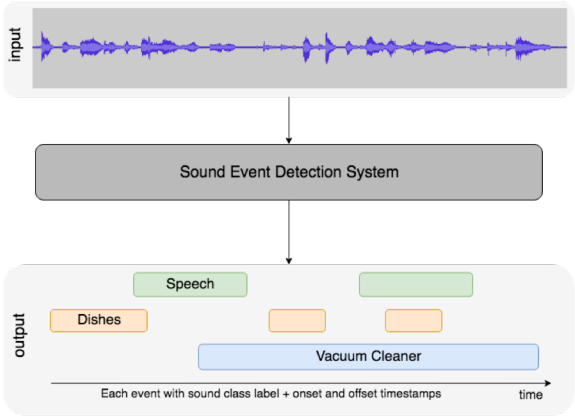
Description and analysis of novelties introduced in DCASE Task 4 2022 on the baseline system

Francesca Ronchini[1], Samuele Cornell[2], Romain Serizel[1], Nicolas Turpault[1], Eduardo Fonseca[3], Daniel P.W. Ellis[3]

[1]Université de Lorraine, CNRS, Inria, Loria [2]Università Politecnica delle Marche
[3]Google, Inc.

Why ?

- Necessity of real-world strongly annotated data ?
- Impact of external data and pre-trained models ?
- Environmental impact of our training models ?



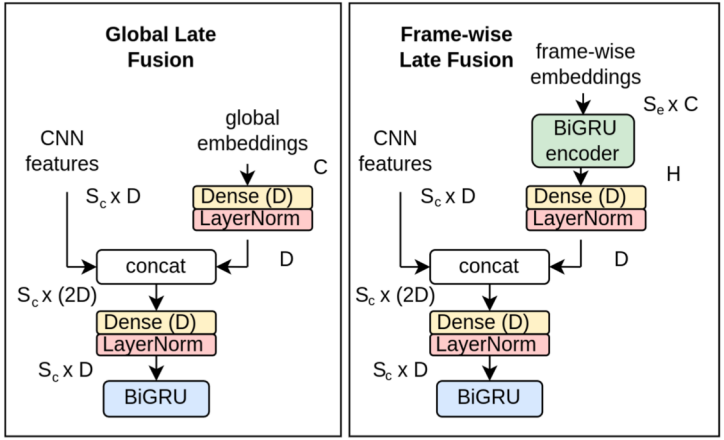
Audioset strong !

- 3470 clips matching our labels

- These annotations are really expensive to get:
  - > Are they really helping ?
  - > Can’t we get the same performance with synthetic data ?

Embeddings from pre-trained models

- Using PANNs and AST:
  - > Global or frame embeddings ?
- Integrating embeddings is not trivial:
  - > Global fusion
  - > Frame-wise fusion
  - > Same problem with sound separation



Energy-based metric ?

$$\mathrm{EW-PSDS} = \mathrm{PSDS} * \frac{\mathrm{kWh_{baseline}}}{\mathrm{kWh_{submission}}}$$

PSDS: polyphonic sound detection scores
$\mathrm{kWh_{baseline}}$: baseline energy consumption
$\mathrm{kWh_{submission}}$: system energy consumption

- How does this impact our models ?
- Are the systems using pre-trained models more efficient ?

This is a simple suggestion, it should be improved → come & discuss !

Thank you

nicolas.turpault@sonaide.fr