karoldvl / **ESC-50**

ESC-50: Dataset for Environmental Sound Classification

#dataset #environmental-sounds #audio

19 commits	1 branch	0 releases	1 contributor

Branch: **master** | **New pull request** | **Create new file** | **Upload files** | **Find file** | **Clone or download**

karoldvl Add tests for mono — Latest commit `82974d4` 9 days ago

.circleci	Update dataset (#2)	a month ago
.github	Add probot-stale	19 days ago
audio	Update dataset (#2)	a month ago
meta	Update dataset (#2)	a month ago
tests	Add tests for mono	9 days ago
.gitignore	Update dataset (#2)	a month ago
LICENSE	Update dataset (#2)	a month ago
README.md	Add "NELS - Never-Ending Learner of Sounds" paper	21 days ago
esc50.gif	Update dataset (#2)	a month ago
pytest.ini	Update dataset (#2)	a month ago
requirements.txt	Update dataset (#2)	a month ago

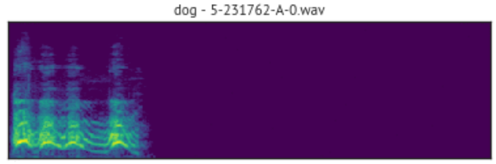
README.md

ESC-50: Dataset for Environmental Sound Classification

> **Overview** | **Download** | **Results** | **Repository content** | **License** | **Citing** | **Caveats** | **Changelog**
>
> PASSED | license CC BY-NC | download .zip

The **ESC-50 dataset** is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification.

dog - 5-231762-A-0.wav



The dataset consists of 5-second-long recordings organized into 50 semantical classes (with 40 examples per class) loosely arranged into 5 major categories:

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

Clips in this dataset have been manually extracted from public field recordings gathered by the **Freesound.org project**. The dataset has been prearranged into 5 folds for comparable cross-validation, making sure that fragments from the same original source file are contained in a single fold.

A more thorough description of the dataset is available in the original paper with some supplementary materials on GitHub: **ESC: Dataset for Environmental Sound Classification - paper replication data**.

Download

The dataset can be downloaded as a single .zip file (~600 MB):

Download ESC-50 dataset

Results

Numerous machine learning & signal processing approaches have been evaluated on the ESC-50 dataset. Most of them are listed here. If you know of some other reference, you can message me or open a Pull Request directly.

> **Terms used in the table:**
>
> • CNN - Convolutional Neural Network
> • CRNN - Convolutional Recurrent Neural Network
> • GMM - Gaussian Mixture Model
> • GTCC - Gammatone Cepstral Coefficients
> • GTSC - Gammatone Spectral Coefficients
> • k-NN - k-Neareast Neighbors
> • MFCC - Mel-Frequency Cepstral Coefficients
> • MLP - Multi-Layer Perceptron
> • RBM - Restricted Boltzmann Machine
> • RNN - Recurrent Neural Network
> • SVM - Support Vector Machine
> • TEO - Teager Energy Operator
> • ZCR - Zero-Crossing Rate

Title	Notes	Accuracy	Paper	Code
Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification	CNN with filterbanks learned using convolutional RBM + fusion with GTSC and mel energies	86.50%	sailor2017	
Learning from Between-class Examples for Deep Sound Recognition	EnvNet-v2 (tokozume2017a) + data augmentation + Between-Class learning	84.90%	tokozume2017b	
Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification	CNN working with phase encoded mel filterbank energies (PEFBEs), fusion with Mel energies	84.15%	tak2017	
Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes	CNN pretrained on AudioSet	83.50%	kumar2017	

Title	Notes	Accuracy	Paper	Code
Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification	CNN with filterbanks learned using convolutional RBM + fusion with GTSC	83.00%	sailor2017	
Novel TEO-based Gammatone Features for Environmental Sound Classification	Fusion of GTSC & TEO-GTSC with CNN	81.95%	agrawal2017	
Learning from Between-class Examples for Deep Sound Recognition	EnvNet-v2 (tokozume2017a) + Between-Class learning	81.80%	tokozume2017b	
Human accuracy	Crowdsourcing experiment in classifying ESC-50 by human listeners	81.30%	piczak2015a	
Objects that Sound	*Look, Listen and Learn* (L3) network (arandjelovic2017a) with stride 2, larger batches and learning rate schedule	79.80%	arandjelovic2017b	
Look, Listen and Learn	8-layer convolutional subnetwork pretrained on an audio-visual correspondence task	79.30%	arandjelovic2017a	
Novel TEO-based Gammatone Features for Environmental Sound Classification	GTSC with CNN	79.10%	agrawal2017	
Learning from Between-class Examples for Deep Sound Recognition	EnvNet-v2 (tokozume2017a) + data augmentation	78.80%	tokozume2017b	
Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification	CNN with filterbanks learned using convolutional RBM	78.45%	sailor2017	
Learning from Between-class Examples for Deep Sound Recognition	Baseline CNN (piczak2015b) + Batch Normalization + Between-Class learning	76.90%	tokozume2017b	
Novel TEO-based Gammatone Features for Environmental Sound Classification	TEO-GTSC with CNN	74.85%	agrawal2017	
Learning from Between-class Examples for Deep Sound Recognition	EnvNet-v2 (tokozume2017a)	74.40%	tokozume2017b	
Soundnet: Learning sound representations from unlabeled video	8-layer CNN (raw audio) with transfer learning from unlabeled videos	74.20%	aytar2016	
Learning from Between-class Examples for Deep Sound Recognition	18-layer CNN on raw waveforms (dai2016) + Between-Class learning	73.30%	tokozume2017b	
Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification	CNN working with phase encoded mel filterbank energies (PEFBEs)	73.25%	tak2017	
Classifying environmental sounds using image recognition networks	16 kHz sampling rate, GoogLeNet on spectrograms (40 ms frame length)	73.20%	boddapati2017	
Learning from Between-class Examples for Deep Sound Recognition	Baseline CNN (piczak2015b) + Batch Normalization	72.40%	tokozume2017b	
Novel TEO-based Gammatone Features for Environmental Sound Classification	Fusion of MFCC & TEO-GTCC with GMM	72.25%	agrawal2017	
Learning environmental sounds with end-to-end convolutional neural network (EnvNet)	Combination of spectrogram and raw waveform CNN	71.00%	tokozume2017a	
Novel TEO-based Gammatone Features for Environmental Sound Classification	TEO-GTCC with GMM	68.85%	agrawal2017	

Title	Notes	Accuracy	Paper	Code
Classifying environmental sounds using image recognition networks	16 kHz sampling rate, AlexNet on spectrograms (30 ms frame length)	68.70%	boddapati2017	
Very Deep Convolutional Neural Networks for Raw Waveforms	18-layer CNN on raw waveforms	68.50%	dai2016, tokozume2017b	
Classifying environmental sounds using image recognition networks	32 kHz sampling rate, GoogLeNet on spectrograms (30 ms frame length)	67.80%	boddapati2017	
WSNet: Learning Compact and Efficient Networks with Weight Sampling	SoundNet 8-layer CNN architecture with 100x model compression	66.25%	jin2017	
Soundnet: Learning sound representations from unlabeled video	5-layer CNN (raw audio) with transfer learning from unlabeled videos	66.10%	aytar2016	
WSNet: Learning Compact and Efficient Networks with Weight Sampling	SoundNet 8-layer CNN architecture with 180x model compression	65.80%	jin2017	
Soundnet: Learning sound representations from unlabeled video	5-layer CNN trained on raw audio of ESC-50 only	65.00%	aytar2016	
Environmental Sound Classification with Convolutional Neural Networks - ***CNN baseline***	CNN with 2 convolutional and 2 fully-connected layers, mel-spectrograms as input, vertical filters in the first layer	64.50%	piczak2015b	
auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks	MLP classifier on features extracted with an RNN autoencoder	64.30%	freitag2017	
Classifying environmental sounds using image recognition networks	32 kHz sampling rate, AlexNet on spectrograms (30 ms frame length)	63.20%	boddapati2017	
Classifying environmental sounds using image recognition networks	CRNN	60.30%	boddapati2017	
Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks	3-layer CNN with vertical filters on wideband mel-STFT (*median accuracy*)	*56.37%*	huzaifah2017	
Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks	3-layer CNN with square filters on wideband mel-STFT (*median accuracy*)	*54.00%*	huzaifah2017	
Soundnet: Learning sound representations from unlabeled video	8-layer CNN trained on raw audio of ESC-50 only	51.10%	aytar2016	
Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks	5-layer CNN with square filters on wideband mel-STFT (*median accuracy*)	*50.87%*	huzaifah2017	
Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks	5-layer CNN with vertical filters on wideband mel-STFT (*median accuracy*)	*46.25%*	huzaifah2017	
Baseline - random forest	Baseline ML approach (MFCC & ZCR + random forest)	44.30%	piczak2015a	
Soundnet: Learning sound representations from unlabeled video	Convolutional autoencoder trained on unlabeled videos	39.90%	aytar2016	
Baseline - SVM	Baseline ML approach (MFCC & ZCR + SVM)	39.60%	piczak2015a	

Title	Notes	Accuracy	Paper	Code
Baseline - k-NN	Baseline ML approach (MFCC & ZCR + k-NN)	32.20%	piczak2015a	
A mixture model-based real-time audio sources classification method	Dictionary of sound models used for classification (*accuracy is computed on segments instead of files*)	*94.00%*	baelde2017	
NELS - Never-Ending Learner of Sounds	Large-scale audio crawling with classifiers trained on AED datasets (including ESC-50)	N/A	elizalde2017	
Utilizing Domain Knowledge in End-to-End Audio Processing	End-to-end CNN with learned mel-spectrogram transformation	N/A	tax2017	
Deep Neural Network based learning and transferring mid-level audio features for acoustic scene classification	Transfer learning from various datasets, including ESC-50	N/A	mun2017	
Features and Kernels for Audio Event Recognition	MFCC, GMM, SVM	N/A	kumar2016b	
A real-time environmental sound recognition system for the Android OS	Real-time sound recognition for Android evaluated on ESC-10	N/A	pillos2016	
Comparing Time and Frequency Domain for Audio Event Recognition Using Deep Learning	Discriminatory effectiveness of different signal representations compared on ESC-10 and Freiburg-106	N/A	hertel2016	
Audio Event and Scene Recognition: A Unified Approach using Strongly and Weakly Labeled Data	Combination of weakly labeled data (YouTube) with strong labeling (ESC-10) for Acoustic Event Detection	N/A	kumar2016a	

Repository content

- `audio/*.wav`

  2000 audio recordings in WAV format (5 seconds, 44.1 kHz, mono) with the following naming convention:

  `{FOLD}-{CLIP_ID}-{TAKE}-{TARGET}.wav`

  - `{FOLD}` - index of the cross-validation fold,
  - `{CLIP_ID}` - ID of the original Freesound clip,
  - `{TAKE}` - letter disambiguating between different fragments from the same Freesound clip,
  - `{TARGET}` - class in numeric format [0, 49].

- `meta/esc50.csv`

  CSV file with the following structure:

filename	fold	target	category	esc10	src_file	take

  The `esc10` column indicates if a given file belongs to the *ESC-10* subset (10 selected classes, CC BY license).

- `meta/esc50-human.xlsx`

  Additional data pertaining to the crowdsourcing experiment (human classification accuracy).

License

The dataset is available under the terms of the Creative Commons Attribution Non-Commercial license.

A smaller subset (clips tagged as *ESC-10*) is distributed under CC BY (Attribution).

Attributions for each clip are available in the LICENSE file.

Citing

If you find this dataset useful in an academic setting please cite:

download paper PDF

> K. J. Piczak. **ESC: Dataset for Environmental Sound Classification**. *Proceedings of the 23rd Annual ACM Conference on Multimedia*, Brisbane, Australia, 2015.
>
> [DOI: http://dx.doi.org/10.1145/2733373.2806390]

```
@inproceedings{piczak2015dataset,
  title = {{ESC}: {Dataset} for {Environmental Sound Classification}},
  author = {Piczak, Karol J.},
  booktitle = {Proceedings of the 23rd {Annual ACM Conference} on {Multimedia}},
  date = {2015-10-13},
  url = {http://dl.acm.org/citation.cfm?doid=2733373.2806390},
  doi = {10.1145/2733373.2806390},
  location = {{Brisbane, Australia}},
  isbn = {978-1-4503-3459-4},
  publisher = {{ACM Press}},
  pages = {1015--1018}
}
```

Caveats

Please be aware of potential information leakage while training models on *ESC-50*, as some of the original Freesound recordings were already preprocessed in a manner that might be class dependent (mostly bandlimiting). Unfortunately, this issue went unnoticed when creating the original version of the dataset. Due to the number of methods already evaluated on *ESC-50*, no changes rectifying this issue will be made in order to preserve comparability.

Changelog

v2.0.0 (2017-12-13)

> - Change to WAV version as default.

v2.0.0-pre (2016-10-10) (wav-files branch)

> - Replace OGG recordings with cropped WAV files for easier loading and frame-level precision (some of the OGG recordings had a slightly different length when loaded).
> - Move recordings to a one directory structure with a meta CSV file.

v1.0.0 (2015-04-15)

> - Initial version of the dataset (OGG format).