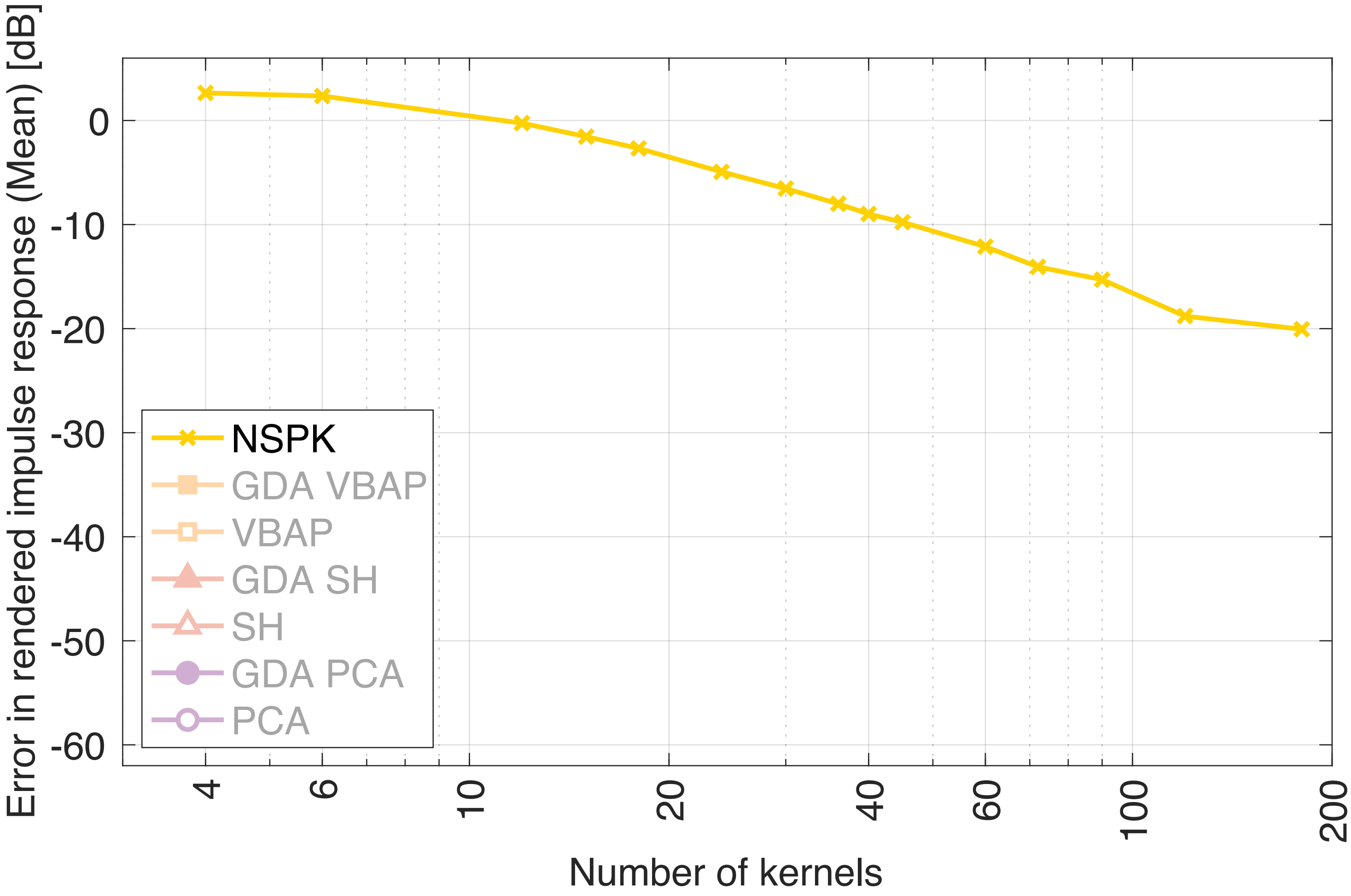


Aligned

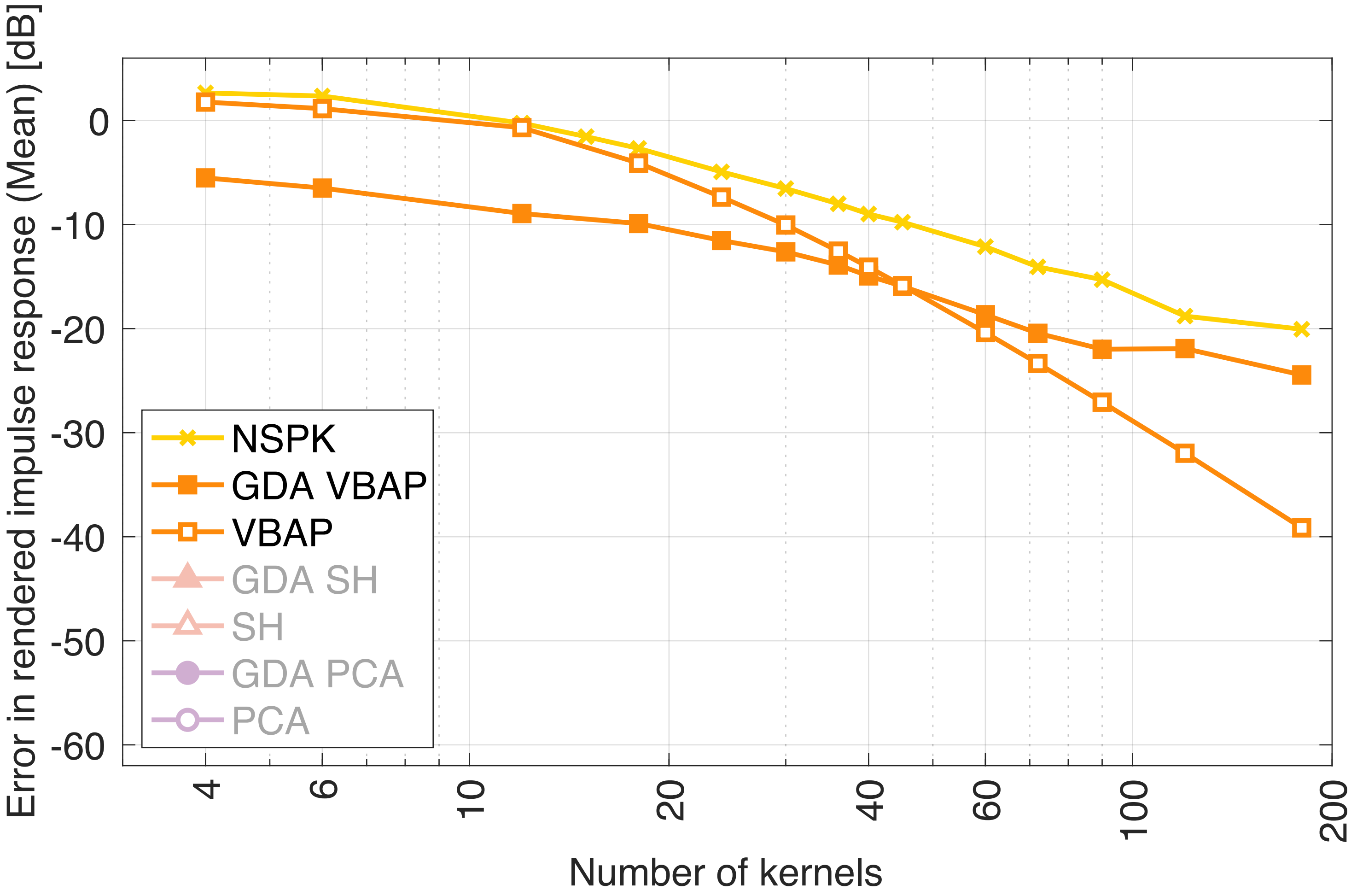
Evaluation - accuracy

- Ground truth defined on 1 degree grid in horizontal plane
- For each method, reconstruct impulse response for each direction of arrival using varying number of kernels
- Compute error with respect to ground truth in each direction

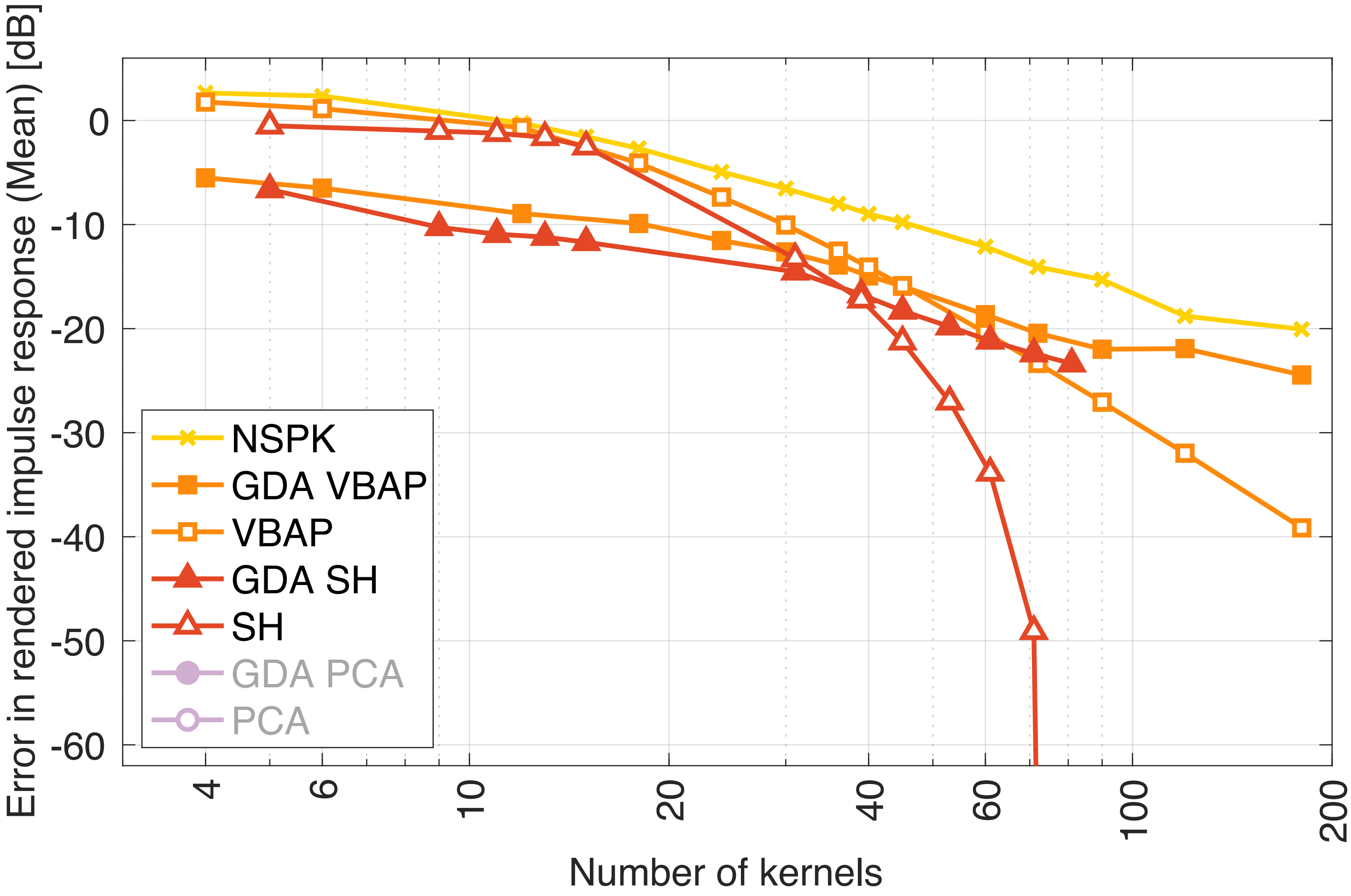
Mean error over all directions



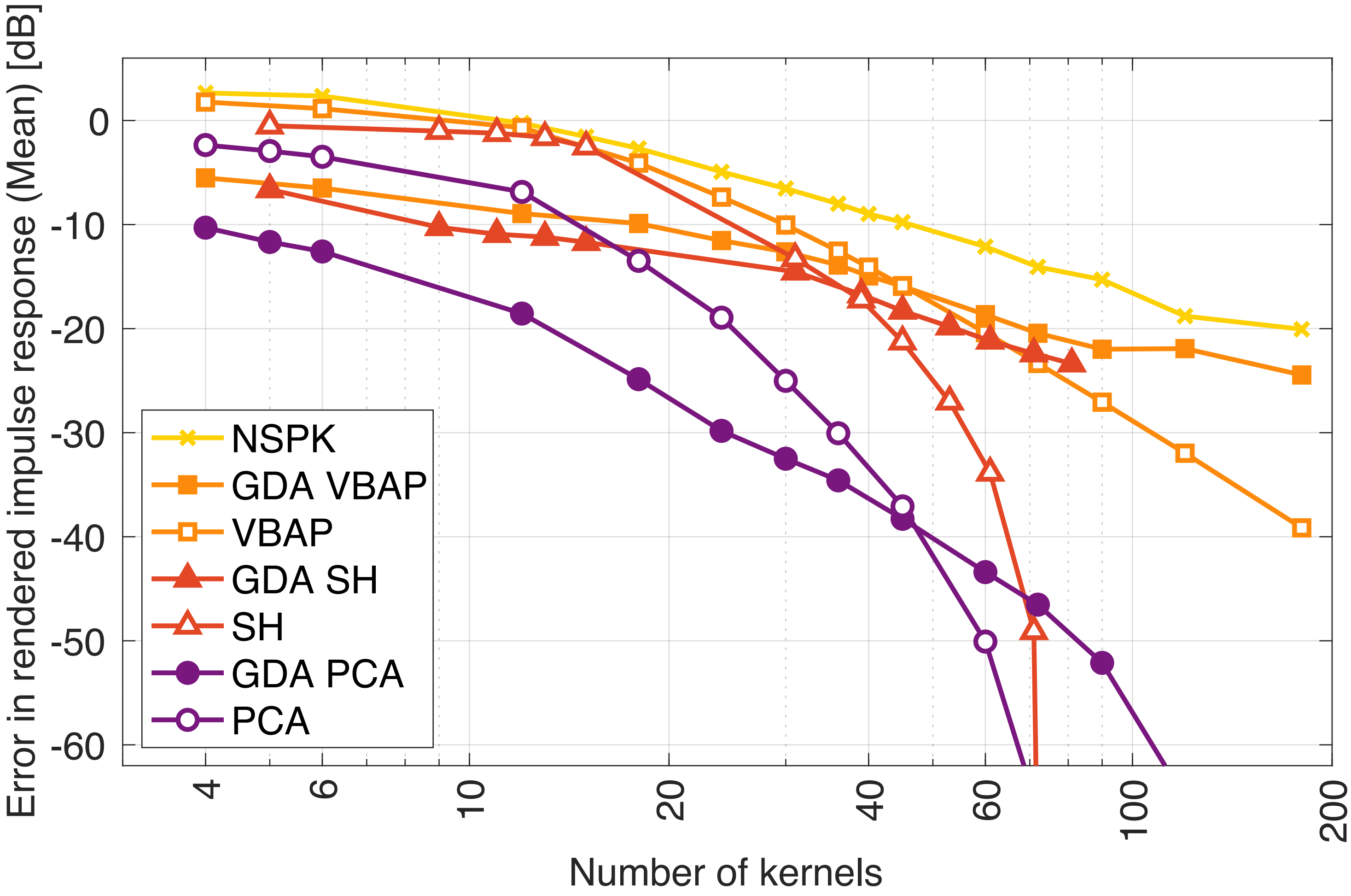
Mean error over all directions



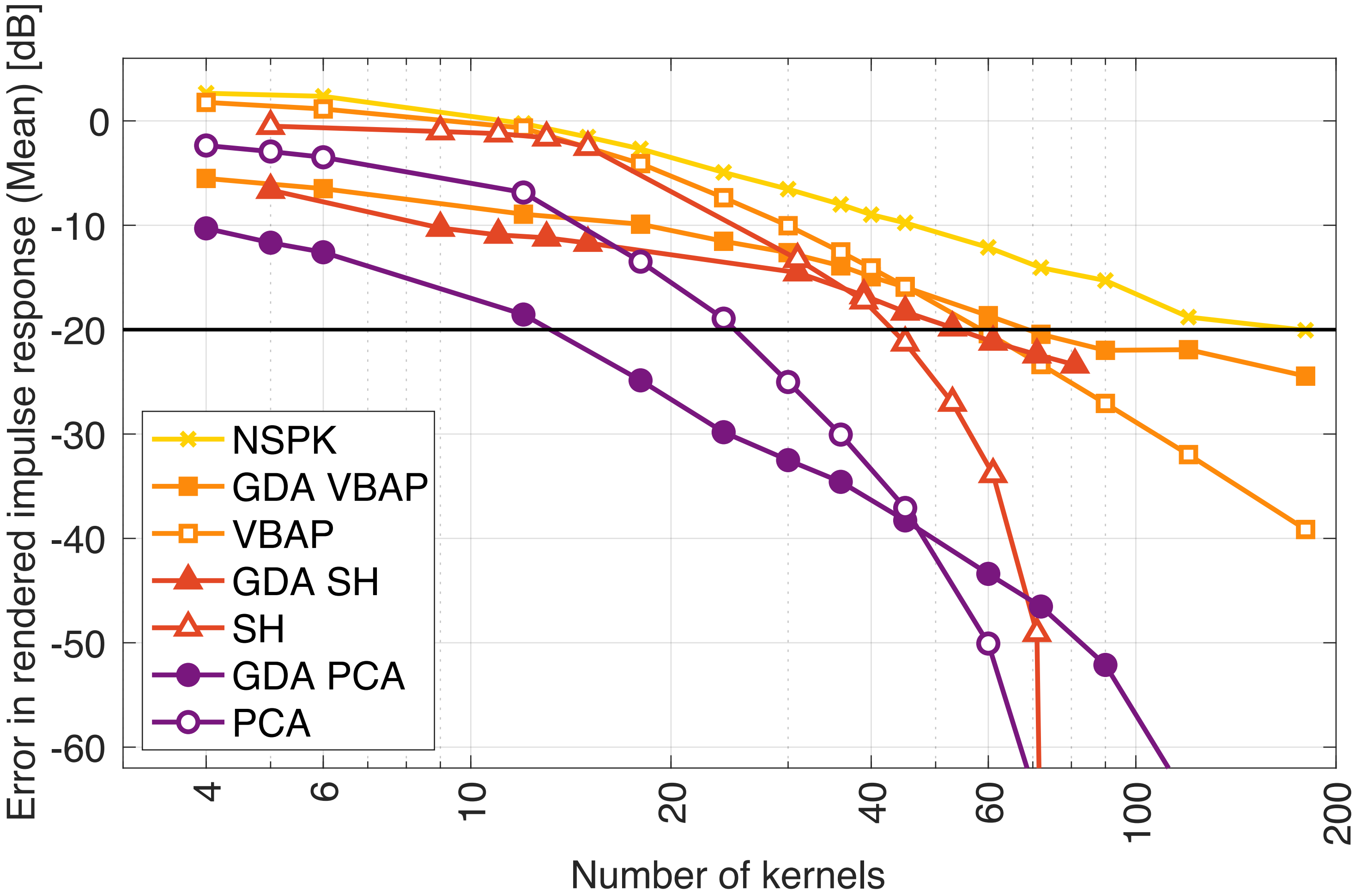
Mean error over all directions



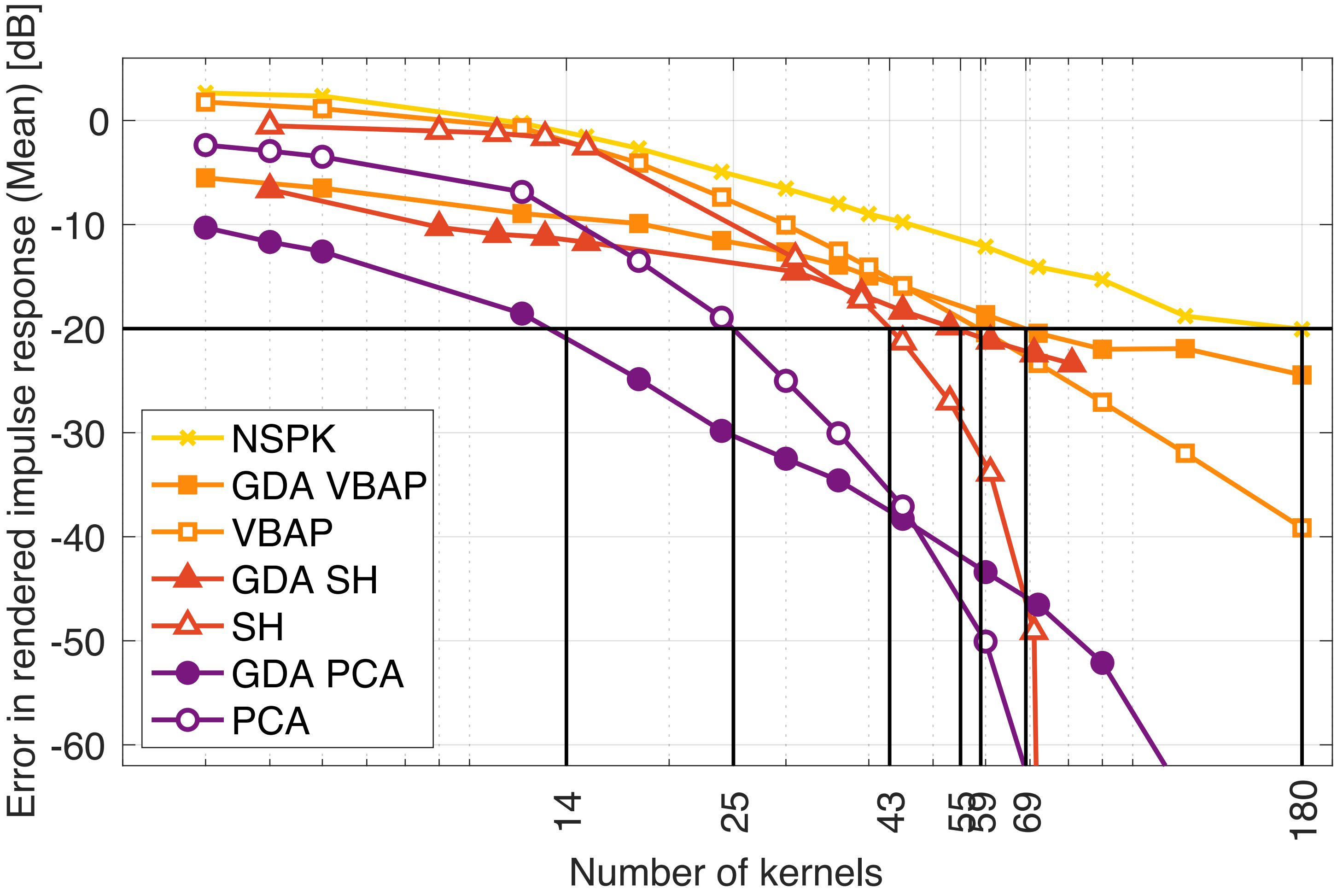
Mean error over all directions



Mean error over all directions

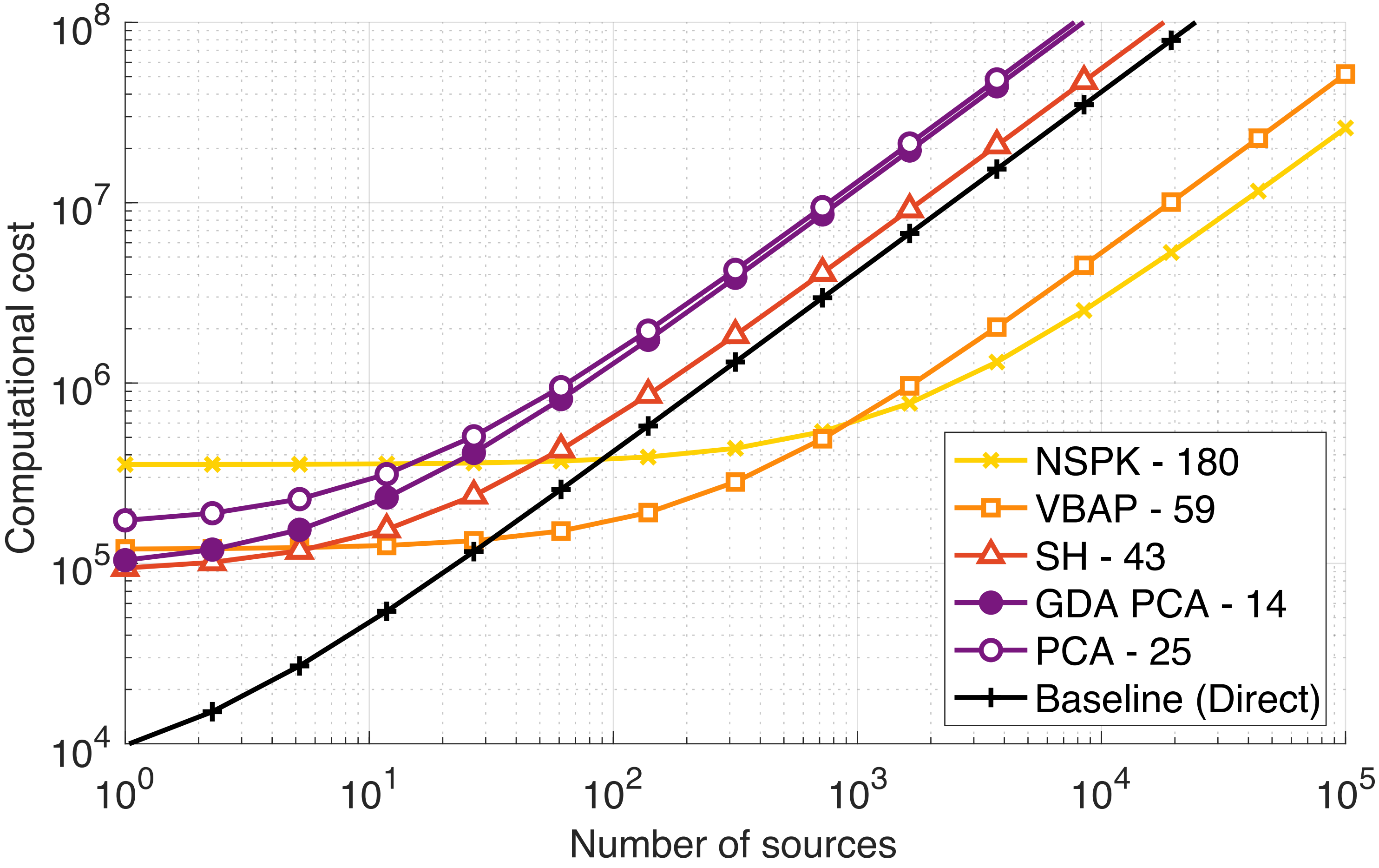


Mean error over all directions



Imperial College London

Computational cost



Summary

- Simulation of dynamic sound scenes for listener-in-the-loop experiments
- Evaluated several pipelines in terms of the accuracy verses number of kernels
- Time aligned pipelines achieve the most accurate performance when a limited number of kernels are available
- Computational cost analysis suggests that microphone independent encoding approaches offer better scalability

Imperial College London

Processing pipelines for efficient, physically-accurate simulation of microphone array signals in dynamic sound scenes

Alastair H. Moore, Rebecca R. Vos,
Patrick A. Naylor and Mike Brookes

AUD-34: Acoustic System Identification and Modeling
Friday, 11 June from 14:00 to 14:45 in Eastern Daylight Time