DCASE2019 CHALLENGE

Summary & Results, Task 5

T5 Urban Sound Tagging

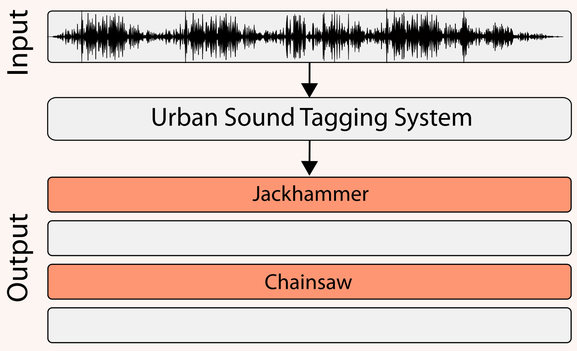
Coordinators

Mark Cartwright, Ana Elisa Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, Juan P. Bello

Results

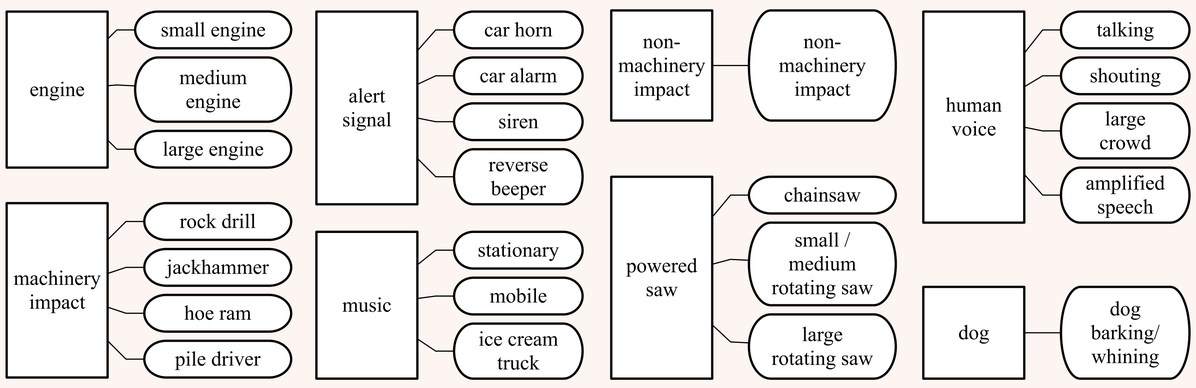
tinyurl.com/dcase2019-t5

Task description



- **Multilabel sound-event tagging of 10-second urban recordings**
- *Motivation*: Urban noise pollution monitoring
- *Examples*: Alert city agencies of noise code violations

Dataset



- **Recordings**: 10s recordings from 44 Sounds of New York City (SONYC) urban acoustic sensors
- **Tags**: 23 fine-level and 8 coarse-level tags developed in consultation with the New York Department of Environmental Protection. If an annotator has uncertainty at the fine-level, they may provide just a coarse-level tag.
- **Additional metadata**: Sensor ID, Annotator ID, Proximity (*near*, *far*, *not sure*)
- **Training set**: 2351 recordings annotated by 3 Zooniverse volunteers
- **Validation set**: 443 recordings annotated by SONYC research team
- **Test set**: 274 recordings annotated by SONYC research team

Evaluation Metrics

Because SONYC-UST has incomplete ground truth at the fine taxonomical level, we evaluate the prediction at the fine level when possible, but fall back to the coarse level if necessary.

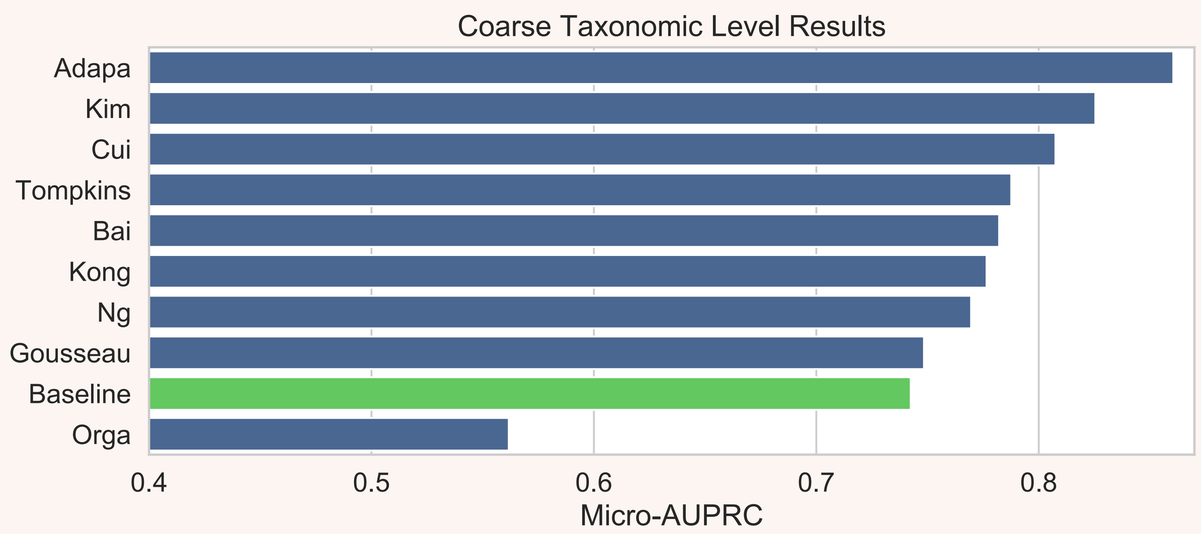
- **Primary metric**: Micro-AUPRC
- **Secondary metrics**: Micro-F1@0.5, Macro-AUPRC

Baseline System

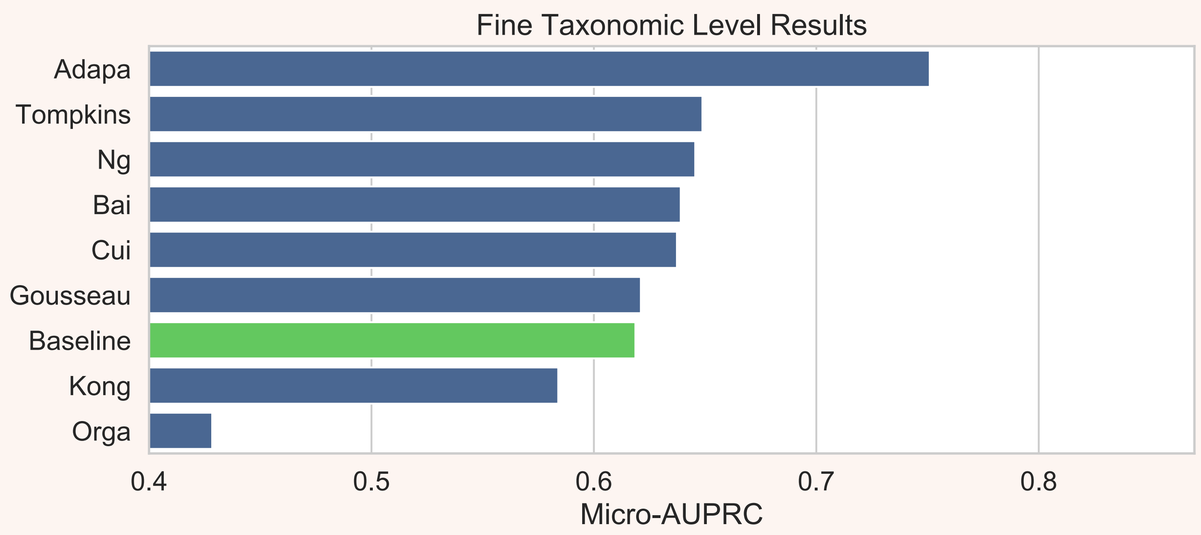
- **Model**: Multilabel logistic regression
- **Input**: 10-frames of 128-d VGGish features
- **Target**: Annotations aggregated with minority vote
- **Temporal aggregation**: Trained at the frame level and averaged output tag probabilities as clip-level tag probabilities

Results

- 24 Systems (10 Teams)

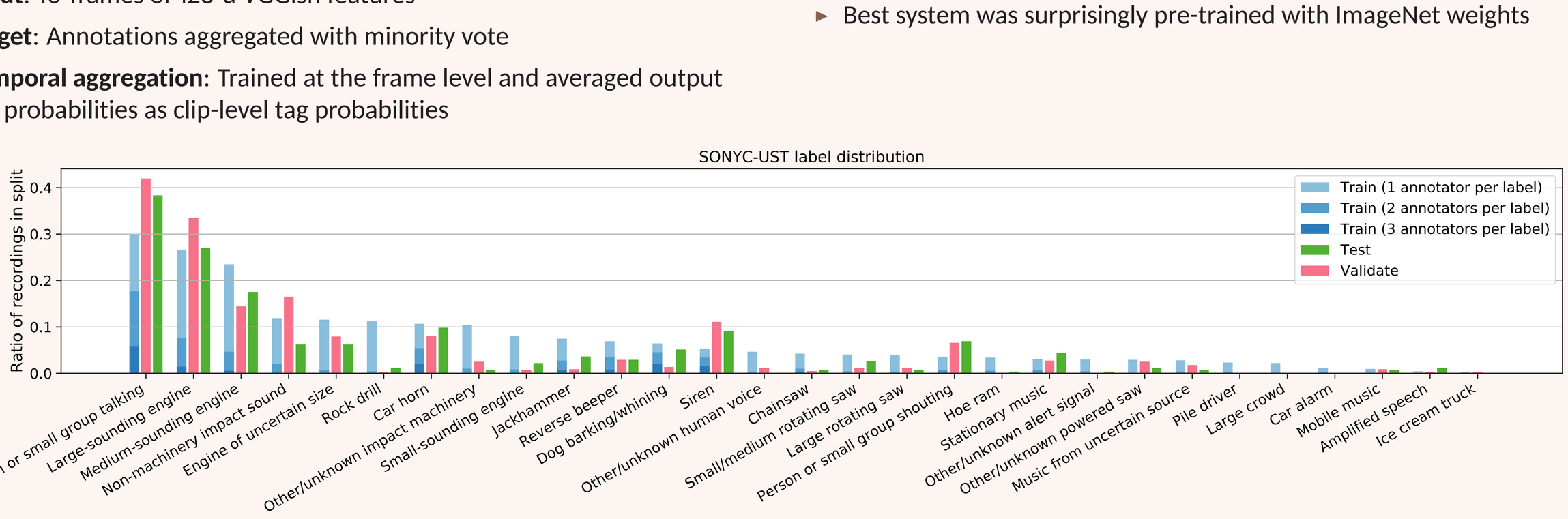


System	Feats.	Aug.	Ext. Data	Class.	Macro-AUPRC	Micro-F1	**Micro-AURPC**
Adapa	MelSpec	mixup, random erase, scaling, shifting	pre-trained model	CNN	0.72	0.63	**0.86**
Kim	MelSpec		pre-trained model	CNN	0.70	0.73	**0.83**
Cui	MelSpec			CNN	0.67	0.52	**0.81**
Tompkins	MelSpec	scaling, shifting, noise	pre-trained model	CNN	0.67	0.55	**0.79**
Bai	MFCC, MelSpec, STFT, HPSS			CNN	0.65	0.71	**0.78**



Discussion

- Systems under-utilized additional metadata
- Best system was surprisingly pre-trained with ImageNet weights



DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS

NYU

Cornell University Founded A.D. 1865

Adobe