Machine Learning for Indoor Acoustics

Antoine Deleforge
Inria (Nancy - Grand Est)
« What is the shape of the room? »
« What is the shape of the room? »

« Is the floor made of tiles or carpet? »
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
1) Intro & Background

Sound Propagation

- What is sound?
1) Intro & Background

Sound Propagation

- What is sound?
  - A Mechanical Vibration
1) Intro & Background

Sound Propagation

• What is sound?
  • A Mechanical Vibration
  • A Variation of Air Pressure
1) Intro & Background

Sound Propagation

• What is sound?
  • A Mechanical Vibration
  • A Variation of Air Pressure
  • A 3D Wave

\[ \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 \]
1) Intro & Background

Sound Propagation

• What is sound?
  • A Mechanical Vibration
  • A Variation of Air Pressure
  • A 3D Wave

• Sound has a **speed**: \( c \approx 343 \text{ m/sec} \)

\[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
\]
1) Intro & Background

**Sound Propagation**

- **What is sound?**
  - A Mechanical Vibration
  - A Variation of Air Pressure
  - A 3D Wave

- **Sound has a speed:** $c \approx 343$ m/sec

- **Sound dissipates:** $\approx -6$ dB every doubling of distance
1) Intro & Background

Sound Propagation

- **What is sound?**
  - A Mechanical Vibration
  - A Variation of Air Pressure
  - A 3D Wave

- **Sound has a speed:** \( c \approx 343 \text{ m/sec} \)

- **Sound dissipates:** \( \approx -6 \text{ dB} \) every doubling of distance

- **Sound Interacts:**

\[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
\]
1) Intro & Background

Sound Propagation

• What is sound?
  • A Mechanical Vibration
  • A Variation of Air Pressure
  • A 3D Wave

• Sound has a **speed**: \( c \approx 343 \text{ m/sec} \)

• Sound **dissipates**: \( \approx -6 \text{ dB every doubling of distance} \)

• Sound **Interacts**: Absorption
1) Intro & Background

Sound Propagation

• What is sound?
  • A Mechanical Vibration
  • A Variation of Air Pressure
  • A 3D Wave

• Sound has a speed: \( c \approx 343 \text{ m/sec} \)

• Sound dissipates: \( \approx -6 \text{ dB} \) every doubling of distance

• Sound Interacts:

  Absorption  Transmission
1) Intro & Background

Sound Propagation

- What is sound?
  - A Mechanical Vibration
  - A Variation of Air Pressure
  - A 3D Wave

- Sound has a speed: \( c \approx 343 \text{ m/sec} \)

- Sound dissipates: \( \approx -6 \text{ dB} \) every doubling of distance

- Sound interacts:
  - Absorption
  - Transmission
  - Reflexion
1) Intro & Background

Sound Propagation

- **What is sound?**
  - A Mechanical Vibration
  - A Variation of Air Pressure
  - A 3D Wave

- Sound has a **speed**: $c \approx 343$ m/sec

- **Sound dissipates**: $\approx -6$ dB every doubling of distance

- **Sound Interacts:**
  - Absorption
  - Transmission
  - Reflexion
  - Specular
  - Diffuse

\[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
\]
1) Introduction
1) Introduction

Absorption/Transmission

Diffusion
1) Introduction

Absorption/Transmission

Diffusion

« Reverberation »
1) Introduction

Absorption/Transmission

Diffusion

« Reverberation »

A signal model of reverberation?
1) Intro & Background

The Room Impulse Response

- **Impulse response**: The response of an LTI system to a perfect impulse (*Dirac*).
- **Room Impulse response (RIR)**: Captures the linear filtering effect due to the propagation of sound from a *point source* to a *microphone* inside a room.
1) Intro & Background

The Room Impulse Response

• **Impulse response**: The response of an LTI system to a perfect impulse (*Dirac*).

• **Room Impulse response (RIR)**: Captures the linear filtering effect due to the propagation of sound from a *point source* to a *microphone* inside a room.

\[ \delta(t) \]

Input: \[ \delta(t) \]

\[ 0 \]

Time
1) Intro & Background

The Room Impulse Response

- **Impulse response**: The response of an LTI system to a perfect impulse (*Dirac*).
- **Room Impulse response (RIR)**: Captures the linear filtering effect due to the propagation of sound from a point source to a microphone inside a room.

**Input:** \( \delta(t) \)

**Output:** \( h(t) \)

\[\text{Pressure (Pascals)}\]

\[\text{Time (seconds)}\]
1) Intro & Background

The Room Impulse Response

- **Impulse response**: The response of an LTI system to a perfect impulse (*Dirac*).
- **Room Impulse response (RIR)**: Captures the linear filtering effect due to the propagation of sound from a **point source** to a **microphone** inside a room.

Input: $\delta(t)$

Output: $h(t)$

Direct path

Early reflections = "ECHOES"

Diffuse/late reverberation
1) Intro & Background

The Room Impulse Response

- **Impulse response**: The response of an LTI system to a perfect impulse (*Dirac*).
- **Room Impulse response (RIR)**: Captures the linear filtering effect due to the propagation of sound from a **point source** to a **microphone** inside a room.

**Input**: $\delta(t)$

**Output**: $h(t)$

The Fourier transform $\tilde{h}(\omega)$ of a RIR is called **Room Transfer Function**. It captures the effect of the room in different **frequency bands**.
The Room Impulse Response

- Can be used to « reverberate » any dry sound source signal $s(t)$:

$$x(t) = (h \ast s)(t) \overset{\text{def}}{=} \int_{-\infty}^{+\infty} h(u)s(t-u)du$$

Fourier

$$\tilde{x}(\omega) = \tilde{h}(\omega)\tilde{s}(\omega)$$
1) Intro & Background

The Room Impulse Response

- Can be used to « reverberate » any dry sound source signal $s(t)$:

\[ x(t) = (h \ast s)(t) = \int_{-\infty}^{+\infty} h(u)s(t - u)du \]

\[ \tilde{x}(\omega) = \tilde{h}(\omega)\tilde{s}(\omega) \]

Fourier

www.openair.hosted.york.ac.uk/
1) Intro & Background

The Room Impulse Response

- Can be used to « reverberate » any dry sound source signal $s(t)$:

$$x(t) = (h * s)(t) \overset{\text{def}}{=} \int_{-\infty}^{+\infty} h(u)s(t - u)du$$

$$\tilde{x}(\omega) = \tilde{h}(\omega)\tilde{s}(\omega)$$

- Generalization to multiple microphones:

$$\begin{align*}
x_1(t) &= (h_1 * s)(t) + n_2(t) \\
x_2(t) &= (h_2 * s)(t) + n_2(t) \\
\vdots & \quad \vdots \\
x_M(t) &= (h_M * s)(t) + n_M(t)
\end{align*}$$

www.openair.hosted.york.ac.uk/
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

RIRs
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

RIRs

Reverberated audio signals
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

Room acoustic simulators

RIRs

Reverberated audio signals
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

Room acoustic simulators ✔

Simple convolutions ✔

RIRs

Reverberated audio signals
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

Room acoustic simulators ✓

Simple convolutions ✓

RIRs

Reverberated audio signals
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

Room acoustic simulators ✓
Simple convolutions ✓

RIRs

(Blind)

Reverberated audio signals
1) Intro & Background

- Source & receivers positions & properties
- Room geometry
- Surface properties

Room acoustic simulators ✓
Simple convolutions ✓

RIRs

(Blind)

Reverberated audio signals

Difficult (interesting) inverse problems!
1) Intro & Background

Why do we care?
1) Intro & Background

Why do we care?

1) Indoor noise disturbance

- Make acoustic diagnosis faster / better [16]
1) Intro & Background

Why do we care?

1) Indoor noise disturbance

→ Make acoustic diagnosis faster / better [16]

2) Audio Augmented Reality [6, 17]
1) Intro & Background

Why do we care?

1) Indoor noise disturbance

   Make acoustic diagnosis faster / better [16]

2) Audio Augmented Reality [6, 17]

   ![Virtual source]

3) “Echo-Aware” Audio Signal Processing [7, 8]
   - Hearing aids
   - Vocal assistant devices
   - …
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
2) Virtually Supervised Learning

Audio Inputs
2) Virtually Supervised Learning

Audio Inputs ➔ Features ➔ Feature Vector ➔ Extraction

- Audio Inputs
- Features Extraction
- Feature Vector
2) Virtually Supervised Learning

Audio Inputs → Features Extraction → Feature Vector → Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Audio Inputs → Features Extraction → Feature Vector → Acoustic / geometric parameters

- Physics-Driven Approaches
- Audio Inputs
- Features Extraction
- Feature Vector
- Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches
2) Virtually Supervised Learning

a) Physics-Driven Approaches
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Audio Inputs → Features Extraction → Feature Vector → Forward Physical Model → Acoustic / geometric parameters

Forward Physical Model

\[ \Delta p = 0 \]
2) Virtually Supervised Learning

a) Physics-Driven Approaches

\[
\cos(\alpha) = c \frac{\tau}{h}
\]

Audio Inputs \rightarrow Features Extraction \rightarrow Feature Vector

Forward Physical Model

\[ \square p = 0 \]

Close-form

Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Sabine’s law:

\[ \cos(\alpha) = \frac{c \tau}{h} \]

\[ RT_{60}(b) \approx 0.16 \frac{V}{S\bar{\alpha}(b)} \]

Audio Inputs → Features Extraction → Feature Vector → Forward Physical Model → Close-form → Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

\[ \cos(\alpha) = \frac{c \, \tau}{h} \]

*Sabine’s law:*

\[ RT_{60}(b) \approx 0.16 \frac{V}{S\bar{\alpha}(b)} \]

Forward Physical Model

- No training data needed

Audio Inputs → Features Extraction → Feature Vector → Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Sabine’s law:

\[ \cos(\alpha) = \frac{c \tau}{h} \]

\[ RT_{60}(b) \approx 0.16 \frac{V}{S\alpha(b)} \]

Audio Inputs

Forward Physical Model

Close-form

Features

Extraction

Feature Vector

Acoustic / geometric parameters

- No training data needed
- Computationally efficient
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Sabine’s law:

\[ \cos(\alpha) = \frac{c \tau}{h} \]

\[ RT_{60}(b) \approx 0.16 \frac{V}{S\tilde{\alpha}(b)} \]

Audio Inputs → Features Extraction → Feature Vector → Forward Physical Model

- No training data needed
- Computationally efficient
- Suffers in complex conditions

Close-form

Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

\[
\cos(\alpha) = \frac{c \tau}{h}
\]

\[
RT_{60}(b) \approx 0.16 \frac{V}{S\bar{\alpha}(b)}
\]

Sabine's law:

- **Audio Inputs**
- **Features Extraction**
- **Feature Vector**
- **Forward Physical Model**
- **Close-form**
- **Acoustic / geometric parameters**

- **☑ No training data needed**
- **☑ Computationally efficient**
- **☒ Suffers in complex conditions**
- **☒ Limited**
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Audio Inputs

Features Extraction

Feature Vector

Forward Physical Model

\[ \Delta p = 0 \]

No close-form

Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Optimization-based inversion

\[ \arg\min_{x \in \Sigma} \| y - A(x) \| \]

Forward Physical Model

\[ \Box p = 0 \]

No close-form

Audio Inputs → Features Extraction → Feature Vector → Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Optimization-based inversion

\[
\arg\min_{x \in \Sigma} \| y - A(x) \|
\]

Forward Physical Model

\[ \Box p = 0 \]

No close-form

No training data needed

Audio Inputs → Features Extraction → Feature Vector → Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Optimization-based inversion

\[ \arg\min_{x \in \Sigma} \| y - A(x) \| \]

[10, 11, 18]

Forward Physical Model

\[ \nabla p = 0 \]

No training data needed

\[ \times \] Non-Convex / Hard to inverse

No close-form

Audio Inputs

Features Extraction

Feature Vector

Acoustic / geometric parameters
2) Virtually Supervised Learning

a) Physics-Driven Approaches

Optimization-based inversion

\[ \arg\min_{x \in \Sigma} \| y - A(x) \| \]

Forward Physical Model

\[ p = 0 \]

- No training data needed
- Non-Convex / Hard to inverse
- Sensitive to model mismatch

Audio Inputs

Features Extraction

Feature Vector

Acoustic / geometric parameters

[10, 11, 18]
b) Real-Data-Driven Approaches [1, 2, 3, 6]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches [1, 2, 3, 6]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches \[1, 2, 3, 6\]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches \([1, 2, 3, 6]\)

- Audio Inputs
- Features Extraction
- Feature Vector
- Acoustic / geometric parameters
- Other Sensors
- Annotations

Audio Inputs

Training Audio Inputs

Features Extraction

Feature Vector
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches\[1, 2, 3, 6\]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches [1, 2, 3, 6]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches [1, 2, 3, 6]
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches \([1, 2, 3, 6]\)

Check on Youtube: https://youtu.be/mhOlCVpY7iA
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches \([1, 2, 3, 6]\)

Check on Youtube: https://youtu.be/mhOlcVpY7iA
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches [1, 2, 3, 6]

Check on Youtube: https://youtu.be/mhOlcVpY7iA

Audio Inputs

Features Extraction

Feature Vector

Training Data

Machine Learning

Learned Model

Annotations

Other Sensors

✓ Very Accurate
✓ Noise-Robust
✗ Room-specific

Acoustic / geometric parameters
2) Virtually Supervised Learning

b) Real-Data-Driven Approaches [1, 2, 3, 6]

Check on Youtube: https://youtu.be/mhOlcVpY7iA

Audio Inputs

Features Extraction

Feature Vector

Training Data

Annotations

Other Sensors

Machine Learning

Learned Model

✓ Very Accurate
✓ Noise-Robust
✗ Room-specific
✗ Costly to acquire

Acoustic / geometric parameters
c) **Virtually-Supervised Learning**

\[4, 5, 9, 16, 17\]
2) Virtually Supervised Learning

c) **Virtually-Supervised Learning**

\[4, 5, 9, 16, 17\]
2) Virtually Supervised Learning

c) Virtually-Supervised Learning
[4, 5, 9, 16, 17]

Audio Inputs

Simulated Training Signals

Acoustic Simulation

Forward Physical Model

Simulated Training Signals

Acoustic / geometric parameters

Audio Inputs

Features Extraction

Feature Vector

Acoustic Simulation
2) Virtually Supervised Learning

c) **Virtually-Supervised Learning**

\[4, 5, 9, 16, 17\]

- **Audio Inputs**
  - Features Extraction
  - Feature Vector
  - Acoustic / geometric parameters

- **Simulated Training Signals**
  - Acoustic Simulation

- **Training Data**
  - Forward Physical Model

- **Annotations**
2) Virtually Supervised Learning

c) Virtually-Supervised Learning

[4, 5, 9, 16, 17]
2) Virtual Supervised Learning

c) **Virtually-Supervised Learning**

[4, 5, 9, 16, 17]

![Diagram of Virtually-Supervised Learning process]

- **Audio Inputs**
  - Features Extraction
  - Feature Vector
- **Simulated Training Signals**
  - Features Extraction
  - Acoustic Simulation
- **Training Data**
  - Annotations
  - Machine Learning
  - Learned Model
  - **Forward Physical Model**
    - $p = 0$
  - **Unlimited data for free**
- **Acoustic / geometric parameters**
2) Virtually Supervised Learning

c) ***Virtually-Supervised Learning***

\[4, 5, 9, 16, 17\]

![Diagram of Virtually-Supervised Learning process]

Audio Inputs -> Features Extraction -> Feature Vector

- **Simulated Training Signals**
- **Acoustic Simulation**
- **Features Extraction**
- **Training Data**
- **Annotations**
- **Forward Physical Model**
  \[\Box p = 0\]

Machine Learning -> Learned Model

- **Unlimited data for free**
- **Robustness**

- **Acoustic / geometric parameters**
2) Virtually Supervised Learning

c) Virtually-Supervised Learning

[4, 5, 9, 16, 17]
2) Virtually Supervised Learning

c) Virtually-Supervised Learning

\[4, 5, 9, 16, 17\]

Audio Inputs

Features Extraction

Feature Vector

Simulated Training Signals

Features Extraction

Training Data

Forward Physical Model

\[ \square p = 0 \]

Annotations

Learned Model

\(\checkmark\) Unlimited data for free

\(\checkmark\) Robustness

\(\checkmark\) Physics-based

\(?\) Real-data generalisation

Machine Learning

Acoustic / geometric parameters
## 2) Virtually Supervised Learning

### RIR Simulation Trade-offs

<table>
<thead>
<tr>
<th>Realism vs. Computational complexity</th>
<th>Diversity vs. Training set size</th>
</tr>
</thead>
</table>
2) Virtually Supervised Learning

### RIR Simulation Trade-offs

<table>
<thead>
<tr>
<th>Realism vs. Computational complexity</th>
<th>Diversity vs. Training set size</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Discretized wave equation solvers (e.g. FDTD)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
\]
### 2) Virtually Supervised Learning

## RIR Simulation Trade-offs

<table>
<thead>
<tr>
<th>Realism vs. Computational complexity</th>
<th>Diversity vs. Training set size</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Discretized wave equation solvers (e.g. FDTD)</td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0$</td>
<td>✓ Solve everything</td>
</tr>
<tr>
<td></td>
<td>☢ Intractable above ~4 kHz</td>
</tr>
</tbody>
</table>

Antoine.Deleforge@inria.fr
## RIR Simulation Trade-offs

### Realism vs. Computational complexity
- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  - Solve everything
  - Intractable above ~4 kHz
- Image source method [13]

### Diversity vs. Training set size
- Solve everything
- Intractable above ~4 kHz
## RIR Simulation Trade-offs

### Realism vs. Computational complexity

<table>
<thead>
<tr>
<th>Discretized wave equation solvers (e.g. FDTD)</th>
<th>Diversity vs. Training set size</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 )</td>
<td><strong>✓</strong> Solve everything</td>
</tr>
<tr>
<td>✘ Intractable above ~4 kHz</td>
<td></td>
</tr>
</tbody>
</table>

### Image source method [13]

- **✓** Fast (for low reflection orders)
- ✘ Doesn’t capture low-freq effects
- ✘ Specular reflections only
2) Virtually Supervised Learning

## RIR Simulation Trade-offs

### Realism vs. Computational complexity

<table>
<thead>
<tr>
<th>Discretized wave equation solvers (e.g. FDTD)</th>
<th>Solve everything</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 )</td>
<td>Intractable above ~4 kHz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image source method [13]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Fast (for low reflection orders)</td>
<td>✓ Doesn’t capture low-freq effects</td>
</tr>
<tr>
<td>✗ Specular reflections only</td>
<td></td>
</tr>
</tbody>
</table>

### Diversity vs. Training set size

<table>
<thead>
<tr>
<th>Energy-based / Ray-based / Particle-based methods</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Image source method [13]:*

- Fast (for low reflection orders)
- Doesn’t capture low-freq effects
- Specular reflections only
2) Virtually Supervised Learning

## RIR Simulation Trade-offs

### Realism vs. Computational complexity
- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  - Solve everything
  - Intractable above ~4 kHz
- Image source method [13]
  - Fast (for low reflection orders)
  - Doesn’t capture low-freq effects
  - Specular reflections only
- Energy-based / Ray-based / Particle-based methods
  - Versatile
  - Doesn’t capture low-freq effects
  - Approx. TOAs

### Diversity vs. Training set size
2) Virtually Supervised Learning

## RIR Simulation Trade-offs

### Realism vs. Computational complexity

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>Solvability</th>
</tr>
</thead>
</table>
| Discretized wave equation solvers (e.g. FDTD)                         | \[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
\]                                                      | ✔ Solve everything                       |
|                                                                      |                                                                        | ✗ Intractable above ~4 kHz                |
|                                                                      |                                                                        | ✗ Doesn’t capture low-freq effects        |
|                                                                      |                                                                        | ✗ Specular reflections only              |
| Energy-based / Ray-based / Particle-based methods                     |                                                                        | ✔ Versatile                              |
|                                                                      |                                                                        | ✗ Doesn’t capture low-freq effects        |
|                                                                      |                                                                        | ✗ Approx. TOAs                           |

### Diversity vs. Training set size

- **Simulators efficiently combining the last two:** RoomSim [14], Pyroomacoustics [15]
### RIR Simulation Trade-offs

#### Realism vs. Computational complexity

- Discretized wave equation solvers (e.g. FDTD)
  \[ \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 \]
  - ✓ Solve everything
  - ✗ Intractable above ~4 kHz

- Image source method [13]
  - ✓ Fast (for low reflection orders)
  - ✗ Doesn’t capture low-freq effects
  - ✗ Specular reflections only

- Energy-based / Ray-based / Particle-based methods
  - ✓ Versatile
  - ✗ Doesn’t capture low-freq effects
  - ✗ Approx. TOAs

#### Diversity vs. Training set size

- Room size? *Toilet, Office, Airport Hall*

---

**Simulators efficiently combining the last two:**

RoomSim [14], Pyroomacoustics [15]
2) Virtually Supervised Learning

RIR Simulation Trade-offs

Realism vs. Computational complexity

- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  ✔️ Solve everything
  ❌ Intractable above ~4 kHz

- Image source method \[13\]
  ✔️ Fast (for low reflection orders)
  ❌ Doesn’t capture low-freq effects
  ❌ Specular reflections only

- Energy-based / Ray-based / Particle-based methods
  ✔️ Versatile
  ❌ Doesn’t capture low-freq effects
  ❌ Approx. TOAs

Diversity vs. Training set size

- Room size? Toilet, Office, Airport Hall
- Room shape? Shoebox, Auditorium, Underground cave

• Simulators efficiently combining the last two:
  RoomSim \[14\], Pyroomacoustics \[15\]
2) Virtually Supervised Learning

RIR Simulation Trade-offs

Realism vs. Computational complexity

- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  ✓ Solve everything
  ✗ Intractable above ~4 kHz

- Image source method [13]
  ✓ Fast (for low reflection orders)
  ✗ Doesn’t capture low-freq effects
  ✗ Specular reflections only

- Energy-based / Ray-based / Particle-based methods
  ✓ Versatile
  ✗ Doesn’t capture low-freq effects
  ✗ Approx. TOAs

Diversity vs. Training set size

- Room size? Toilet, Office, Airport Hall
- Room shape? Shoebox, Auditorium, Underground cave
- Room acoustics? Abbey Road studio, Cathedral

Simulators efficiently combining the last two:
RoomSim [14], Pyroomacoustics [15]
2) Virtually Supervised Learning

RIR Simulation Trade-offs

<table>
<thead>
<tr>
<th>Realism vs. Computational complexity</th>
<th>Diversity vs. Training set size</th>
</tr>
</thead>
</table>
| - Discretized wave equation solvers (e.g. FDTD)  
  \( \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 \)  
  ✓ Solve everything  
  ❌ Intractable above ~4 kHz |
| - Image source method [13]  
  ✓ Fast (for low reflection orders)  
  ❌ Doesn’t capture low-freq effects  
  ❌ Specular reflections only |
| - Energy-based / Ray-based / Particle-based methods  
  ✓ Versatile  
  ❌ Doesn’t capture low-freq effects  
  ❌ Approx. TOAs |
| - Simulators efficiently combining the last two:  
  RoomSim [14], Pyroomacoustics [15] |
| - Room size? Toilet, Office, Airport Hall |
| - Room shape? Shoebox, Auditorium, Underground cave |
| - Room acoustics? Abbey Road studio, Cathedral |
| - Source/receiver types? Omnidirectional, Cardiod, Human speaker, Hearing aids |
## RIR Simulation Trade-offs

### Realism vs. Computational complexity

- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  - Solve everything
  - Intractable above \(~4\) kHz

- Image source method \([13]\)
  - Fast (for low reflection orders)
  - Doesn’t capture low-freq effects
  - Specular reflections only

- Energy-based / Ray-based / Particle-based methods
  - Versatile
  - Doesn’t capture low-freq effects
  - Approx. TOAs

- Simulators efficiently combining the last two:
  - RoomSim \([14]\), Pyroomacoustics \([15]\)

### Diversity vs. Training set size

- Room size? *Toilet, Office, Airport Hall*
- Room shape? *Shoebox, Auditorium, Underground cave*
- Room acoustics? *Abbey Road studio, Cathedral*
- Source/receiver types? *Omnidirectional, Cardiod, Human speaker, Hearing aids*
- Random shoebox rooms with:
  - length/width in [2m, 10m]
  - height in [2m, 6m]
2) Virtually Supervised Learning

RIR Simulation Trade-offs

Realism vs. Computational complexity
- Discretized wave equation solvers (e.g. FDTD)
  \[ \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0 \]
  ✓ Solve everything
  ✗ Intractable above ~4 kHz
- Image source method [13]
  ✓ Fast (for low reflection orders)
  ✗ Doesn’t capture low-freq effects
  ✗ Specular reflections only
- Energy-based / Ray-based / Particle-based methods
  ✓ Versatile
  ✗ Doesn’t capture low-freq effects
  ✗ Approx. TOAs
- Simulators efficiently combining the last two: RoomSim [14], Pyroomacoustics [15]

Diversity vs. Training set size
- Room size? Toilet, Office, Airport Hall
- Room shape? Shoebox, Auditorium, Underground cave
- Room acoustics? Abbey Road studio, Cathedral
- Source/receiver types? Omnidirectional, Cardiod, Human speaker, Hearing aids
- Random shoebox rooms with:
  - length/width in [2m, 10m]
  - height in [2m, 6m]
- Omnidirectional sources and receivers placed uniformly at random in the room with 1m « safe distance »
2) Virtually Supervised Learning

RIR Simulation Trade-offs

Realism vs. Computational complexity

- Discretized wave equation solvers (e.g. FDTD)
  \[
  \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0
  \]
  ✓ Solve everything
  ✗ Intractable above ~4 kHz

- Image source method [13]
  ✓ Fast (for low reflection orders)
  ✗ Doesn’t capture low-freq effects
  ✗ Specular reflections only

- Energy-based / Ray-based / Particle-based methods
  ✓ Versatile
  ✗ Doesn’t capture low-freq effects
  ✗ Approx. TOAs

Simulators efficiently combining the last two:
RoomSim [14], Pyroomacoustics [15]

Diversity vs. Training set size

- Room size? Toilet, Office, Airport Hall
- Room shape? Shoebox, Auditorium, Underground cave
- Room acoustics? Abbey Road studio, Cathedral
- Source/receiver types? Omnidirectional, Cardiod, Human speaker, Hearing aids
- Random shoebox rooms with:
  - length/width in [2m, 10m]
  - height in [2m, 6m]
- Omnidirectional sources and receivers placed uniformly at random in the room with 1m “safe distance”
- 10k – 100k RIRs
RIR Simulation Trade-offs

What about the surface acoustic properties?
RIR Simulation Trade-offs

What about the surface acoustic properties?

- Diffusion coefficients:
  - Same random value for all surfaces
  - In \([0, 0.3]\) < 500 Hz, in \([0.2, 1]\) > 500 Hz
RIR Simulation Trade-offs

What about the surface acoustic properties?

- Diffusion coefficients:
  - Same random value for all surfaces
  - \( \text{in } [0, 0.3] < 500 \text{ Hz, in } [0.2, 1] > 500 \text{ Hz} \)

\( \text{Meant to emulate typical surface diffusivity and room furnishing [X]} \)
2) Virtually Supervised Learning

RIR Simulation Trade-offs

What about the surface acoustic properties?

- Diffusion coefficients:
  - Same random value for all surfaces
  - In [0, 0.3] < 500 Hz, in [0.2, 1] > 500 Hz
  
  Meant to emulate typical surface diffusivity and room furnishing [X]

- Absorption coefficients:
  - Typically defined in octave bands \( b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz} \)
2) Virtually Supervised Learning

RIR Simulation Trade-offs

What about the surface acoustic properties?

- Diffusion coefficients:
  - Same random value for all surfaces
  - \[ \text{In } [0, 0.3] < 500 \text{ Hz, in } [0.2, 1] > 500 \text{ Hz} \]

- Absorption coefficients:
  - Typically defined in octave bands \( b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz} \)

\[ \text{A « reflectivity-biased » acoustic sampling strategy [16]} \]

For each surface type (wall, ceiling, floor) toss a coin:
- **On heads**: frequency-independent absorption coefficient in \([0, 0.12]\) for all (hard surfaces)
- **On tails**: random absorption profile inside realistic ranges (treated surface)
RIR Simulation Trade-offs

What about the surface acoustic properties?

- Diffusion coefficients:
  - Same random value for all surfaces
  - \( \ln [0, 0.3] < 500 \text{ Hz}, \text{ in } [0.2, 1] > 500 \text{ Hz} \)

- Absorption coefficients:
  - Typically defined in octave bands (\( b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz} \))

A « reflectivity-biased » acoustic sampling strategy [16]

For each surface type (wall, ceiling, floor) toss a coin:
- **On heads**: frequency-independent absorption coefficient in \([0, 0.12]\) for all (hard surfaces)
- **On tails**: random absorption profile inside realistic ranges (treated surface)
2) Virtually Supervised Learning

RIR Simulation Trade-offs

What about the surface acoustic properties?

• Diffusion coefficients:
  • Same random value for all surfaces
  • In [0, 0.3] < 500 Hz, in [0.2, 1] > 500 Hz

• Absorption coefficients:
  • Typically defined in octave bands \( b \in \{125, 250, 500, \ldots, 4000\} \) Hz

A « reflectivity-biased » acoustic sampling strategy [16]

For each surface type (wall, ceiling, floor) toss a coin:
  • **On heads**: frequency-independent absorption coefficient in [0, 0.12] for all (hard surfaces)
  • **On tails**: random absorption profile inside realistic ranges (treated surface)

\[
\begin{align*}
\text{RT60 (s)} & \\
0 & 1000 \quad 1 \quad 2 \quad 2.5
\end{align*}
\]
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
3) Examples and Results

Example 1: RIR $\rightarrow$ Mean absorption profile of surfaces \[16\]

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b)S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]
Example 1: RIR $\rightarrow$ Mean absorption profile of surfaces \cite{16}

$$\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})$$

Absorption coefficient in $[0,1]$
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces \([16]\)

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b)S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces \[^{16}\]

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

\[ \bar{\alpha}(b) \text{ def } = \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz}) \]

1) MLP

Input RIR → Pre-processing → FC (8000 x 128) → ELU → FC (128 x 64) → ELU → FC (64 x 32) → ELU → FC (32 x 16) → ELU → FC (16 x 6) → Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \), \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \), \ldots , \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)
3) Examples and Results

**Example 1: RIR -> Mean absorption profile of surfaces [16]**

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]

**1) MLP**

![Diagram of the MLP model with pre-processing steps and fully connected layers](image)

- Resample to 16 kHz
- Crop to 0.5 sec
- Normalize to max = 1
- Additive white Gaussian noise (SNR= 30 dB)

→ Input vector in \( \mathbb{R}^{8000} \)

\[
\begin{align*}
\bar{\alpha}_{\text{NN}}(125\text{Hz}) \\
\bar{\alpha}_{\text{NN}}(250\text{Hz}) \\
\vdots \\
\bar{\alpha}_{\text{NN}}(4000\text{Hz})
\end{align*}
\]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces \[16\]

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{i} \alpha_i(b) S_i \quad \left( b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz} \right)
\]

1) MLP

Input RIR

Pre-processing → FC (8000 x 128) → ELU → FC (128 x 64) → ELU → FC (64 x 32) → ELU → FC (32 x 16) → ELU → FC (16 x 6) → Sigmoid

Fully Connected layer:

\[ h_{i+1} = W h_i + b \]
3) Examples and Results

Example 1: RIR \(-\rightarrow\) Mean absorption profile of surfaces \[16\]

\[
\tilde{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]

1) MLP

Input RIR

Pre-processing

FC (8000 x 128) \[\text{ELU} \]

FC (128 x 64) \[\text{ELU} \]

FC (64 x 32) \[\text{ELU} \]

FC (32 x 16) \[\text{ELU} \]

FC (16 x 6) \[\text{Sigmoid} \]

\[\tilde{\alpha}_{\text{NN}}(125\text{Hz}) \]

\[\tilde{\alpha}_{\text{NN}}(250\text{Hz}) \]

\[\vdots\]

\[\tilde{\alpha}_{\text{NN}}(4000\text{Hz}) \]

Exponential Linear Unit:

\[
y = \begin{cases} 
  x, & x \geq 0 \\
  \alpha(e^x - 1), & x < 0 
\end{cases}
\]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

\[
\bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]

1) MLP

Input RIR

Pre-processing

FC (8000 x 128)

ELU

FC (128 x 64)

ELU

FC (64 x 32)

ELU

FC (32 x 16)

ELU

FC (16 x 6)

Sigmoid

\[
\bar{\alpha}_{NN}(125\text{Hz})
\]

\[
\bar{\alpha}_{NN}(250\text{Hz})
\]

\[\vdots\]

\[
\bar{\alpha}_{NN}(4000\text{Hz})
\]

Sigmoid:

\[
y = \frac{1}{1 + e^{-x}}
\]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

\[ \bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz}) \]

1) MLP

Input RIR → Pre-processing → FC (8000 x 128) ELU → FC (128 x 64) ELU → FC (64 x 32) ELU → FC (32 x 16) ELU → FC (16 x 6) Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \)

- \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \)
- \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)

- Output vector in \([0, 1]^6\)
- Loss Function = Mean Squared Error
- Optimal parameters on dev. set over 200 epochs
3) Examples and Results

Example 1: RIR $\rightarrow$ Mean absorption profile of surfaces [16]

\[ \bar{\alpha}(b) \xdef \frac{1}{S_{tot}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz}) \]

1) MLP

```
Input RIR
    Pre-processing
    FC (8000 x 128)
        ELU
    FC (128 x 64)
        ELU
    FC (64 x 32)
        ELU
    FC (32 x 16)
        ELU
    FC (16 x 6)
        Sigmoid
```

\[ \bar{\alpha}_{NN}(125\text{Hz}) \]
\[ \bar{\alpha}_{NN}(250\text{Hz}) \]
\[ \vdots \]
\[ \bar{\alpha}_{NN}(4000\text{Hz}) \]

2) CNN

```
Input RIR
    Pre-processing
    Conv1D (1x64)x33
        ReLU
    Conv1D (64x32)x17
        ReLU
    Conv1D (32x16)x9
        ReLU
    FC (2000 x 32)
        ReLU
    FC (32 x 6)
        Sigmoid
```

\[ \bar{\alpha}_{NN}(125\text{Hz}) \]
\[ \bar{\alpha}_{NN}(250\text{Hz}) \]
\[ \vdots \]
\[ \bar{\alpha}_{NN}(4000\text{Hz}) \]
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

\[ \bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz}) \]

1) MLP

Input RIR → Pre-processing → Conv1D (1x64)x33 → ReLU → Conv1D (64x32)x9 → ReLU → MaxPool (4) → FC (2000 x 32) → ReLU → FC (32 x 6) → Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \) → … → \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)

2) CNN

Input RIR → Pre-processing → Conv1D (1x64)x33 → ReLU + MaxPool (4) → Conv1D (64x32)x17 → ReLU + MaxPool (4) → Conv1D (32x16)x9 → ReLU + MaxPool (4) → FC (2000 x 32) → ReLU → FC (32 x 6) → Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \) → … → \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)

1D convolutional layer:
- 64 input channels
- 32 output channels
- Kernel size: 17
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

\[ \bar{\alpha}(b) \overset{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b)S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz}) \]

1) MLP

Input RIR → Pre-processing → FC (8000 x 128) → ELU → Rectified Linear Unit: \( y = \max(0, x) \) → MaxPool(4) → FC (32 x 16) → ELU → FC (16 x 6) → Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)

2) CNN

Input RIR → Pre-processing → Conv1D (1x64)x33 → ReLU → MaxPool(4) → Conv1D (64x32)x17 → ReLU → MaxPool(4) → Conv1D (32x16)x9 → ReLU → MaxPool(4) → FC (2000 x 32) → ReLU → FC (32 x 6) → Sigmoid → \( \bar{\alpha}_{\text{NN}}(125\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(250\text{Hz}) \) → \( \bar{\alpha}_{\text{NN}}(4000\text{Hz}) \)
3) Examples and Results

**Example 1: RIR -> Mean absorption profile of surfaces** [16]

\[
\bar{\alpha}(b) \stackrel{\text{def}}{=} \frac{1}{S_{\text{tot}}} \sum_{\text{surface } i} \alpha_i(b) S_i \quad (b \in \{125, 250, 500, \ldots, 4000\} \text{ Hz})
\]

1) **MLP**

- **Input RIR**
- Pre-processing
- FC (8000 x 128)
- ELU
- FC (128 x 64)
- ELU
- FC (64 x 32)
- ELU
- FC (32 x 16)
- ELU
- FC (16 x 6)
- Sigmoid

\[
\begin{align*}
\bar{\alpha}_{\text{NN}}(125\text{Hz}) \\
\bar{\alpha}_{\text{NN}}(250\text{Hz}) \\
\vdots \\
\bar{\alpha}_{\text{NN}}(4000\text{Hz})
\end{align*}
\]

2) **CNN**

- **Input RIR**
- Pre-processing
- Conv1D (1x64)x33
  - ReLU + MaxPool(4)
- Conv1D (64x32)x17
  - ReLU + MaxPool(4)
- Conv1D (32x16)x9
  - ReLU + MaxPool(4)
- Conv1D (2000 x 32)
  - ReLU
- FC (32 x 6)
  - Sigmoid

\[
\begin{align*}
\bar{\alpha}_{\text{NN}}(125\text{Hz}) \\
\bar{\alpha}_{\text{NN}}(250\text{Hz}) \\
\vdots \\
\bar{\alpha}_{\text{NN}}(4000\text{Hz})
\end{align*}
\]
3) Examples and Results

**Example 1: RIR -> Mean absorption profile of surfaces** [16]

- **Simulated test results**: RoomSim, real absorption profiles, 5 room geometries, 500 RIRs

- Comparing two training sets (Unif., RB) and the two neural networks (MLP, CNN) against Sabine and Eyring’s laws (given true $S_{tot}$ and $V$)
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- **Simulated test results**: RoomSim, real absorption profiles, 5 room geometries, 500 RIRs

- Comparing two training sets (Unif., RB) and the two neural networks (MLP, CNN) against Sabine and Eyring’s laws (given true $S_{tot}$ and $V$)
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- **Simulated test results**: RoomSim, real absorption profiles, 5 room geometries, 500 RIRs

- Comparing two training sets (Unif., RB) and the two neural networks (MLP, CNN) against Sabine and Eyring’s laws (given true $S_{tot}$ and $V$)

→ Training on uniformly sampled acoustics fails to outperform reverberation theory
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- **Simulated test results**: RoomSim, real absorption profiles, 5 room geometries, 500 RIRs

- Comparing two training sets (Unif., RB) and the two neural networks (MLP, CNN) against Sabine and Eyring’s laws (given true $S_{tot}$ and $V$)

  - Training on uniformly sampled acoustics fails to outperform reverberation theory

  - Training on the reflectivity-biased set significantly outperforms both baselines
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- Encouraging generalizability to real data (900 RIRs, 10 room configurations [12])
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces \[16\]

- Encouraging generalizability to real data (900 RIRs, 10 room configurations \[12\])

\[A\] : RIR featuring « nice » reverberation decay

\[B\] : RIR with « unusual » reverberation decay
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- Encouraging generalizability to real data (900 RIRs, 10 room configurations [12])

\[ A \]: RIR featuring « nice » reverberation decay

\[ B \]: RIR with « unusual » reverberation decay

![Image of a room with equipment]

\( \bar{\alpha}(1000\text{Hz}), \) only RIRs in \( A \)

<table>
<thead>
<tr>
<th>( )</th>
<th>Absolute error on ( \bar{\alpha} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>( + )</td>
</tr>
<tr>
<td>R2</td>
<td>( + )</td>
</tr>
<tr>
<td>R3</td>
<td>( + )</td>
</tr>
<tr>
<td>R4</td>
<td>( + )</td>
</tr>
<tr>
<td>R5</td>
<td>( + )</td>
</tr>
<tr>
<td>R6</td>
<td>( + )</td>
</tr>
<tr>
<td>R7</td>
<td>( + )</td>
</tr>
<tr>
<td>R8</td>
<td>( + )</td>
</tr>
<tr>
<td>R9</td>
<td>( + )</td>
</tr>
<tr>
<td>R10</td>
<td>( + )</td>
</tr>
</tbody>
</table>
3) Examples and Results

Example 1: RIR -> Mean absorption profile of surfaces [16]

- Encouraging generalizability to real data (900 RIRs, 10 room configurations [12])

$\mathcal{A}$: RIR featuring « nice » reverberation decay

$\mathcal{B}$: RIR with « unusual » reverberation decay

$\alpha(1000\text{Hz})$, only RIRs in $\mathcal{A}$

$\bar{\alpha}(1000\text{Hz})$, RIRs in $\mathcal{A}$ vs $\mathcal{B}$
Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms
Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms

Room impulse responses look like this:
3) Examples and Results

Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms

Room impulse responses look like this:

![Room impulse responses](image)

- Direct path
- First echo
- Early echoes
- Diffuse tail

A « pic-nic » dataset
3) Examples and Results

Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms

Room impulse responses look like this:

![Room impulse responses](image)

- Direct path
- First echo
- Early echoes
- Diffuse tail
- TDOA
3) Examples and Results

Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms

Room impulse responses look like this:

![Room impulse responses diagram](image-url)

- Direct path
- First echo
- Early echoes
- Diffuse tail
- TDOA
- iTDOA
3) Examples and Results

Example 2: Blind echo estimation [4]

A « pic-nic » dataset

- One Source
- Two microphones
- Nearest surface is most reflective
- Random shoe-box rooms

Room impulse responses look like this:
3) Examples and Results

Example 2: Blind echo estimation [4]

Simulated 2-channel white noise

Level Difference Spectrogram

Phase Difference Spectrogram

Avg.

1534-dim. feature vector

3 fully-connected 128-units hidden layers

TDOA

iTDOA

TDOE
3) Examples and Results

**Example 2: Blind echo estimation** [4]

- Simulated 2-channel white noise
- Level Difference Spectrogram
- Phase Difference Spectrogram

**Results on test set**

<table>
<thead>
<tr>
<th>Input</th>
<th>TDOA</th>
<th>iTDOA</th>
<th>TDOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRAGE wn</td>
<td>0.18</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>MIRAGE wn+n</td>
<td>0.68</td>
<td>0.69</td>
<td>0.89</td>
</tr>
<tr>
<td>MIRAGE sp</td>
<td>0.31</td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td>MIRAGE sp+n</td>
<td>0.99</td>
<td>0.98</td>
<td>1.48</td>
</tr>
<tr>
<td>GCC-PHAT wn</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT wn+n</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp</td>
<td>0.32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp+n</td>
<td>1.38</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3) Examples and Results

**Example 2: Blind echo estimation** [4]

Simulated 2-channel white noise

Level Difference Spectrogram

Phase Difference Spectrogram

**Results on test set**

- Good results with white noise

<table>
<thead>
<tr>
<th>Input</th>
<th>TDOA</th>
<th>iTDOA</th>
<th>TDOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRAGE wn</td>
<td>0.18</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>MIRAGE wn+n</td>
<td>0.68</td>
<td>0.69</td>
<td>0.89</td>
</tr>
<tr>
<td>MIRAGE sp</td>
<td>0.31</td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td>MIRAGE sp+n</td>
<td>0.99</td>
<td>0.98</td>
<td>1.48</td>
</tr>
<tr>
<td>GCC-PHAT wn</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT wn+n</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp</td>
<td>0.32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp+n</td>
<td>1.38</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3) Examples and Results

Example 2: Blind echo estimation [4]

Simulated 2-channel white noise

Level Difference Spectrogram

Phase Difference Spectrogram

Avg.

1534-dim. feature vector

3 fully-connected 128-units hidden layers

TDOA
iTDOA
TDOE

Results on test set

✔️ Good results with white noise
❌ Poor generalization to noisy speech and real data

<table>
<thead>
<tr>
<th>Input</th>
<th>TDOA</th>
<th>iTDOA</th>
<th>TDOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRAGE wn</td>
<td>0.18</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>MIRAGE wn+n</td>
<td>0.68</td>
<td>0.69</td>
<td>0.89</td>
</tr>
<tr>
<td>MIRAGE sp</td>
<td>0.31</td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td>MIRAGE sp+n</td>
<td>0.99</td>
<td>0.98</td>
<td>1.48</td>
</tr>
<tr>
<td>GCC-PHAT wn</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT wn+n</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp</td>
<td>0.32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCC-PHAT sp+n</td>
<td>1.38</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3) Examples and Results

Example 3: Blind room parameter estimation [17]

- Joint estimation of **volume**, **total surface**, $RT_{60}(b)$ and $\alpha(b)$ from multiple, multichannel noisy speech recordings
3) Examples and Results

**Example 3: Blind room parameter estimation** [17]

- Joint estimation of **volume**, **total surface**, $RT_{60}(b)$ and $\tilde{\alpha}(b)$ from multiple, multichannel noisy speech recordings
3) Examples and Results

Example 3: Blind room parameter estimation [17]

- Joint estimation of **volume**, **total surface**, \(RT_{60}(b)\) and \(\bar{\alpha}(b)\) from multiple, multichannel noisy speech recordings

- A maximum-likelihood cost-function:
  \[ L_{\theta}(x, y) = - \log p_{\theta}(y|x) = - \log \mathcal{N}(y; \mu_{\theta}(x), \sigma_{\theta}^2(x)) \]
  \[ = \frac{1}{2} \sum_{d=1}^{D} \log \sigma_{d,\theta}(x) + \frac{(y_d - \mu_{d,\theta}(x))^2}{\sigma_{d,\theta}^2(x)} \]
3) Examples and Results

**Example 3: Blind room parameter estimation** [17]

- Joint estimation of **volume**, **total surface**, $\text{RT}_{60}(b)$ and $\bar{\alpha}(b)$ from multiple, multichannel noisy speech recordings

  - A maximum-likelihood cost-function:
    \[
    L_\theta(x, y) = -\log p_\theta(y | x) = -\log \mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2(x))
    \]
    \[
    = \frac{1}{2} \sum_{d=1}^{D} \log \sigma_{d, \theta}^2(x) + \frac{(y_d - \mu_{d, \theta}(x))^2}{\sigma_{d, \theta}^2(x)}
    \]

- Allows aggregating multiple source-receiver recordings via Bayes’ theorem:

  \[
  p_\theta(y_d | \bar{x} = [x_1, \ldots, x_J]) = \mathcal{N}(y_d; \bar{\mu}_{d, \theta}(\bar{x}), 1/\gamma_{d, \theta}^2(\bar{x}))
  \]
  \[
  \bar{\mu}_{d, \theta}(\bar{x}) = \sum_{j=1}^{J} \gamma_{d, \theta}(x_j) \mu_{d, \theta}(x_j), \quad \gamma_{d, \theta}(\bar{x}) = \sum_{j=1}^{J} \gamma_{d, \theta}(x_j)
  \]
3) Examples and Results

Example 3: Blind room parameter estimation [17]

![Graphs showing α, RT₆₀(s), S (m²), and V (m³) vs. #pos for different methods.]

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th># pos</th>
<th>α</th>
<th>RT₆₀(s)</th>
<th>S</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6]</td>
<td>SC</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>SC</td>
<td>1</td>
<td>0.061</td>
<td>0.134</td>
<td>129.6</td>
<td>154.5</td>
</tr>
<tr>
<td>Ours</td>
<td>SC</td>
<td>5</td>
<td>0.060</td>
<td>0.097</td>
<td>125.8</td>
<td>149.1</td>
</tr>
<tr>
<td>Ours</td>
<td>SC+IC</td>
<td>1</td>
<td>0.084</td>
<td>0.101</td>
<td>89.4</td>
<td>107.6</td>
</tr>
<tr>
<td>Ours</td>
<td>SC+IC</td>
<td>5</td>
<td>0.094</td>
<td>0.062</td>
<td>50.2</td>
<td>68.8</td>
</tr>
</tbody>
</table>

- 2-channels help V and S
- Multiple observations help
- Poor results for mean absorption below 1000 Hz

Mean results on 3 real rooms [12] (30 rec. per room)
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
OUTLINE

1) Intro & Background
2) Virtually-Supervised Learning
3) Examples and Results
4) Conclusions and Outlook
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is .... far from solved! In particular at low-freq (<1000 Hz)
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is …. far from solved! In particular at low-freq (<1000 Hz)
- Next challenge: estimation of individual surfaces/objects in the room?
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is .... far from solved! In particular at low-freq (<1000 Hz)
- Next challenge: estimation of individual surfaces/objects in the room?
- A variety of open and useful inverse problems in acoustics
4) Conclusions & Outlook

• Recovering the geometry and acoustic profiles of a room from a clap recording is …. far from solved! In particular at low-freq (<1000 Hz)

• Next challenge: estimation of individual surfaces/objects in the room?

• A variety of open and useful inverse problems in acoustics

• *How to develop « hybrid » models that are jointly driven by data, signal processing and physics?*
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is …. far from solved! In particular at low-freq (<1000 Hz)

- Next challenge: estimation of individual surfaces/objects in the room?

- A variety of open and useful inverse problems in acoustics

- *How to develop « hybrid » models that are jointly driven by data, signal processing and physics?*

- Promising directions: domain adaptation, transfer learning, generative models, adversarial networks, neural RIR generation, …
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is .... far from solved! In particular at low-freq (<1000 Hz)
- Next challenge: estimation of individual surfaces/objects in the room?
- A variety of open and useful inverse problems in acoustics

*How to develop « hybrid » models that are jointly driven by data, signal processing and physics?*

- Promising directions: domain adaptation, transfer learning, generative models, adversarial networks, neural RIR generation, ...
- *Coming soon:* extension of Pyroomacoustics that loads measured source and receiver directivity profiles
4) Conclusions & Outlook

- Recovering the geometry and acoustic profiles of a room from a clap recording is ... far from solved! In particular at low-freq (<1000 Hz)
- Next challenge: estimation of individual surfaces/objects in the room?
- A variety of open and useful inverse problems in acoustics
- *How to develop « hybrid » models that are jointly driven by data, signal processing and physics?*
- Promising directions: domain adaptation, transfer learning, generative models, adversarial networks, neural RIR generation, ...
- *Coming soon*: extension of Pyroomacoustics that loads measured source and receiver directivity profiles

![Venn diagram with intersections of Acoustics, Signal processing, and Machine learning]

Thank You!  Questions?


