Formula-Supervised Sound Event Detection: Pre-Training Without Real Data

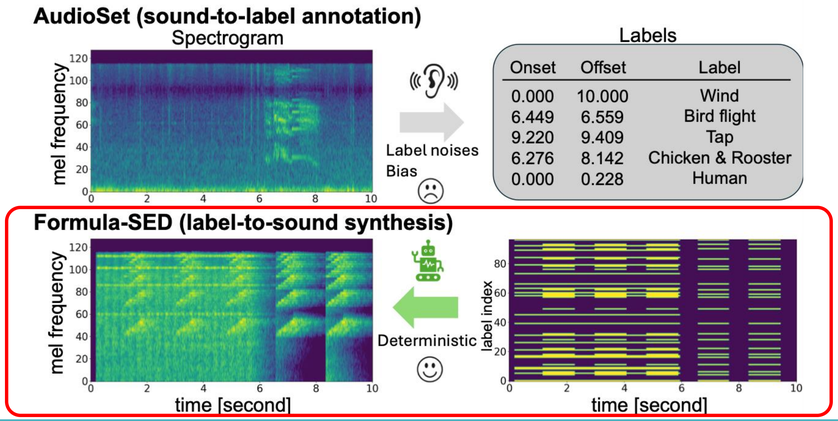
Yuto Shibata*†, Keitaro Tanaka†‡, Yoshiaki Bando†, Keisuke Imoto†§, Hirokatsu Kataoka†¶, and Yoshimitsu Aoki

*Keio University, Japan †National Institute of Advanced Industrial Science and Technology (AIST), Japan ‡Waseda University, Japan §Doshisha University, Japan ¶University of Oxford, United Kingdom

Project page

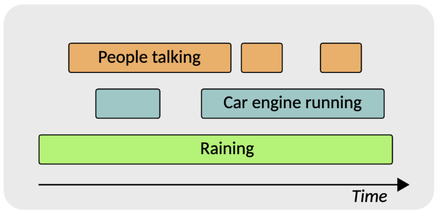
1. Summary

- We propose an effective pre-training method based on ***Formula-SED***, audio data synthesized **solely from mathematical formulas**.
- Supervised pre-training significantly improved downstream accuracy.



2. Background

Sound Event Detection (SED):
Predict the time intervals in which specific acoustic events occur.



The overview of SED [5]

- Clip level weak label
- Timestamped strong label
  → ☹ High cost
  ☹ Annotator's subjectivity
- Real data → Copyright / Privacy

Large-scale pre-training with strong labels have not achieved.

3. Formula-Driven Acoustic Supervised Learning

3-A. Parametric Sound Event Synthesis

- Synthesize source signals by summing harmonic and inharmonic components and finally convolving reverberation (Spectral Modeling Synthesis).

$$x(n) = A(n) \sum_{k=1}^{K} c_k(n) \sin(\phi_k(n)) + F_t(n) * v(n)$$

$A(n)$: Amplitude; $c_k(n)$: Normalized amplitude of k-th harmonic; $F_t(n)$: Linear Time-Invariant Finite Impulse Response filter; $v(n)$: Noise (uniform)

$$\phi_k(n) = 2\pi \sum_{s=0}^{n} k f_0(s) + \phi_{0,k}$$

$\phi_{0,k}$: Initial phase (random); $f_0(s)$: Fundamental frequency

3-B. Parameter Generation with Gaussian Processes

- Gaussian processes are used to stochastically represent smoothness.
- Correlation between harmonic and inharmonic component are expressed by ICM (Intrinsic Coregionalization Model) [1]

$$\begin{pmatrix} v_{\mathrm{har}}(n) \\ v_{\mathrm{noise}}(n) \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} \bar{v}_{\mathrm{har}} \\ \bar{v}_{\mathrm{noise}} \end{pmatrix}, \mathbf{B} \otimes \mathbf{K}(n, n') \right)$$

3-C. Ground-Truth Label Generation

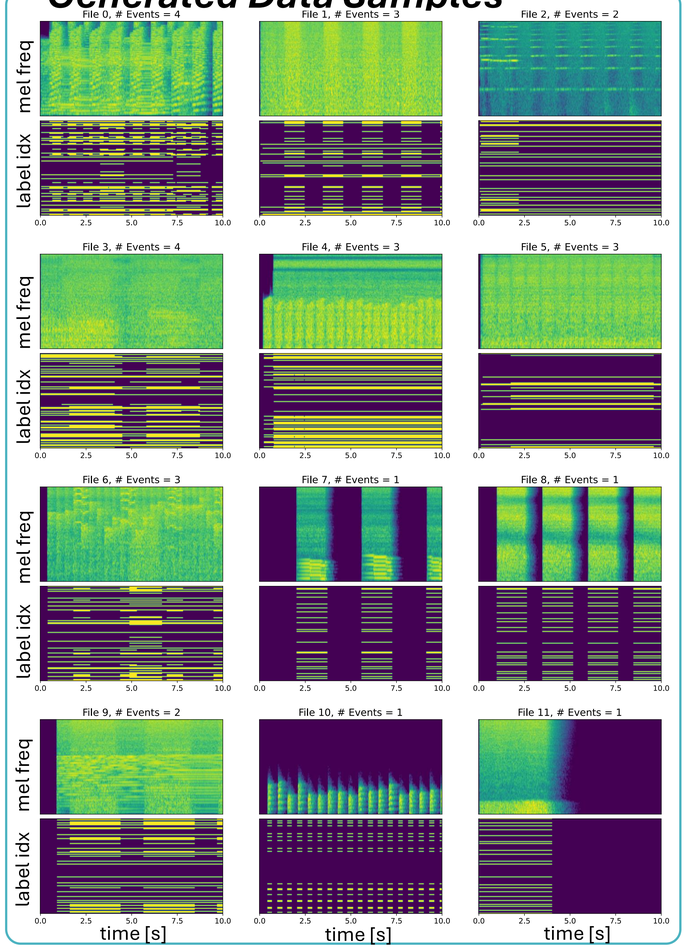
- 21 synthesis parameters serve as ground-truth labels for pre-training. These capture short- and long-term sound changes crucial for SED.
- Continuous parameters are discretized based on predetermined thresholds.
- At each time step, these labels are represented as a multi-hot vector.

Automatic generation of **Accurate** (time & event label) and **Unbiased** labels ☺

List of synthesis parameters (label)

Scale	Name and number of classes
Global	• Voiced segment duration (3) • Harmonic volume (3) and inharmonic volume (3) • F0 variance (4) and bias (4) • Number of harmonics (3), envelope variance (4), length scale (4), sharpness (4), and kernel (8) • Inharmonic distribution sharpness (4), mode (10), and kernel (8) • Discrete or continuous pitch (2) • Reverb strength (-)
Local	• Harmonic-Inharmonic volume correlation (2), volume variance (4) and kernel (8) • F0 variance (4), length scale (4), and kernel (8)

Generated Data Samples



4. Experimental Evaluation

Experimental setting

Pre-training Dataset:
- Audioset data w/ Strong labels (80k) [2]
- **Formula-SED (ours) {50k, 100k, 1M}**

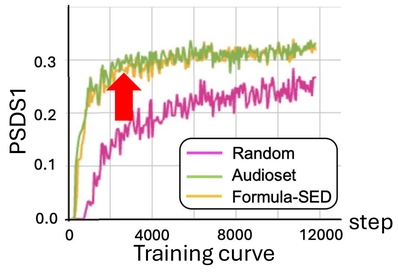
Downstream Task: DCASE 2023, task 4 (DESED dataset)
Evaluation metrics: PSDS1, 2, Event-F1, Inersection-F1

Quantitative Comparison

- Downstream accuracy is improved by pre-training using the proposed dataset (Tab 1).
- For CRNN baseline, our pre-training method outperformed real but noisy AudioSet.
- The training curve shows our pre-training method speeds up fine-tuning convergence.
- In the baseline CRNN, accuracy improved monotonically with the increase in the pre-training data scale (Tab 2).

Table1. Accuracy comparison

Model	PSDS1	PSDS2	E-F1(%)	I-F1(%)
CRNN baseline [3]	0.352	0.579	45.7	65.8
w/ Formula-SED (100k)	**0.405**	**0.641**	**49.6**	**72.3**
w/ Audioset Strong (79k) [2]	0.387	0.618	47.7	70.7
Paderborn CRNN [4]	0.262	0.506	34.9	57.3
w/ Formula-SED (100k)	0.278	0.539	35.3	59.4
w/ Audioset Strong (79k) [2]	**0.355**	**0.622**	**43.9**	**67.8**



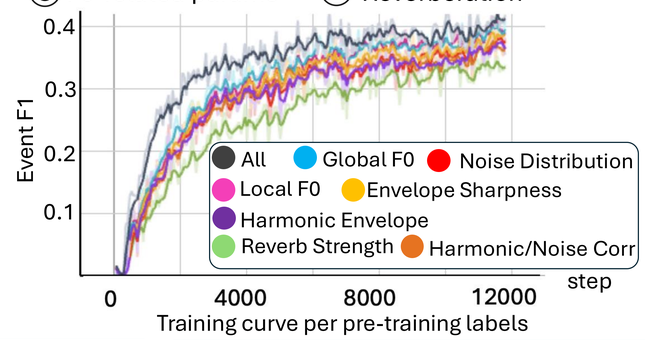
Training curve

Table2. The impact of pre-training dataset size

Model	Size	PSDS1	PSDS2	E-F1(%)	I-F1(%)
CRNN baseline	50k	0.380	0.620	49.4	70.8
CRNN baseline	100k	0.405	0.641	49.6	72.3
CRNN baseline	1M	**0.420**	**0.653**	**51.4**	**72.8**
Paderborn CRNN	50k	0.288	**0.553**	35.8	**62.4**
Paderborn CRNN	100k	0.278	0.539	35.3	59.4
Paderborn CRNN	1M	**0.304**	0.552	**37.1**	61.4

Pre-training Label Analysis

- Formula-SED consists of finite set of synthesis params.
- We investigate which audio parameters are crucial.



Training curve per pre-training labels

5. Future work

- Evaluate the effectiveness of our Formula-SED for various tasks beyond sound event detection.
- Investigate the relationship between label thresholds and downstream accuracy.

6. Reference

[1] E. V. Bonilla, K. Chai, and C. Williams, "Multi-task gaussian process prediction," NIPS, vol. 20, 2007.
[2] S. Hershey et al, "The benefit of temporally-strong labels in audio event classification," ICASSP, pp. 366–370, 2021.
[3] M. Fuentes et al, DCASE 2023. Tampere, Finland: Tampere University, September 2023.
[4] J.EbbersandR.Haeb-Umbach,"Pre-trainingandself-trainingforsound event detection in domestic environments," Tech. Rep. of DCASE 2022 Challenge Task 4, 2022.
[5] Mesaros et al. "Sound event detection: A tutorial." *IEEE Signal Processing Magazine* 38.5 (2021): 67-83.

7. Contact

Yuto Shibata : yuto071508@keio.jp