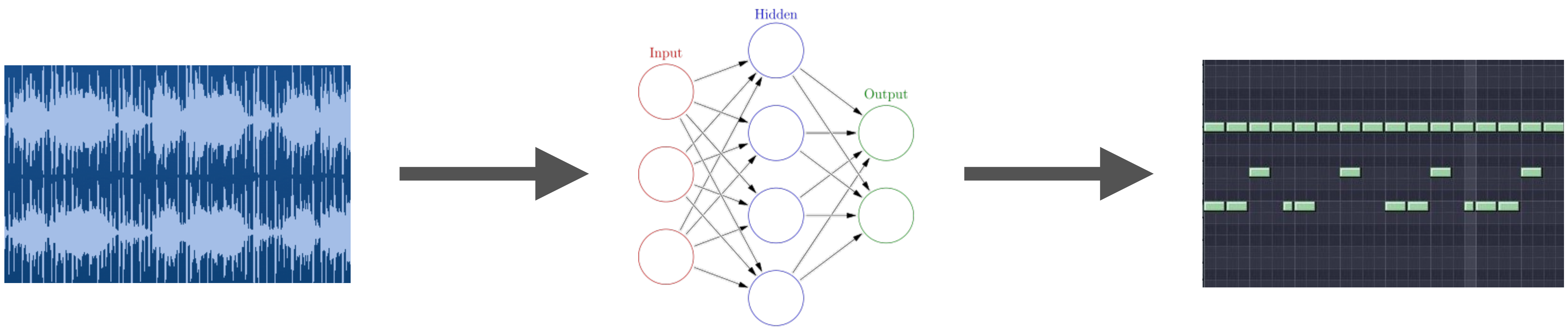
DRUM TRANSCRIPTION FROM POLYPHONIC MUSIC WITH RECURRENT NEURAL NETWORKS

Richard Vogl[1,2], Matthias Dorfer[1], Peter Knees[2]
richard.vogl@tuwien.ac.at, matthias.dorfer@jku.at, peter.kness@tuwien.ac.at

1 JOHANNES KEPLER UNIVERSITÄT LINZ
Department of Computational Perception

2 TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria
ifs
mir group

INTRODUCTION

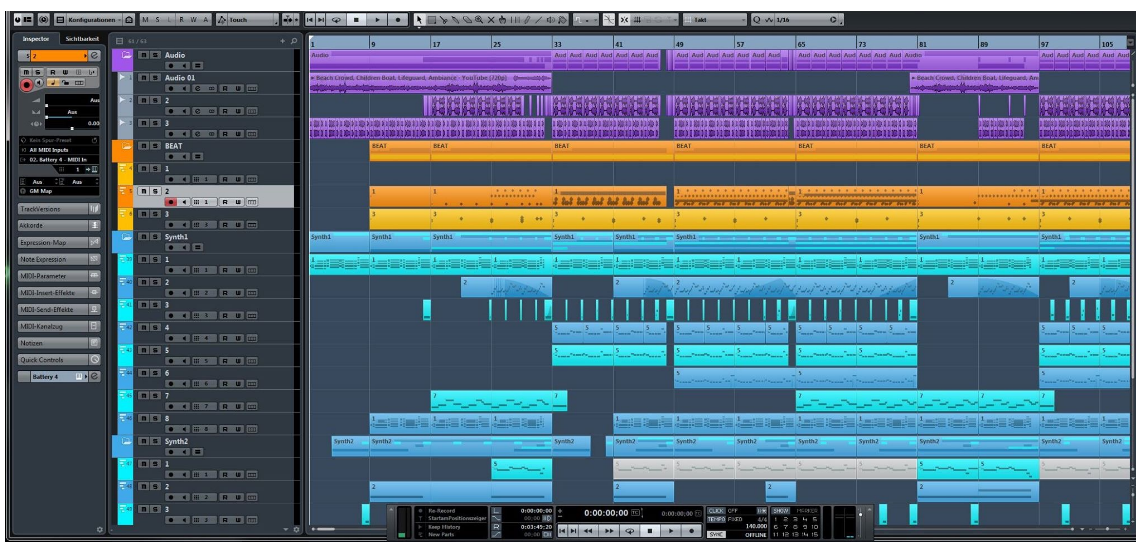


- Goal: **model** for drum note detection in polyphonic music
  - **In:** Western popular **music containing drums**
  - **Out:** **Symbolic representation** of notes played by drum instruments

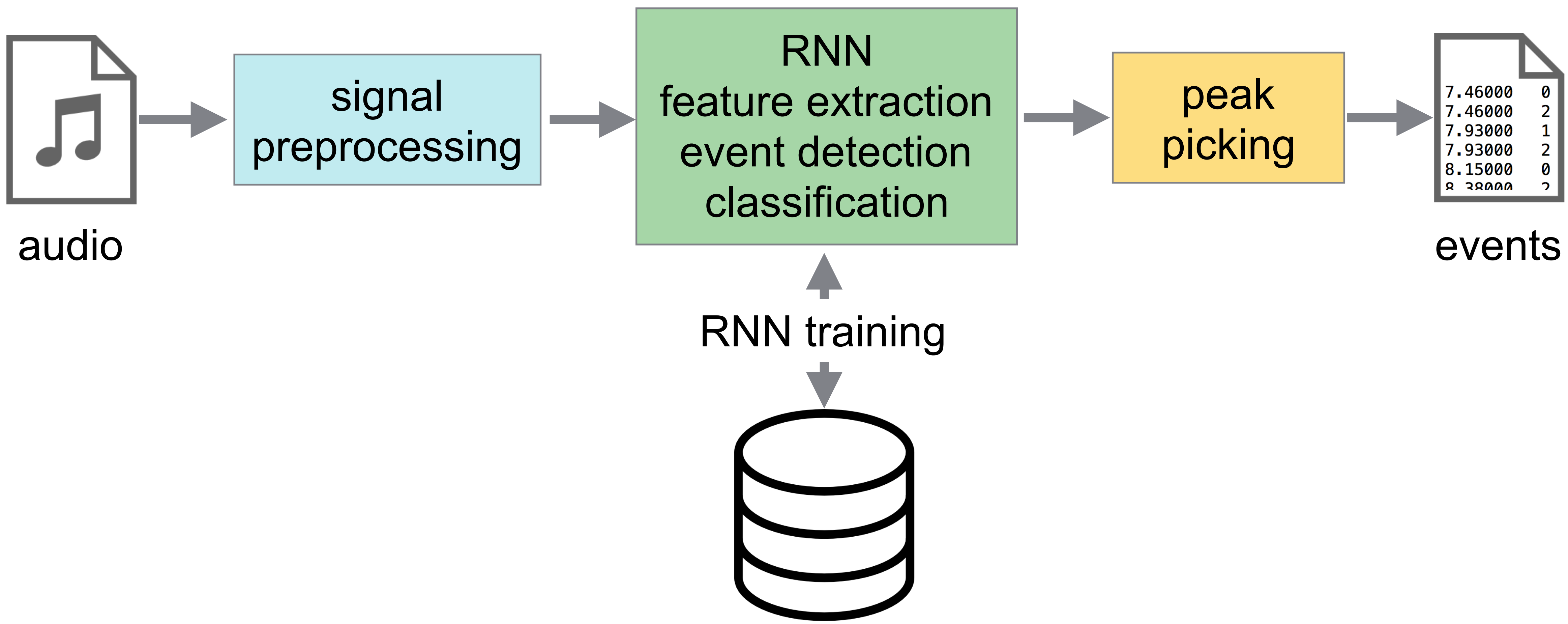
- Focus on **three major drum instruments**: snare, bass drum, hi-hat

INTRODUCTION

- Wide range of **applications**
  - **Sheet music** generation
  - **Re-synthesis** for music production
  - Higher level **MIR** tasks



SYSTEM ARCHITECTURE

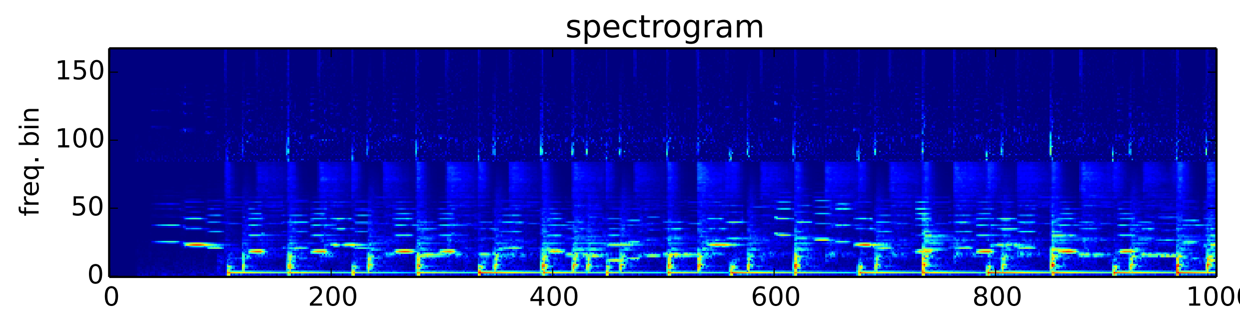


ADVANTAGES OF RNNS

- Relatively **easy to fit** large and diverse datasets
- Once trained, **computational complexity** of transcription relatively **low**
- **Online capable**
- **Generalize well**
- **Easy to adapt** to new data
- **End-to-end:** learn features, event detection, and classification at once
- **Scale better** with number of instruments (rank problem in NMF)
- Trending topic: lots of **theoretical work** to benefit from

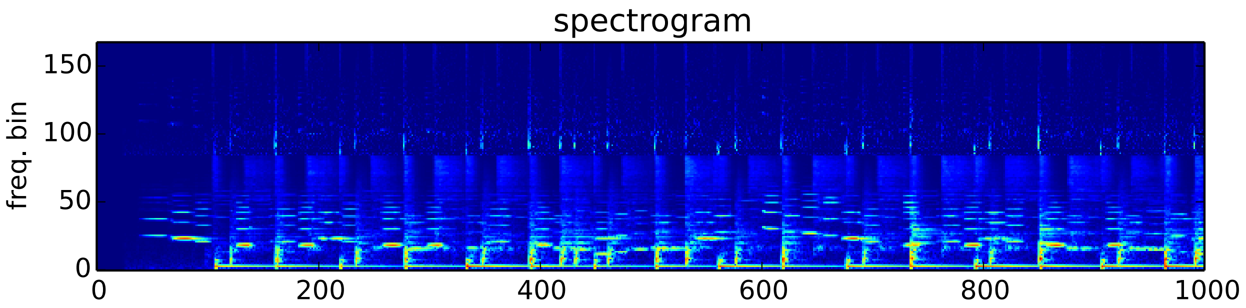
DATA PREPARATION

- Signal preprocessing
  - Log magnitude spectrogram @ 100Hz
  - Log frequency scale, 84 frequency bins
  - Additionally 1st order differential
  - 168 value input vector for RNN

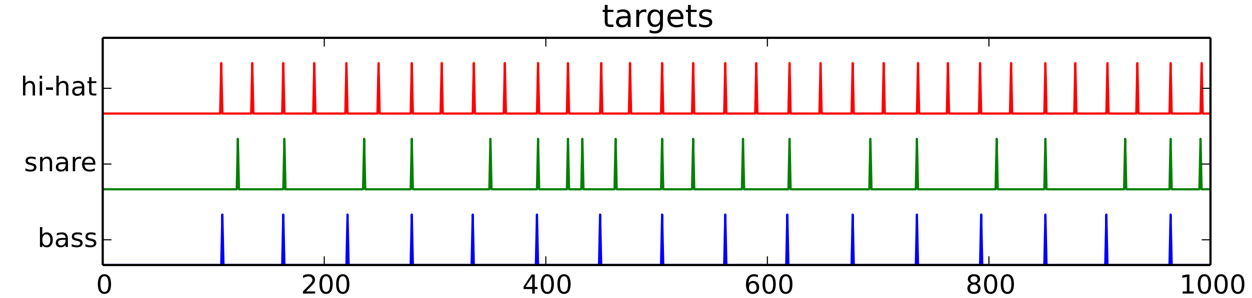


DATA PREPARATION

- Signal preprocessing
  - Log magnitude spectrogram @ 100Hz
  - Log frequency scale, 84 frequency bins
  - Additionally 1st order differential
  - 168 value input vector for RNN

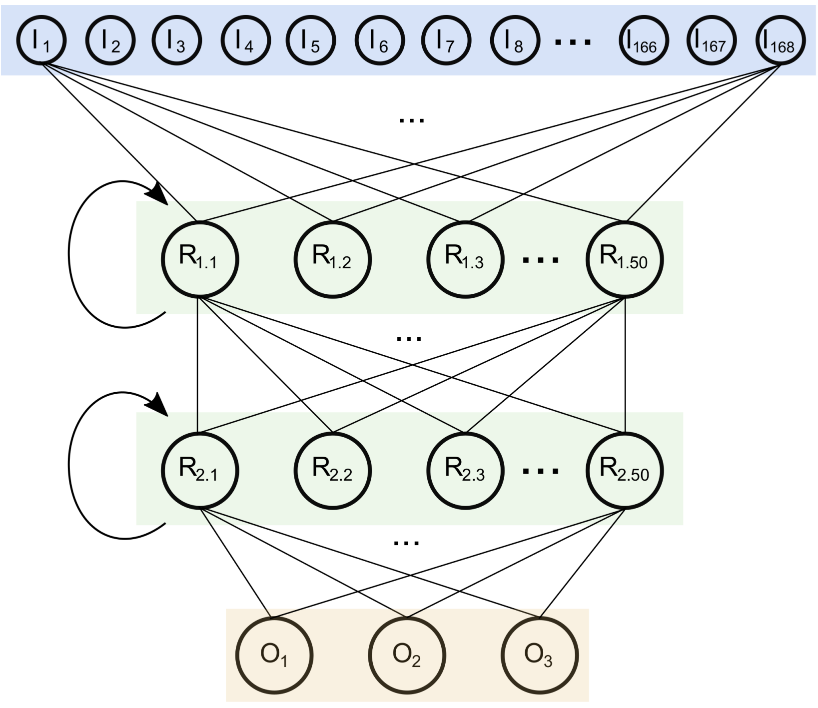


- RNN targets
  - Annotations from training examples
  - Target vectors @ 100Hz frame rate

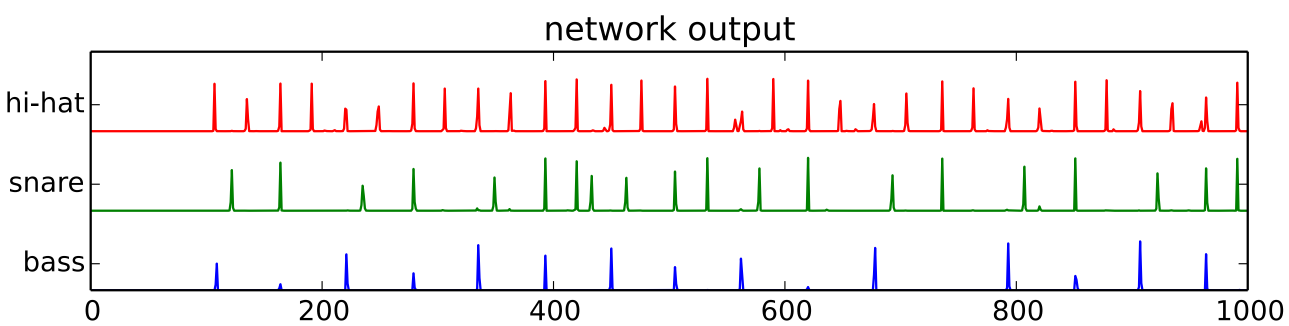


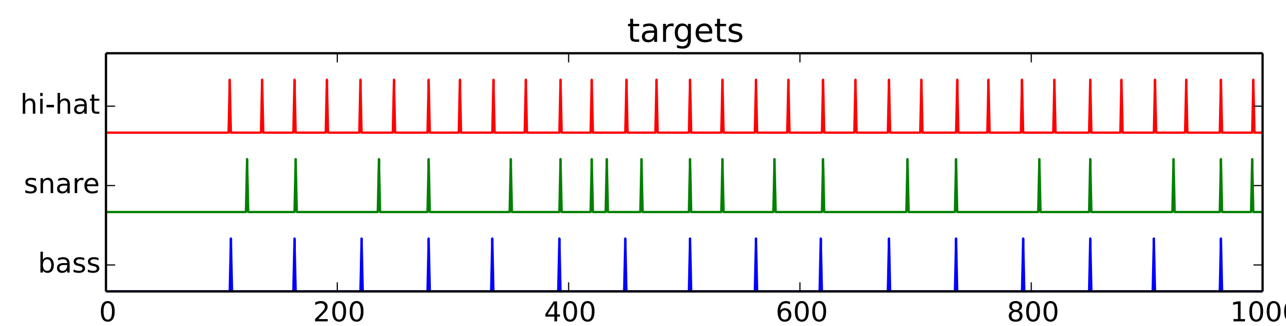
RNN ARCHITECTURE

- Two layers containing 50 **GRU**s each
  - Recurrent connections
- Output: dense layer with three **sigmoid units**
  - No softmax: events are independent
  - Value represent certainty/pseudo-probability of drum onset
  - Does not model intensity/velocity



PEAK PICKING





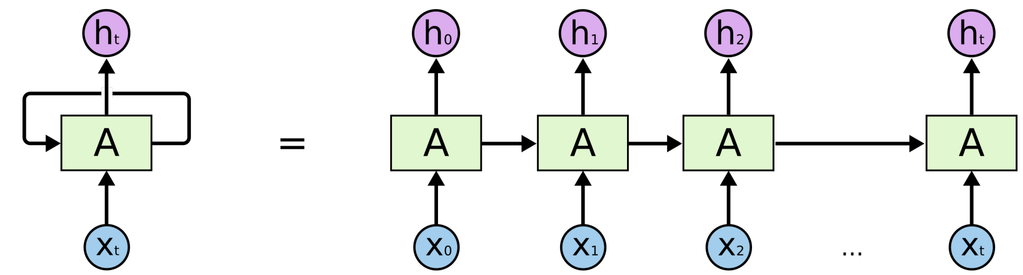
Select onsets at position n in activation function F(n) if:

$$F(n) = max(F(n-m), \cdots, F(n)),$$
$$F(n) \geq mean(F(n-a), \cdots, F(n)) + \delta,$$
$$n - n_{lp} > w,$$

[Böck et. al 2012]

RNN TRAINING

- Backpropagation through time (**BPTT**)
- **Unfold** RNN **in time** for training



[Olah 2015]

- Loss ($\mathscr{L}$): mean **cross-entropy** between output ($\hat{y}_n$) and targets (y_n) for each instrument

$$\mathcal{L}(\Theta) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n))]$$

- Mean over instruments with **different weighting** (w_i) per instrument (~+3% f-measure)

$$\mathcal{L}(\Theta) = \sum_{i=1}^{I} w_i \mathcal{L}_i(\Theta) \qquad \sum_{i=1}^{I} w_i = 1$$

- Update model parameters (θ) using gradient ($\mathscr{G}$) calculated on **mini-batch** and learn rate (η)

$$\mathcal{G}_t = \nabla_\Theta \mathcal{L}(\Theta_t)$$

$$\Theta_{t+1} = \Theta_t - \eta \mathcal{G}_t$$

RNN TRAINING (2)

- **RMSprop**
  - uses weight for learn rate based on moving mean squared gradient E[$\mathcal{G}^2$]

$$\cancel{\Theta_{t+1} = \Theta_t - \eta \mathcal{G}_t}$$

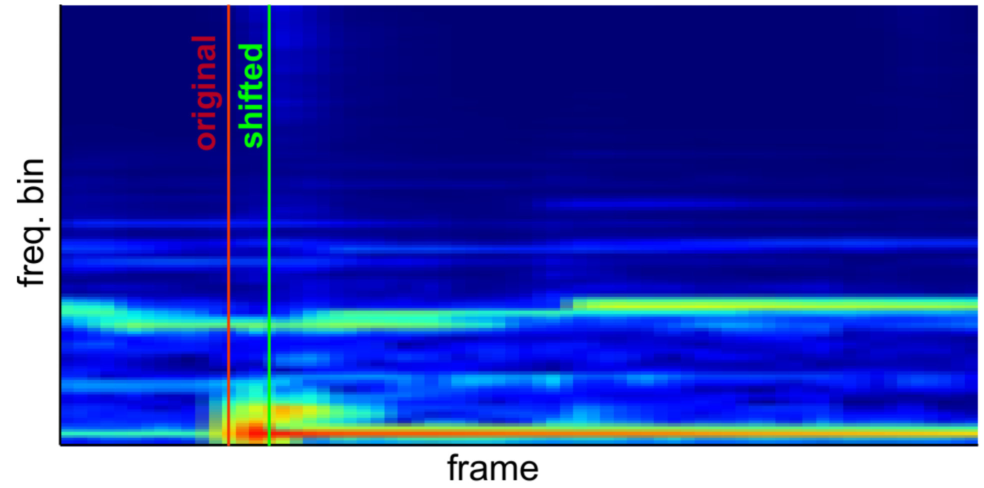
$$E[\mathcal{G}^2]_t = 0.9E[\mathcal{G}^2]_{t-1} + 0.1\mathcal{G}_t^2$$

$$\Theta_{t+1} = \Theta_t - \frac{\eta}{\sqrt{E[\mathcal{G}^2]_t + \epsilon}} \mathcal{G}_t$$

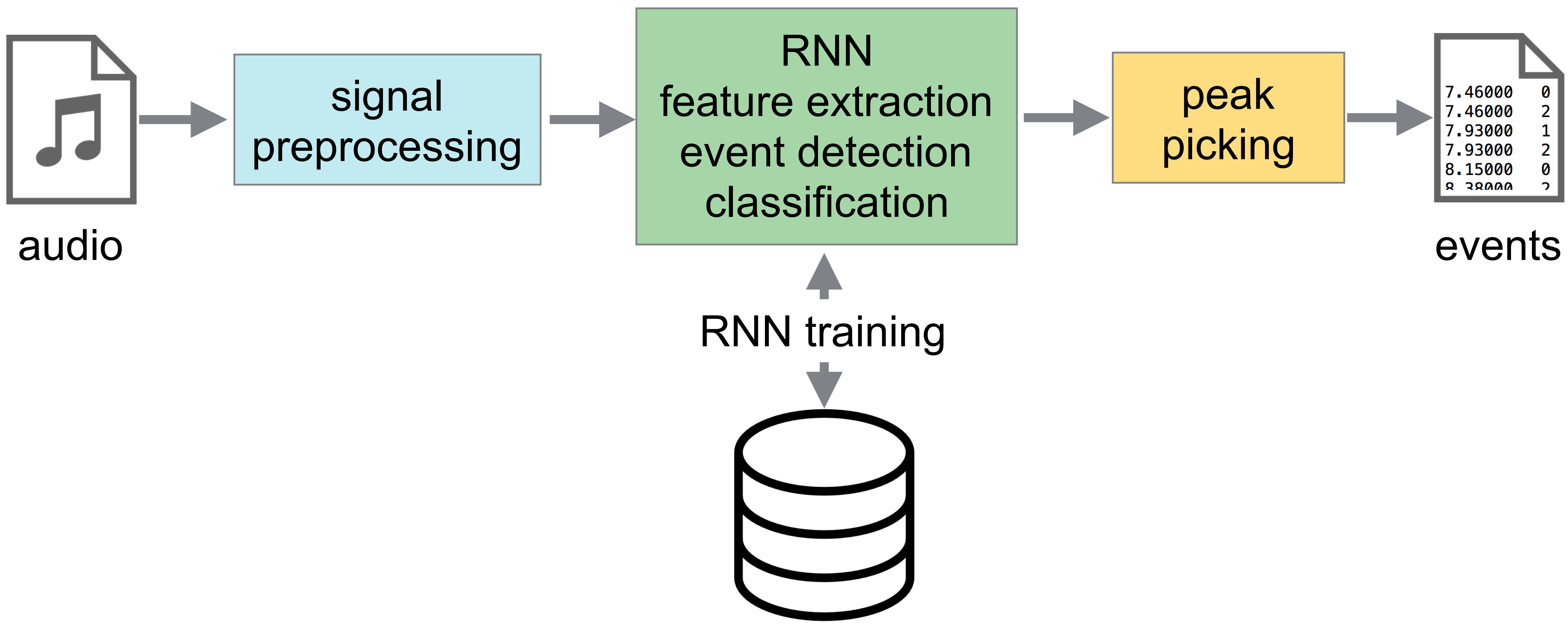
- **Data augmentation**
  - Random transformations of training samples (pitch shift, time stretch)

- **Drop-out**
  - Randomly disable connections between second GRU layer and dense layer

- **Label time shift** instead of BDRNN



SYSTEM ARCHITECTURE



DATA / EVALUATION

- IDMT-SMT-Drums [Dittmar and Gärtner 2014]
  - Three classes (Real, Techno, and Wave / recorded/synthesized/ sampled)
  - 95 simple solo drum tracks (30sec), plus training and single instrument tracks

- ENST-Drums [Gillet and Richard 2006]
  - Drum recordings, three drummers on three different drum kits
  - ~75 min per drummer, training, solo tracks plus accompaniment

- Precision, Recall, F-measure for drum note onsets
- Tolerance: 20ms

EXPERIMENTS

- **SMT optimized**
  - Six fold cross-validation on randomized split of solo drum tracks
- **SMT solo**
  - Three fold cross-validation on different types of solo drum tracks
- **ENST solo**
  - Three fold cross-validation on solo drum tracks of different drummers / drum kits
- **ENST accompanied**
  - Three fold cross-validation on tracks with accompaniment

RESULTS

Method	SMT opt.	SMT solo	ENST solo	ENST acc.
NMF-SAB **[Dittmar and Gärtner 2014]**	95.0	—	—	—
PFNMF **[Wu and Lerch 2015]**	—	81.6	77.9	72.2
HMM **[Paulus and Klapuri 2009]**	—	—	81.5	74.7
BDRNN **[Southall et al. 2016]**	96.1	83.3	73.2	66.9
tsRNN	**96.6**	**92.5**	**83.3**	**75.0**
	δ = 0.10	δ = 0.15	δ = 0.15	δ = 0.10

RESULTS

Method	SMT opt.	SMT solo	ENST solo	ENST acc.
NMF-SAB **[Dittmar and Gärtner 2014]**	95.0	—	—	—
PFNMF **[Wu and Lerch 2015]**	—	81.6	77.9	72.2
HMM **[Paulus and Klapuri 2009]**	—	—	81.5	74.7
BDRNN **[Southall et al. 2016]**	96.1	83.3	73.2	66.9
tsRNN	**96.6**	**92.5**	**83.3**	**75.0**
	$\delta = 0.10$	$\delta = 0.15$	$\delta = 0.15$	$\delta = 0.10$

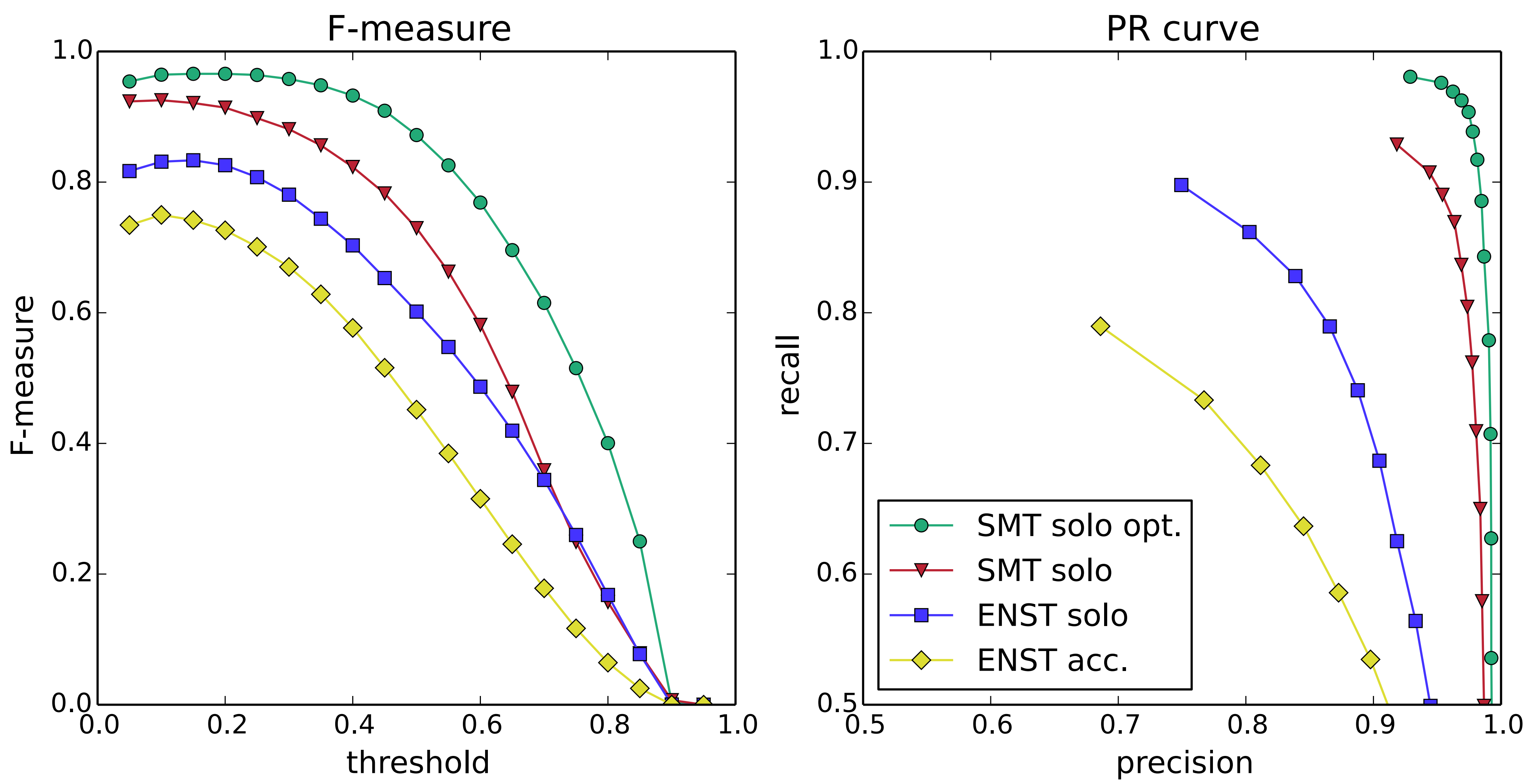
RESULTS

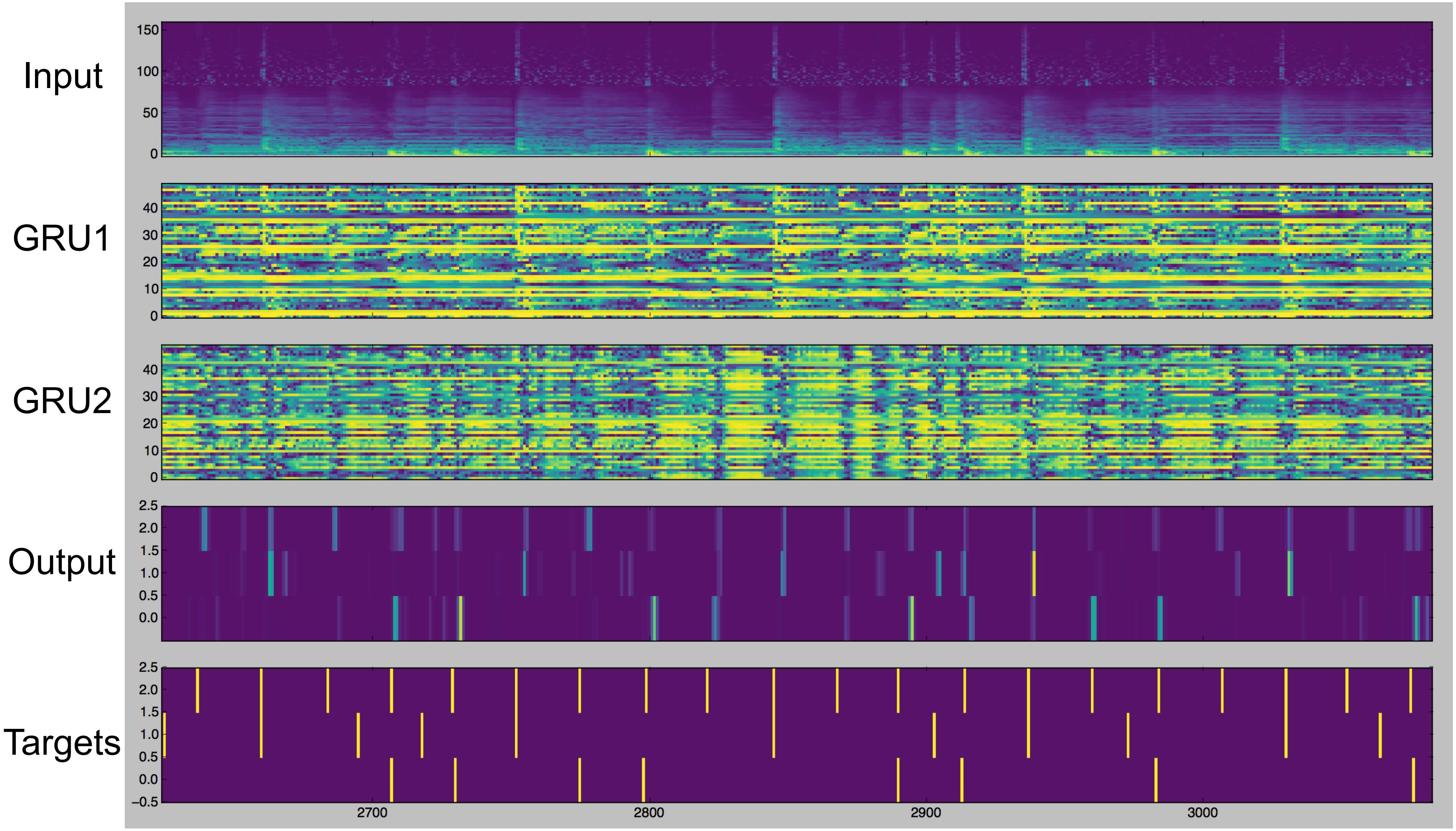
Method	SMT opt.	SMT solo	ENST solo	ENST acc.
NMF-SAB **[Dittmar and Gärtner 2014]**	95.0	—	—	—
PFNMF **[Wu and Lerch 2015]**	—	81.6	77.9	72.2
HMM **[Paulus and Klapuri 2009]**	—	—	81.5	74.7
BDRNN **[Southall et al. 2016]**	96.1	83.3	73.2	66.9
tsRNN	**96.6**	**92.5**	**83.3**	**75.0**
	$\delta = 0.10$	$\delta = 0.15$	$\delta = 0.15$	$\delta = 0.10$

RESULTS

Method	SMT opt.	SMT solo	ENST solo	ENST acc.
NMF-SAB **[Dittmar and Gärtner 2014]**	95.0	—	—	—
PFNMF **[Wu and Lerch 2015]**	—	81.6	77.9	72.2
HMM **[Paulus and Klapuri 2009]**	—	—	81.5	74.7
BDRNN **[Southall et al. 2016]**	96.1	83.3	73.2	66.9
tsRNN	**96.6**	**92.5**	**83.3**	**75.0**
	$\delta = 0.10$	$\delta = 0.15$	$\delta = 0.15$	$\delta = 0.10$

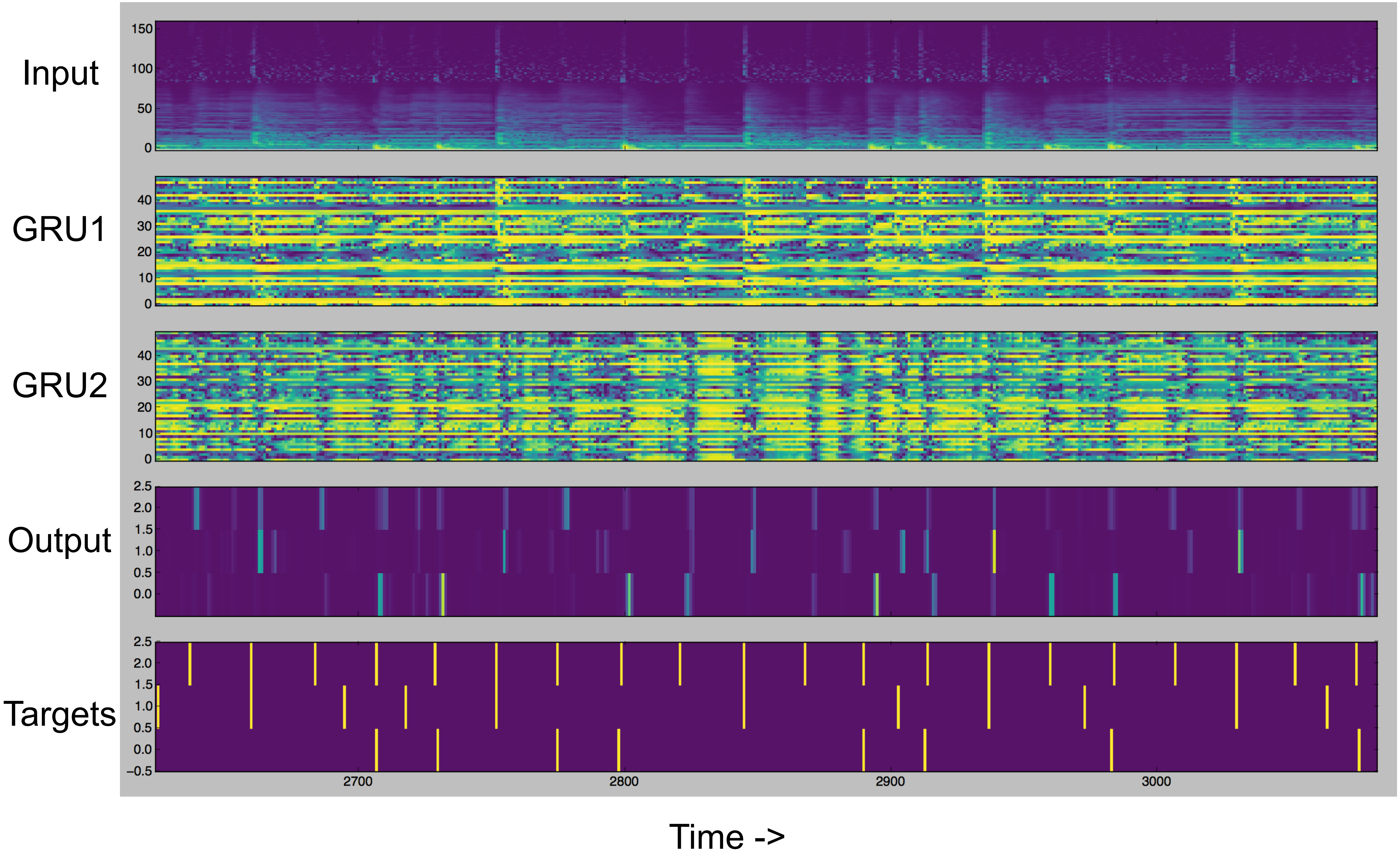
RESULTS

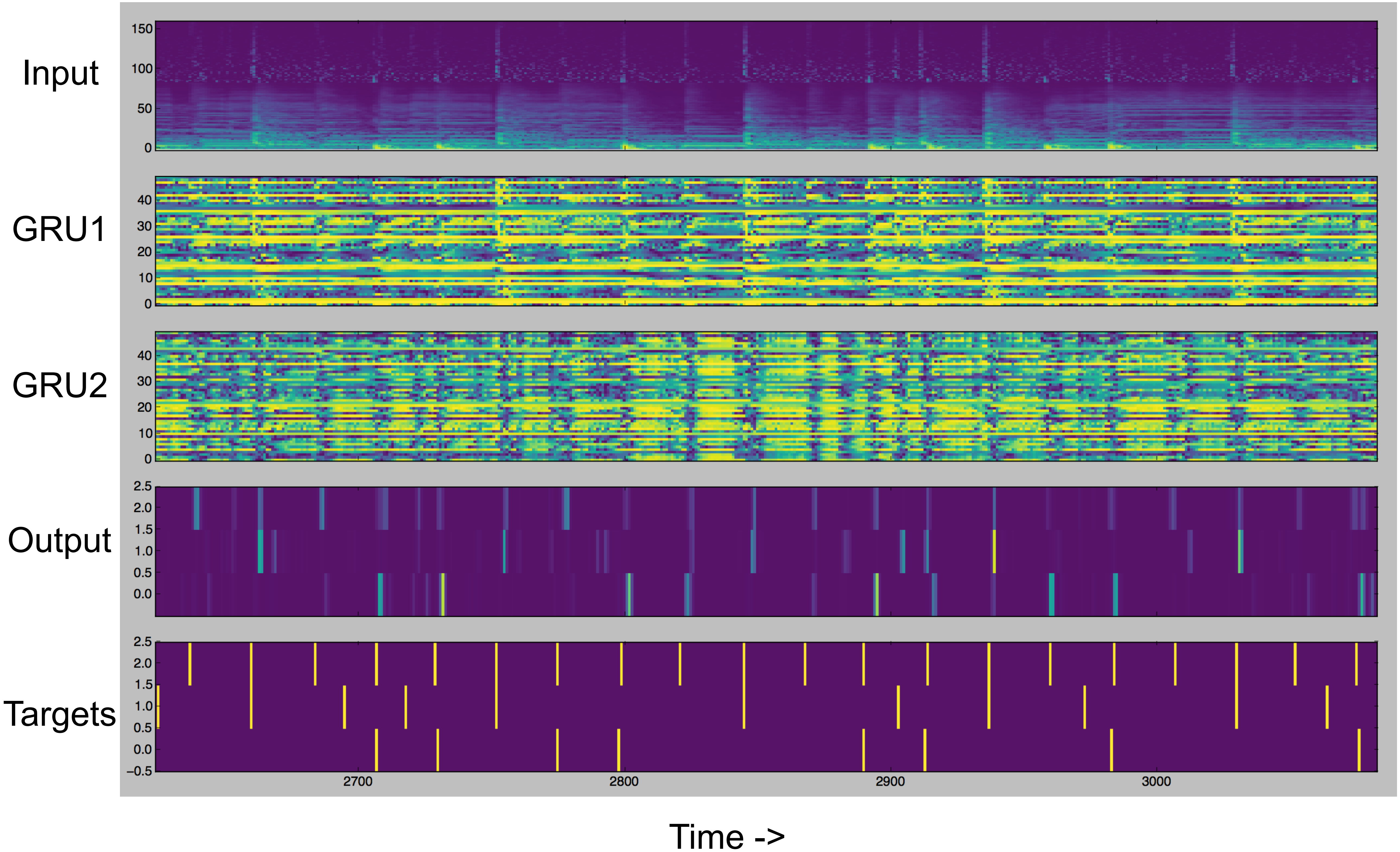

F-measure
PR curve
F-measure
threshold
recall
precision
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
SMT solo opt.
SMT solo
ENST solo
ENST acc.



Input
GRU1
GRU2
Output
Targets
150
100
50
0
40
30
20
10
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2700
2800
2900
3000

Time ->


Input
GRU1
GRU2
Output
Targets
150
100
50
0
40
30
20
10
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2700
2800
2900
3000
Time ->


Input
GRU1
GRU2
Output
Targets
150
100
50
0
40
30
20
10
0
40
30
20
10
0
2.5
2.0
1.5
1.0
0.5
0.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2700
2800
2900
3000
Time ->

CONCLUSIONS

- Towards a **generic** end-to-end **acoustic model for drum detection** using **RNN**s
- **Data augmentation** greatly improves generalization
- **Weighting loss functions** helps to improve detection of difficult instruments
- RNNs with label **time shift** perform equal to BDRNN
- Simple RNN architecture **performs better or similarly well** as handcrafted techniques

  while using a **smaller tolerance** window (20ms)