Graph Theory for the Discovery of Non-Parametric Audio Objects

Christopher Srinivasa*, Martin Bouchard*, Ramin Pichevar^, and Hossein Najaf-Zadeh^

University of Ottawa*
Communications Research Centre Canada^
Outline

I. Problem Statement
II. Pre-processing
III. Proposed Object Extraction Framework
IV. Performance
V. Summary and Future Work
Section I. Problem Statement
Section I: Audio Coding

- Current object-based audio coding methods are parametric: describe a signal as a set of concise audio objects using prior knowledge about the structure of audio

- Example is the harmonic sound: combination of sinusoids described by single parameter (i.e. the fundamental frequency)

- Set of parameters available to model signal must be defined a priori

- Choice of parameters available may not always be optimal
Section I: Proposed Method

- Proposed method defines new type of objects: Non-Parametric Objects (NPO)
- NPOs: combinations of coefficients occurring more than once in time-frequency representation
- Allows extraction of broader class of audio objects (no predefined parameterization space imposed on extracted objects)
- Object: sound which minimizes external cost function unrelated to shape
Section I: Proposed Method

- Proposed object extraction framework if used in full audio coder

- Focus: audio object extraction framework, pre-processing and reconstruction modules
Section II. Pre-processing
Section II: Signal Decomposition

- Ear deciphers signal based on its frequency content via hair cells on membrane
- Achieved by evaluating signal against set of kernels representing different frequencies
  - Algorithm: Perceptual Matching Pursuit
  - Kernels: Gammatones
Section II: Representation

- Resulting representation is a spikegram
- Each spike described by four components
  \[ s^i = \langle s^i_{mag}, s^i_{ang}, s^i_{pos}, s^i_{chan} \rangle \]
- Signal reconstructed by summing all spikes
Section III. Proposed Object Extraction Framework
Section III: Audio Object Extraction

- Each spike a vertex, linked to other vertices via labelled edges based on similarities of vertices and their relationships

- Objects mined as frequent subgraphs where each subgraph is an instance of an object
Section III: Audio Object Extraction

- Each object characterized by the recurring edge labels involved in each of its instances
- For object recording intent, only stars are valid

- Framework summarized as follows
Section III: Graph Formation

- Can form graph with any spike components
- As example, use only frequency and time
- Each edge described by a feature vector
  \[ z_{e(i,j)}^{(i,j)} = \langle s_{\text{chan}}^{v_i}, s_{\text{chan}}^{v_j}, s_{\text{pos}}^{v_j} - s_{\text{pos}}^{v_i} \rangle \]
- Edges are labelled by clustering their vectors
- Quality Threshold clustering ensures all vectors assigned to same cluster do not deviate from each other by more than predefined threshold(s)
Section III: Subgraph Extraction

- Frequent stars mined with iterative star extraction algorithm and cost function
- Cost function based on fixed-length bit scheme
- Star search finds all frequent stars in graph and computes extraction cost for each one
- Frequent star with minimal cost is recognized and its instances are recorded and removed
- Each minimal frequent star becomes an object
Section III: Audio Object Recording

- Each object is recorded with the anchor point in each instance and other spikes recorded as representative relationships.

[Diagram showing object relationships]

- In the example, only the time and frequency spike components involved in the objects are recorded.

- Objects can be reconstructed by placing copies of the representative relationships at each instance.
Section IV. Performance
Section IV: Setup

- Tested on five audio excerpts
- Gain evaluated using a cumulative cost function
- All spike components considered for overall gain
- Relative gain factors in only the compressed spike components (i.e. time and frequency)
- Quality evaluated using Signal to Noise Ratio, Segmental Signal to Noise Ratio, and Perceptual Evaluation of Audio Quality model
Results show average overall and relative gains of 15.90% and 23.53% with a PEAQ score of -0.395

<table>
<thead>
<tr>
<th>Audio</th>
<th>Objects</th>
<th>Gain (%)</th>
<th>SNR (dB)</th>
<th>SSNR (dB)</th>
<th>PEAQ Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castanet</td>
<td>1178</td>
<td>O 28.67</td>
<td>10.41</td>
<td>12.18</td>
<td>-0.752</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 43.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vibraphone</td>
<td>2305</td>
<td>O 24.44</td>
<td>21.18</td>
<td>16.06</td>
<td>-0.734</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 36.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Speech</td>
<td>1318</td>
<td>O 8.71</td>
<td>16.16</td>
<td>17.96</td>
<td>-0.191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 12.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Speech</td>
<td>1061</td>
<td>O 8.39</td>
<td>19.47</td>
<td>20.45</td>
<td>-0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 12.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Piano</td>
<td>1395</td>
<td>O 9.28</td>
<td>20.60</td>
<td>21.03</td>
<td>-0.175</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 13.54</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section IV: Results

- New types of objects also found
Section V. Summary and Future Work
Section V: Summary

- Novel graph theoretic framework applied to discover new types of audio objects: NPOs
- Shape of NPOs not restricted by any a priori psychoacoustic knowledge
- New types of objects discovered while achieving compression and maintaining a high audio quality
Section V: Future Work

- Further evaluate performance with informal/formal listening tests and larger corpus
- Create a full end-to-end audio coder centered around the proposed framework
- Make quantitative comparisons with other coders in industry
Questions