Closing the Gap Between Probabilities and Decisions in Temporal Detection and Classification

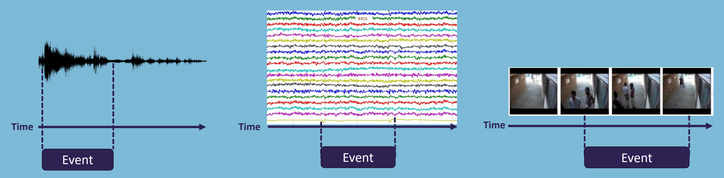
By **Dr Çağdaş Bilen**, Head of Research, representing Audio Analytic's research team

ID:15

1 The basic premise

Temporal detection and classification is relevant to a wide range of applications:

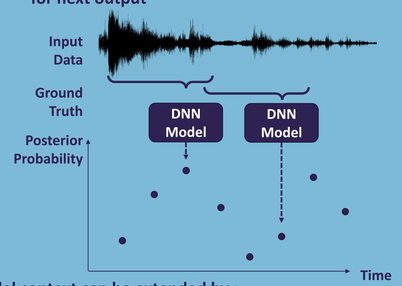
- (Audio) Sound event detection
- (Audio) Keyword spotting
- (Time-series) ECG/EEG
- (Video) Anomaly detection in surveillance
- etc.



Given time domain data, detect and classify events in time

2 Typical models

- The model observes a limited temporal context
- It predicts posterior probability for the time instant
- It then observes next overlapping temporal context for next output

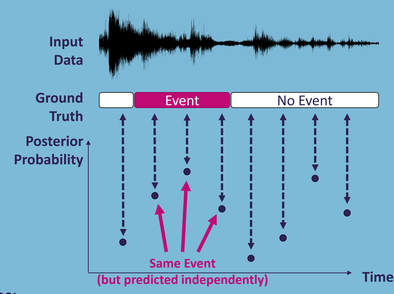


Model context can be extended by :

- Bigger models
- Multiple Instance Learning
- RNNs / CRNNs
- Attention, etc.

Typical training

- Supervised training with strong and/or weak labels
- Trained with log-loss (or variations)
  - Binary XE, Categorical XE, focal loss, etc.

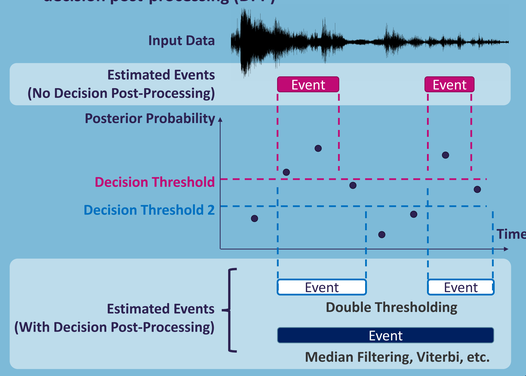


Observations:

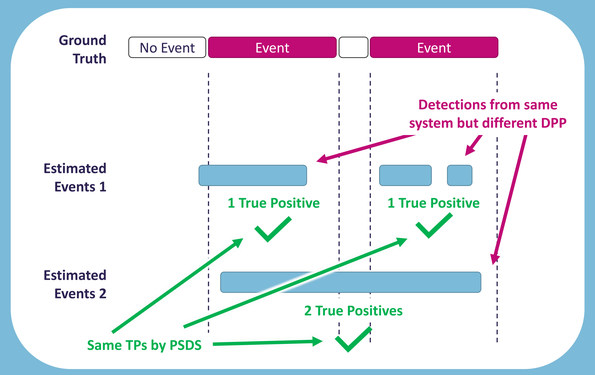
- Model predictions are instantaneous: no awareness of the concept of events
- Predictions could be temporally inconsistent / noisy.

Typical inference

- Thresholding can be used to estimate events
- Inconsistent/noisy model outputs could be improved through decision post-processing (DPP)

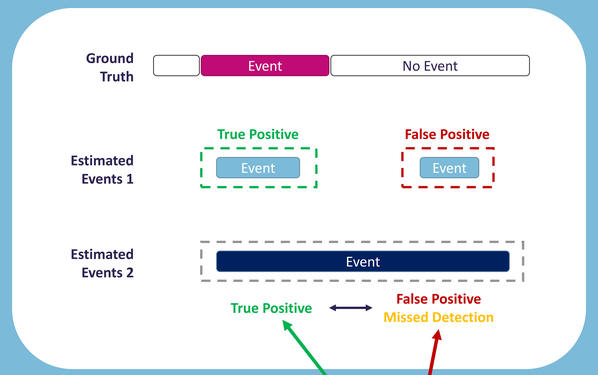


3 Difficulty of comparing ML models



- The impact of DPP may complicate comparisons between models
- We have developed PSDS metric [3] to enable comparing systems with different temporal behaviour
- FPs would still be quite different.

Evaluating performance

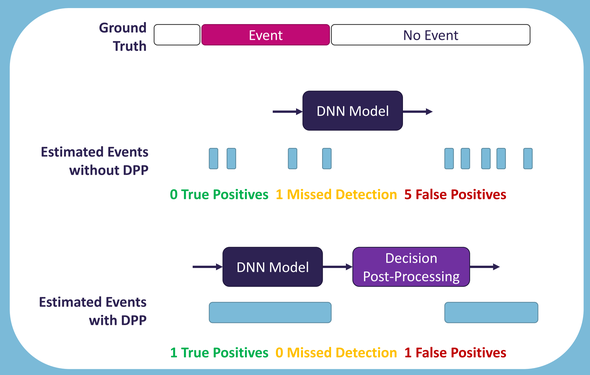


- Performance measured in event-based TPs and FPs
- The decision post-processing can significantly alter the final performance and success in the market.

Same event may be TP or FP based on application needs

- **Detection task** – relaxed timing constraints
- **Segmentation task** – strict timing constraints

Decision post-processing and final performance



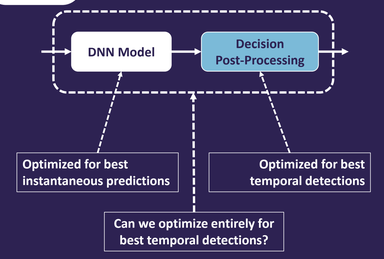
- In industry, a good DPP can reduce the FPs by orders of magnitude
- However, optimizing the temporal decisions is less explored in public research despite the performance.

4 Real-world impact: An example in sound event detection

- YAMNet [1] is a pretrained deep net that predicts audio event classes based on the AudioSet [2] dataset.
- We evaluated YAMNet performance
  - On a challenging test set for one class
  - With high TP target performance
- YAMNet output compared with two DPP options:
  - Simple DPP (temporal smoothing of five outputs)
  - Sophisticated DPP.

	TPs	FPs	Improvement
YAMNet → Direct output of YAMNet	99.7%	1500 per hr	1x
YAMNet → Simple DPP → Simple DPP applied to YAMNet	99.6%	500 per hr	3x
YAMNet → Sophisticated DPP → Sophisticated DPP applied to YAMNet	99.5%	10 per hr	150x

5 In conclusion: The gap between probabilities and decisions



- The performance improves significantly with DPP but the result can still be sub-optimal
- It is possible to achieve big improvement through DPP because models are not optimized for the final decisions, but for instantaneous probabilities.

Take-home questions:

- We can do event-based evaluation, can we do event-based training?
- Can we design "event-aware" models?
- Can we enable true end-to-end event detection and classification?

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet
[2] Gemmeke, Jort F., et al. "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events", ICASSP 2017
[3] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection", ICASSP 2020

Audio Analytic – 2 Quayside, Cambridge, CB5 8AB, UK
audioanalytic.com cagdas.bilen@audioanalytic.com