Wavelets Revisited for the Classification of Acoustic Scenes

Kun Qian[1,2,3], Zhao Ren[1,3], Vedhas Pandit[1,3], Zijiang Yang[1,3], Zixing Zhang[3] and Björn Schuller[1,3,4]

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

Introduction

- **Motivation:**
  - Acoustic Scene Classification (ASC) is challenging and useful
  - Wavelets are efficient in analysis of non-stationary signals
- **Contributions:**
  - Explore the performance of optimised features extracted by ***Wavelet Transformation*** (WT) and ***Wavelet Packet Transformation*** (WPT)

Wavelet Features

- The WPT Energy (WPTE) is defined as:

$$\mathbf{E}_{\mathbf{\Omega}_{j,k}} = \log \sqrt{\frac{\sum_{n=1}^{N_{j,k}} (\mathbf{w}_{j,k,n})^2}{N_{j,k}}},$$

where $\mathbf{w}_{j,k,n}$ are the coefficients calculated by WPT from the analysed signal at the subspace $\mathbf{\Omega}_{j,k}$. $N_{j,k}$ 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:

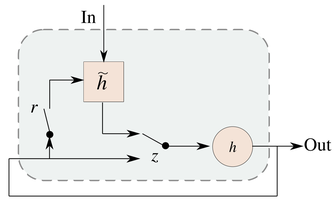
$$\bar{\mathbf{E}}_{\mathbf{\Omega}_j} = \frac{(\mathbf{w}_j)^2}{\sum_{j=1}^{\boldsymbol{J_{max}}} (\mathbf{w}_j)^2} \times 100,$$

where $\mathbf{w}_j$ are the coefficients generated by DWT at the j-th decomposition level. Furthermore, the *mean*, *variance*, *waveform length* (the sum of the absolute differences), and *entropy* are calculated from the above vector as *low level descriptors* (LLDs).

- Totatlly, there are $2^{\boldsymbol{J_{max}}+1} - 1$ WPTE based LLDs, and $4 \times (\boldsymbol{J_{max}} + 1)$ WTE based LLDs. $\boldsymbol{J_{max}}$ is the maximum level for wavelet decomposition.
- Wavelet Energy Features (WEF): WPTE+WTE.

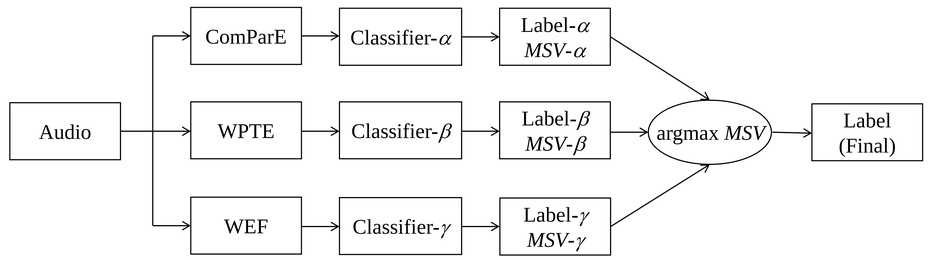
Classifiers

Figure: Diagram of a Grated Recurrent Unit.



- *Support Vector Machines* (SVMs)
- *Gated Recurrent Neural Networks* (GRNNs)
- Decision Fusion by *Margin Sampling Value* (MSV)

Figure: Diagram of a Decision Fusion Process.



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*

Experimental Setup

- features (*functionals* applied to *LLDs*):
  - ComParE: 6 373 features
  - WPTE: 1 020 features
  - WEF: 1 148 features
- standardisation
- maximum wavelet decomposition level $\boldsymbol{J_{max}}$: 7
- SVMs:
  - *linear kernel*
  - C-value is optimised to 0.01, 10 and 0.1 for ComParE, WPTE, and WEF, respectively
- GRNNs:
  - two-layer: 120-60
  - *learning rate*: 0.0002, *drop out rate*: 0.1, *epoch*: 50

Experimental Results

Table: Performance comparison between different feature set by SVMs.

accuracy [%]	Fold1	Fold2	Fold3	Fold4	Mean
ComParE	76.8	76.8	75.7	82.5	77.9
WPTE	76.1	75.9	72.8	78.3	75.7
WEF	79.9	79.0	75.2	77.1	77.8
ComParE+WPTE	80.6	82.3	79.9	85.5	82.1
ComParE+WEF	82.3	83.9	81.7	83.7	82.9
WPTE+WEF	80.1	79.8	76.4	80.0	79.1
ComParE+WPTE+WEF	82.4	83.9	81.7	84.7	**83.2**

Table: Performance comparison between different feature sets by GRNNs.

accuracy [%]	Fold1	Fold2	Fold3	Fold4	Mean
ComParE	79.3	74.8	77.0	81.0	78.0
WPTE	73.6	71.8	71.1	74.1	72.6
WEF	77.7	76.6	73.1	76.8	76.0
ComParE+WPTE	82.1	79.0	80.1	84.8	81.5
ComParE+WEF	83.2	81.2	81.3	84.7	**82.6**
WPTE+WEF	78.5	77.2	74.3	77.6	76.9
ComParE+WPTE+WEF	82.6	81.8	81.0	85.0	**82.6**

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

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).