Identifying Actions for Sound Event Classification

Carnegie Mellon University

Benjamin Elizalde† Radu Revutchi, Samarjit Das⋆ Bhiksha Raj, Ian Lane, Laurie M. Heller

Emails: benjaminm@microsoft.com, radurevutchi@cmu.edu, bhiksha@cmu.edu, ianlane@cmu.edu, laurieheller@cmu.edu, samarjit.das@us.bosch.com
Carnegie Mellon University, ⋆Bosch Research Pittsburgh, †Submitted while at Microsoft

Poster in 4 sentences!

1. We studied how identifying actions can benefit Sound Event Classification (SEC).
2. We used crowdsourcing to relate 20 actions to 50 sound events in 2,000 recordings (ESC-50).
3. Annotations were used to create Action Vectors (AVs) as features.
4. Stacking AVs with SoTA audio embeddings resulted in one of the highest accuracies, 88%.

Annotations and code available on GitHub.

1. Why use actions for sound events?

- Sound events occur only when actions are performed.
- Each type of action is causally connected to the acoustics in the sound event.
- Categories that are semantically similar may be produced by different actions, and therefore have different acoustics.

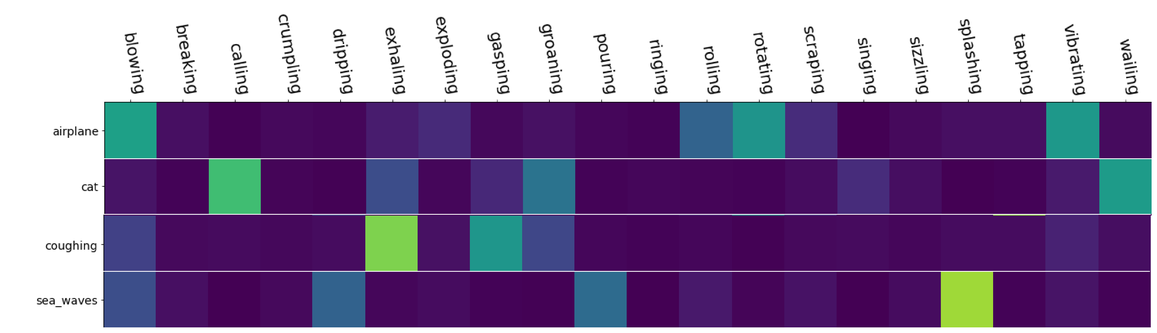
2. Relating Actions to Sound Events

- We augmented the annotations of the ESC-50 sound event dataset.
- We chose actions that did not result in one-to-one relabelling a sound event class.
- For example, we did not replace the "hand sawing" class with action "sawing". Instead, we used verbs that were less tied to specific objects ("saw"), such as "scraping".
- We used crowdsourcing to relate 20 actions to 50 sound events in 2,000 recordings.
- We derived one Action Vector for each recording and used them as features.

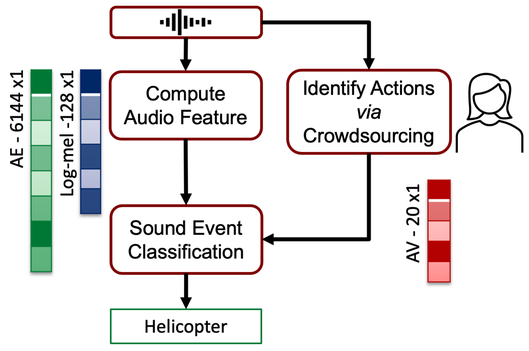
3. Action Vectors (AVs)

- An AV is a 20-dimensional feature where each dimension is an action and the score in each dimension determines how likely the action was to have contributed to the sound event.
- Although the AVs varied among exemplars within a class, the average pattern of ratings for each class was unique (in a 50x20 matrix).
- A multidimensional AV for each file does not equate to a single verb or class.

Figure 1: Rows: average of AVs across all exemplars in a class. Columns: action ratings, brighter green equals more likely.



4. Sound Event Classification (SEC)



- Our goal was to compare SEC for systems that used two common audio feature sets, with and without stacking our AVs.
- Features: Librosa's log-mel spectrograms have 128 dims summarized with their mean across time frames. SoTA OpenL3's Audio Embeddings have 6144 dims similarly summarized.
- Classifiers: Linear SVM used one-vs-rest multiclass classification. DNN had 3 hidden layers: dim. of features to 800 neurons, to 500, to 200, and to 50 classes.

	Input Features	linear SVM	DNN
1	log-mel spectrograms	30.70%	34.00%
2	AVs (Action Vectors)	48.25%	51.81%
3	AEs (Audio Embeddings)	80.90%	81.46%
4	AVs + log-mel	55.05%	69.50%
5	AEs + log-mel	74.35%	78.77%
6	AVs + AEs + log-mel	78.35%	83.31%
7	**AVs + AEs**	**86.60%**	**88.00%**

- The finding that the information was complementary was not known a priori.
- It is remarkable that by adding 0.3% dims (20/6144) to the AEs we can improve performance by an absolute 7%.
- AEs extracted with the alternative parameter of 512 dims underperformed the AVs with 20 dims (40% vs 51.81%).

5. Insights of AVs

- **AVs provide novel information via the graded combination of multiple actions per sound event.**
- Our approach offers more nuanced information than would a one-to-one relabelling with verbs.
- AVs provide an interpretable representation for understanding the confusability and heterogeneity of certain sounds.
- With a vocabulary of 20 actions we can express the diversity in 2,000 different recordings of 50 sounds events.
- We expect that our selection of actions should work with other relevant sound event datasets.
- Sound classes could be broken up or grouped together depending on the actions.
- For example, "airplane" is a category in most databases despite the fact that the actions and acoustics differ from propeller and jet engine, making it a challenge in SEC.

6. Future Work for AVs

- To increase SEC accuracy and interpretability of AVs, we should increase the number of actions.
- Annotations will be used to train models that can automatically generate AVs given an audio file.
- Actions and physical properties can be used to 1) build knowledge of sound events, 2) label audio samples without forcing them into a single class, and 3) describe content of unlabeled audio.