Language Modelling for Sound Event Detection with Teacher Forcing & Scheduled Sampling

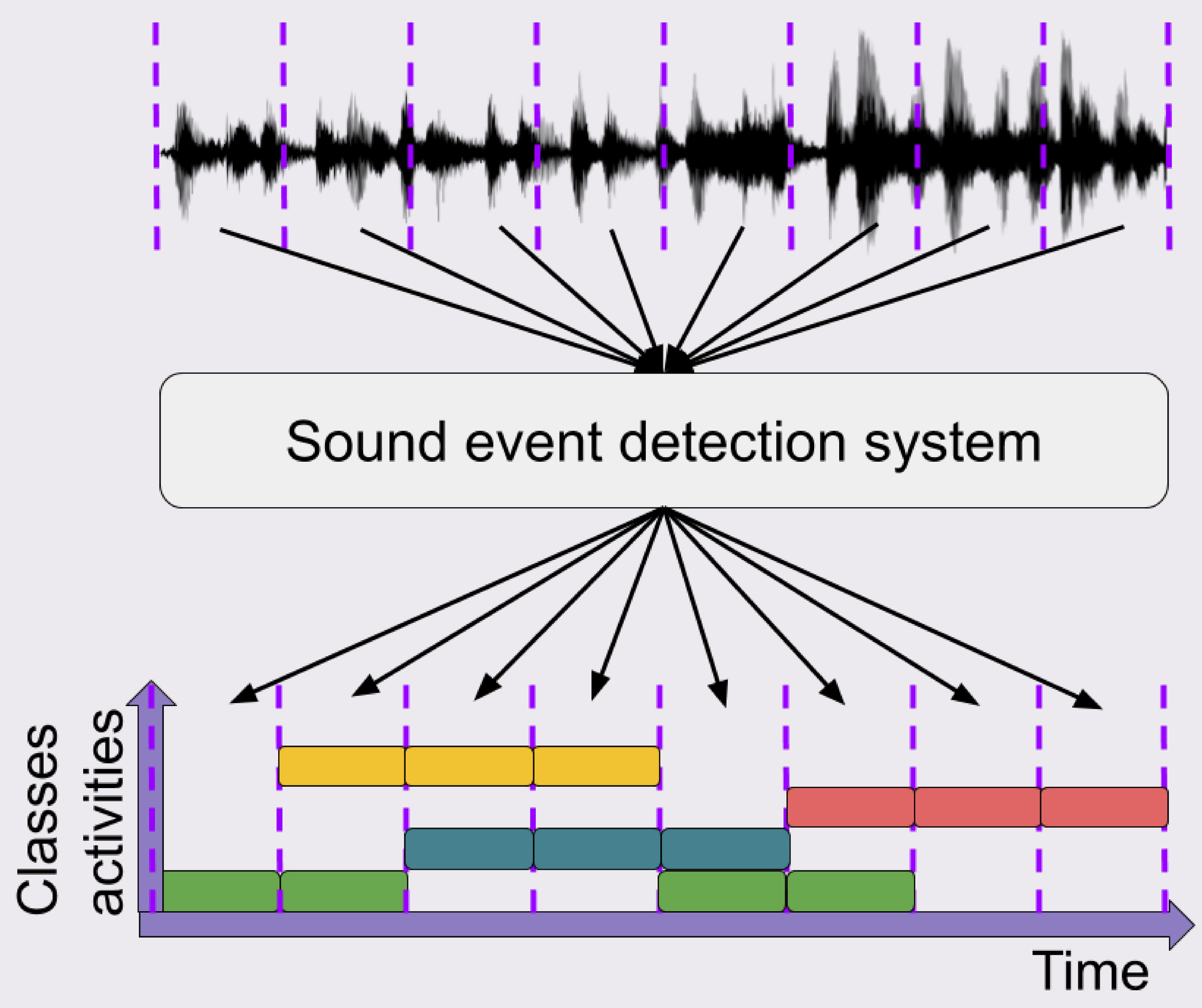
Konstantinos Drossos, Shayan Gharib, Paul Magron, and Tuomas Virtanen

Audio Research Group, Tampere University, Finland

Tampere University

Introduction

- **Sound event detection (SED) & SED methods**
  - Modelling of temporal audio patterns
    - Usually involving recurrent neural networks (RNNs)
    - *No temporal structure of class activities*



Problem

- **Temporal structure of sound events in real-life**
  - Intra-structure, e.g. “footsteps”
  - Inter-structure, e.g. “car horn” → “car passing by”
- **Exploitation of temporal structure of classes activities**
  - Jointly learnt language model in SED
  - Widely used in machine translation → language model

Previous approaches

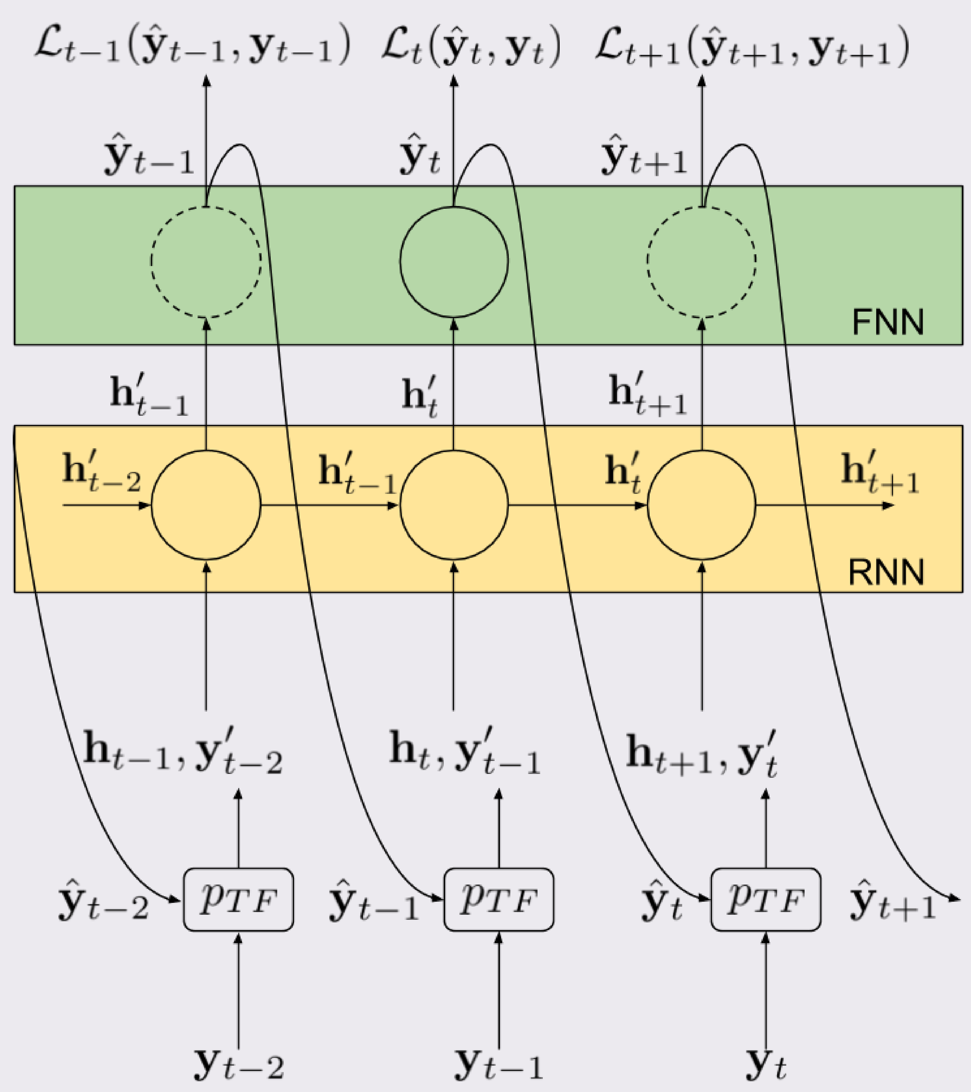
- HMM for sound event duration modelling [1]
- *n*-gram modelling [2]
- Connectionist temporal classification (CTC) loss [2]
- Time-shifted class activities as extra input to system [2]

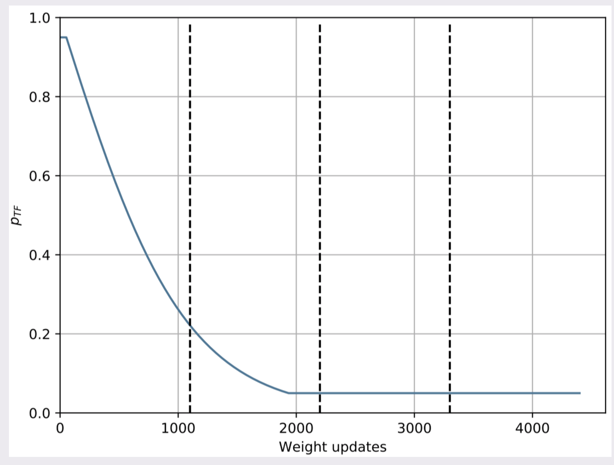
Teacher forcing

- **Teacher forcing**
  - Jointly learnable language model
  - Condition RNN input → next time-step class activities
  - Using **predicted** activities
    - ☺ Learn to recover from errors
    - ☹ Unstable training
    - ☹ Not correct conditioning on initial epochs
  - Using **ground truth** activities
    - ☺ Stable learning and correct conditioning on initial epochs
    - ☹ Prone to errors

Teacher forcing & scheduled sampling

- **Scheduled sampling** → best of both worlds
  - Gradually switch from ground truths to predictions





Evaluation & Results

- **Datasets:** Two real-life recordings, one synthetic
  - Real-life: DCASE 2016 & 2017
  - Synthetic: TUTSED-Synthetic 2017
- **Baseline** [3]: 3 CNN blocks, 1 GRU, 1 Classifier (CRNN)
- **Previous SOTA** [2]: *n*-grams for language model learning

Table 1: Mean/STD of F_1 score (higher is better) and error rate (*ER*) (lower is better). For [2] only the mean is available

	Baseline	**[2]**	**Proposed**
	TUT Sound Events 2016 dataset		
$\mathbf{F_1}$	0.28/0.01	0.29	0.37/0.02
ER	0.86/0.02	0.94	0.79/0.01
	TUT Sound Events 2017 dataset		
$\mathbf{F_1}$	0.48/0.01	–	0.50/0.02
ER	0.72/0.01	–	0.70/0.01
	TUT-SED Synthetic 2016 dataset		
$\mathbf{F_1}$	0.58/0.01	–	0.54/0.01
ER	0.54/0.01	–	0.61/0.02

Conclusions & future work

- Clear benefit on real-life recordings
- Performance drop on synthetic dataset
- Future work: explore learned language model

References

[1] T. Hayashi et al., “Duration-controlled LSTM for polyphonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2059–2070, November 2017.

[2] G. Huang, T. Heittola, and T. Virtanen, “Using sequential information in polyphonic sound event detection,” in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Sep. 2018, pp. 291–295.

[3] E. Çakir et. al, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291–1303, June 2017.

https://github.com/dr-costas/SEDLM

konstantinos.drossos@tuni.fi