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 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

\[ x_s(t) \rightarrow * \rightarrow \sum \rightarrow y_1(t) \]

\[ \vdots \]

\[ x_s(t) \rightarrow * \rightarrow \sum \rightarrow y_M(t) \]
S wavefronts, M microphones - dynamic

\[ x_s(t) \rightarrow > \rightarrow * \rightarrow \sum \rightarrow y_1(t) \]

Requires 2SM filters – lots of convolution!
Fast convolution

\[ x_s(t) \rightarrow \text{FFT} \rightarrow H \rightarrow \text{FFT}^{-1} \rightarrow y_1(t) \]
S wavefronts, M microphones - dynamic

\[ x_s(t) \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow y_1(t) \]

\[ x_s(t) \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow y_M(t) \]
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

\[ b(t) \rightarrow FFT \rightarrow H \rightarrow \sum \rightarrow FFT^{-1} \rightarrow y_1(t) \]

- **Encoded signal**
- **K channels**

\[ b(t) \rightarrow FFT \rightarrow H \rightarrow \sum \rightarrow FFT^{-1} \rightarrow y_M(t) \]

- **K FFTs**
- **KM filter kernels**
- **M inverse FFTs**
Virtual speaker encoding (1)

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

$$x_s(t) \rightarrow \begin{cases} > \\ < \end{cases} \rightarrow b(t)$$

- 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 wavetable

Time-aligned PCA spatialization is novel approach
Time alignment example

• Front left channel of hearing aid array
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

![Graph showing the mean error over all directions for different methods (NSPK, GDA VBAP, VBAP, GDA SH, SH, GDA PCA, PCA) as the number of kernels increases from 4 to 200. The error is measured in dB. The graph shows a decreasing trend in error as the number of kernels increases.](image-url)
Mean error over all directions

![Graph showing error in rendered impulse response (Mean) over number of kernels for different algorithms. The y-axis represents error in dB ranging from -60 to 0, and the x-axis represents the number of kernels ranging from 4 to 200. Different lines and markers represent different algorithms: NSPK, GDA VBAP, VBAP, GDA SH, SH, GDA PCA, and PCA.]
Mean error over all directions
Mean error over all directions

Graph showing error in rendered impulse response (Mean) [dB] against the number of kernels for different methods: NSPK, GDA VBAP, VBAP, GDA SH, SH, GDA PCA, and PCA.
Mean error over all directions

Error in rendered impulse response (Mean) [dB]

Number of kernels
Mean error over all directions

![Graph showing the mean error over all directions for different kernels and algorithms. The x-axis represents the number of kernels, ranging from 14 to 180, and the y-axis represents the error in rendered impulse response (Mean) in dB. The graph compares NSPK, GDA VBAP, VBAP, GDA SH, SH, GDA PCA, and PCA. Each algorithm has a distinct line and marker style for easy identification.]
Computational cost

- NSPK - 180
- VBAP - 59
- SH - 43
- GDA PCA - 14
- PCA - 25
- Baseline (Direct)
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
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