Feature Engineering and Stacked Echo State Networks for Musical Onset Detection

Peter Steiner peter.steiner@tu-dresden.de // Azarakhsh Jalalvand // Simon Stone // Peter Birkholz

Musical Onset Detection

Musical Onset Detection is one of the most fundamental tasks in music analysis. Here, we clarify frequent questions regarding feature design and standardization. We also propose a novel way of stacking Echo State Networks (ESNs) [2] for onset detection.

Echo State Networks

ESNs are Recurrent Neural Networks (RNNs) that consist of input, reservoir and output weight matrices. [2]

ESNs have achieved comparable results to CNNs for example in image recognition. [3]

Some beneficial advantages of ESNs:
- Suitable for processing temporal data
- Simple training
- Robust against noise

In various cases, multiple ESNs are stacked. Typically, the output of the first ESN is the input of the second ESN.

The second ESN can learn dependencies between the outputs of the first ESN.

Here, we propose a novel way of stacking ESNs:
- The first ESN computes an ODF from features
- The second ESN receives the features as input and the ODF from the first ESN as bias.

Utilized Dataset

1024 minutes of audio files sampled at 44.1 kHz with 27700 annotated and manually corrected onsets. [4]

The dataset is split into eight folds for 8-fold cross-validation. Hyper-parameters were tuned using six folds for training and the seventh for validation. The folds were rotated until every subset has been used for validation exactly one time.

After fixing the hyper-parameters, final models for each fold were trained using 7 folds, and the test scores were computed on the last subset.

Reported scores are the mean values across eight repetitions.

Results

We have slightly outperformed the bidirectional LSTM network [5] by 0.013 F-Measure but fell slightly short by 0.017 to the CNN [6].

Compared to the CNN (289,406), our proposed ESN (66,002) has significantly fewer trainable parameters.

Architecture | Precision | Recall | F-Score
--- | --- | --- | ---
ESN (7) | 0.854 | 0.774 | 0.812
Bidirectional LSTM [5] | 0.892 | 0.855 | 0.873
CNN [6] | 0.917 | 0.889 | 0.903

Feature Engineering

No standardization and transforming feature into bipolar features performed best.

Not all variance of a feature actually carries exploitable information. Thus, statistical normalization in any way decreased the performance.

Different window sizes are able to deal with different kinds of onsets.

Short windows can capture high \( f_1 \) values and fast onsets, whereas long windows are more useful to detect very soft onsets.

The second derivative is beneficial for small reservoirs. Large reservoirs learn it themselves.
- Small reservoirs: Small improvements by incorporating the second derivative.
- Large reservoirs: The performance with and without the second derivative is similar.