Feature Engineering and Stacked Echo State Networks for Musical Onset Detection

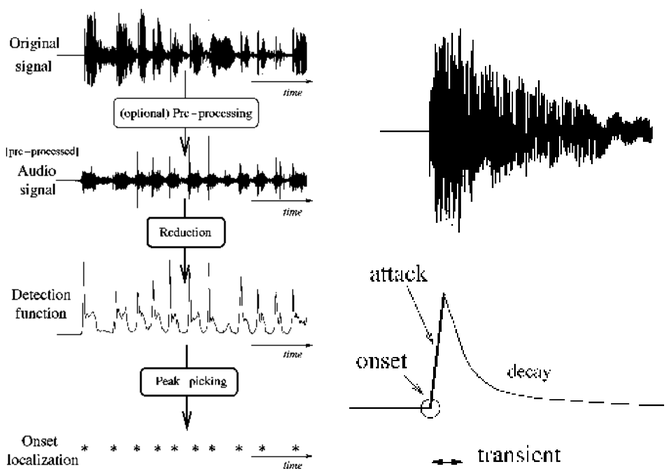
Peter Steiner peter.steiner@tu-dresden.de // Azarakhsh Jalalvand // Simon Stone // Peter Birkholz

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.



Utilized Dataset

102 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.

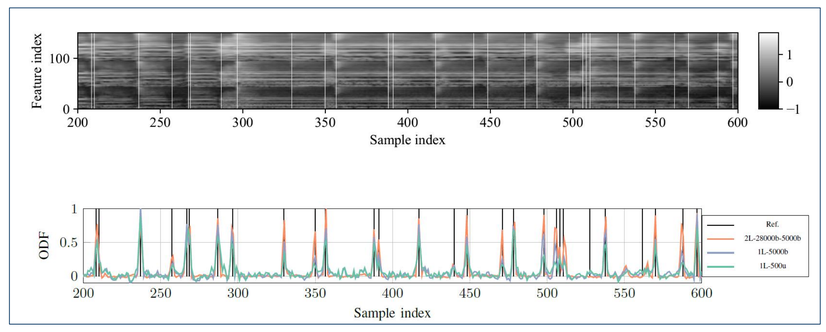
Reported scores are the mean values across eight repetitions.

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 parameter.

Architecture	Precision	Recall	F-Score
1L-24000b	0.881	0.804	0.840
2L-28000b-5000b	0.920	0.855	0.886
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



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.

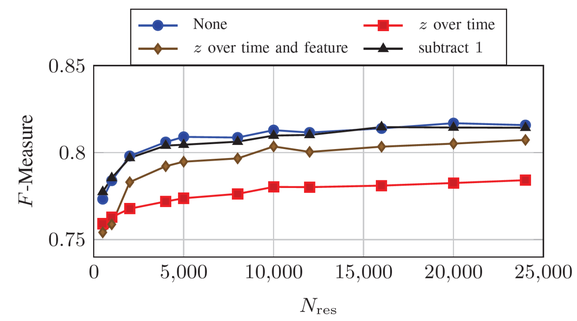
Here, we propose a novel way of stacking ESNs:

- 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.

Feature Engineering

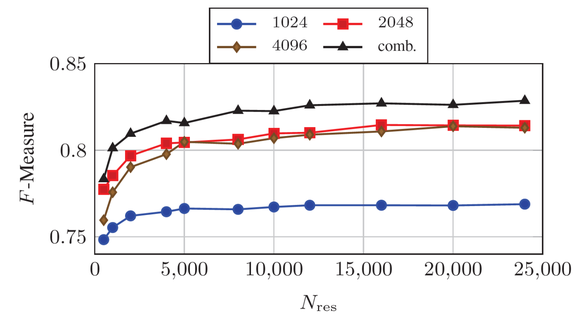
No standardization and transforming feature into bipolar features performed best.

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



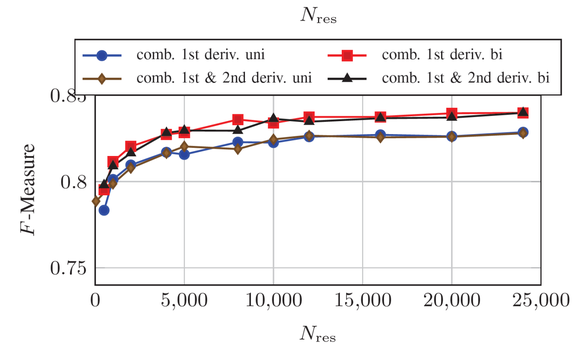
Different window sizes are able to deal with different kinds of onsets.

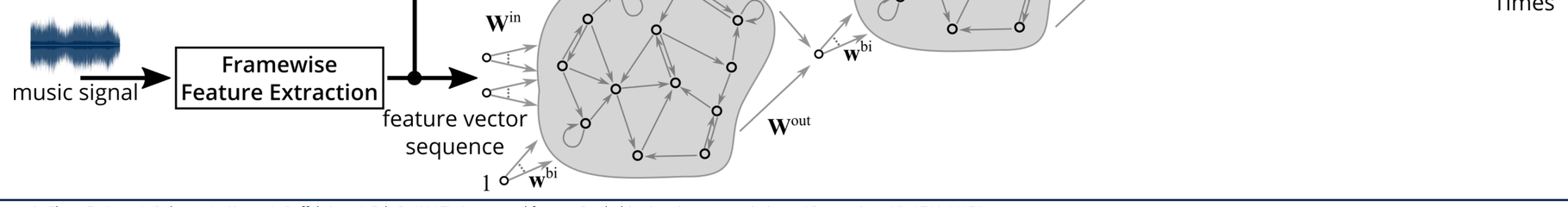
Short windows can capture high f_0 values and fast onsets, whereas long windows are more useful to detect very soft onsets.



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





[1] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., ... & Eck, D. (2017). Onsets and frames: Dual-objective piano transcription. arXiv preprint arXiv:1710.11153.
[2] Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148(34), 13.
[3] Jalalvand, A., Demuynck, K., De Neve, W., & Martens, J. P. (2018). On the application of reservoir computing networks for noisy image recognition. Neurocomputing, 277, 237-248.
[4] Böck, S., Krebs, F., & Schedl, M. (2012, October). Evaluating the Online Capabilities of Onset Detection Methods. In Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Porto, Portugal (pp. 49-54).
[5] Eyben, F., Böck, S., Schuller, B., & Graves, A. (2010). Universal onset detection with bidirectional long-short term memory neural networks. In . In Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands (pp. 89-594).
[6] Schlüter, J., & Böck, S. (2014, May). Improved musical onset detection with convolutional neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6979-6983). IEEE.
[7] Steiner, P., Stone, S., & Birkholz, P. (2020). Note onset detection using echo state networks. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020, 157-164.

Europäische Union
Europa fördert Sachsen.
ESF
Europäischer Sozialfonds
Freistaat Sachsen
Diese Maßnahme wird mitfinanziert durch Steuermittel auf Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.