IDENTIFYING ACTIONS FOR SOUND EVENT CLASSIFICATION

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

Carnegie Mellon University, ⋆ Bosch Research Pittsburgh
Email: bmartin1,radurevutchi,bhiksha,lane,laurieheller@cmu.edu, samarjit.das@us.bosch.com

ABSTRACT

In Psychology, actions are pivotal for humans to perceive and separate sound events. In Machine Learning, action recognition achieves high accuracy; however, it has not been asked if identifying actions can benefit Sound Event Classification (SEC), as opposed to semi-direct processing of the audio signal. Therefore, we propose a new Psychology-inspired approach for SEC that includes identification of actions *via* human listeners. To achieve this goal, we used crowdsourcing to have listeners identify 20 actions that in isolation or in combination may have produced any of the 50 sound events in the well-studied dataset ESC-50. The resulting annotations for each audio recording relate actions to an entire database of sound events for the first time [1]. The annotations were used to create interpretable representations called Action Vectors (AVs). We evaluated SEC by comparing the AV with two types of audio features – log-mel spectrograms and state of the art audio embeddings. Because audio features and AVs capture different abstractions of the acoustic content, we combined them and achieved the highest reported accuracy (86.75%) in ESC-50, showing that Psychology-inspired approaches can improve SEC.

Index Terms— Sound Event Classification, Psychology, Crowdsourcing, Audio Signal Processing, Audio Tagging

1. INTRODUCTION

Sound Event Classification (or Audio Tagging) aims to assign a sound event class label to an audio clip (e.g. "dog", "car beeping"). Typically, SEC consists of taking an input audio file, computing features inspired by auditory perception, and used them to train Machine Learning classifiers. Motivated by the success of understanding how humans hear in order to build acoustic intelligence, we looked further into human perception and identification of sound events.

Psychomechanics is the study of the perception of physical properties of sound-sources [1], which has found that actions distinguish sound events better than the source materials, size and shape. Gaver [2] argued that listening to everyday sounds focuses mainly on the physical aspects producing the sound. Lemaitre and Heller [3] concluded that the listener's ability to identify materials based on audio is in general not accurate, except for water and resonant materials (e.g. glass, metal). Moreover, perception of size and shape are in general less reliable than materials [4, 5, 6, 7]. In contrast, Lemaitre and Heller [8] showed that actions are the pivotal principle for perceiving and separating sounds. They also found that actions producing simple sound events were better identified than the materials [3]. Verbs can describe actions and interactions between one or several objects and sometimes also the material the objects are made of [9, 10, 11, 12, 13]. VanDerveer [14] found that people who were asked to identify sounds made by common objects would spontaneously describe the actions involved in generating the sounds. Hence, it is expected that Machine Learning models trained on physical properties of sound-sources will exhibit similar accuracy.

In fact, despite the scarce literature of Machine Learning models trained on physical properties of sound-sources, models trained on audio labeled with actions achieved significantly higher accuracy than materials. Owen et al [15] recorded videos of themselves hitting and scratching eleven surfaces of different materials (e.g. grass, metal, water, wood). They evaluated audio-based classification of the eleven materials and achieved 40% accuracy. Then, they evaluated audio-based action classification of hitting and scratching and achieved 84% accuracy. Previously in [16], we collected audio labeled with 400 suffixed nouns derived from verbs referring to the action generation (e.g "clapping crowd"). Then, we evaluated classification and achieved an overall 71% accuracy, with classes performing as high as "ringing alert" with 92%. Psychology and Machine Learning have independently demonstrated that actions can be reliably identified. In Psychology, identifying actions have been a successful intermediate step to recognizing sound events, yet Machine Learning had not tested a similar approach for SEC.

Identifying actions for SEC can provide an interpretable intermediate step and can be used across different sound event datasets. Typically, SEC processes the audio signal semi-directly and maps it to a sound event. Actions can serve to bridge the audio with the sound event providing an interpretable step to explain SEC. This is important because sound event datasets have many categories named after sound-sources and overlook the fact that the same object can produce different types of sounds. For example, the dataset of ESC-50 [17] evidenced that for listeners and Machine Learn-

Thanks to Bosch Research and the SOWG for their financial support.

[1] Dataset will be released after revision.

ing models, classification of "airplane" and "helicopter" were highly confused because they both share similar sounds produced by the propeller and rotor engine respectively. In this case, identifying the sound produced by the action of rotating could serve to explain the acoustic confusion, because actions tend to produce consistent acoustics that can cut across the semantics of sound events labels. It would also help to determine if we can combine training audio of the same class, but from different datasets or annotation processes [18]. We aim for a set of actions fundamental enough that it could describe a larger set of sound events in different contexts.

Therefore, we propose a new Psychology-inspired approach for SEC that includes identification of actions. We used crowdsourcing to have humans identify 20 actions that in isolation or in combination produced any of the 50 sound events in the well-studied dataset ESC-50. The resulting annotations for each audio recording relate actions to sound events for the first time (available online). Combining audio features and action identification improved SEC. We demonstrated the benefit of drawing from domain knowledge of Psychology to design Machine Learning algorithms for SEC.

2. RELATING ACTIONS TO SOUND EVENTS

In order to relate actions to sound events, we chose a well-studied sound event dataset called ESC-50 [17]. ESC-50 has 50 classes from five broad categories: animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior sounds. Each class has 40 files of five seconds for a total of 2,000 audio files. The sound categories do not necessarily have a strong intra-class acoustic consistency. The sound events are generally exposed in the foreground with limited background noise when possible. Field recordings may exhibit overlap of competing sound sources in the background. The audio comes from the Internet, meaning the recording process is unknown. Categories consist of one or two words, about 14% are labeled with a specific action and most of them with nouns. The next step was to select relevant actions.

Selecting actions that in isolation or combination produced at least part of the ESC-50 sound events was challenging. The recording process and definition of the sound event classes were unknown, thus we listened to audio recordings corresponding to each class and chose actions that could have produced at least part of the audio content. We had to choose actions that did not imply relabelling a sound event class with a dominant action. For example, choosing the action "sawing" implied relabelling the sound event class "hand sawing". In practice, this is not an issue, but we wanted to draw new insights from using different action classes. The number of actions represented the number of options that would be given to the annotators, the higher the number of options the lower the annotation quality. Listening studies [8] to identify actions in audio clips considered about 20 category-options at a time, so we kept our selection of actions around that number.

The selection of actions was inspired by listening experiments in the literature of Psychology and our own listening experiments in our auditory lab. We drew actions from a previously collected dataset by Heller [19], and two taxonomies of everyday sounds, Gaver's [2] and Houix's [20]. The taxonomies included different variations of actions, such as interactions (friction), specific actions (scraping), manners of actions (scraping rapidly), and objects of actions (scraping a board). Lemaitre and Heller [8] carried on listening experiments with these variations and concluded that specific actions achieved the highest accuracy and fastest identification response. Thus, we constrained our study to specific actions. The revised selection had about 25 actions, which was then tested in our own listening experiments with a small subset of the dataset. The final selection had 20 actions shown in Table 1, which were used to annotate all the audio in ESC-50.

dripping	splashing	pouring	breaking
rolling	scraping	exhaling	vibrating
groaning	gasping	singing	tapping
crumpling	blowing	exploding	rotating
wailing	calling	ringing	sizzling

Table 1. We selected 20 actions that in isolation or combination could have produced at least part (of most) of the 50 sound events.

Due to the large number of audio files (2,000) in ESC-50 dataset, we identified actions using crowdsourcing (Mechanical Turk). We designed an interface that included a playable audio clip (without showing the sound event label), the question "For each action below, judge how likely it is to have produced at least part of the sound event.", and the 20 actions to be scored. The scores were inspired by a five-point Likert scale ranging from 0-4 where 0 meant that the action contribution was very unlikely and 4 that it was very likely. We asked for annotators to be fluent in English, wear headphones, have no hearing impairments, and be between 18 and 60 years old. Although we also restricted participation to annotators with high success rates in Mechanical Turk, we rejected about 17% of the annotations for having conflicting actions co-occurring with high-scores. For example, for an audio of "toilet flushing" some participants would choose water-related actions and vocalization actions, even though there were no vocalizations. Each audio file was annotated by three different participants for a total of 6,000 annotations for the entire dataset. An extended explanation of identifying actions via crowdsourcing and the collected annotations will be included in an online repository.

3. EXPERIMENTS AND RESULTS

3.1. Creating Action Vectors

After having collected the annotations derived from identifying actions in the audio of ESC-50, we proceeded to create Action Vectors. To create an AV for each audio file, we

summed the scores independently for each of the 20 actions across the three annotators to have accumulated scores from 0 to 12. An AV is an interpretable audio representation where each dimension is an action and the score in each dimension determines how much that action (was likely to have) contributed to produce the sound event. Although the vector has only one dimension to represent how the sound unfolds in time, the temporal information is implied in the action. Once we created the 2,000 AV for ESC-50, we used them for SEC.

3.2. Comparing Sound Event Classification using Audio Features and Action Vectors

In this paper, we study for the first time how identifying actions can benefit SEC, as opposed to semi-direct processing of the audio signal. Therefore, we compared SEC utilizing AVs versus two common audio features. The two main SEC pipelines for our experiments are illustrated in Figure 1. On the left is the typical approach that takes the input audio, computes audio features, and assigns a class label. On the right, we depict our addition of an intermediate step where listeners identify actions in the audio.

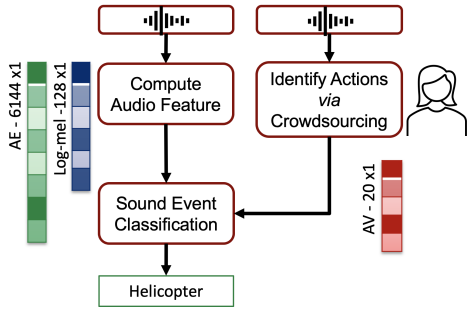

Fig. 1. Typically, SEC takes the input audio, computes audio features and assigns a class label. We proposed to add an intermediate step where listeners identify actions in the audio. The identified actions are transformed into Action Vectors and are used for automatic SEC.

For the input audio, we computed two commonly-used types of audio features, log-mel spectrograms and state-of-the-art data-driven audio embeddings (AE). We computed log-mel spectrograms using the librosa package [21] with default settings resulting in 128 mel-filters summarized with their mean across time frames. We computed AE using a deep learning architecture called OpenL3 [22], which was pre-trained with $60M$ videos. The parameters used were, content_type: $music$, input_repr: $mel256$, embedding_size: 6144. The AE were summarized with their mean across time frames. The audio features and the AV were normalized ($L2$) and/or scaled (removed the mean and scaled to unit variance) before passing them to the classifier.

The audio features and the AV were used to train two types of Machine Learning classifiers for SEC, a linear Support Vector Machine (SVM) and a Deep Neural Network (DNN). A linear SVM classifier provides fast computation time and is often employed with large dimensionality features. We included a DNN as a non-linear classifier alternative that often performs better than linear classifiers. The parameters of both classifiers and the architecture of the DNN are quite similar regardless of the input feature, but we tuned the parameters to maximize performance. For the SVM we set $C = 35$. The DNN had 3 hidden layers $800 : 500 : 200$ with $tanh$ activation. The output layer had a $softmax$ activation combined with a Categorical Cross-Entropy Loss. We used SGD optimizer ($lr = 0.008$) and 100 epochs. ESC-50 distributes the audio files into five folds, so we ran the pipeline using the five combinations in which each fold has to be the test set once. Overall accuracy was computed across the five folds to evaluate SEC performance. The SVM produces the same accuracy in every run, but the DNN was run 10 times to compute a standard deviation.

Features	linear SVM	DNN
log-mel spectrogram	33.30%	34.00%
AE (Audio Embeddings)	79.35%	79.75%
log-mel+AE	81.05%	80.77%
AV (Action Vectors)	48.25%	51.81%
AV+log-mel	55.05%	69.50%
AV+AE	**86.60%**	**86.75%**

Table 2. Overall classification accuracy (%) with different inputs. When AE are combined with AV, we achieved the highest accuracy reported in the ESC-50 dataset.

SEC using AV resulted in good performance, suggesting that AVs carry useful information to bridge audio and sound events. Table 2 shows SEC accuracy using AVs and both types of audio features. SEC with AV yielded 48.25% with the SVM and 51.81% (sd=0.4%) with the DNN. SEC with log-mel spectrograms yielded 33.30% accuracy with the SVM and 34.00% (sd=0.3%) using the DNN, which is consistent with other papers that use log-mel-based features [17]. SEC with AE yielded 79.35% with the SVM and 79.75% (sd=0.3%) with the DNN, which is consistent with the OpenL3 paper [22]. The AV outperformed log-mel spectrograms by an absolute 15% to 17% accuracy, but AV under performed AE by an absolute 30 to 28% accuracy.

Audio features capture low-level spectro-temporal patterns of the audio content, whereas AVs capture a higher-level combination of spectro-temporal patterns. Hence, we combined AVs and audio features independently to evaluate how they complement each other for SEC. We used a SEC pipeline similar to the one used in the first set of experiments and tried three combinations: AV with log-mel spectrograms, AV with AE, and log-mel spectrograms with AE.

SEC combining identification of actions and audio features resulted in the highest accuracy reported in ESC-50. Table 2 shows SEC using the three stacked combinations. SEC with AV and AE yielded 86.6% accuracy with SVM

and 86.75% (sd=0.3%) with DNN, which are the highest accuracy reported in ESC-50. In the literature, the accuracy range above human performance goes from 81.8% [23] to 86.5% [24], where human performance is 81.3% [17]. These numbers are produced by complex systems that combine features and classifiers, use transfer learning, and add data augmentation. SEC with AV and log-mel spectrograms yielded 55.05% with SVM and 69.50% (sd=0.3%) with DNN. In both cases the combination resulted in better performance than in isolation. On the other hand, SEC with stacked AE and log-mel spectrograms yielded 81.05% with SVM and 80.77% (sd=0.5%) with DNN, suggesting that the audio features have minimal complementary information. Although the pre-trained network used to extract the AE uses log-mel spectrograms as the input feature, AE are data-driven features that do not resemble the initial log-mel spectrograms.

4. DISCUSSION

Identification of actions represented by AV have different benefits, such as: providing pivotal information in the combination of actions, providing an interpretable semantic representation, and providing a new way to organize sound events.

The main reason why AV works well is because the actions are combined to characterize a sound event. We considered multiple actions, instead of just relabelling sound events with a dominant action, and each action can have a degree of contribution instead of a binary contribution. To confirm this, we scaled the scores of the AV to be from 0 to 1 (instead of 0 to 12) and then quantized them with a threshold of 0.5, scores under the threshold were set to 0, and scores greater or equal were set to 1. Then we ran SEC stacking AE and the new quantized AV and achieved 82% instead of 86.75%. On average, AV have non-zero scores in 6 out of 20 dimensions.

Actions are fundamental categories that can acoustically describe a larger number of sound events. AV with only 20 dimensions can acoustically describe 50 sound events, more than twice the number of categories, whereas the log-mel spectrograms have 128 dimensions and the AE have 6144. In fact, when extracting the AE with 512 dimensions (an alternative parameter value) SEC accuracy dropped from 79.75%to about 40% (AV have 51.81%). More importantly, each dimension in the AV is interpretable.

AVs are an interpretable semantic representation, as opposed to AE and most data-driven representations. Whether AVs are part of a SEC pipeline or not, they can help to explain inter and intra-class confusion. For instance, the high confusion occurring between "airplane" and "helicopter" can be explained by a subset of "airplane" recordings having sounds produced by a propeller engine, which are similar to the sounds produced by the rotor engine of a helicopter. These recordings scored high for the action"rotating". Actions tend to produce consistent acoustics that can cut across the semantics of sound events. Thus, a set of actions can describe sound events with similar acoustic content, but with different labels and in different contexts. For example, we could also use "rotating" for the sound event of "washing machine" in domestic sounds. To increase the SEC accuracy and interpretable capabilities of AV, we could increase the number of actions.

SEC accuracy using AV depends on the selection and number of actions. Not all of the actions affect SEC equally. This is determined by how an action, isolation or in combination, can distinguish sound events. For example, one of the actions was "calling", which scored high in 11 sounds produced by vocal tracts, such as "frog", "dog", "crow", resulting in inter-class confusion. When removing those 11 classes from the dataset, SEC accuracy with AE remained around 80%, but with AV increased from 51% to 57%. Thus, by adding more actions that distinguish vocalizations, the better AV can discriminate between such sound events. The selection of actions also affects how we organize sound events.

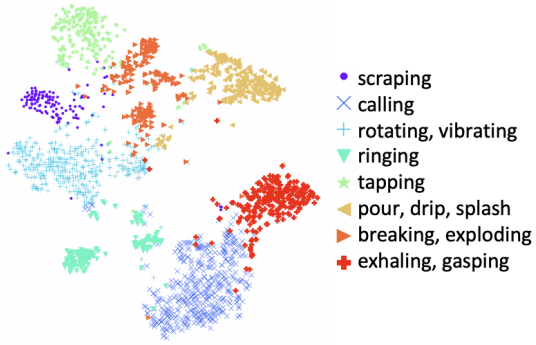

Fig. 2. AV provide a new way to organize sound events based on shared actions. The plot shows how ESC-50 clustered in 8 groups and the corresponding dominant actions.

AVs provide a new way to organize sound events based on shared actions. We grouped the $2,000$ AV of ESC-50 using K-means (k=8) and plot them (tsne, perplexity=50) in Figure 2. Each of the 8 groups was assigned a label corresponding to its dominant action(s): scraping, calling, rotating and vibrating, ringing, tapping, pouring and dripping and splashing, breaking and exploding, exhaling and gasping. Varying the number of clusters created groups with different dominant actions. Organizing sound events by their shared actions rather than their shared objects may be helpful to automatic classification in ways that have yet to be explored.

5. CONCLUSIONS

We demonstrated how our proposed Psychology-inspired approach of identifying actions improved SEC. AV were derived from humans identifying actions, but we are in the process of using our annotations to train models that can automatically generate the AV given an audio file. Drawing domain knowledge from Psychology can help to refer to sound events by description and not force them into one name. Descriptions can help in aspects such as building knowledge or ameliorate the problem of large-scale annotations.

6. REFERENCES

[1] Guillaume Lemaitre, Nicolas Grimault, and Clara Suied, "Acoustics and psychoacoustics of sound scenes and events," in *Computational Analysis of Sound Scenes and Events*, pp. 41–67. Springer, 2018.

[2] William W Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[3] Guillaume Lemaitre and Laurie M Heller, "Auditory perception of material is fragile while action is strikingly robust," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1337–1348, 2012.

[4] Massimo Grassi, "Do we hear size or sound? balls dropped on plates," *Perception & psychophysics*, vol. 67, no. 2, pp. 274–284, 2005.

[5] Massimo Grassi, Massimiliano Pastore, and Guillaume Lemaitre, "Looking at the world with your ears: How do we get the size of an object from its sound?," *Acta psychologica*, vol. 143, no. 1, pp. 96–104, 2013.

[6] Mark MJ Houben, Armin Kohlrausch, and Dik J Hermes, "The contribution of spectral and temporal information to the auditory perception of the size and speed of rolling balls," *Acta acustica united with acustica*, vol. 91, no. 6, pp. 1007–1015, 2005.

[7] Robert A Lutfi, "Auditory detection of hollowness," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1010–1019, 2001.

[8] Guillaume Lemaitre and Laurie M Heller, "Evidence for a basic level in a taxonomy of everyday action sounds," *Experimental brain research*, vol. 226, no. 2, 2013.

[9] Alireza Darvishi, Eugen Munteanu, Valentin Guggiana, Helmut Schauer, M Motavalli, and Matthias Rauterberg, "Designing environmental sounds based on the results of interaction between objects in the real world," in *Human—Computer Interaction*. Springer, 1995.

[10] R.M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*, Inner Traditions, 1993.

[11] Brian Gygi, Gary R. Kidd, and Charles S. Watson, "Spectral-temporal factors in the identification of environmental sounds," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1252–1265, 2004.

[12] James A Ballas and James H Howard Jr, "Interpreting the language of environmental sounds," *Environment and behavior*, vol. 19, no. 1, pp. 91–114, 1987.

[13] Danièle Dubois, Catherine Guastavino, and Manon Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 865–874, 2006.

[14] Nancy J VanDerveer, *Ecological acoustics: Human perception of environmental sounds.*, Cornell University, 1980.

[15] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recog.*, 2016.

[16] Sebastian Säger, Benjamin Elizalde, Damian Borth, Christian Schulze, Bhiksha Raj, and Ian Lane, "Audiopairbank: towards a large-scale tag-pair-based audio content analysis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 12, 2018.

[17] Karol J. Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30*, 2015.

[18] Benjamin Elizalde, Mirco Ravanelli, and Gerald Friedland, "Audio concept ranking for video event detection on user-generated content," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[19] Laurie Heller and Asadali Sheikh, "Acoustic features of environmental sounds that convey actions.," *Auditory Perception, Cognition, and Action Meeting*, November 2019, Montreal, Canada.

[20] Olivier Houix, Guillaume Lemaitre, Nicolas Misdariis, Patrick Susini, and Isabel Urdapilleta, "A lexical analysis of environmental sound categories.," *Journal of Experimental Psychology: Applied*, vol. 18, pp. 52, 2012.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[22] J. Cramer, H-.H. Wu, J. Salamon, and J. P. Bello, "Look, listen and learn more: Design choices for deep audio embeddings," in *IEEE ICASSP*, Brighton, UK, 2019.

[23] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[24] Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification.," in *INTERSPEECH*, 2017.