Deep Unsupervised Representation Learning for Abnormal Heart Sound Classification

**Shahin Amiriparian**\(^1,2\), Maximilian Schmitt\(^1\), Nicholas Cummins\(^1\), Kun Qian\(^1,2\), Fengquan Dong\(^3\), Björn Schuller\(^1,4\)

1. Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
2. Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
3. Shenzhen University General Hospital, Shenzhen, P. R. China.
4. Group on Language, Audio, and Music, Imperial College London, UK

shahin.amiriparian@tum.de
Problem Description

• A myriad of acoustic sounds
  – Reflecting our **physiological** and **pathological states**

• Classification of **abnormal heart sounds**

• Feature **engineering** vs. deep representation **learning**
Research Aims of Paper

- **Extraction** of expert-designed features

- **Quantisation** of expert-designed features

- **Learning** task-dependent deep representation
Heart Sound Dataset

- Heart Sounds Shenzhen (HSS) corpus
- 845 recordings (30 seconds on average)
- Total length: 7 hours

<table>
<thead>
<tr>
<th>Partition</th>
<th>normal</th>
<th>mild</th>
<th>moderate/severe</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train.</td>
<td>84</td>
<td>276</td>
<td>142</td>
<td>502</td>
</tr>
<tr>
<td>Devel.</td>
<td>32</td>
<td>98</td>
<td>50</td>
<td>180</td>
</tr>
<tr>
<td>Test</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>163</td>
</tr>
</tbody>
</table>
Heart Sound Dataset

- Recording equipment
  - electronic stethoscope

- Recording from one of the:
  - auscultatory mitral area
  - aortic valve auscultation area
  - pulmonary valve auscultation area, and
  - auscultatory area of the tricuspid valve.

- 170 independent subjects (55 f and 115 m)
  - Mean age: 65.4 years
  - Standard deviation: 13.2 years
Feature Engineering

• ComParE feature set (6373 dimensional):
  – Prosodic
  – Spectral
  – Cepstral, and
  – Voice quality low-level descriptors (LLDs)
Feature Engineering

• **Bag-of-Audio-Words:**
  
  – Quantisation of ComParE features
  
  – openXBOW [1]
  
  – Forming **sparse** fixed-length **histogram** representation of an audio clip

---

• Using auDeep [2]:
  – Deep representation learning from raw audio

---


• Hann windows with width $w$ and overlap $0.5w$

• Computing $N_{mel}$ of log-scaled Mel frequency bands

• Normalising the Mel-spectra $[-1, 1]$

• Amplitude clipping $\{-30, -45, -60, -75\} \text{dB}$
Recurrent Sequence to Sequence Autoencoders

Amiriparian, et. al

Deep Unsupervised Representation Learning for Abnormal Heart Sound Classification
Autoencoder Hyperparameters

• Tested
  - $N_{\text{layer}} \in \{2, 3, 4\}$
  - $N_{\text{unit}} \in \{64, 128, 256, 512\}$
  - All combinations of unidirectional and bidirectional encoder and decoder RNN

• Best configuration
  - $N_{\text{layer}} = 2$
  - $N_{\text{unit}} = 256$
  - Unidirectional encoder
  - Bidirectional decoder
## Results

### Engineered Features

<table>
<thead>
<tr>
<th>System</th>
<th>Dimensionality</th>
<th>UAR [%]</th>
<th>C</th>
<th>Devel.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPARE</td>
<td>6373</td>
<td>$10^{-6}$</td>
<td>41.1</td>
<td>44.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$10^{-5}$</td>
<td>44.5</td>
<td>45.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$10^{-4}$</td>
<td>50.3</td>
<td>46.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$10^{-3}$</td>
<td>44.5</td>
<td>40.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$10^{-2}$</td>
<td>43.2</td>
<td>41.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>Dimensionality</th>
<th>UAR [%]</th>
<th>C</th>
<th>Devel.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoAW</td>
<td>250</td>
<td>$10^{-3}$</td>
<td>43.1</td>
<td>43.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>$10^{-3}$</td>
<td>42.3</td>
<td>47.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>$10^{-2}$</td>
<td>43.7</td>
<td>41.0</td>
<td></td>
</tr>
</tbody>
</table>

### Learnt Deep Representations

<table>
<thead>
<tr>
<th>System</th>
<th>Dimensionality</th>
<th>UAR [%]</th>
<th>C</th>
<th>Devel.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUDEEP: Individual Feature Sets</td>
<td>1024</td>
<td>$2 \cdot 10^{-2}$</td>
<td>32.8</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>$5 \cdot 10^{-4}$</td>
<td>38.4</td>
<td>40.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>$6 \cdot 10^{-2}$</td>
<td>39.6</td>
<td>45.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>$8 \cdot 10^{-3}$</td>
<td>36.9</td>
<td>41.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>$4 \cdot 10^{-3}$</td>
<td>35.2</td>
<td><strong>47.9</strong></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

• Promising results with **sequence to sequence autoencoders**

• Effective alternative to **expert-designed** feature sets

• Fully **unsupervised** autoencoder training

• **Variable-length** input
Future Research

- Applying **data augmentation** techniques
- Comparison and/or fusion with Deep Convolutional Generative **Adversarial Networks**
- Feature selection and **dimensionality reduction**