Wavelets Revisited for the Classification of Acoustic Scenes

Kun Qian\textsuperscript{1,2,3}, Zhao Ren\textsuperscript{1,2}, Vedhas Pandit\textsuperscript{1,2}, Zijing Yang\textsuperscript{1,2}, Zixing Zhang\textsuperscript{1} and Björn Schuller\textsuperscript{1,3,4}

\textsuperscript{1} Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, Germany
\textsuperscript{2} MISP Group, MMK, Technische Universität München, Germany
\textsuperscript{3} Chair of Complex & Intelligent Systems, Universität Passau, Germany
\textsuperscript{4} GLAM – Group on Language, Audio & Music, Imperial College London, UK

\section*{Introduction}

\textbf{Motivation:}
- Acoustic Scene Classification (ASC) is challenging and useful.
- Wavelets are efficient in analysis of non-stationary signals.

\textbf{Contributions:}
- Explore the performance of optimized features extracted by Wavelet Transformation (WT) and Wavelet Packet Transformation (WPT).

\section*{Wavelet Features}

The WPT Energy (WPTE) is defined as:
\[ E_{WPTE} = \log \frac{\sum_{j,k} (w_{j,k})^2}{N_{j,k}}, \]
where \( w_{j,k} \) are the coefficients calculated by WPT from the analyzed signal at the subspace \( \Omega_{j,k} \). \( N_{j,k} \) is the total number of wavelet coefficients in the \( k \)-th subband at the \( j \)-th decomposition level.

The WT Energy (WTE) is defined as:
\[ E_{WTE} = \frac{\left[ \frac{1}{N_j} \sum_j (w_j)^2 \right]^2}{100}, \]
where \( w_j \) are the coefficients generated by DWT at the \( j \)-th decomposition level. Furthermore, the \textit{mean}, \textit{variance}, \textit{waveform length} (the sum of the absolute differences), and \textit{entropy} are calculated from the above vector as low-level descriptors (LLDs).

Totally, there are \( 2^{J_{max}+1} - 1 \) WPTE based LLDs, and \( 4 \times (J_{max} + 1) \) WTE based LLDs. \( J_{max} \) is the maximum level for wavelet decomposition.

Wavelet Energy Features (WEF): WPTE+WTE.

\section*{Experimental Setup}

\textbf{Dataset:}
- DCASE 2017 Database:
  - 312 segments of 10 seconds in each of the 15 classes
  - total duration is 13 hours
  - 15 acoustic scene classes: beach, bus, cafe/restaurant, car, city centre, forest path, grocery store, home, library, metro station, office, park, residential area, train, and tram

\textbf{Classifiers:}
- Support Vector Machines (SVMs)
- Gated Recurrent Neural Networks (GRNNs)
- Decision Fusion by Margin Sampling Value (MSV)

\section*{Experimental Results}

Table: Performance comparison between different feature set by SVMs.

<table>
<thead>
<tr>
<th></th>
<th>Fold1</th>
<th>Fold2</th>
<th>Fold3</th>
<th>Fold4</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComParE</td>
<td>76.8</td>
<td>76.8</td>
<td>75.7</td>
<td>82.5</td>
<td>77.9</td>
</tr>
<tr>
<td>WPTE</td>
<td>75.8</td>
<td>78.3</td>
<td>75.7</td>
<td>77.8</td>
<td>78.0</td>
</tr>
<tr>
<td>WEF</td>
<td>79.9</td>
<td>79.0</td>
<td>72.8</td>
<td>73.2</td>
<td>77.8</td>
</tr>
<tr>
<td>ComParE+WPTE</td>
<td>80.6</td>
<td>82.3</td>
<td>79.9</td>
<td>85.5</td>
<td>82.1</td>
</tr>
<tr>
<td>ComParE+WEF</td>
<td>82.3</td>
<td>83.9</td>
<td>81.7</td>
<td>83.7</td>
<td>82.9</td>
</tr>
<tr>
<td>WPTE+WEF</td>
<td>80.1</td>
<td>79.8</td>
<td>76.4</td>
<td>80.0</td>
<td>79.1</td>
</tr>
<tr>
<td>ComParE+WPTE+WEF</td>
<td>82.4</td>
<td>81.9</td>
<td>81.7</td>
<td>84.7</td>
<td>83.2</td>
</tr>
</tbody>
</table>

Table: Performance comparison between different feature sets by GRNNs.

<table>
<thead>
<tr>
<th></th>
<th>Fold1</th>
<th>Fold2</th>
<th>Fold3</th>
<th>Fold4</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComParE</td>
<td>79.3</td>
<td>74.8</td>
<td>77.0</td>
<td>81.0</td>
<td>78.0</td>
</tr>
<tr>
<td>WPTE</td>
<td>73.6</td>
<td>71.8</td>
<td>71.1</td>
<td>74.4</td>
<td>72.6</td>
</tr>
<tr>
<td>WEF</td>
<td>77.7</td>
<td>76.6</td>
<td>73.1</td>
<td>76.8</td>
<td>76.0</td>
</tr>
<tr>
<td>ComParE+WPTE</td>
<td>82.1</td>
<td>79.0</td>
<td>80.1</td>
<td>84.8</td>
<td>81.5</td>
</tr>
<tr>
<td>ComParE+WEF</td>
<td>83.2</td>
<td>81.2</td>
<td>81.3</td>
<td>84.7</td>
<td>82.6</td>
</tr>
<tr>
<td>WPTE+WEF</td>
<td>75.8</td>
<td>72.7</td>
<td>74.3</td>
<td>77.6</td>
<td>76.9</td>
</tr>
<tr>
<td>ComParE+WPTE+WEF</td>
<td>82.6</td>
<td>81.8</td>
<td>81.0</td>
<td>85.0</td>
<td>82.6</td>
</tr>
</tbody>
</table>

\section*{Conclusion}

- Wavelet features can perform well for ASC.
- Wavelet features help improve the final performance of ASC when fused with temporal and spectral features.
- Future work:
  - Evaluate system in noisy conditions.
  - Feature selection and enhancement.
  - Use more sophisticated deep models.

\section*{Acknowledgements}

This work was partially supported by the China Scholarship Council (CSC), the European Union’s Seventh Framework under grant agreements No.338164 (ERC StG iHEARU), and the EU’s Horizon 2020 Programme through the Innovation Action No.645094 (SEWA).