Immersive Spatial Audio Reproduction for VR/AR
Using Room Acoustic Modelling from 360° Images

Hansung Kim, Luca Remaggi, Philip J.B. Jackson and Adrian Hilton
Centre for Vision, Speech and Signal Processing
University of Surrey, UK
Introduction

• Spatial Audio for Immersive Virtual and Augmented reality
  - Human perception relies on both audio and visual information
  - Spatio-temporal synchronisation of sound with visual information improves the sense of presence in VR/AR environments (Larsson 2010)
Introduction

• Best way to reproduce the acoustic of spaces
  - Measuring Room Impulse Response (RIR)

• Problem of RIR measurement for practical applications
  - Too invasive
  - RIR is valid only at a single point of measurement for a static scene
Introduction

• How to evaluate?
  - Coherency – audio-visual information
  - Plausibility (internal reference) – for VR applications
  - Authenticity (external reference) – for AR applications
Overview

• Goal
  - Simple and practical system to estimate room acoustic for plausible reproduction of spatial audio using 360° cameras

• Assumptions
  - Human audio perception is not sensitive enough to recognise differences of sound from the change of geometrical details (JUDD 1932)

• Contributions
  - Approximated room geometry estimation
  - Acoustic room modelling using visual semantic segmentation
  - Objective evaluation of estimated room acoustics
  - VR implementation
Overview

360 stereo image pair

Depth estimation

Object segmentation

3D room layout reconstruction

VR scene with spatial audio

Acoustic material mapping
Capture system

- **Vertical 360 Stereo Capture**
  - Simple 1D matching for depth estimation
  - Real-world scale depth without calibration
  - Less occlusion between cameras
  - Higher accuracy for side regions
Geometry Reconstruction

- Spherical stereo geometry
  - Feature-based dense block matching method*

\[
\begin{align*}
  r_t &= \frac{B}{\left( \frac{\sin \theta_t}{\tan(\theta_t + d)} - \cos \theta_t \right)} \\
  d(\theta) &= \theta_t - \theta_b
\end{align*}
\]

Object and material recognition

- Semantic object segmentation and mapping to acoustic materials
  - SegNet* for semantic image segmentation
  - Mapping materials and acoustic attributes**

3D room modelling

- Final 3D room geometry reconstruction
  - Cuboid fitting and labelling
  - Fitting with point cloud occupancy
VR Scene with Spatial Audio

• Metadata format
  - OBJ for geometry
  - JSON for scene and acoustics information

• VR Platform
  - Unity with Google Resonance Audio package
  - Alternative option: Unreal / Steam Audio
Experiments

• Datasets and estimated depth maps

Listening room (LR)

Usability Lab (UL)
Experiments

- Datasets and estimated depth maps

Meeting room (MR)

Studio (ST)
Experiments

• Computational load
  - Geometry reconstruction
    • Processor: Intel Core i7 3.40 GHz CPU with 32G RAM
    • Processing time: around 5 mins
  - Semantic segmentation
    • Processor: NVIDIA Tesla M2090 GPU with 5GB RAM
    • Processing time: around 3 mins

• Evaluation of room layout reconstruction

<table>
<thead>
<tr>
<th>Ground-truth (m³)</th>
<th>Estimated (m³)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.61×4.28×2.33</td>
<td>5.52×4.35×2.36</td>
<td>1.3</td>
</tr>
<tr>
<td>5.57×5.20×2.91</td>
<td>5.92×4.95×2.95</td>
<td>27.0</td>
</tr>
<tr>
<td>5.64×5.05×2.90</td>
<td>5.77×5.17×2.98</td>
<td>7.6</td>
</tr>
<tr>
<td>17.08×14.55×6.50</td>
<td>16.53×14.87×5.70</td>
<td>13.2</td>
</tr>
</tbody>
</table>
Experiments

- Object recognition and segmentation results

Eigen (ICCV 2015)

SegNet
Experiments

- Object recognition result (heat map) for MR

![Heat map for Eigen (ICCV 2015)](image1)

![Heat map for SegNet](image2)
Experiments

• Room Impulse Response (RIR)
  - Ground-truth RIRs vs. Estimated RIRs
  - Evaluation
    • Early Decay Time (EDT) – early reflections
    • RT60 – late reverberation
  - Just-Noticeable Difference (JND) level
    • 5% for the EDT (Vorlander 1995) and 20% for the RT60 (Meng 2006)
Experiments

• Sound Rendering
  - Ground-truth RIR vs. Estimated RIR
VR Scene Rendering

• VR Demo on Unity with Google Resonance Audio
  - Interactive real-time spatial audio rendering
  - Comparison of Open space vs. Room only vs. Room with objects
VR Scene Rendering

• VR Demo on HTC VIVE Pro headset
  - Audio in this video has been recorded using an external speaker
Conclusion

• Summary
  - Vision-based 3D structure and acoustic property estimation system
  - Reproduction of plausible spatial audio in VR/AR environment
  - VR implementation

• Future work
  - Robust material detection
  - Subjective evaluation of plausibility in VR reproductions
  - Perception with/without visual cue
Thank you very much!

Hansung Kim
h.kim@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey, UK