lhotse

Release 0.1

Lhotse development team

Nov 19, 2020

CONTENTS:

CHAPTER

ONE

GETTING STARTED

Lhotse is a Python library aiming to make speech and audio data preparation flexible and accessible to a wider community. Alongside k2, it is a part of the next generation Kaldi speech processing library.

1.1 About

1.1.1 Main goals

- Attract a wider community to speech processing tasks with a **Python-centric design**.
- Accommodate experienced Kaldi users with an **expressive command-line interface**.
- Provide **standard data preparation recipes** for commonly used corpora.
- Provide **PyTorch Dataset classes** for speech and audio related tasks.
- Flexible data preparation for model training with the notion of **audio cuts**.
- **Efficiency**, especially in terms of I/O bandwidth and storage capacity.

1.1.2 Main ideas

Like Kaldi, Lhotse provides standard data preparation recipes, but extends that with a seamless PyTorch integration through task-specific Dataset classes. The data and meta-data are represented in human-readable text manifests and exposed to the user through convenient Python classes.

Lhotse introduces the notion of audio cuts, designed to ease the training data construction with operations such as mixing, truncation and padding that are performed on-the-fly to minimize the amount of storage required. Data augmentation and feature extraction are supported both in pre-computed mode, with highly-compressed feature matrices stored on disk, and on-the-fly mode that computes the transformations upon request. Additionally, Lhotse introduces feature-space cut mixing to make the best of both worlds.

1.2 Installation

Lhotse supports Python version 3.7 and later.

1.2.1 Pip

Lhotse is available on PyPI:

```
pip install lhotse
```

To install the latest, unreleased version, do:

```
pip install git+https://github.com/lhotse-speech/lhotse
```

1.2.2 Development installation

For development installation, you can fork/clone the GitHub repo and install with pip:

```
git clone https://github.com/lhotse-speech/lhotse
cd lhotse
pip install -e '.[dev]'

# Running unit tests
pytest test
```

This is an editable installation (`-e` option), meaning that your changes to the source code are automatically reflected when importing lhotse (no re-install needed). The `[dev]` part means you're installing extra dependencies that are used to run tests, build documentation or launch jupyter notebooks.

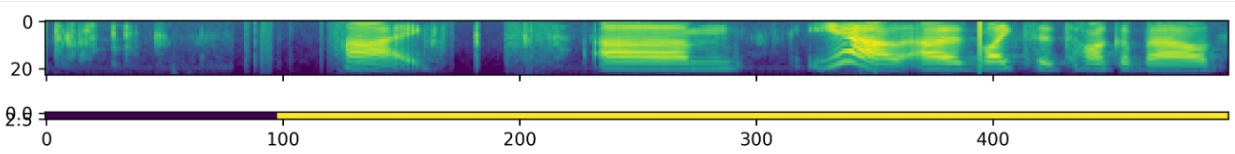
1.3 Examples

We have example recipes showing how to prepare data and load it in Python as a PyTorch `Dataset`. They are located in the `examples` directory.

A short snippet to show how Lhotse can make audio data prepartion quick and easy:

```
from lhotse import CutSet, Fbank, LilcomFilesWriter
from lhotse.dataset import VadDataset
from lhotse.recipes import prepare_switchboard

# Prepare data manifests from a raw corpus distribution.
# The RecordingSet describes the metadata about audio recordings;
# the sampling rate, number of channels, duration, etc.
# The SupervisionSet describes metadata about supervision segments:
# the transcript, speaker, language, and so on.
swbd = prepare_switchboard('/export/corpora3/LDC/LDC97S62')

# CutSet is the workhorse of Lhotse, allowing for flexible data manipulation.
# We create 5-second cuts by traversing SWBD recordings in windows.
# No audio data is actually loaded into memory or stored to disk at this point.
cuts = CutSet.from_manifests(
    recordings=swbd['recordings'],
    supervisions=swbd['supervisions']
).cut_into_windows(duration=5)

# We compute the log-Mel filter energies and store them on disk;
# Then, we pad the cuts to 5 seconds to ensure all cuts are of equal length,
# as the last window in each recording might have a shorter duration.
# The padding will be performed once the features are loaded into memory.
with LilcomFilesWriter('feats') as storage:
    cuts = cuts.compute_and_store_features(
        extractor=Fbank(),
        storage=storage,
    ).pad(duration=5.0)

# Construct a Pytorch Dataset class for Voice Activity Detection task:
dataset = VadDataset(cuts)
dataset[0]
```

The `VadDataset` will yield a pair of input and supervision tensors such as the following - the speech starts roughly at the first second (100 frames):



CHAPTER

TWO

REPRESENTING A CORPUS

In Lhotse, we represent the data using YAML (more readable) or JSON (faster) manifests. For most audio corpora, we will need two types of manifests to fully describe them: a recording manifest and a supervision manifest.

Caution: We show all the examples in YAML format for improved readability. However, when processing medium/large datasets, we recommend to use JSON, which is much quicker to load and save.

2.1 Recording manifest

The recording manifest describes the recordings in a given corpus. It only contains information about the recording itself - this manifest does not specify any segmentation information or supervision such as the transcript or the speaker. It means that when a recording is a 1 hour long file, it is a single item in this manifest.

When coming from Kaldi, think of it as *wav.scp* on steroids, that also contains *reco2dur*, *reco2num_samples* and some extra information.

This is a YAML manifest for a corpus with two recordings:

```
---
- id: 'recording-1'
  sampling_rate: 8000
  num_samples: 4000
  duration: 0.5
  sources:
    - type: file
      channels: [0]
      source: 'test/fixtures/mono_c0.wav'
    - type: file
      channels: [1]
      source: 'test/fixtures/mono_c1.wav'
- id: 'recording-2'
  sampling_rate: 8000
  num_samples: 8000
  duration: 1.0
  sources:
    - type: file
      channels: [0, 1]
      source: 'test/fixtures/stereo.wav'
```

Each recording is described by:

- a unique id,

- its sampling rate,
- the number of samples,
- the duration in seconds,
- a list of audio sources.

Audio source is a useful abstraction for cases when the user has an audio format not supported by the library, or wants to use shell tools such as SoX to perform some additional preprocessing. An audio source has the following properties:

- type: either *file* or *command*
- channel_ids: a list of integer identifiers for each channel in the recording
- source: in case of a *file*, it's a path; in case of a *command*, its a shell command that will be expected to write a WAVE file to stdout.

2.1.1 Python

In Python, the recording manifest is represented by classes `RecordingSet`, `Recording`, and `AudioSource`. Example usage:

```
recordings = RecordingSet.from_yaml('audio.yml')
for recording in recordings:
    # Note: all time units in Lhotse are seconds
    if recording.duration >= 7.5:
        samples = recording.load_audio(
            channels=0,
            offset=2.5,
            duration=5.0
        )
        # Further sample processing
```

2.2 Supervision manifest

The supervision manifest contains the supervision information that we have about the recordings. In particular, it involves the segmentation - there might be a single segment for a single utterance recording, and multiple segments for a recording of a converstion.

When coming from Kaldi, think of it as a *segments* file on steroids, that also contains *utt2spk*, *utt2gender*, *utt2dur*, etc.

This is a YAML supervision manifest:

```
---
- id: 'segment-1'
  recording_id: 'recording-2'
  channel: 0
  start: 0.1
  duration: 0.3
  text: 'transcript of the first segment'
  language: 'english'
  speaker: 'Norman Dyhrentfurth'

- id: 'segment-2'
  recording_id: 'recording-2'
```

(continues on next page)

(continued from previous page)

```
  start: 0.5
  duration: 0.4
```

Each segment is characterized by the following attributes:

- a unique id,
- a corresponding recording id,
- start time in seconds, relative to the beginning of the recording,
- the duration in seconds

Each segment may be assigned optional supervision information. In this example, the first segment contains the transcription text, the language of the utterance and a speaker name. The second segment contains only the minimal amount of information, which should be interpreted as: "this is some area of interest in the recording that we know nothing else about."

2.2.1 Python

In Python, the supervision manifest is represented by classes `SupervisionSet` and `SupervisionSegment`. Example usage:

```
supervisions = SupervisionSet.from_segments([
    SupervisionSegment(
        id='segment-1',
        recording_id='recording-1',
        start=0.5,
        duration=10.7,
        text='quite a long utterance'
    )
])
print(f'There is {len(supervisions)} supervision in the set.')
```

2.3 Standard data preparation recipes

We provide a number of standard data preparation recipes. By that, we mean a collection of a Python function + a CLI tool that create the manifests given a corpus directory.

Currently supported corpora:

- AMI `lhotse.recipes.prepare_ami()`
- English Broadcast News 1997 `lhotse.recipes.prepare_broadcast_news()`
- Full or Mini LibriSpeech `lhotse.recipes.prepare_librispeech()`
- Heroico `lhotse.recipes.prepare_heroico()`
- LJ Speech `lhotse.recipes.prepare_ljspeech()`
- Mini LibriMix `lhotse.recipes.prepare_librimix()`
- Switchboard `lhotse.recipes.prepare_switchboard()`
- TED-LIUM v3 `lhotse.recipes.prepare_tedlium()`

2.4 Adding new corpora

General pointers:

- Each corpus has a dedicated Python file in `lhotse/recipes`.
- For publicly available corpora that can be freely downloaded, we usually define a function called `download`, `download_and_untar`, etc.
- Each data preparation recipe should expose a single function called `prepare_X`, with X being the name of the corpus, that produces dicts like: `{'recordings': <RecordingSet>, 'supervisions': <SupervisionSet>}` for the data in that corpus.
- When a corpus defines standard split (e.g. train/dev/test), we return a dict with the following structure: `{'train': {'recordings': <RecordingSet>, 'supervisions': <SupervisionSet>}, 'dev': ...}`
- Some corpora (like LibriSpeech) come with pre-segmented recordings. In these cases, the `SupervisionSegment` will exactly match the `Recording` duration (and there will likely be exactly one segment corresponding to any recording).
- Corpora with longer recordings (e.g. conversational, like Switchboard) should have exactly one `Recording` object corresponding to a single conversation/session, that spans its whole duration. Each speech segment in that recording should be represented as a `SupervisionSegment` with the same `recording_id` value.
- Corpora with multiple channels for each session (e.g. AMI) should have a single `Recording` with multiple `AudioSource` objects - each corresponding to a separate channel.

CHAPTER

THREE

CUTS

3.1 Overview

Audio cuts are one of the main Lhotse features. Cut is a part of a recording, but it can be longer than a supervision segment, or even span multiple segments. The regions without a supervision are just audio that we don't assume we know anything about - there may be silence, noise, non-transcribed speech, etc. Task-specific datasets can leverage this information to generate masks for such regions.

Currently, cuts are created after the feature extraction step (we might still change that). It means that every cut also represents the extracted features for the part of recording it represents.

Cuts can be modified using three basic operations: truncation, mixing and appending. These operations are not immediately performed on the audio or features. Instead, we create new `Cut` objects, possibly of different types, that represent a cut after modification. We only modify the actual audio and feature matrices once the user calls `load_features()` or `load_audio()`.

This design allows for quick on-the-fly data augmentation. In each training epoch, we may mix the cuts with different noises, SNRs, etc. We also do not need to re-compute and store the features for different mixes, as the mixing process consists of element-wise addition of the spectral energies (possibly with additional exp and log operations for log-energies). As of now, we only support this dynamic mix on log Mel energy (_fbank_) features. We anticipate to add support for other types of features as well.

The common attributes for all cut objects are the following:

- id
- duration
- supervisions
- num_frames
- num_features
- load_features()
- truncate()
- mix()
- append()
- from_dict()

3.2 Types of cuts

There are three cut classes:

- `Cut`, also referred to as "simple cut", can be traced back to a single particular recording (and channel).
- `PaddingCut` is an "artificial" recording used for padding other Cuts through mixing to achieve uniform duration.
- `MixedCut` is a collection of `Cut` and `PaddingCut` objects, together with mix parameters: offset and desired sound-to-noise ratio (SNR) for each track. Both the offset and the SNR are relative to the first cut in the mix.

Each of these types has additional attributes that are not common - e.g., it makes sense to specify *start* for `Cut` to locate it in the source recording, but it is undefined for `MixedCut` and `PaddingCut`.

3.3 Cut manifests

All cut types can be stored in the YAML manifests. An example manifest with simple cuts might look like:

```
- duration: 10.0
  features:
    channels: 0
    duration: 16.04
    num_features: 23
    num_frames: 1604
    recording_id: recording-1
    start: 0.0
    storage_path: test/fixtures/libri/storage/dc2e0952-f2f8-423c-9b8c-f5481652ee1d.llc
    storage_type: lilcom
    type: fbank
  id: 849e13d8-61a2-4d09-a542-dac1aee1b544
  start: 0.0
  supervisions: []
  type: Cut
```

Notice that the cut type is specified in YAML. The supervisions list might be empty - some tasks do not need them, e.g. unsupervised training, source separation, or speech enhancement.

Mixed cuts look differently in the manifest:

```
- id: mixed-cut-id
  tracks:
    - cut:
        duration: 7.78
        features:
          channels: 0
          duration: 7.78
          type: fbank
          num_frames: 778
          num_features: 23
          recording_id: 7850-286674-0014
          start: 0.0
          storage_path: test/fixtures/mix_cut_test/feats/storage/9dc645db-cbe4-4529-
↪85e4-b6ed4f59c340.llc
          storage_type: lilcom
        id: 0c5fdf79-efe7-4d45-b612-3d90d9af8c4e
```

(continues on next page)

(continued from previous page)

```
        start: 0.0
        supervisions:
          - channel: 0
            duration: 7.78
            gender: f
            id: 7850-286674-0014
            language: null
            recording_id: 7850-286674-0014
            speaker: 7850-286674
            start: 0.0
            text: SURE ENOUGH THERE HE CAME THROUGH THE SHALLOW WATER HIS WET BACK␣
↪SHELL PARTLY
              OUT OF IT AND SHINING IN THE SUNLIGHT
      offset: 0.0
    - cut:
        duration: 9.705
        features:
          channels: 0
          duration: 9.705
          type: fbank
          num_frames: 970
          num_features: 23
          recording_id: 2412-153948-0014
          start: 0.0
          storage_path: test/fixtures/mix_cut_test/feats/storage/5078e7eb-57a6-4000-
↪b0f2-fa4bf9c52090.llc
          storage_type: lilcom
        id: 78bef88d-e62e-4cfa-9946-a1311442c6f7
        start: 0.0
        supervisions:
          - channel: 0
            duration: 9.705
            gender: f
            id: 2412-153948-0014
            language: null
            recording_id: 2412-153948-0014
            speaker: 2412-153948
            start: 0.0
            text: THERE WAS NO ONE IN THE WHOLE WORLD WHO HAD THE SMALLEST IDEA SAVE␣
↪THOSE
              WHO WERE THEMSELVES ON THE OTHER SIDE OF IT IF INDEED THERE WAS ANY ONE␣
↪AT ALL
              COULD I HOPE TO CROSS IT
      offset: 3.89
      snr: 20.0
  type: MixedCut
```

Mixed cuts literally consist of simple cuts, their feature descriptions, and their supervisions. These are combined together when a user queries `MixedCut` for supervisions, features, or duration. Note that the first simple cut is missing an SNR field - it is optional (i.e. *None*). That is because the semantics of 0 SNR are: re-scale one of the signals, so that the SNR between two signals is zero. We denote no re-scaling by not specifying the SNR at all.

The amount of text in these manifests can be considerable in larger datasets, but they are highly compressible. We support their automated (de-)compression with gzip - it's sufficient to add ".gz" at the end of filename when writing or reading, both in Python classes and the CLI tools.

3.4 Python

Some examples of how cuts can be manipulated to create a desired dataset for model training.

```
cuts = CutSet.from_yaml('cuts.yml')
# Reject too short segments
cuts = cuts.filter(lambda cut: cut.duration >= 3.0)
# Pad short segments with silence to 5 seconds.
cuts = cuts.pad(desired_duration=5.0)
# Truncate longer segments to 5 seconds.
cuts = cuts.truncate(max_duration=5.0, offset_type='random')
# Save cuts
cuts.to_yaml('cuts-5s.yml')
```

3.5 CLI

Analogous examples of how to perform the same operations in the terminal:

```
# Reject short segments
lhotse yaml filter duration>=3.0 cuts.yml cuts-3s.yml
# Pad short segments to 5 seconds.
lhotse cut pad --duration 5.0 cuts-3s.yml cuts-5s-pad.yml
# Truncate longer segments to 5 seconds.
lhotse cut truncate --max-duration 5.0 --offset-type random cuts-5s-pad.yml cuts-5s.
↪yml
```

CHAPTER

FOUR

FEATURE EXTRACTION

Feature extraction in Lhotse is currently based exclusively on the Torchaudio library. We support spectrograms, log-Mel energies (*fbank*) and MFCCs. *Fbank* are the default features. We also support custom defined feature extractors via a Python API (which won't be available in the CLI, unless there is a popular demand for that).

We are striving for a simple relation between the audio duration, the number of frames, and the frame shift. You only need to know two of those values to compute the third one, regardless of the frame length. This is equivalent of having Kaldi's `snip_edges` parameter set to False.

4.1 Storing features

Features in Lhotse are stored as numpy matrices with shape `(num_frames, num_features)`. By default, we use lilcom for lossy compression and reduce the size on the disk by about 3x. The lilcom compression method uses a fixed precision that doesn't depend on the magnitude of the thing being compressed, so it's better suited to log-energy features than energy features. We currently support two kinds of storage:

- HDF5 files with multiple feature matrices
- directory with feature matrix per file

We retrieve the arrays by loading the whole feature matrix from disk and selecting the relevant region (e.g. specified by a cut). Therefore it makes sense to cut the recordings first, and then extract the features for them to avoid loading unnecessary data from disk (especially for very long recordings).

There are two types of manifests:

- one describing the feature extractor;
- one describing the extracted feature matrices.

The feature extractor manifest is mapped to a Python configuration dataclass. An example for *spectrogram*:

```
dither: 0.0
energy_floor: 1e-10
frame_length: 0.025
frame_shift: 0.01
min_duration: 0.0
preemphasis_coefficient: 0.97
raw_energy: true
remove_dc_offset: true
round_to_power_of_two: true
window_type: povey
type: spectrogram
```

And the corresponding configuration class:

class lhotse.features.**SpectrogramConfig**(*dither: float = 0.0*, *window_type: str = 'povey'*, *frame_length: float = 0.025*, *frame_shift: float = 0.01*, *remove_dc_offset: bool = True*, *round_to_power_of_two: bool = True*, *energy_floor: float = 1e-10*, *min_duration: float = 0.0*, *preemphasis_coefficient: float = 0.97*, *raw_energy: bool = True*)

dither: float = 0.0

window_type: str = 'povey'

frame_length: float = 0.025

frame_shift: float = 0.01

remove_dc_offset: bool = True

round_to_power_of_two: bool = True

energy_floor: float = 1e-10

min_duration: float = 0.0

preemphasis_coefficient: float = 0.97

raw_energy: bool = True

__init__(*dither=0.0*, *window_type='povey'*, *frame_length=0.025*, *frame_shift=0.01*, *remove_dc_offset=True*, *round_to_power_of_two=True*, *energy_floor=1e-10*, *min_duration=0.0*, *preemphasis_coefficient=0.97*, *raw_energy=True*)

Initialize self. See help(type(self)) for accurate signature.

The feature matrices manifest is a list of documents. These documents contain the information necessary to tie the features to a particular recording: `start`, `duration`, `channel` and `recording_id`. They currently do not have their own IDs. They also provide some useful information, such as the type of features, number of frames and feature dimension. Finally, they specify how the feature matrix is stored with `storage_type` (currently `numpy` or `lilcom`), and where to find it with the `storage_path`. In the future there might be more storage types.

```
- channels: 0
  duration: 16.04
  num_features: 23
  num_frames: 1604
  recording_id: recording-1
  start: 0.0
  storage_path: test/fixtures/libri/storage/dc2e0952-f2f8-423c-9b8c-f5481652ee1d.llc
  storage_type: lilcom
  type: fbank
```

4.2 Creating custom feature extractor

There are two components needed to implement a custom feature extractor: a configuration and the extractor itself. We expect the configuration class to be a dataclass, so that it can be automatically mapped to dict and serialized. The feature extractor should inherit from `FeatureExtractor`, and implement a small number of methods/properties. The base class takes care of initialization (you need to pass a config object), serialization to YAML, etc. A minimal, complete example of adding a new feature extractor:

```
from scipy.signal import stft

@dataclass
class ExampleFeatureExtractorConfig:
    frame_len: Seconds = 0.025
    frame_shift: Seconds = 0.01


class ExampleFeatureExtractor(FeatureExtractor):
    """
    A minimal class example, showing how to implement a custom feature extractor in␣
↪Lhotse.
    """
    name = 'example-feature-extractor'
    config_type = ExampleFeatureExtractorConfig

    def extract(self, samples: np.ndarray, sampling_rate: int) -> np.ndarray:
        f, t, Zxx = stft(
            samples,
            sampling_rate,
            nperseg=round(self.config.frame_len * sampling_rate),
            noverlap=round(self.frame_shift * sampling_rate)
        )
        # Note: returning a magnitude of the STFT might interact badly with lilcom␣
↪compression,
        # as it performs quantization of the float values and works best with log-
↪scale quantities.
        # It's advised to turn lilcom compression off, or use log-scale, in such␣
↪cases.
        return np.abs(Zxx)

    @property
    def frame_shift(self) -> Seconds:
        return self.config.frame_shift

    def feature_dim(self, sampling_rate: int) -> int:
        return (sampling_rate * self.config.frame_len) / 2 + 1
```

The overridden members include:

- `name` for easy debuggability/automatic re-creation of an extractor;
- `config_type` which specifies the complementary configuration class type;
- `extract()` where the actual computation takes place;
- `frame_shift` property, which is key to know the relationship between the duration and the number of frames.
- `feature_dim()` method, which accepts the `sampling_rate` as its argument, as some types of features (e.g. spectrogram) will depend on that.

Additionally, there are two extra methods than when overridden, allow to perform dynamic feature-space mixing (see Cuts):

```
@staticmethod
def mix(features_a: np.ndarray, features_b: np.ndarray, gain_b: float) -> np.ndarray:
    raise ValueError(f'The feature extractor\'s "mix" operation is undefined.')

@staticmethod
```

(continues on next page)

(continued from previous page)

```
def compute_energy(features: np.ndarray) -> float:
    raise ValueError(f'The feature extractor\'s "compute_energy" is undefined.')
```

They are:

- `mix()` which specifies how to mix two feature matrices to obtain a new feature matrix representing the sum of signals;
- `compute_energy()` which specifies how to obtain a total energy of the feature matrix, which is needed to mix two signals with a specified SNR. E.g. for a power spectrogram, this could be the sum of every time-frequency bin. It is expected to never return a zero.

During the feature-domain mix with a specified signal-to-noise ratio (SNR), we assume that one of the signals is a reference signal - it is used to initialize the `FeatureMixer` class. We compute the energy of both signals and scale the non-reference signal, so that its energy satisfies the requested SNR. The scaling factor (gain) is computed using the following formula:

```

        reference_feats = self.tracks[0]
        num_frames_offset = compute_num_frames(duration=offset, frame_shift=self.
↪frame_shift)
        current_num_frames = reference_feats.shape[0]
        incoming_num_frames = feats.shape[0] + num_frames_offset
        mix_num_frames = max(current_num_frames, incoming_num_frames)

        feats_to_add = feats

```

Note that we interpret the energy and the SNR in a power quantity context (as opposed to root-power/field quantities).

4.3 Storage backend details

Lhotse can be extended with additional storage backends via two abstractions: `FeaturesWriter` and `FeaturesReader`. We currently implement the following writers (and their corresponding readers):

- `lhotse.features.io.LilcomFilesWriter`
- `lhotse.features.io.NumpyFilesWriter`
- `lhotse.features.io.LilcomHdf5Writer`
- `lhotse.features.io.NumpyHdf5Writer`

The `FeaturesWriter` and `FeaturesReader` API is as follows:

class lhotse.features.io.**FeaturesWriter**

`FeaturesWriter` defines the interface of how to store numpy arrays in a particular storage backend. This backend could either be:

- separate files on a local filesystem;
- a single file with multiple arrays;
- cloud storage;
- etc.

Each class inheriting from `FeaturesWriter` must define:

- **the `write()` method, which defines the storing operation** (accepts a `key` used to place the `value` array in the storage);
- **the `storage_path()` property, which is either a common directory for the files,** the name of the file storing multiple arrays, name of the cloud bucket, etc.
- **the `name()` property that is unique to this particular storage mechanism -** it is stored in the features manifests (metadata) and used to automatically deduce the backend when loading the features.

Each `FeaturesWriter` can also be used as a context manager, as some implementations might need to free a resource after the writing is finalized. By default nothing happens in the context manager functions, and this can be modified by the inheriting subclasses.

Example:

with MyWriter('some/path') as storage: extractor.extract_from_recording_and_store(recording, storage)

The features loading must be defined separately in a class inheriting from `FeaturesReader`.

abstract property name

Return type `str`

abstract property storage_path

Return type `str`

abstract write(*key*, *value*)

Return type `str`

class lhotse.features.io.**FeaturesReader**

`FeaturesReader` defines the interface of how to load numpy arrays from a particular storage backend. This backend could either be:

- separate files on a local filesystem;
- a single file with multiple arrays;
- cloud storage;
- etc.

Each class inheriting from `FeaturesReader` must define:

- **the `read()` method, which defines the loading operation** (accepts the `key` to locate the array in the storage and return it). The read method should support selecting only a subset of the feature matrix, with the bounds expressed as arguments `left_offset_frames` and `right_offset_frames`. It's up to the Reader implementation to load only the required part or trim it to that range only after loading. It is assumed that the time dimension is always the first one.
- **the `name()` property that is unique to this particular storage mechanism -** it is stored in the features manifests (metadata) and used to automatically deduce the backend when loading the features.

The features writing must be defined separately in a class inheriting from `FeaturesWriter`.

abstract property name

Return type `str`

abstract read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type `ndarray`

4.4 Python usage

The feature manifest is represented by a `FeatureSet` object. Feature extractors have a class that represents both the extract and its configuration, named `FeatureExtractor`. We provide a utility called `FeatureSetBuilder` that can process a `RecordingSet` in parallel, store the feature matrices on disk and generate a feature manifest.

For example:

```
from lhotse import RecordingSet, Fbank, LilcomFilesWriter

# Read a RecordingSet from disk
recording_set = RecordingSet.from_yaml('audio.yml')
# Create a log Mel energy filter bank feature extractor with default settings
feature_extractor = Fbank()
# Create a feature set builder that uses this extractor and stores the results in a␣
↪directory called 'features'
with LilcomFilesWriter('features') as storage:
    builder = FeatureSetBuilder(feature_extractor=feature_extractor, storage=storage)
    # Extract the features using 8 parallel processes, compress, and store them on in
↪'features/storage/' directory.
    # Then, return the feature manifest object, which is also compressed and
    # stored in 'features/feature_manifest.json.gz'
    feature_set = builder.process_and_store_recordings(
        recordings=recording_set,
        num_jobs=8
    )
```

4.5 CLI usage

An equivalent example using the terminal:

```
lhotse write-default-feature-config feat-config.yml
lhotse make-feats -j 8 --storage-type lilcom_files -f feat-config.yml audio.yml␣
↪features/
```

4.6 Kaldi compatibility caveats

We are relying on Torchaudio Kaldi compatibility module, so most of the spectrogram/fbank/mfcc parameters are the same as in Kaldi. However, we are not fully compatible - Kaldi computes energies from a signal scaled between -32,768 to 32,767, while Torchaudio scales the signal between -1.0 and 1.0. It results in Kaldi energies being significantly greater than in Lhotse. By default, we turn off dithering for deterministic feature extraction.

CHAPTER

FIVE

AUGMENTATION

We support time-domain data augmentation via WavAugment and torchaudio libraries. They both leverage libsox to provide about 50 different audio effects like reverb, speed perturbation, pitch, etc.

Since `WavAugment` depends on libsox, it is an optional depedency for Lhotse, which can be installed using `tools/install_wavaugment.sh` (for convenience, the script will also compile libsox from source - note that the `WavAugment` authors warn their library is untested on Mac).

Torchaudio also depends on libsox, but seems to provide it when installed via anaconda. This functionality is only available with PyTorch 1.7+ and torchaudio 0.7+.

Using Lhotse's Python API, you can compose an arbitrary effect chain. On the other hand, for the CLI we provide a small number of predefined effect chains, such as `pitch` (pitch shifting), `reverb` (reverberation), and `pitch_reverb_tdrop` (pitch shift + reverberation + time dropout of a 50ms chunk).

5.1 Python usage

Warning: When using WavAugment or torchaudio data augmentation together with a multiprocessing executor (i.e. `ProcessPoolExecutor`), it is necessary to start it using the "spawn" context. Otherwise the process may hang (or terminate) on some systems due to libsox internals not handling forking well. Use: `ProcessPoolExecutor(..., mp_context=multiprocessing.get_context("spawn"))`.

Lhotse's `FeatureExtractor` and `Cut` offer convenience functions for feature extraction with data augmentation performed before that. These functions expose an optional parameter called `augment_fn` that has a signature like:

```
def augment_fn(audio: Union[np.ndarray, torch.Tensor], sampling_rate: int) -> np.
↪ndarray: ...
```

For `torchaudio` we define a `SoxEffectTransform` class:

class lhotse.augmentation.**SoxEffectTransform**(*effects*)

Class-style wrapper for torchaudio SoX effect chains. It should be initialized with a config-like list of items that define SoX effect to be applied. It supports sampling randomized values for effect parameters through the `RandomValue` wrapper.

Example:

```
>>> audio = np.random.rand(16000)
>>> augment_fn = SoxEffectTransform(effects=[
>>>     ['reverb', 50, 50, RandomValue(0, 100)],
>>>     ['speed', RandomValue(0.9, 1.1)],
```

(continues on next page)

(continued from previous page)

```
>>>     ['rate', 16000],
>>> ])
>>> augmented = augment_fn(audio, 16000)
```

See SoX manual or `torchaudio.sox_effects.effect_names()` for the list of possible effects. The parameters and the meaning of the values are explained in SoX manual/help.

__init__(*effects*)
Initialize self. See help(type(self)) for accurate signature.

sample_effects()
Resolve a list of effects, replacing random distributions with samples from them. It converts every number to string to match the expectations of torchaudio.

Return type `List[List[str]]`

We define a `WavAugmenter` class that is a thin wrapper over `WavAugment`. It can either be created with a predefined, or a user-supplied effect chain.

class lhotse.augmentation.**WavAugmenter**(*effect_chain*)
A wrapper class for WavAugment's effect chain. You should construct the `augment.EffectChain` beforehand and pass it on to this class.

This class is only available when WavAugment is installed, as it is an optional dependency for Lhotse. It can be installed using the script in "<main-repo-directory>/tools/install_wavaugment.sh"

For more details on how to augment, see https://github.com/facebookresearch/WavAugment

__init__(*effect_chain*)
Initialize self. See help(type(self)) for accurate signature.

static create_predefined(*name*, *sampling_rate*, ***kwargs*)
Create a WavAugmenter class with one of the predefined augmentation setups available in Lhotse. Some examples are: "pitch", "reverb", "pitch_reverb_tdrop".

Parameters

- **name** (`str`) – the name of the augmentation setup.
- **sampling_rate** (`int`) – expected sampling rate of the input audio.

Return type `WavAugmenter`

5.2 CLI usage

To extract features from augmented audio, you can pass an extra `--augmentation` argument to `lhotse feat extract`.

```
lhotse feat extract -a pitch ...
```

You can create a dataset with both clean and augmented features by combining different variants of extracted features, e.g.:

```
lhotse feat extract audio.yml clean_feats/
lhotse feat extract -a pitch audio.yml pitch_feats/
lhotse feat extract -a reverb audio.yml reverb_feats/
lhotse yaml combine {clean,pitch,reverb}_feats/feature_manifest.yml.gz combined_feats.
↪yml
```

CHAPTER

SIX

PYTORCH DATASETS

Caution: Lhotse datasets are still very much in the works and are subject to breaking changes.

We supply subclasses of the `torch.data.Dataset` for various audio/speech tasks. These datasets are created from `CutSet` objects and load the features from disk into memory on-the-fly. Each dataset accepts an optional `root_dir` argument which is used as a prefix for the paths to features and audio.

Currently, we provide the following:

class lhotse.dataset.diarization.**DiarizationDataset**(*cuts*, *min_speaker_dim=None*, *global_speaker_ids=False*)

A PyTorch Dataset for the speaker diarization task. Our assumptions about speaker diarization are the following:

- **we assume a single channel input (for now), which could be either a true mono signal** or a beamforming result from a microphone array.
- **we assume that the supervision used for model training is a speech activity matrix, with one** row dedicated to each speaker (either in the current cut or the whole dataset, depending on the settings). The columns correspond to feature frames. Each row is effectively a Voice Activity Detection supervision for a single speaker. This setup is somewhat inspired by the TS-VAD paper: https://arxiv.org/abs/2005.07272

Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor
    'speaker_activity': (num_speaker x T) tensor
}
```

Constructor arguments:

Parameters

- **cuts** (*CutSet*) – a `CutSet` used to create the dataset object.
- **min_speaker_dim** (`Optional[int]`) – optional int, when specified it will enforce that the matrix shape is at least that value (useful for datasets like CHiME 6 where the number of speakers is always 4, but some cuts might have less speakers than that).
- **global_speaker_ids** (`bool`) – a bool, indicates whether the same speaker should always retain the same row index in the speaker activity matrix (useful for speaker-dependent systems)
- **root_dir** – a prefix path to be attached to the feature files paths.

__init__(*cuts*, *min_speaker_dim=None*, *global_speaker_ids=False*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.unsupervised.**UnsupervisedDataset**(*cuts*)
Dataset that contains no supervision - it only provides the features extracted from recordings. The returned features are a `torch.Tensor` of shape `(T x F)`, where T is the number of frames, and F is the feature dimension.

__init__(*cuts*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.unsupervised.**UnsupervisedWaveformDataset**(*cuts*)
A variant of UnsupervisedDataset that provides waveform samples instead of features. The output is a tensor of shape (C, T), with C being the number of channels and T the number of audio samples. In this implemenation, there will always be a single channel.

class lhotse.dataset.unsupervised.**DynamicUnsupervisedDataset**(*feature_extractor*, *cuts*, *augment_fn=None*)
An example dataset that shows how to use on-the-fly feature extraction in Lhotse. It accepts two additional inputs - a FeatureExtractor and an optional WavAugmenter for time-domain data augmentation.. The output is approximately the same as that of the `UnsupervisedDataset` - there might be slight differences for `MixedCut``s, because this dataset mixes them in the time domain, and ``UnsupervisedDataset` does that in the feature domain. Cuts that are not mixed will yield identical results in both dataset classes.

__init__(*feature_extractor*, *cuts*, *augment_fn=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.speech_recognition.**SpeechRecognitionDataset**(*cuts*)
The PyTorch Dataset for the speech recognition task. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor,
    'text': string,
    'supervisions_mask': (T) tensor
}
```

The `supervisions_mask` field is a mask that specifies which frames are covered by a supervision by assigning a value of 1 (in this case: segments with transcribed speech contents), and which are not by asigning a value of 0 (in this case: padding, contextual noise, or in general the acoustic context without transcription).

In the future, will be extended by graph supervisions.

__init__(*cuts*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.speech_recognition.**K2SpeechRecognitionIterableDataset**(*cuts*, *max_frames=26000*, *max_cuts=None*, *shuffle=False*, *concat_cuts=True*, *concat_cuts_gap=1.0*, *concat_cuts_duration_fact*
The PyTorch Dataset for the speech recognition task using K2 library.

This dataset internally batches and collates the Cuts and should be used with PyTorch DataLoader with argument batch_size=None to work properly. The batch size is determined automatically to satisfy the constraints of `max_frames` and `max_cuts`.

This dataset will automatically partition itself when used with a multiprocessing DataLoader (i.e. the same cut will not appear twice in the same epoch).

By default, we "pack" the batches to minimize the amount of padding - we achieve that by concatenating the cuts' feature matrices with a small amount of silence (padding) in between.

Each item in this dataset is a dict of:

```
{
    'features': float tensor of shape (B, T, F)
    'supervisions': [
        {
            'cut_id': List[str] of len S
            'sequence_idx': Tensor[int] of shape (S,)
            'text': List[str] of len S
            'start_frame': Tensor[int] of shape (S,)
            'num_frames': Tensor[int] of shape (S,)
        }
    ]
}
```

Dimension symbols legend: * `B` - batch size (number of Cuts) * `S` - number of supervision segments (greater or equal to B, as each Cut may have multiple supervisions) * `T` - number of frames of the longest Cut * `F` - number of features

The 'sequence_idx' field is the index of the Cut used to create the example in the Dataset.

__init__(*cuts*, *max_frames=26000*, *max_cuts=None*, *shuffle=False*, *concat_cuts=True*, *concat_cuts_gap=1.0*, *concat_cuts_duration_factor=2*)

K2 ASR IterableDataset constructor.

Parameters

- **cuts** (*CutSet*) – the `CutSet` to sample data from.
- **max_frames** (`int`) – The maximum number of feature frames that we're going to put in a single batch. The padding frames do not contribute to that limit, since we pack the batch by default to minimze the amount of padding.
- **max_cuts** (`Optional[int]`) – The maximum number of cuts sampled to form a mini-batch. By default, this constraint is off.
- **shuffle** (`bool`) – When `True`, the cuts will be shuffled at the start of iteration. Convenient when mini-batch loop is inside an outer epoch-level loop, e.g.: *for epoch in range(10): for batch in dataset: ...* as every epoch will see a different cuts order.
- **concat_cuts** (`bool`) – When `True`, we will concatenate the cuts to minimize the total amount of padding; e.g. instead of creating a batch with 40 examples, we will merge some of the examples together adding some silence between them to avoid a large number of padding frames that waste the computation. Enabled by default.
- **concat_cuts_gap** (`float`) – The duration of silence in seconds that is inserted between the cuts; it's goal is to let the model "know" that there are separate utterances in a single example.
- **concat_cuts_duration_factor** (`float`) – Determines the maximum duration of the concatenated cuts; by default it's twice the duration of the longest cut in the batch.

lhotse.dataset.speech_recognition.**concat_cuts**(*cuts*, *gap=1.0*, *max_duration=None*)

We're going to concatenate the cuts to minimize the amount of total padding frames used. This is actually solving a knapsack problem. In this initial implementation we're using a greedy approach: going from the back (i.e. the shortest cuts) we'll try to concat them to the longest cut that still has some "space" at the end.

Parameters

- **cuts** (List[Union[*Cut*, *MixedCut*, *PaddingCut*]]) – a list of cuts to pack.
- **gap** (float) – the duration of silence inserted between concatenated cuts.
- **max_duration** (Optional[float]) – the maximum duration for the concatenated cuts (by default set to the duration of the first cut).

:return a list of packed cuts.

Return type List[Union[*Cut*, *MixedCut*, *PaddingCut*]]

class lhotse.dataset.speech_recognition.**K2SpeechRecognitionDataset**(*cuts*)

The PyTorch Dataset for the speech recognition task using K2 library. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor,
    'supervisions': List[Dict] -> [
        {
            'sequence_idx': int
            'text': string,
            'start_frame': int,
            'num_frames': int
        } (multiplied N times, for each of the N supervisions present in the Cut)
    ]
}
```

The 'sequence_idx' field is the index of the Cut used to create the example in the Dataset. It is mapped to the batch index later in the DataLoader.

__init__(*cuts*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.speech_recognition.**K2DataLoader**(**args*, ***kwds*)

A PyTorch DataLoader that has a custom collate_fn that complements the K2SpeechRecognitionDataset.

The 'features' tensor is collated in a standard way to return a tensor of shape (B, T, F).

The 'supervisions' dict contains the same fields as in K2SpeechRecognitionDataset, except that each sub-field (like 'start_frame') is a 1D PyTorch tensor with shape (B,). The 'text' sub-field is an exception - it's a list of strings with length equal to batch size.

The 'sequence_idx' sub-field in 'supervisions', which originally points to index of the example in the Dataset, is remapped to the index of the corresponding features matrix in the collated 'features'. Multiple supervisions coming from the same cut will share the same 'sequence_idx'.

For an example, see test/dataset/test_speech_recognition_dataset.py::test_k2_dataloader().

__init__(**args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

dataset

batch_size

num_workers

pin_memory

drop_last

timeout

sampler

prefetch_factor

lhotse.dataset.speech_recognition.**multi_supervision_collate_fn**(*batch*)
Custom collate_fn for K2SpeechRecognitionDataset.

It merges the items provided by K2SpeechRecognitionDataset into the following structure:

```
{
    'features': float tensor of shape (B, T, F)
    'supervisions': [
        {
            'sequence_idx': Tensor[int] of shape (S,)
            'text': List[str] of len S
            'start_frame': Tensor[int] of shape (S,)
            'num_frames': Tensor[int] of shape (S,)
        }
    ]
}
```

Dimension symbols legend: * `B` - batch size (number of Cuts), * `S` - number of supervision segments (greater or equal to B, as each Cut may have multiple supervisions), * `T` - number of frames of the longest Cut * `F` - number of features

Return type `Dict`

lhotse.dataset.**speech_synthesis**
alias of `lhotse.dataset.speech_synthesis`

class lhotse.dataset.source_separation.**DynamicallyMixedSourceSeparationDataset**(*sources_set*, *mixtures_set*, *nonsources_set=Non*

A PyTorch Dataset for the source separation task. It's created from a number of CutSets:

- `sources_set`: provides the audio cuts for the sources that (the targets of source separation),
- `mixtures_set`: provides the audio cuts for the signal mix (the input of source separation),
- `nonsources_set`: *(optional)* provides the audio cuts for other signals that are in the mix, but are not the targets of source separation. Useful for adding noise.

When queried for data samples, it returns a dict of:

```
{
    'sources': (N x T x F) tensor,
    'mixture': (T x F) tensor,
    'real_mask': (N x T x F) tensor,
    'binary_mask': (T x F) tensor
}
```

This Dataset performs on-the-fly feature-domain mixing of the sources. It expects the mixtures_set to contain MixedCuts, so that it knows which Cuts should be mixed together.

__init__(*sources_set*, *mixtures_set*, *nonsources_set=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.source_separation.**PreMixedSourceSeparationDataset**(*sources_set*, *mixtures_set*)
A PyTorch Dataset for the source separation task. It's created from two CutSets - one provides the audio cuts for the sources, and the other one the audio cuts for the signal mix. When queried for data samples, it returns a dict of:

```
{
    'sources': (N x T x F) tensor,
    'mixture': (T x F) tensor,
    'real_mask': (N x T x F) tensor,
    'binary_mask': (T x F) tensor
}
```

It expects both CutSets to return regular Cuts, meaning that the signals were mixed in the time domain. In contrast to DynamicallyMixedSourceSeparationDataset, no on-the-fly feature-domain-mixing is performed.

__init__(*sources_set*, *mixtures_set*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.vad.**VadDataset**(*cuts*)
The PyTorch Dataset for the voice activity detection task. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor
    'is_voice': (T x 1) tensor
}
```

__init__(*cuts*)
Initialize self. See help(type(self)) for accurate signature.

CHAPTER

SEVEN

KALDI INTEROPERABILITY

We support importing Kaldi data directories that contain at least the `wav.scp` file, required to create the `RecordingSet`. Other files, such as `segments`, `utt2spk`, etc. are used to create the `SupervisionSet`.

We currently do not support the following (but may start doing so in the future):

- Importing Kaldi's extracted features (`feats.scp` is ignored)
- Exporting Lhotse manifests as Kaldi data directories.

7.1 Python

Python methods related to Kaldi support:

`lhotse.kaldi.`**`load_kaldi_data_dir`**(*path*, *sampling_rate*)

Load a Kaldi data directory and convert it to a Lhotse RecordingSet and SupervisionSet manifests. For this to work, at least the wav.scp file must exist. SupervisionSet is created only when a segments file exists. All the other files (text, utt2spk, etc.) are optional, and some of them might not be handled yet. In particular, feats.scp files are ignored.

Return type `Tuple[`*RecordingSet*`, Optional[`*SupervisionSet*`]]`

`lhotse.kaldi.`**`load_kaldi_text_mapping`**(*path*, *must_exist=False*)

Load Kaldi files such as utt2spk, spk2gender, text, etc. as a dict.

Return type `Dict[str, Optional[str]]`

7.2 CLI

Converting Kaldi data directory called `data/train`, with 16kHz sampling rate recordings, to a directory with Lhotse manifests called `train_manifests`:

```
lhotse convert-kaldi data/train 16000 train_manifests
```

CHAPTER

EIGHT

COMMAND-LINE INTERFACE

8.1 lhotse obtain

Command group for download and extract data.

```
lhotse obtain [OPTIONS] COMMAND [ARGS]...
```

8.1.1 heroico

heroico download.

```
lhotse obtain heroico [OPTIONS] TARGET_DIR
```

Arguments

TARGET_DIR
Required argument

8.1.2 librimix

Mini LibriMix download.

```
lhotse obtain librimix [OPTIONS] TARGET_DIR
```

Arguments

TARGET_DIR
Required argument

8.1.3 mini-librispeech

Mini Librispeech download.

```
lhotse obtain mini-librispeech [OPTIONS] TARGET_DIR
```

Arguments

TARGET_DIR
Required argument

8.1.4 tedlium

TED-LIUM v3 download (approx. 11GB).

```
lhotse obtain tedlium [OPTIONS] TARGET_DIR
```

Arguments

TARGET_DIR
Required argument

8.2 lhotse prepare

Command group with data preparation recipes.

```
lhotse prepare [OPTIONS] COMMAND [ARGS]...
```

8.2.1 broadcast-news

English Broadcast News 1997 data preparation. It will output three manifests: for recordings, topic sections, and speech segments. It supports the following LDC distributions:

* 1997 English Broadcast News Train (HUB4)
Speech LDC98S71
Transcripts LDC98T28

This data is not available for free - your institution needs to have an LDC subscription.

```
lhotse prepare broadcast-news [OPTIONS] AUDIO_DIR TRANSCRIPT_DIR OUTPUT_DIR
```

Arguments

AUDIO_DIR
Required argument

TRANSCRIPT_DIR
Required argument

OUTPUT_DIR
Required argument

8.2.2 heroico

heroico Answers ASR data preparation.

```
lhotse prepare heroico [OPTIONS] SPEECH_DIR TRANSCRIPT_DIR OUTPUT_DIR
```

Arguments

SPEECH_DIR
Required argument

TRANSCRIPT_DIR
Required argument

OUTPUT_DIR
Required argument

8.2.3 librimix

LibrMix source separation data preparation.

```
lhotse prepare librimix [OPTIONS] LIBRIMIX_CSV OUTPUT_DIR
```

Options

--sampling-rate <sampling_rate>
Sampling rate to set in the RecordingSet manifest.

--min-segment-seconds <min_segment_seconds>
Remove segments shorter than MIN_SEGMENT_SECONDS.

--with-precomputed-mixtures, --no-precomputed-mixtures
Optionally create an RecordingSet manifest including the precomputed LibriMix mixtures.

Arguments

LIBRIMIX_CSV
Required argument

OUTPUT_DIR
Required argument

8.2.4 mini-librispeech

Mini Librispeech ASR data preparation.

```
lhotse prepare mini-librispeech [OPTIONS] CORPUS_DIR OUTPUT_DIR
```

Arguments

CORPUS_DIR
Required argument

OUTPUT_DIR
Required argument

8.2.5 switchboard

The Switchboard corpus preparation.

This is conversational telephone speech collected as 2-channel, 8kHz-sampled
data. We are using just the Switchboard-1 Phase 1 training data.
The catalog number LDC97S62 (Switchboard-1 Release 2) corresponds, we believe,
to what we have. We also use the Mississippi State transcriptions, which
we download separately from
http://www.isip.piconepress.com/projects/switchboard/releases/switchboard_word_alignments.tar.gz

This data is not available for free - your institution needs to have an LDC subscription.

```
lhotse prepare switchboard [OPTIONS] AUDIO_DIR OUTPUT_DIR
```

Options

--transcript-dir <transcript_dir>

--sentiment-dir <sentiment_dir>
Optional path to LDC2020T14 package with sentiment annotations for SWBD.

--omit-silence, --retain-silence
Should the [silence] segments be kept.

Arguments

AUDIO_DIR
Required argument

OUTPUT_DIR
Required argument

8.2.6 tedlium

TED-LIUM v3 recording and supervision manifest preparation.

```
lhotse prepare tedlium [OPTIONS] TEDLIUM_DIR OUTPUT_DIR
```

Arguments

TEDLIUM_DIR
Required argument

OUTPUT_DIR
Required argument

8.3 lhotse cut

Group of commands used to create CutSets.

```
lhotse cut [OPTIONS] COMMAND [ARGS]...
```

8.3.1 append

Create a new CutSet by appending the cuts in CUT_MANIFESTS. CUT_MANIFESTS are iterated position-wise (the cuts on i'th position in each manfiest are appended to each other). The cuts are appended in the order in which they appear in the input argument list. If CUT_MANIFESTS have different lengths, the script stops once the shortest CutSet is depleted.

```
lhotse cut append [OPTIONS] [CUT_MANIFESTS]... OUTPUT_CUT_MANIFEST
```

Arguments

CUT_MANIFESTS
Optional argument(s)

OUTPUT_CUT_MANIFEST
Required argument

8.3.2 mix-by-recording-id

Create a CutSet stored in OUTPUT_CUT_MANIFEST by matching the Cuts from CUT_MANIFESTS by their recording IDs and mixing them together.

```
lhotse cut mix-by-recording-id [OPTIONS] [CUT_MANIFESTS]...
                               OUTPUT_CUT_MANIFEST
```

Arguments

CUT_MANIFESTS
Optional argument(s)

OUTPUT_CUT_MANIFEST
Required argument

8.3.3 mix-sequential

Create a CutSet stored in OUTPUT_CUT_MANIFEST by iterating jointly over CUT_MANIFESTS and mixing the Cuts on the same positions. E.g. the first output cut is created from the first cuts in each input manifest. The mix is performed by summing the features from all Cuts. If the CUT_MANIFESTS have different number of Cuts, the mixing ends when the shorter manifest is depleted.

```
lhotse cut mix-sequential [OPTIONS] [CUT_MANIFESTS]... OUTPUT_CUT_MANIFEST
```

Arguments

CUT_MANIFESTS
Optional argument(s)

OUTPUT_CUT_MANIFEST
Required argument

8.3.4 pad

Create a new CutSet by padding the cuts in CUT_MANIFEST. The cuts will be right-padded, i.e. the padding is placed after the signal ends.

```
lhotse cut pad [OPTIONS] CUT_MANIFEST OUTPUT_CUT_MANIFEST
```

Options

-d, --duration <duration>
Desired duration of cuts after padding. Cuts longer than this won't be affected. By default, pad to the longest cut duration found in CUT_MANIFEST.

Arguments

CUT_MANIFEST
Required argument

OUTPUT_CUT_MANIFEST
Required argument

8.3.5 random-mixed

Create a CutSet stored in OUTPUT_CUT_MANIFEST that contains supervision regions from SUPERVISION_MANIFEST and features supplied by FEATURE_MANIFEST. It first creates a trivial CutSet, splits it into two equal, randomized parts and mixes their features. The parameters of the mix are controlled via SNR_RANGE and OFFSET_RANGE.

```
lhotse cut random-mixed [OPTIONS] SUPERVISION_MANIFEST FEATURE_MANIFEST
                        OUTPUT_CUT_MANIFEST
```

Options

-s, --snr-range <snr_range>
Range of SNR values (in dB) that will be uniformly sampled in order to mix the signals.

-o, --offset-range <offset_range>
Range of relative offset values (0 - 1), which will offset the "right" signal by this many times the duration of the "left" signal. It is uniformly sampled for each mix operation.

Arguments

SUPERVISION_MANIFEST
Required argument

FEATURE_MANIFEST
Required argument

OUTPUT_CUT_MANIFEST
Required argument

8.3.6 simple

Create a CutSet stored in OUTPUT_CUT_MANIFEST. Depending on the provided options, it may contain any combination of recording, feature and supervision manifests. Either RECORDING_MANIFEST or FEATURE_MANIFEST has to be provided. When SUPERVISION_MANIFEST is provided, the cuts time span will correspond to that of the supervision segments. Otherwise, that time span corresponds to the one found in features, if available, otherwise recordings.

```
lhotse cut simple [OPTIONS] OUTPUT_CUT_MANIFEST
```

Options

-r, --recording-manifest <recording_manifest>
Optional recording manifest - will be used to attach the recordings to the cuts.

-f, --feature-manifest <feature_manifest>
Optional feature manifest - will be used to attach the features to the cuts.

-s, --supervision_manifest <supervision_manifest>
Optional supervision manifest - will be used to attach the supervisions to the cuts.

Arguments

OUTPUT_CUT_MANIFEST
Required argument

8.3.7 truncate

Truncate the cuts in the CUT_MANIFEST and write them to OUTPUT_CUT_MANIFEST. Cuts shorter than MAX_DURATION will not be modified.

```
lhotse cut truncate [OPTIONS] CUT_MANIFEST OUTPUT_CUT_MANIFEST
```

Options

--preserve-id
Should the cuts preserve IDs (by default, they will get new, random IDs)

-d, --max-duration <max_duration>
The maximum duration in seconds of a cut in the resulting manifest. [required]

-o, --offset-type <offset_type>
Where should the truncated cut start: "start" - at the start of the original cut, "end" - MAX_DURATION before the end of the original cut, "random" - randomly choose somewhere between "start" and "end" options.

Options start|end|random

--keep-overflowing-supervisions, --discard-overflowing-supervisions
When a cut is truncated in the middle of a supervision segment, should the supervision be kept.

Arguments

CUT_MANIFEST
Required argument

OUTPUT_CUT_MANIFEST
Required argument

8.3.8 windowed

Create a CutSet stored in OUTPUT_CUT_MANIFEST from feature regions in FEATURE_MANIFEST. The feature matrices are traversed in windows with CUT_SHIFT increments, creating cuts of constant CUT_DURATION.

```
lhotse cut windowed [OPTIONS] FEATURE_MANIFEST OUTPUT_CUT_MANIFEST
```

Options

-d, --cut-duration <cut_duration>
How long should the cuts be in seconds.

-s, --cut-shift <cut_shift>
How much to shift the cutting window in seconds (by default the shift is equal to CUT_DURATION).

--keep-shorter-windows, --discard-shorter-windows
When true, the last window will be used to create a Cut even if its duration is shorter than CUT_DURATION.

Arguments

FEATURE_MANIFEST
Required argument

OUTPUT_CUT_MANIFEST
Required argument

8.4 lhotse manifest

Generic commands working with all or most manifest types.

```
lhotse manifest [OPTIONS] COMMAND [ARGS]...
```

8.4.1 combine

Load MANIFESTS, combine them into a single one, and write it to OUTPUT_MANIFEST.

```
lhotse manifest combine [OPTIONS] [MANIFESTS]... OUTPUT_MANIFEST
```

Arguments

MANIFESTS
Optional argument(s)

OUTPUT_MANIFEST
Required argument

8.4.2 filter

Filter a MANIFEST according to the rule specified in PREDICATE, and save the result to OUTPUT_MANIFEST. It is intended to work generically with most manifest types - it supports RecordingSet, SupervisionSet and CutSet.

The PREDICATE specifies which attribute is used for item selection. Some examples:
lhotse manifest filter 'duration>4.5' supervision.json output.json
lhotse manifest filter 'num_frames<600' cuts.json output.json
lhotse manifest filter 'start=0' cuts.json output.json
lhotse manifest filter 'channel!=0' audio.json output.json

It currently only supports comparison of numerical manifest item attributes, such as: start, duration, end, channel, num_frames, num_features, etc.

```
lhotse manifest filter [OPTIONS] PREDICATE MANIFEST OUTPUT_MANIFEST
```

Arguments

PREDICATE
Required argument

MANIFEST
Required argument

OUTPUT_MANIFEST
Required argument

8.4.3 split

Load MANIFEST, split it into NUM_SPLITS equal parts and save as separate manifests in OUTPUT_DIR.

```
lhotse manifest split [OPTIONS] NUM_SPLITS MANIFEST OUTPUT_DIR
```

Options

--randomize
Optionally randomize the sequence before splitting.

Arguments

NUM_SPLITS
Required argument

MANIFEST
Required argument

OUTPUT_DIR
Required argument

8.5 lhotse feat

Feature extraction related commands.

```
lhotse feat [OPTIONS] COMMAND [ARGS]...
```

8.5.1 extract

Extract features for recordings in a given AUDIO_MANIFEST. The features are stored in OUTPUT_DIR, with one file per recording (or segment).

```
lhotse feat extract [OPTIONS] RECORDING_MANIFEST OUTPUT_DIR
```

Options

-a, --augmentation <augmentation>
Optional time-domain data augmentation effect chain to apply.

Options pitch|speed|reverb|pitch_reverb_tdrop

-f, --feature-manifest <feature_manifest>
Optional manifest specifying feature extractor configuration.

--storage-type <storage_type>
Select a storage backend for the feature matrices.

Options lilcom_files|lilcom_hdf5|numpy_files|numpy_hdf5

-t, --lilcom-tick-power <lilcom_tick_power>
Determines the compression accuracy; the input will be compressed to integer multiples of 2^tick_power

-r, --root-dir <root_dir>
Root directory - all paths in the manifest will use this as prefix.

-j, --num-jobs <num_jobs>
Number of parallel processes.

Arguments

RECORDING_MANIFEST
Required argument

OUTPUT_DIR
Required argument

8.5.2 write-default-config

Save a default feature extraction config to OUTPUT_CONFIG.

```
lhotse feat write-default-config [OPTIONS] OUTPUT_CONFIG
```

Options

-f, --feature-type <feature_type>
Which feature extractor type to use.

Options fbank|mfcc|spectrogram

Arguments

OUTPUT_CONFIG
Required argument

8.6 lhotse convert-kaldi

Convert a Kaldi data dir DATA_DIR into a directory MANIFEST_DIR of lhotse manifests. Ignores feats.scp. The SAMPLING_RATE has to be explicitly specified as it is not available to read from DATA_DIR.

```
lhotse convert-kaldi [OPTIONS] DATA_DIR SAMPLING_RATE MANIFEST_DIR
```

Arguments

DATA_DIR
Required argument

SAMPLING_RATE
Required argument

MANIFEST_DIR
Required argument

CHAPTER

NINE

API REFERENCE

This page contains a comprehensive list of all classes and functions within *lhotse*.

9.1 Datasets

PyTorch Dataset wrappers for common tasks.

9.1.1 Speech Recognition

class lhotse.dataset.speech_recognition.**SpeechRecognitionDataset**(*cuts*)

The PyTorch Dataset for the speech recognition task. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor,
    'text': string,
    'supervisions_mask': (T) tensor
}
```

The `supervisions_mask` field is a mask that specifies which frames are covered by a supervision by assigning a value of 1 (in this case: segments with transcribed speech contents), and which are not by asigning a value of 0 (in this case: padding, contextual noise, or in general the acoustic context without transcription).

In the future, will be extended by graph supervisions.

__init__(*cuts*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.speech_recognition.**K2SpeechRecognitionIterableDataset**(*cuts*, *max_frames=26000*, *max_cuts=None*, *shuffle=False*, *concat_cuts=True*, *concat_cuts_gap=1.0*, *concat_cuts_duration_fact*

The PyTorch Dataset for the speech recognition task using K2 library.

This dataset internally batches and collates the Cuts and should be used with PyTorch DataLoader with argument batch_size=None to work properly. The batch size is determined automatically to satisfy the constraints of `max_frames` and `max_cuts`.

This dataset will automatically partition itself when used with a multiprocessing DataLoader (i.e. the same cut will not appear twice in the same epoch).

By default, we "pack" the batches to minimize the amount of padding - we achieve that by concatenating the cuts' feature matrices with a small amount of silence (padding) in between.

Each item in this dataset is a dict of:

```
{
    'features': float tensor of shape (B, T, F)
    'supervisions': [
        {
            'cut_id': List[str] of len S
            'sequence_idx': Tensor[int] of shape (S,)
            'text': List[str] of len S
            'start_frame': Tensor[int] of shape (S,)
            'num_frames': Tensor[int] of shape (S,)
        }
    ]
}
```

Dimension symbols legend: * `B` - batch size (number of Cuts) * `S` - number of supervision segments (greater or equal to B, as each Cut may have multiple supervisions) * `T` - number of frames of the longest Cut * `F` - number of features

The 'sequence_idx' field is the index of the Cut used to create the example in the Dataset.

__init__(*cuts*, *max_frames=26000*, *max_cuts=None*, *shuffle=False*, *concat_cuts=True*, *concat_cuts_gap=1.0*, *concat_cuts_duration_factor=2*)

K2 ASR IterableDataset constructor.

Parameters

- **cuts** (*CutSet*) – the `CutSet` to sample data from.
- **max_frames** (`int`) – The maximum number of feature frames that we're going to put in a single batch. The padding frames do not contribute to that limit, since we pack the batch by default to minimze the amount of padding.
- **max_cuts** (`Optional[int]`) – The maximum number of cuts sampled to form a mini-batch. By default, this constraint is off.
- **shuffle** (`bool`) – When `True`, the cuts will be shuffled at the start of iteration. Convenient when mini-batch loop is inside an outer epoch-level loop, e.g.: *for epoch in range(10): for batch in dataset: ...* as every epoch will see a different cuts order.
- **concat_cuts** (`bool`) – When `True`, we will concatenate the cuts to minimize the total amount of padding; e.g. instead of creating a batch with 40 examples, we will merge some of the examples together adding some silence between them to avoid a large number of padding frames that waste the computation. Enabled by default.
- **concat_cuts_gap** (`float`) – The duration of silence in seconds that is inserted between the cuts; it's goal is to let the model "know" that there are separate utterances in a single example.
- **concat_cuts_duration_factor** (`float`) – Determines the maximum duration of the concatenated cuts; by default it's twice the duration of the longest cut in the batch.

lhotse.dataset.speech_recognition.**concat_cuts**(*cuts*, *gap=1.0*, *max_duration=None*)

We're going to concatenate the cuts to minimize the amount of total padding frames used. This is actually solving a knapsack problem. In this initial implementation we're using a greedy approach: going from the back (i.e. the shortest cuts) we'll try to concat them to the longest cut that still has some "space" at the end.

Parameters

- **cuts** (List[Union[ForwardRef, ForwardRef, ForwardRef]]) – a list of cuts to pack.
- **gap** (float) – the duration of silence inserted between concatenated cuts.
- **max_duration** (Optional[float]) – the maximum duration for the concatenated cuts (by default set to the duration of the first cut).

:return a list of packed cuts.

Return type List[Union[ForwardRef, ForwardRef, ForwardRef]]

class lhotse.dataset.speech_recognition.**K2SpeechRecognitionDataset**(*cuts*)

The PyTorch Dataset for the speech recognition task using K2 library. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor,
    'supervisions': List[Dict] -> [
        {
            'sequence_idx': int
            'text': string,
            'start_frame': int,
            'num_frames': int
        } (multiplied N times, for each of the N supervisions present in the Cut)
    ]
}
```

The 'sequence_idx' field is the index of the Cut used to create the example in the Dataset. It is mapped to the batch index later in the DataLoader.

__init__(*cuts*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.speech_recognition.**K2DataLoader**(**args*, ***kwds*)

A PyTorch DataLoader that has a custom collate_fn that complements the K2SpeechRecognitionDataset.

The 'features' tensor is collated in a standard way to return a tensor of shape (B, T, F).

The 'supervisions' dict contains the same fields as in K2SpeechRecognitionDataset, except that each sub-field (like 'start_frame') is a 1D PyTorch tensor with shape (B,). The 'text' sub-field is an exception - it's a list of strings with length equal to batch size.

The 'sequence_idx' sub-field in 'supervisions', which originally points to index of the example in the Dataset, is remapped to the index of the corresponding features matrix in the collated 'features'. Multiple supervisions coming from the same cut will share the same 'sequence_idx'.

For an example, see test/dataset/test_speech_recognition_dataset.py::test_k2_dataloader().

__init__(**args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

dataset

batch_size

num_workers

pin_memory

drop_last

timeout

sampler

prefetch_factor

lhotse.dataset.speech_recognition.**multi_supervision_collate_fn**(*batch*)

Custom collate_fn for K2SpeechRecognitionDataset.

It merges the items provided by K2SpeechRecognitionDataset into the following structure:

```
{
    'features': float tensor of shape (B, T, F)
    'supervisions': [
        {
            'sequence_idx': Tensor[int] of shape (S,)
            'text': List[str] of len S
            'start_frame': Tensor[int] of shape (S,)
            'num_frames': Tensor[int] of shape (S,)
        }
    ]
}
```

Dimension symbols legend: * `B` - batch size (number of Cuts), * `S` - number of supervision segments (greater or equal to B, as each Cut may have multiple supervisions), * `T` - number of frames of the longest Cut * `F` - number of features

Return type `Dict`

9.1.2 Source Separation

class lhotse.dataset.source_separation.**SourceSeparationDataset**(*sources_set*, *mixtures_set*)

An abstract base class, implementing PyTorch Dataset for the source separation task. It's created from two CutSets - one provides the audio cuts for the sources, and the other one the audio cuts for the signal mix. When queried for data samples, it returns a dict of:

```
{
    'sources': (N x T x F) tensor,
    'mixture': (T x F) tensor,
    'real_mask': (N x T x F) tensor,
    'binary_mask': (T x F) tensor
}
```

__init__(*sources_set*, *mixtures_set*)

Initialize self. See help(type(self)) for accurate signature.

validate()

class lhotse.dataset.source_separation.**DynamicallyMixedSourceSeparationDataset**(*sources_set*, *mixtures_set*, *non-sources_set=Non*

A PyTorch Dataset for the source separation task. It's created from a number of CutSets:

- `sources_set`: provides the audio cuts for the sources that (the targets of source separation),
- `mixtures_set`: provides the audio cuts for the signal mix (the input of source separation),
- `nonsources_set`: *(optional)* provides the audio cuts for other signals that are in the mix, but are not the targets of source separation. Useful for adding noise.

When queried for data samples, it returns a dict of:

```
{
    'sources': (N x T x F) tensor,
    'mixture': (T x F) tensor,
    'real_mask': (N x T x F) tensor,
    'binary_mask': (T x F) tensor
}
```

This Dataset performs on-the-fly feature-domain mixing of the sources. It expects the mixtures_set to contain MixedCuts, so that it knows which Cuts should be mixed together.

__init__(*sources_set*, *mixtures_set*, *nonsources_set=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.source_separation.**PreMixedSourceSeparationDataset**(*sources_set*, *mixtures_set*)
A PyTorch Dataset for the source separation task. It's created from two CutSets - one provides the audio cuts for the sources, and the other one the audio cuts for the signal mix. When queried for data samples, it returns a dict of:

```
{
    'sources': (N x T x F) tensor,
    'mixture': (T x F) tensor,
    'real_mask': (N x T x F) tensor,
    'binary_mask': (T x F) tensor
}
```

It expects both CutSets to return regular Cuts, meaning that the signals were mixed in the time domain. In contrast to DynamicallyMixedSourceSeparationDataset, no on-the-fly feature-domain-mixing is performed.

__init__(*sources_set*, *mixtures_set*)
Initialize self. See help(type(self)) for accurate signature.

9.1.3 Unsupervised

class lhotse.dataset.unsupervised.**UnsupervisedDataset**(*cuts*)
Dataset that contains no supervision - it only provides the features extracted from recordings. The returned features are a `torch.Tensor` of shape `(T x F)`, where T is the number of frames, and F is the feature dimension.

__init__(*cuts*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.dataset.unsupervised.**UnsupervisedWaveformDataset**(*cuts*)
A variant of UnsupervisedDataset that provides waveform samples instead of features. The output is a tensor of shape (C, T), with C being the number of channels and T the number of audio samples. In this implemenation, there will always be a single channel.

class lhotse.dataset.unsupervised.**DynamicUnsupervisedDataset**(*feature_extractor*, *cuts*, *augment_fn=None*)

An example dataset that shows how to use on-the-fly feature extraction in Lhotse. It accepts two additional inputs - a FeatureExtractor and an optional WavAugmenter for time-domain data augmentation.. The output is approximately the same as that of the `UnsupervisedDataset` - there might be slight differences for `MixedCut``s, because this dataset mixes them in the time domain, and ``UnsupervisedDataset` does that in the feature domain. Cuts that are not mixed will yield identical results in both dataset classes.

__init__(*feature_extractor*, *cuts*, *augment_fn=None*)

Initialize self. See help(type(self)) for accurate signature.

9.1.4 Voice Activity Detection

class lhotse.dataset.vad.**VadDataset**(*cuts*)

The PyTorch Dataset for the voice activity detection task. Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor
    'is_voice': (T x 1) tensor
}
```

__init__(*cuts*)

Initialize self. See help(type(self)) for accurate signature.

9.1.5 Diarization (experimental)

class lhotse.dataset.diarization.**DiarizationDataset**(*cuts*, *min_speaker_dim=None*, *global_speaker_ids=False*)

A PyTorch Dataset for the speaker diarization task. Our assumptions about speaker diarization are the following:

- **we assume a single channel input (for now), which could be either a true mono signal** or a beamforming result from a microphone array.
- **we assume that the supervision used for model training is a speech activity matrix, with one** row dedicated to each speaker (either in the current cut or the whole dataset, depending on the settings). The columns correspond to feature frames. Each row is effectively a Voice Activity Detection supervision for a single speaker. This setup is somewhat inspired by the TS-VAD paper: https://arxiv.org/abs/2005.07272

Each item in this dataset is a dict of:

```
{
    'features': (T x F) tensor
    'speaker_activity': (num_speaker x T) tensor
}
```

Constructor arguments:

Parameters

- **cuts** (*CutSet*) – a `CutSet` used to create the dataset object.
- **min_speaker_dim** (`Optional[int]`) – optional int, when specified it will enforce that the matrix shape is at least that value (useful for datasets like CHiME 6 where the number of speakers is always 4, but some cuts might have less speakers than that).

- **global_speaker_ids** (bool) – a bool, indicates whether the same speaker should always retain the same row index in the speaker activity matrix (useful for speaker-dependent systems)
- **root_dir** – a prefix path to be attached to the feature files paths.

__init__(*cuts*, *min_speaker_dim=None*, *global_speaker_ids=False*)
Initialize self. See help(type(self)) for accurate signature.

9.2 Recording manifests

Data structures used for describing audio recordings in a dataset.

class lhotse.audio.**AudioSource**(*type: str*, *channels: List[int]*, *source: str*)
AudioSource represents audio data that can be retrieved from somewhere. Supported sources of audio are currently: - 'file' (formats supported by librosa, possibly multi-channel) - 'command' [unix pipe] (must be WAVE, possibly multi-channel)

type: str

channels: List[int]

source: str

load_audio(*offset_seconds=0.0*, *duration_seconds=None*)
Load the AudioSource (both files and commands) with librosa, accounting for many audio formats and multi-channel inputs. Returns numpy array with shapes: (n_samples) for single-channel, (n_channels, n_samples) for multi-channel.

Return type ndarray

with_path_prefix(*path*)

Return type *AudioSource*

static from_dict(*data*)

Return type *AudioSource*

__init__(*type*, *channels*, *source*)
Initialize self. See help(type(self)) for accurate signature.

lhotse.audio.**read_audio**(*path*, *offset*, *duration*)

Return type Tuple[ndarray, int]

class lhotse.audio.**Recording**(*id: str*, *sources: List[*lhotse.audio.AudioSource*]*, *sampling_rate: int*, *num_samples: int*, *duration: float*)
Recording represents an AudioSource along with some metadata.

id: str

sources: List[AudioSource]

sampling_rate: int

num_samples: int

duration: Seconds

static from_sphere(*sph_path*, *relative_path_depth=None*)
Read a SPHERE file's header and create the corresponding Recording.

Parameters

- **sph_path** (Union[Path, str]) – Path to the sphere (.sph) file.
- **relative_path_depth** (Optional[int]) – optional int specifying how many last parts of the file path should be retained in the AudioSource. By default writes the path as is.

Return type *Recording*

Returns a new Recording instance pointing to the sphere file.

property num_channels

property channel_ids

load_audio(*channels=None*, *offset_seconds=0.0*, *duration_seconds=None*)

Return type ndarray

with_path_prefix(*path*)

Return type *Recording*

static from_dict(*data*)

Return type *Recording*

__init__(*id*, *sources*, *sampling_rate*, *num_samples*, *duration*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.audio.**RecordingSet**(**args*, ***kwds*)
RecordingSet represents a dataset of recordings. It does not contain any annotation - just the information needed to retrieve a recording (possibly multi-channel, from files or from shell commands and pipes) and some metadata for each of them.

It also supports (de)serialization to/from YAML and takes care of mapping between rich Python classes and YAML primitives during conversion.

recordings: Dict[str, Recording]

static from_recordings(*recordings*)

Return type *RecordingSet*

static from_dicts(*data*)

Return type *RecordingSet*

to_dicts()

Return type List[dict]

filter(*predicate*)
Return a new RecordingSet with the Recordings that satisfy the *predicate*.

Parameters **predicate** (Callable[[*Recording*], bool]) – a function that takes a recording as an argument and returns bool.

Return type *RecordingSet*

Returns a filtered RecordingSet.

split(*num_splits*, *randomize=False*)
Split the RecordingSet into num_splits pieces of equal size.

Parameters

- **num_splits** (int) – Requested number of splits.

- **randomize** (bool) – Optionally randomize the recordings order first.

Return type List[*RecordingSet*]

Returns A list of RecordingSet pieces.

load_audio(*recording_id*, *channels=None*, *offset_seconds=0.0*, *duration_seconds=None*)

Return type ndarray

with_path_prefix(*path*)

Return type *RecordingSet*

num_channels(*recording_id*)

Return type int

sampling_rate(*recording_id*)

Return type int

num_samples(*recording_id*)

Return type int

duration(*recording_id*)

Return type float

__init__(*recordings*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.audio.**AudioMixer**(*base_audio*, *sampling_rate*)
Utility class to mix multiple waveforms into a single one. It should be instantiated separately for each mixing session (i.e. each MixedCut will create a separate AudioMixer to mix its tracks). It is initialized with a numpy array of audio samples (typically float32 in [-1, 1] range) that represents the "reference" signal for the mix. Other signals can be mixed to it with different time offsets and SNRs using the add_to_mix method. The time offset is relative to the start of the reference signal (only positive values are supported). The SNR is relative to the energy of the signal used to initialize the AudioMixer.

__init__(*base_audio*, *sampling_rate*)

Parameters

- **base_audio** (ndarray) – A numpy array with the audio samples for the base signal (all the other signals will be mixed to it).
- **sampling_rate** (int) – Sampling rate of the audio.

property unmixed_audio
Return a numpy ndarray with the shape (num_tracks, num_samples), where each track is zero padded and scaled adequately to the offsets and SNR used in add_to_mix call.

Return type ndarray

property mixed_audio
Return a numpy ndarray with the shape (1, num_samples) - a mono mix of the tracks supplied with add_to_mix calls.

Return type ndarray

add_to_mix(*audio*, *snr=None*, *offset=0.0*)
Add audio (only support mono-channel) of a new track into the mix. :type audio: ndarray :param audio: An array of audio samples to be mixed in. :type snr: Optional[float] :param snr: Signal-to-noise ratio, assuming *audio* represents noise (positive SNR - lower *audio* energy, negative SNR - higher *audio*

energy) :type offset: `float` :param offset: How many seconds to shift *audio* in time. For mixing, the signal will be padded before the start with low energy values. :return:

`lhotse.audio.`**`audio_energy`**(*audio*)

Return type `float`

9.3 Supervision manifests

Data structures used for describing supervisions in a dataset.

class `lhotse.supervision.`**SupervisionSegment**(*id: str*, *recording_id: str*, *start: float*, *duration: float*, *channel: int = 0*, *text: Union[str, NoneType] = None*, *language: Union[str, NoneType] = None*, *speaker: Union[str, NoneType] = None*, *gender: Union[str, NoneType] = None*, *custom: Union[Dict[str, Any], NoneType] = None*)

id: str

recording_id: str

start: Seconds

duration: Seconds

channel: int = 0

text: Optional[str] = None

language: Optional[str] = None

speaker: Optional[str] = None

gender: Optional[str] = None

custom: Optional[Dict[str, Any]] = None

property end

Return type `float`

with_offset(*offset*)
Return an identical `SupervisionSegment`, but with the `offset` added to the `start` field.

Return type *SupervisionSegment*

trim(*end*)
Return an identical `SupervisionSegment`, but ensure that `self.start` is not negative (in which case it's set to 0) and `self.end` does not exceed the `end` parameter.

This method is useful for ensuring that the supervision does not exceed a cut's bounds, in which case pass `cut.duration` as the `end` argument, since supervision times are relative to the cut.

Return type *SupervisionSegment*

map(*transform_fn*)
Return a copy of the current segment, transformed with `transform_fn`.

Parameters **transform_fn** (`Callable[[`*SupervisionSegment*`], `*SupervisionSegment*`]`) – a function that takes a segment as input, transforms it and returns a new segment.

Return type *SupervisionSegment*

Returns a modified `SupervisionSegment`.

transform_text(*transform_fn*)
Return a copy of the current segment with transformed `text` field. Useful for text normalization, phonetic transcription, etc.

Parameters **transform_fn** (`Callable[[str], str]`) – a function that accepts a string and returns a string.

Return type *SupervisionSegment*

Returns a `SupervisionSegment` with adjusted text.

static from_dict(*data*)

Return type *SupervisionSegment*

__init__(*id*, *recording_id*, *start*, *duration*, *channel=0*, *text=None*, *language=None*, *speaker=None*, *gender=None*, *custom=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.supervision.**SupervisionSet**(**args*, ***kwds*)
SupervisionSet represents a collection of segments containing some supervision information. The only required fields are the ID of the segment, ID of the corresponding recording, and the start and duration of the segment in seconds. All other fields, such as text, language or speaker, are deliberately optional to support a wide range of tasks, as well as adding more supervision types in the future, while retaining backwards compatibility.

segments: Dict[str, SupervisionSegment]

static from_segments(*segments*)

Return type *SupervisionSet*

static from_dicts(*data*)

Return type *SupervisionSet*

to_dicts()

Return type `List[dict]`

split(*num_splits*, *randomize=False*)
Split the `SupervisionSet` into `num_splits` pieces of equal size.

Parameters

- **num_splits** (`int`) – Requested number of splits.
- **randomize** (`bool`) – Optionally randomize the supervisions order first.

Return type `List[SupervisionSet]`

Returns A list of `SupervisionSet` pieces.

filter(*predicate*)
Return a new SupervisionSet with the SupervisionSegments that satisfy the *predicate*.

Parameters **predicate** (`Callable[[SupervisionSegment], bool]`) – a function that takes a supervision as an argument and returns bool.

Return type *SupervisionSet*

Returns a filtered SupervisionSet.

map(*transform_fn*)
Map a `transform_fn` to the SupervisionSegments and return a new `SupervisionSet`.

Parameters **transform_fn** (Callable[[*SupervisionSegment*], *SupervisionSegment*]) – a function that modifies a supervision as an argument.

Return type *SupervisionSet*

Returns a new SupervisionSet with modified segments.

transform_text(*transform_fn*)
Return a copy of the current SupervisionSet with the segments having a transformed text field. Useful for text normalization, phonetic transcription, etc.

Parameters **transform_fn** (Callable[[str], str]) – a function that accepts a string and returns a string.

Return type *SupervisionSet*

Returns a SupervisionSet with adjusted text.

find(*recording_id*, *channel=None*, *start_after=0*, *end_before=None*, *adjust_offset=False*)
Return an iterable of segments that match the provided recording_id.

Parameters

- **recording_id** (str) – Desired recording ID.
- **channel** (Optional[int]) – When specified, return supervisions in that channel - otherwise, in all channels.
- **start_after** (float) – When specified, return segments that start after the given value.
- **end_before** (Optional[float]) – When specified, return segments that end before the given value.
- **adjust_offset** (bool) – When true, return segments as if the recordings had started at start_after. This is useful for creating Cuts. Fom a user perspective, when dealing with a Cut, it is no longer helpful to know when the supervisions starts in a recording - instead, it's useful to know when the supervision starts relative to the start of the Cut. In the anticipated use-case, start_after and end_before would be the beginning and end of a cut; this option converts the times to be relative to the start of the cut.

Return type Iterable[*SupervisionSegment*]

Returns An iterator over supervision segments satisfying all criteria.

__init__(*segments*, *_segments_by_recording_id=None*)
Initialize self. See help(type(self)) for accurate signature.

9.4 Feature extraction and manifests

Data structures and tools used for feature extraction and description.

9.4.1 Features API - extractor and manifests

class lhotse.features.base.**FeatureExtractor**(*config=None*)

The base class for all feature extractors in Lhotse. It is initialized with a config object, specific to a particular feature extraction method. The config is expected to be a dataclass so that it can be easily serialized.

All derived feature extractors must implement at least the following:

- a `name` class attribute (how are these features called, e.g. 'mfcc')
- a `config_type` class attribute that points to the configuration dataclass type
- the `extract` method,
- the `frame_shift` property.

Feature extractors that support feature-domain mixing should additionally specify two static methods:

- `compute_energy`, and
- `mix`.

By itself, the `FeatureExtractor` offers the following high-level methods that are not intended for overriding:

- `extract_from_samples_and_store`
- `extract_from_recording_and_store`

These methods run a larger feature extraction pipeline that involves data augmentation and disk storage.

name = None

config_type = None

__init__(*config=None*)

Initialize self. See help(type(self)) for accurate signature.

abstract extract(*samples*, *sampling_rate*)

Defines how to extract features using a numpy ndarray of audio samples and the sampling rate.

Return type `ndarray`

Returns a numpy ndarray representing the feature matrix.

abstract property frame_shift

Return type `float`

abstract feature_dim(*sampling_rate*)

Return type `int`

static mix(*features_a*, *features_b*, *energy_scaling_factor_b*)

Perform feature-domain mix of two singals, `a` and `b`, and return the mixed signal.

Parameters

- **features_a** (`ndarray`) – Left-hand side (reference) signal.
- **features_b** (`ndarray`) – Right-hand side (mixed-in) signal.
- **energy_scaling_factor_b** (`float`) – A scaling factor for `features_b` energy. It is used to achieve a specific SNR. E.g. to mix with an SNR of 10dB when both `features_a` and `features_b` energies are 100, the `features_b` signal energy needs to be scaled by 0.1. Since different features (e.g. spectrogram, fbank, MFCC) require different combination of transformations (e.g. exp, log, sqrt, pow) to allow mixing

of two signals, the exact place where to apply `energy_scaling_factor_b` to the signal is determined by the implementer.

Return type `ndarray`

Returns A mixed feature matrix.

static compute_energy(*features*)

Compute the total energy of a feature matrix. How the energy is computed depends on a particular type of features. It is expected that when implemented, `compute_energy` will never return zero.

Parameters **features** (`ndarray`) – A feature matrix.

Return type `float`

Returns A positive float value of the signal energy.

extract_from_samples_and_store(*samples*, *storage*, *sampling_rate*, *offset=0*, *augment_fn=None*)

Extract the features from an array of audio samples in a full pipeline:

- optional audio augmentation;
- extract the features;
- save them to disk in a specified directory;
- return a `Features` object with a description of the extracted features.

Note, unlike in `extract_from_recording_and_store`, the returned `Features` object might not be suitable to store in a `FeatureSet`, as it does not reference any particular `Recording`. Instead, this method is useful when extracting features from cuts - especially `MixedCut` instances, which may be created from multiple recordings and channels.

Parameters

- **samples** (`ndarray`) – a numpy ndarray with the audio samples.
- **sampling_rate** (`int`) – integer sampling rate of `samples`.
- **storage** (*FeaturesWriter*) – a `FeaturesWriter` object that will handle storing the feature matrices.
- **offset** (`float`) – an offset in seconds for where to start reading the recording - when used for `Cut` feature extraction, must be equal to `Cut.start`.
- **augment_fn** (`Optional[Callable[[ndarray, int], ndarray]]`) – an optional `WavAugmenter` instance to modify the waveform before feature extraction.

Returns a `Features` manifest item for the extracted feature matrix (it is not written to disk).

extract_from_recording_and_store(*recording*, *storage*, *offset=0*, *duration=None*, *channels=None*, *augment_fn=None*)

Extract the features from a `Recording` in a full pipeline:

- load audio from disk;
- optionally, perform audio augmentation;
- extract the features;
- save them to disk in a specified directory;
- return a `Features` object with a description of the extracted features and the source data used.

Parameters

- **recording** (*Recording*) – a `Recording` that specifies what's the input audio.
- **storage** (*FeaturesWriter*) – a `FeaturesWriter` object that will handle storing the feature matrices.
- **offset** (`float`) – an optional offset in seconds for where to start reading the recording.
- **duration** (`Optional[float]`) – an optional duration specifying how much audio to load from the recording.
- **channels** (`Union[int, List[int], None]`) – an optional int or list of ints, specifying the channels; by default, all channels will be used.
- **augment_fn** (`Optional[Callable[[ndarray, int], ndarray]]`) – an optional `WavAugmenter` instance to modify the waveform before feature extraction.

Returns a `Features` manifest item for the extracted feature matrix.

classmethod from_dict(*data*)

Return type *FeatureExtractor*

classmethod from_yaml(*path*)

Return type *FeatureExtractor*

to_yaml(*path*)

lhotse.features.base.**get_extractor_type**(*name*)

Return the feature extractor type corresponding to the given name.

Parameters **name** (`str`) – specifies which feature extractor should be used.

Return type `Type`

Returns A feature extractors type.

lhotse.features.base.**create_default_feature_extractor**(*name*)

Create a feature extractor object with a default configuration.

Parameters **name** (`str`) – specifies which feature extractor should be used.

Return type `Optional[`*FeatureExtractor*`]`

Returns A new feature extractor instance.

lhotse.features.base.**register_extractor**(*cls*)

This decorator is used to register feature extractor classes in Lhotse so they can be easily created just by knowing their name.

An example of usage:

@register_extractor class MyFeatureExtractor: ...

Parameters **cls** – A type (class) that is being registered.

Returns Registered type.

class lhotse.features.base.**TorchaudioFeatureExtractor**(*config=None*)

Common abstract base class for all torchaudio based feature extractors.

feature_fn = None

extract(*samples*, *sampling_rate*)

Defines how to extract features using a numpy ndarray of audio samples and the sampling rate.

Return type `ndarray`

Returns a numpy ndarray representing the feature matrix.

property frame_shift

Return type float

class lhotse.features.base.**Features**(*type: str*, *num_frames: int*, *num_features: int*, *sampling_rate: int*, *start: float*, *duration: float*, *storage_type: str*, *storage_path: str*, *storage_key: str*, *recording_id: Optional[str] = None*, *channels: Optional[Union[int, List[int]]] = None*)

Represents features extracted for some particular time range in a given recording and channel. It contains metadata about how it's stored: storage_type describes "how to read it", for now it supports numpy arrays serialized with np.save, as well as arrays compressed with lilcom; storage_path is the path to the file on the local filesystem.

type: str

num_frames: int

num_features: int

sampling_rate: int

start: Seconds

duration: Seconds

storage_type: str

storage_path: str

storage_key: str

recording_id: Optional[str] = None

channels: Optional[Union[int, List[int]]] = None

property end

Return type float

property frame_shift

Return type float

load(*start=None*, *duration=None*)

Return type ndarray

with_path_prefix(*path*)

Return type *Features*

static from_dict(*data*)

Return type *Features*

__init__(*type*, *num_frames*, *num_features*, *sampling_rate*, *start*, *duration*, *storage_type*, *storage_path*, *storage_key*, *recording_id=None*, *channels=None*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.features.base.**FeatureSet**(**args*, ***kwds*)

Represents a feature manifest, and allows to read features for given recordings within particular channels and time ranges. It also keeps information about the feature extractor parameters used to obtain this set. When a given recording/time-range/channel is unavailable, raises a KeyError.

features: List[Features]

static from_features(*features*)

Return type *FeatureSet*

static from_dicts(*data*)

Return type *FeatureSet*

to_dicts()

Return type List[dict]

with_path_prefix(*path*)

Return type *FeatureSet*

split(*num_splits*, *randomize=False*)
Split the FeatureSet into num_splits pieces of equal size.

Parameters

- **num_splits** (int) – Requested number of splits.
- **randomize** (bool) – Optionally randomize the features order first.

Return type List[*FeatureSet*]

Returns A list of FeatureSet pieces.

find(*recording_id*, *channel_id=0*, *start=0.0*, *duration=None*, *leeway=0.05*)
Find and return a Features object that best satisfies the search criteria. Raise a KeyError when no such object is available.

Parameters

- **recording_id** (str) – str, requested recording ID.
- **channel_id** (int) – int, requested channel.
- **start** (float) – float, requested start time in seconds for the feature chunk.
- **duration** (Optional[float]) – optional float, requested duration in seconds for the feature chunk. By default, return everything from the start.
- **leeway** (float) – float, controls how strictly we have to match the requested start and duration criteria. It is necessary to keep a small positive value here (default 0.05s), as there might be differneces between the duration of recording/supervision segment, and the duration of features. The latter one is constrained to be a multiple of frame_shift, while the former can be arbitrary.

Return type *Features*

Returns a Features object satisfying the search criteria.

load(*recording_id*, *channel_id=0*, *start=0.0*, *duration=None*)
Find a Features object that best satisfies the search criteria and load the features as a numpy ndarray. Raise a KeyError when no such object is available.

Return type ndarray

__init__(*features=<factory>*, *_features_by_recording_id=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.features.base.**FeatureSetBuilder**(*feature_extractor*, *storage*, *augment_fn=None*)
An extended constructor for the FeatureSet. Think of it as a class wrapper for a feature extraction script. It

consumes an iterable of Recordings, extracts the features specified by the FeatureExtractor config, and saves stores them on the disk.

Eventually, we plan to extend it with the capability to extract only the features in specified regions of recordings and to perform some time-domain data augmentation.

__init__ (*feature_extractor*, *storage*, *augment_fn=None*)
Initialize self. See help(type(self)) for accurate signature.

process_and_store_recordings (*recordings*, *output_manifest=None*, *num_jobs=1*)

Return type *FeatureSet*

lhotse.features.base.**store_feature_array** (*feats*, *storage*)
Store `feats` array on disk, using `lilcom` compression by default.

Parameters

- **feats** (`ndarray`) – a numpy ndarray containing features.
- **storage** (*FeaturesWriter*) – a `FeaturesWriter` object to use for array storage.

Return type `str`

Returns a path to the file containing the stored array.

9.4.2 Torchaudio feature extractors

class lhotse.features.fbank.**FbankConfig** (*dither: float = 0.0*, *window_type: str = 'povey'*, *frame_length: float = 0.025*, *frame_shift: float = 0.01*, *remove_dc_offset: bool = True*, *round_to_power_of_two: bool = True*, *energy_floor: float = 1e-10*, *min_duration: float = 0.0*, *preemphasis_coefficient: float = 0.97*, *raw_energy: bool = True*, *low_freq: float = 20.0*, *high_freq: float = - 400.0*, *num_mel_bins: int = 40*, *use_energy: bool = False*, *vtln_low: float = 100.0*, *vtln_high: float = - 500.0*, *vtln_warp: float = 1.0*)

dither: float = 0.0

window_type: str = 'povey'

frame_length: float = 0.025

frame_shift: float = 0.01

remove_dc_offset: bool = True

round_to_power_of_two: bool = True

energy_floor: float = 1e-10

min_duration: float = 0.0

preemphasis_coefficient: float = 0.97

raw_energy: bool = True

low_freq: float = 20.0

high_freq: float = -400.0

num_mel_bins: int = 40

use_energy: bool = False

vtln_low: float = 100.0

vtln_high: float = -500.0

vtln_warp: float = 1.0

__init__(*dither=0.0*, *window_type='povey'*, *frame_length=0.025*, *frame_shift=0.01*, *remove_dc_offset=True*, *round_to_power_of_two=True*, *energy_floor=1e-10*, *min_duration=0.0*, *preemphasis_coefficient=0.97*, *raw_energy=True*, *low_freq=20.0*, *high_freq=- 400.0*, *num_mel_bins=40*, *use_energy=False*, *vtln_low=100.0*, *vtln_high=-500.0*, *vtln_warp=1.0*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.features.fbank.**Fbank**(*config=None*)

Log Mel energy filter bank feature extractor based on `torchaudio.compliance.kaldi.fbank` function.

name = 'fbank'

config_type

alias of *FbankConfig*

feature_dim(*sampling_rate*)

Return type int

static mix(*features_a*, *features_b*, *energy_scaling_factor_b*)

Perform feature-domain mix of two singals, `a` and `b`, and return the mixed signal.

Parameters

- **features_a** (ndarray) – Left-hand side (reference) signal.
- **features_b** (ndarray) – Right-hand side (mixed-in) signal.
- **energy_scaling_factor_b** (float) – A scaling factor for `features_b` energy. It is used to achieve a specific SNR. E.g. to mix with an SNR of 10dB when both `features_a` and `features_b` energies are 100, the `features_b` signal energy needs to be scaled by 0.1. Since different features (e.g. spectrogram, fbank, MFCC) require different combination of transformations (e.g. exp, log, sqrt, pow) to allow mixing of two signals, the exact place where to apply `energy_scaling_factor_b` to the signal is determined by the implementer.

Return type ndarray

Returns A mixed feature matrix.

static compute_energy(*features*)

Compute the total energy of a feature matrix. How the energy is computed depends on a particular type of features. It is expected that when implemented, `compute_energy` will never return zero.

Parameters **features** (ndarray) – A feature matrix.

Return type float

Returns A positive float value of the signal energy.

class lhotse.features.mfcc.**MfccConfig**(*dither: float = 0.0*, *window_type: str = 'povey'*, *frame_length: float = 0.025*, *frame_shift: float = 0.01*, *remove_dc_offset: bool = True*, *round_to_power_of_two: bool = True*, *energy_floor: float = 1e-10*, *min_duration: float = 0.0*, *preemphasis_coefficient: float = 0.97*, *raw_energy: bool = True*, *low_freq: float = 20.0*, *high_freq: float = 0.0*, *num_mel_bins: int = 23*, *use_energy: bool = False*, *vtln_low: float = 100.0*, *vtln_high: float = - 500.0*, *vtln_warp: float = 1.0*, *cepstral_lifter: float = 22.0*, *num_ceps: int = 13*)

dither: float = 0.0

window_type: str = 'povey'

frame_length: float = 0.025

frame_shift: float = 0.01

remove_dc_offset: bool = True

round_to_power_of_two: bool = True

energy_floor: float = 1e-10

min_duration: float = 0.0

preemphasis_coefficient: float = 0.97

raw_energy: bool = True

low_freq: float = 20.0

high_freq: float = 0.0

num_mel_bins: int = 23

use_energy: bool = False

vtln_low: float = 100.0

vtln_high: float = -500.0

vtln_warp: float = 1.0

cepstral_lifter: float = 22.0

num_ceps: int = 13

__init__(*dither=0.0*, *window_type='povey'*, *frame_length=0.025*, *frame_shift=0.01*, *remove_dc_offset=True*, *round_to_power_of_two=True*, *energy_floor=1e-10*, *min_duration=0.0*, *preemphasis_coefficient=0.97*, *raw_energy=True*, *low_freq=20.0*, *high_freq=0.0*, *num_mel_bins=23*, *use_energy=False*, *vtln_low=100.0*, *vtln_high=- 500.0*, *vtln_warp=1.0*, *cepstral_lifter=22.0*, *num_ceps=13*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.features.mfcc.**Mfcc**(*config=None*)

MFCC feature extractor based on `torchaudio.compliance.kaldi.mfcc` function.

name = 'mfcc'

config_type

alias of *MfccConfig*

feature_dim(*sampling_rate*)

Return type int

class lhotse.features.spectrogram.**SpectrogramConfig**(*dither: float = 0.0*, *window_type: str = 'povey'*, *frame_length: float = 0.025*, *frame_shift: float = 0.01*, *remove_dc_offset: bool = True*, *round_to_power_of_two: bool = True*, *energy_floor: float = 1e-10*, *min_duration: float = 0.0*, *preemphasis_coefficient: float = 0.97*, *raw_energy: bool = True*)

dither: float = 0.0

window_type: str = 'povey'

frame_length: float = 0.025

frame_shift: float = 0.01

remove_dc_offset: bool = True

round_to_power_of_two: bool = True

energy_floor: float = 1e-10

min_duration: float = 0.0

preemphasis_coefficient: float = 0.97

raw_energy: bool = True

__init__(*dither=0.0*, *window_type='povey'*, *frame_length=0.025*, *frame_shift=0.01*, *remove_dc_offset=True*, *round_to_power_of_two=True*, *energy_floor=1e-10*, *min_duration=0.0*, *preemphasis_coefficient=0.97*, *raw_energy=True*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.features.spectrogram.**Spectrogram**(*config=None*)
Log spectrogram feature extractor based on torchaudio.compliance.kaldi.spectrogram function.

name = 'spectrogram'

config_type
alias of *SpectrogramConfig*

feature_dim(*sampling_rate*)

Return type int

static mix(*features_a*, *features_b*, *energy_scaling_factor_b*)
Perform feature-domain mix of two singals, a and b, and return the mixed signal.

Parameters

- **features_a** (ndarray) – Left-hand side (reference) signal.
- **features_b** (ndarray) – Right-hand side (mixed-in) signal.
- **energy_scaling_factor_b** (float) – A scaling factor for features_b energy. It is used to achieve a specific SNR. E.g. to mix with an SNR of 10dB when both features_a and features_b energies are 100, the features_b signal energy needs to be scaled by 0.1. Since different features (e.g. spectrogram, fbank, MFCC) require different combination of transformations (e.g. exp, log, sqrt, pow) to allow mixing

of two signals, the exact place where to apply `energy_scaling_factor_b` to the signal is determined by the implementer.

Return type `ndarray`

Returns A mixed feature matrix.

static compute_energy(*features*)

Compute the total energy of a feature matrix. How the energy is computed depends on a particular type of features. It is expected that when implemented, `compute_energy` will never return zero.

Parameters **features** (`ndarray`) – A feature matrix.

Return type `float`

Returns A positive float value of the signal energy.

9.4.3 Feature storage

class lhotse.features.io.**FeaturesWriter**

`FeaturesWriter` defines the interface of how to store numpy arrays in a particular storage backend. This backend could either be:

- separate files on a local filesystem;
- a single file with multiple arrays;
- cloud storage;
- etc.

Each class inheriting from `FeaturesWriter` must define:

- **the `write()` method, which defines the storing operation** (accepts a `key` used to place the `value` array in the storage);
- **the `storage_path()` property, which is either a common directory for the files,** the name of the file storing multiple arrays, name of the cloud bucket, etc.
- **the `name()` property that is unique to this particular storage mechanism -** it is stored in the features manifests (metadata) and used to automatically deduce the backend when loading the features.

Each `FeaturesWriter` can also be used as a context manager, as some implementations might need to free a resource after the writing is finalized. By default nothing happens in the context manager functions, and this can be modified by the inheriting subclasses.

Example:

with MyWriter('some/path') as storage: extractor.extract_from_recording_and_store(recording, storage)

The features loading must be defined separately in a class inheriting from `FeaturesReader`.

abstract property name

Return type `str`

abstract property storage_path

Return type `str`

abstract write(*key*, *value*)

Return type `str`

class lhotse.features.io.**FeaturesReader**

FeaturesReader defines the interface of how to load numpy arrays from a particular storage backend. This backend could either be:

- separate files on a local filesystem;
- a single file with multiple arrays;
- cloud storage;
- etc.

Each class inheriting from FeaturesReader must define:

- **the read() method, which defines the loading operation** (accepts the key to locate the array in the storage and return it). The read method should support selecting only a subset of the feature matrix, with the bounds expressed as arguments left_offset_frames and right_offset_frames. It's up to the Reader implementation to load only the required part or trim it to that range only after loading. It is assumed that the time dimension is always the first one.
- **the name() property that is unique to this particular storage mechanism -** it is stored in the features manifests (metadata) and used to automatically deduce the backend when loading the features.

The features writing must be defined separately in a class inheriting from FeaturesWriter.

abstract property name

Return type str

abstract read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type ndarray

lhotse.features.io.**available_storage_backends**()

Return type List[str]

lhotse.features.io.**register_reader**(*cls*)

Decorator used to add a new FeaturesReader to Lhotse's registry.

Example:

@register_reader class MyFeatureReader(FeatureReader):

...

lhotse.features.io.**register_writer**(*cls*)

Decorator used to add a new FeaturesWriter to Lhotse's registry.

Example:

@register_writer class MyFeatureWriter(FeatureWriter):

...

lhotse.features.io.**get_reader**(*name*)

Find a FeaturesReader sub-class that corresponds to the provided name and return its type.

Example:

reader_type = get_reader("lilcom_files") reader = reader_type("/storage/features/")

Return type Type[*FeaturesReader*]

lhotse.features.io.**get_writer**(*name*)
Find a FeaturesWriter sub-class that corresponds to the provided name and return its type.

Example:

writer_type = get_writer("lilcom_files") writer = writer_type("/storage/features/")

Return type Type[*FeaturesWriter*]

class lhotse.features.io.**LilcomFilesReader**(*storage_path*, **args*, ***kwargs*)
Reads Lilcom-compressed files from a directory on the local filesystem. storage_path corresponds to the directory path; storage_key for each utterance is the name of the file in that directory.

name = 'lilcom_files'

__init__(*storage_path*, **args*, ***kwargs*)
Initialize self. See help(type(self)) for accurate signature.

read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type ndarray

class lhotse.features.io.**LilcomFilesWriter**(*storage_path*, *tick_power=- 5*, **args*, ***kwargs*)
Writes Lilcom-compressed files to a directory on the local filesystem. storage_path corresponds to the directory path; storage_key for each utterance is the name of the file in that directory.

name = 'lilcom_files'

__init__(*storage_path*, *tick_power=- 5*, **args*, ***kwargs*)
Initialize self. See help(type(self)) for accurate signature.

property storage_path

Return type str

write(*key*, *value*)

Return type str

class lhotse.features.io.**NumpyFilesReader**(*storage_path*, **args*, ***kwargs*)
Reads non-compressed numpy arrays from files in a directory on the local filesystem. storage_path corresponds to the directory path; storage_key for each utterance is the name of the file in that directory.

name = 'numpy_files'

__init__(*storage_path*, **args*, ***kwargs*)
Initialize self. See help(type(self)) for accurate signature.

read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type ndarray

class lhotse.features.io.**NumpyFilesWriter**(*storage_path*, **args*, ***kwargs*)
Writes non-compressed numpy arrays to files in a directory on the local filesystem. storage_path corresponds to the directory path; storage_key for each utterance is the name of the file in that directory.

name = 'numpy_files'

__init__(*storage_path*, **args*, ***kwargs*)
Initialize self. See help(type(self)) for accurate signature.

property storage_path

Return type str

write(*key*, *value*)

Return type str

lhotse.features.io.**lookup_cache_or_open**(*storage_path*)

Helper internal function used in HDF5 readers. It opens the HDF files and keeps their handles open in a global program cache to avoid excessive amount of syscalls when the *Reader class is instantiated and destroyed in a loop repeatedly (frequent use-case).

The file handles can be freed at any time by calling close_cached_file_handles().

lhotse.features.io.**close_cached_file_handles**()

Closes the cached file handles in lookup_cache_or_open (see its docs for more details).

Return type None

class lhotse.features.io.**NumpyHdf5Reader**(*storage_path*, **args*, ***kwargs*)

Reads non-compressed numpy arrays from a HDF5 file with a "flat" layout. Each array is stored as a separate HDF Dataset because their shapes (numbers of frames) may vary. storage_path corresponds to the HDF5 file path; storage_key for each utterance is the key corresponding to the array (i.e. HDF5 "Group" name).

name = 'numpy_hdf5'

__init__(*storage_path*, **args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type ndarray

class lhotse.features.io.**NumpyHdf5Writer**(*storage_path*, **args*, ***kwargs*)

Writes non-compressed numpy arrays to a HDF5 file with a "flat" layout. Each array is stored as a separate HDF Dataset because their shapes (numbers of frames) may vary. storage_path corresponds to the HDF5 file path; storage_key for each utterance is the key corresponding to the array (i.e. HDF5 "Group" name).

name = 'numpy_hdf5'

__init__(*storage_path*, **args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

property storage_path

Return type str

write(*key*, *value*)

Return type str

close()

Return type None

class lhotse.features.io.**LilcomHdf5Reader**(*storage_path*, **args*, ***kwargs*)

Reads lilcom-compressed numpy arrays from a HDF5 file with a "flat" layout. Each array is stored as a separate HDF Dataset because their shapes (numbers of frames) may vary. storage_path corresponds to the HDF5 file path; storage_key for each utterance is the key corresponding to the array (i.e. HDF5 "Group" name).

name = 'lilcom_hdf5'

__init__(*storage_path*, **args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

read(*key*, *left_offset_frames=0*, *right_offset_frames=None*)

Return type ndarray

class lhotse.features.io.**LilcomHdf5Writer**(*storage_path*, *tick_power=- 5*, **args*, ***kwargs*)

Writes lilcom-compressed numpy arrays to a HDF5 file with a "flat" layout. Each array is stored as a separate HDF Dataset because their shapes (numbers of frames) may vary. storage_path corresponds to the HDF5 file path; storage_key for each utterance is the key corresponding to the array (i.e. HDF5 "Group" name).

name = 'lilcom_hdf5'

__init__(*storage_path*, *tick_power=- 5*, **args*, ***kwargs*)

Initialize self. See help(type(self)) for accurate signature.

property storage_path

Return type str

write(*key*, *value*)

Return type str

close()

Return type None

9.4.4 Feature-domain mixing

class lhotse.features.mixer.**FeatureMixer**(*feature_extractor*, *base_feats*, *frame_shift*, *padding_value=- 1000.0*)

Utility class to mix multiple feature matrices into a single one. It should be instantiated separately for each mixing session (i.e. each MixedCut will create a separate FeatureMixer to mix its tracks). It is initialized with a numpy array of features (typically float32) that represents the "reference" signal for the mix. Other signals can be mixed to it with different time offsets and SNRs using the add_to_mix method. The time offset is relative to the start of the reference signal (only positive values are supported). The SNR is relative to the energy of the signal used to initialize the FeatureMixer.

It relies on the FeatureExtractor to have defined mix and compute_energy methods, so that the FeatureMixer knows how to scale and add two feature matrices together.

__init__(*feature_extractor*, *base_feats*, *frame_shift*, *padding_value=- 1000.0*)

Parameters

- **feature_extractor** (*FeatureExtractor*) – The FeatureExtractor instance that specifies how to mix the features.
- **base_feats** (ndarray) – The features used to initialize the FeatureMixer are a point of reference in terms of energy and offset for all features mixed into them.
- **frame_shift** (float) – Required to correctly compute offset and padding during the mix.
- **padding_value** (float) – The value used to pad the shorter features during the mix. This value is adequate only for log space features. For non-log space features, e.g. energies, use either 0 or a small positive value like 1e-5.

property num_features

property unmixed_feats

Return a numpy ndarray with the shape (num_tracks, num_frames, num_features), where each track's feature matrix is padded and scaled adequately to the offsets and SNR used in add_to_mix call.

Return type ndarray

property mixed_feats
Return a numpy ndarray with the shape (num_frames, num_features) - a mono mixed feature matrix of the tracks supplied with add_to_mix calls.

Return type ndarray

add_to_mix (*feats*, *snr=None*, *offset=0.0*)
Add feature matrix of a new track into the mix. :type feats: ndarray :param feats: A 2D feature matrix to be mixed in. :type snr: Optional[float] :param snr: Signal-to-noise ratio, assuming feats represents noise (positive SNR - lower feats energy, negative SNR - higher feats energy) :type offset: float :param offset: How many seconds to shift feats in time. For mixing, the signal will be padded before the start with low energy values.

9.5 Augmentation

9.6 Cuts

Data structures and tools used to create training/testing examples.

class lhotse.cut.**CutUtilsMixin**
A mixin class for cuts which contains all the methods that share common implementations.

Note: Ideally, this would've been an abstract base class specifying the common interface, but ABC's do not mix well with dataclasses in Python. It is possible we'll ditch the dataclass for cuts in the future and make this an ABC instead.

property trimmed_supervisions
Return the supervisions in this Cut that have modified time boundaries so as not to exceed the Cut's start or end.

Note that when cut.supervisions is called, the supervisions may have negative start values that indicate the supervision actually begins before the cut, or end values that exceed the Cut's duration (it means the supervision continued in the original recording after the Cut's ending).

Return type List[*SupervisionSegment*]

mix (*other*, *offset_other_by=0.0*, *snr=None*)
Refer to mix() documentation.

Return type *MixedCut*

append (*other*, *snr=None*)
Append the other Cut after the current Cut. Conceptually the same as mix but with an offset matching the current cuts length. Optionally scale down (positive SNR) or scale up (negative SNR) the other cut. Returns a MixedCut, which only keeps the information about the mix; actual mixing is performed during the call to load_features.

Return type *MixedCut*

compute_features (*extractor*, *augment_fn=None*)
Compute the features from this cut. This cut has to be able to load audio.

Parameters

- **extractor** (*FeatureExtractor*) – a FeatureExtractor instance used to compute the features.

- **augment_fn** (Optional[Callable[[ndarray, int], ndarray]]) – optional WavAugmenter instance for audio augmentation.

Return type ndarray

Returns a numpy ndarray with the computed features.

plot_audio()
Display a plot of the waveform. Requires matplotlib to be installed.

play_audio()
Display a Jupyter widget that allows to listen to the waveform. Works only in Jupyter notebook/lab or similar (e.g. Colab).

plot_features()
Display the feature matrix as an image. Requires matplotlib to be installed.

speakers_feature_mask(*min_speaker_dim=None*, *speaker_to_idx_map=None*)
Return a matrix of per-speaker activity in a cut. The matrix shape is (num_speakers, num_frames), and its values are 0 for nonspeech **frames** and 1 for speech **frames** for each respective speaker.

This is somewhat inspired by the TS-VAD setup: https://arxiv.org/abs/2005.07272

Parameters

- **min_speaker_dim** (Optional[int]) – optional int, when specified it will enforce that the matrix shape is at least that value (useful for datasets like CHiME 6 where the number of speakers is always 4, but some cuts might have less speakers than that).
- **speaker_to_idx_map** (Optional[Dict[str, int]]) – optional dict mapping speaker names (strings) to their global indices (ints). Useful when you want to preserve the order of the speakers (e.g. speaker XYZ is always mapped to index 2)

Return type ndarray

speakers_audio_mask(*min_speaker_dim=None*, *speaker_to_idx_map=None*)
Return a matrix of per-speaker activity in a cut. The matrix shape is (num_speakers, num_samples), and its values are 0 for nonspeech **samples** and 1 for speech **samples** for each respective speaker.

This is somewhat inspired by the TS-VAD setup: https://arxiv.org/abs/2005.07272

Parameters

- **min_speaker_dim** (Optional[int]) – optional int, when specified it will enforce that the matrix shape is at least that value (useful for datasets like CHiME 6 where the number of speakers is always 4, but some cuts might have less speakers than that).
- **speaker_to_idx_map** (Optional[Dict[str, int]]) – optional dict mapping speaker names (strings) to their global indices (ints). Useful when you want to preserve the order of the speakers (e.g. speaker XYZ is always mapped to index 2)

Return type ndarray

supervisions_feature_mask()
Return a 1D numpy array with value 1 for **frames** covered by at least one supervision, and 0 for **frames** not covered by any supervision.

Return type ndarray

supervisions_audio_mask()
Return a 1D numpy array with value 1 for **samples** covered by at least one supervision, and 0 for **samples** not covered by any supervision.

Return type ndarray

with_id(*id_*)
Return a copy of the Cut with a new ID.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

class lhotse.cut.**Cut**(*id: str*, *start: float*, *duration: float*, *channel: int*, *supervisions: List[lhotse.supervision.SupervisionSegment] = <factory>*, *features: Optional[lhotse.features.base.Features] = None*, *recording: Optional[lhotse.audio.Recording] = None*)

A Cut is a single "segment" that we'll train on. It contains the features corresponding to a piece of a recording, with zero or more SupervisionSegments.

The SupervisionSegments indicate which time spans of the Cut contain some kind of supervision information: e.g. transcript, speaker, language, etc. The regions without a corresponding SupervisionSegment may contain anything - usually we assume it's either silence or some kind of noise.

Note: The SupervisionSegment time boundaries are relative to the beginning of the cut. E.g. if the underlying Recording starts at 0s (always true), the Cut starts at 100s, and the SupervisionSegment starts at 3s, it means that in the Recording the supervision actually started at 103s. In some cases, the supervision might have a negative start, or a duration exceeding the duration of the Cut; this means that the supervision in the recording extends beyond the Cut.

id: str

start: Seconds

duration: Seconds

channel: int

supervisions: List[SupervisionSegment]

features: Optional[*lhotse.features.base.Features*] = None

recording: Optional[*lhotse.audio.Recording*] = None

property recording_id

Return type str

property end

Return type float

property has_features

Return type bool

property has_recording

Return type bool

property frame_shift

Return type Optional[float]

property num_frames

Return type Optional[int]

property num_samples

Return type Optional[int]

property num_features

Return type Optional[int]

property features_type

Return type Optional[str]

property sampling_rate

Return type int

load_features()
Load the features from the underlying storage and cut them to the relevant [begin, duration] region of the current Cut.

Return type Optional[ndarray]

load_audio()
Load the audio by locating the appropriate recording in the supplied RecordingSet. The audio is trimmed to the [begin, end] range specified by the Cut.

Return type Optional[ndarray]

Returns a numpy ndarray with audio samples, with shape (1 <channel>, N <samples>)

compute_and_store_features(*extractor*, *storage*, *augment_fn=None*, **args*, ***kwargs*)
Compute the features from this cut, store them on disk, and attach a feature manifest to this cut. This cut has to be able to load audio.

Parameters

- **extractor** (*FeatureExtractor*) – a FeatureExtractor instance used to compute the features.
- **output_dir** – the directory where the computed features will be stored.
- **augment_fn** (Optional[Callable[[ndarray, int], ndarray]]) – an optional callable used for audio augmentation.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

Returns a new Cut instance with a Features manifest attached to it.

truncate(***, *offset=0.0*, *duration=None*, *keep_excessive_supervisions=True*, *preserve_id=False*)
Returns a new Cut that is a sub-region of the current Cut.

Note that no operation is done on the actual features - it's only during the call to load_features() when the actual changes happen (a subset of features is loaded).

Parameters

- **offset** (float) – float (seconds), controls the start of the new cut relative to the current Cut's start. E.g., if the current Cut starts at 10.0, and offset is 2.0, the new start is 12.0.
- **duration** (Optional[float]) – optional float (seconds), controls the duration of the resulting Cut. By default, the duration is (end of the cut before truncation) - (offset).
- **keep_excessive_supervisions** (bool) – bool. Since trimming may happen inside a SupervisionSegment, the caller has an option to either keep or discard such supervisions.
- **preserve_id** (bool) – bool. Should the truncated cut keep the same ID or get a new, random one.

Return type *Cut*

Returns a new Cut instance. If the current Cut is shorter than the duration, return None.

pad(*duration*)
Return a new MixedCut, padded to `duration` seconds with zeros in the recording, and low-energy values in each feature bin.

Parameters **duration** (`float`) – The cut's minimal duration after padding.

Return type `Union[Cut, MixedCut, PaddingCut]`

Returns a padded MixedCut if `duration` is greater than this cut's duration, otherwise `self`.

map_supervisions(*transform_fn*)
Modify the SupervisionSegments by *transform_fn* of this Cut.

Parameters **transform_fn** (`Callable[[SupervisionSegment], SupervisionSegment]`) – a function that modifies a supervision as an argument.

Return type `Union[Cut, MixedCut, PaddingCut]`

Returns a modified Cut.

static from_dict(*data*)

Return type *Cut*

with_features_path_prefix(*path*)

Return type *Cut*

with_recording_path_prefix(*path*)

Return type *Cut*

__init__(*id*, *start*, *duration*, *channel*, *supervisions=<factory>*, *features=None*, *recording=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.cut.**PaddingCut**(*id: str*, *duration: float*, *sampling_rate: int*, *use_log_energy: bool*, *num_frames: Optional[int] = None*, *num_features: Optional[int] = None*, *num_samples: Optional[int] = None*)
This represents a cut filled with zeroes in the time domain, or low energy/log-energy values in the frequency domain. It's used to make training samples evenly sized (same duration/number of frames).

id: str

duration: Seconds

sampling_rate: int

use_log_energy: bool

num_frames: Optional[int] = None

num_features: Optional[int] = None

num_samples: Optional[int] = None

property start

Return type `float`

property end

Return type `float`

property supervisions

property has_features

Return type `bool`

property has_recording

Return type bool

property frame_shift

load_features(**args*, ***kwargs*)

Return type Optional[ndarray]

load_audio(**args*, ***kwargs*)

Return type Optional[ndarray]

truncate(*, *offset=0.0*, *duration=None*, *keep_excessive_supervisions=True*, *preserve_id=False*)

Return type *PaddingCut*

pad(*duration*)

Create a new PaddingCut with `duration` when its longer than this Cuts duration. Helper function used in batch cut padding.

Parameters **duration** (float) – The cuts minimal duration after padding.

Return type *PaddingCut*

Returns `self` or a new PaddingCut, depending on `duration`.

compute_and_store_features(*extractor*, **args*, ***kwargs*)

Returns a new PaddingCut with updates information about the feature dimension and number of feature frames, depending on the `extractor` properties.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

map_supervisions(*transform_fn*)

Just for consistency with *Cut* and *MixedCut*.

Parameters **transform_fn** (Callable[[Any], Any]) – a dummy function that would be never called actually.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

Returns the PaddingCut itself.

static from_dict(*data*)

Return type *PaddingCut*

with_features_path_prefix(*path*)

Return type *PaddingCut*

with_recording_path_prefix(*path*)

Return type *PaddingCut*

__init__(*id*, *duration*, *sampling_rate*, *use_log_energy*, *num_frames=None*, *num_features=None*, *num_samples=None*)

Initialize self. See help(type(self)) for accurate signature.

class lhotse.cut.**MixTrack**(*cut: Union[*lhotse.cut.Cut, lhotse.cut.PaddingCut*]*, *offset: float = 0.0*, *snr: Optional[float] = None*)

Represents a single track in a mix of Cuts. Points to a specific Cut and holds information on how to mix it with other Cuts, relative to the first track in a mix.

cut: Union[Cut, PaddingCut]

offset: float = 0.0

snr: Optional[float] = None

static from_dict(*data*)

__init__(*cut*, *offset=0.0*, *snr=None*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.cut.**MixedCut**(*id: str*, *tracks: List[*lhotse.cut.MixTrack*]*)
Represents a Cut that's created from other Cuts via mix or append operations. The actual mixing operations are performed upon loading the features into memory. In order to load the features, it needs to access the CutSet object that holds the "ingredient" cuts, as it only holds their IDs ("pointers"). The SNR and offset of all the tracks are specified relative to the first track.

id: str

tracks: List[MixTrack]

property supervisions
Lists the supervisions of the underlying source cuts. Each segment start time will be adjusted by the track offset.

Return type List[*SupervisionSegment*]

property start

Return type float

property end

Return type float

property duration

Return type float

property has_features

Return type bool

property has_recording

Return type bool

property num_frames

Return type Optional[int]

property frame_shift

Return type Optional[float]

property sampling_rate

Return type Optional[int]

property num_samples

Return type Optional[int]

property num_features

Return type Optional[int]

property features_type

Return type Optional[str]

truncate(*, *offset=0.0*, *duration=None*, *keep_excessive_supervisions=True*, *preserve_id=False*)
Returns a new MixedCut that is a sub-region of the current MixedCut. This method truncates the underlying Cuts and modifies their offsets in the mix, as needed. Tracks that do not fit in the truncated cut are removed.

Note that no operation is done on the actual features - it's only during the call to load_features() when the actual changes happen (a subset of features is loaded).

Parameters

- **offset** (float) – float (seconds), controls the start of the new cut relative to the current MixedCut's start.
- **duration** (Optional[float]) – optional float (seconds), controls the duration of the resulting MixedCut. By default, the duration is (end of the cut before truncation) - (offset).
- **keep_excessive_supervisions** (bool) – bool. Since trimming may happen inside a SupervisionSegment, the caller has an option to either keep or discard such supervisions.
- **preserve_id** (bool) – bool. Should the truncated cut keep the same ID or get a new, random one.

Return type *MixedCut*

Returns a new MixedCut instance.

pad(*duration*)
Return a new MixedCut, padded to `duration` seconds with zeros in the recording, and low-energy values in each feature bin.

Parameters **duration** (float) – The cut's minimal duration after padding.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

Returns a padded MixedCut if duration is greater than this cut's duration, otherwise `self`.

load_features(*mixed=True*)
Loads the features of the source cuts and mixes them on-the-fly.

Parameters **mixed** (bool) – when True (default), returns a 2D array of features mixed in the feature domain. Otherwise returns a 3D array with the first dimension equal to the number of tracks.

Return type Optional[ndarray]

Returns A numpy ndarray with features and with shape `(num_frames, num_features)`, or `(num_tracks, num_frames, num_features)`

load_audio(*mixed=True*)
Loads the audios of the source cuts and mix them on-the-fly.

Parameters **mixed** (bool) – When True (default), returns a mono mix of the underlying tracks. Otherwise returns a numpy array with the number of channels equal to the number of tracks.

Return type Optional[ndarray]

Returns A numpy ndarray with audio samples and with shape `(num_channels, num_samples)`

plot_tracks_features()
Display the feature matrix as an image. Requires matplotlib to be installed.

plot_tracks_audio()
Display plots of the individual tracks' waveforms. Requires matplotlib to be installed.

compute_and_store_features(*extractor*, *storage*, *augment_fn=None*, *mix_eagerly=True*)
Compute the features from this cut, store them on disk, and create a new *Cut* object with the feature manifest attached. This cut has to be able to load audio.

Parameters

- **extractor** (*FeatureExtractor*) – a `FeatureExtractor` instance used to compute the features.
- **storage** (*FeaturesWriter*) – a `FeaturesWriter` instance used to store the features.
- **augment_fn** (`Optional[Callable[[ndarray, int], ndarray]]`) – an optional callable used for audio augmentation.
- **mix_eagerly** (`bool`) – when False, extract and store the features for each track separately, and mix them dynamically when loading the features. When True, mix the audio first and store the mixed features, returning a new `Cut` instance with the same ID. The returned `Cut` will not have a `Recording` attached.

Return type `Union[Cut, MixedCut, PaddingCut]`

Returns a new `Cut` instance if `mix_eagerly` is True, or returns `self` with each of the tracks containing the `Features` manifests.

map_supervisions(*transform_fn*)
Modify the SupervisionSegments by *transform_fn* of this MixedCut.

Parameters **transform_fn** (`Callable[[SupervisionSegment], SupervisionSegment]`) – a function that modifies a supervision as an argument.

Return type `Union[Cut, MixedCut, PaddingCut]`

Returns a modified MixedCut.

static from_dict(*data*)

Return type *MixedCut*

with_features_path_prefix(*path*)

Return type *MixedCut*

with_recording_path_prefix(*path*)

Return type *MixedCut*

__init__(*id*, *tracks*)
Initialize self. See help(type(self)) for accurate signature.

class lhotse.cut.**CutSet**(**args*, ***kwds*)
CutSet combines features with their corresponding supervisions. It may have wider span than the actual supervisions, provided the features for the whole span exist. It is the basic building block of PyTorch-style Datasets for speech/audio processing tasks.

cuts: Dict[str, AnyCut]

property mixed_cuts

Return type `Dict[str, MixedCut]`

property simple_cuts

Return type `Dict[str, Cut]`

property ids

Return type Iterable[str]

property speakers

Return type FrozenSet[str]

static from_cuts(*cuts*)

Return type *CutSet*

static from_manifests(*recordings=None*, *supervisions=None*, *features=None*)
Create a CutSet from any combination of supervision, feature and recording manifests. At least one of recording_set or feature_set is required. The Cut boundaries correspond to those found in the feature_set, when available, otherwise to those found in the recording_set When a supervision_set is provided, we'll attach to the Cut all supervisions that have a matching recording ID and are fully contained in the Cut's boundaries.

Return type *CutSet*

static from_dicts(*data*)

Return type *CutSet*

to_dicts()

Return type List[dict]

describe()
Print a message describing details about the CutSet - the number of cuts and the duration statistics, including the total duration and the percentage of speech segments.

Example output: Cuts count: 547 Total duration (hours): 326.4 Speech duration (hours): 79.6 (24.4%) **** Duration statistics (seconds): mean 2148.0 std 870.9 min 477.0 25% 1523.0 50% 2157.0 75% 2423.0 max 5415.0 dtype: float64

Return type None

split(*num_splits*, *randomize=False*)
Split the CutSet into num_splits pieces of equal size.

Parameters

- **num_splits** (int) – Requested number of splits.
- **randomize** (bool) – Optionally randomize the cuts order first.

Return type List[*CutSet*]

Returns A list of CutSet pieces.

filter(*predicate*)
Return a new CutSet with the Cuts that satisfy the *predicate*.

Parameters **predicate** (Callable[[Union[*Cut*, *MixedCut*, *PaddingCut*]], bool]) – a function that takes a cut as an argument and returns bool.

Return type *CutSet*

Returns a filtered CutSet.

trim_to_supervisions()
Return a new CutSet with Cuts that have identical spans as their supervisions.

Return type *CutSet*

Returns a CutSet.

trim_to_unsupervised_segments()
Return a new CutSet with Cuts created from segments that have no supervisions (likely silence or noise).

Return type *CutSet*

Returns a CutSet.

mix_same_recording_channels()
Find cuts that come from the same recording and have matching start and end times, but represent different channels. Then, mix them together (in matching groups) and return a new CutSet that contains their mixes. This is useful for processing microphone array recordings.

It is intended to be used as the first operation after creating a new CutSet (but might also work in other circumstances, e.g. if it was cut to windows first).

Example:

```
>>> ami = prepare_ami('path/to/ami')
>>> cut_set = CutSet.from_manifests(recordings=ami['train']['recordings'])
>>> multi_channel_cut_set = cut_set.mix_same_recording_channels()
```

In the AMI example, the multi_channel_cut_set will yield MixedCuts that hold all single-channel Cuts together.

Return type *CutSet*

sort_by_duration(*ascending=False*)
Sort the CutSet according to cuts duration. Descending by default.

Return type *CutSet*

pad(*duration=None*)
Return a new CutSet with Cuts padded to duration in seconds. Cuts longer than duration will not be affected. Cuts will be padded to the right (i.e. after the signal). :type duration: Optional[float] :param duration: The cuts minimal duration after padding. When not specified, we'll choose the duration of the longest cut in the CutSet. :rtype: *CutSet* :return: A padded CutSet.

truncate(*max_duration*, *offset_type*, *keep_excessive_supervisions=True*, *preserve_id=False*)
Return a new CutSet with the Cuts truncated so that their durations are at most *max_duration*. Cuts shorter than *max_duration* will not be changed. :type max_duration: float :param max_duration: float, the maximum duration in seconds of a cut in the resulting manifest. :type offset_type: str :param offset_type: str, can be: - 'start' => cuts are truncated from their start; - 'end' => cuts are truncated from their end minus max_duration; - 'random' => cuts are truncated randomly between their start and their end minus max_duration :type keep_excessive_supervisions: bool :param keep_excessive_supervisions: bool. When a cut is truncated in the middle of a supervision segment, should the supervision be kept. :type preserve_id: bool :param preserve_id: bool. Should the truncated cut keep the same ID or get a new, random one. :rtype: *CutSet* :return: a new CutSet instance with truncated cuts.

cut_into_windows(*duration*, *keep_excessive_supervisions=True*)
Return a new CutSet, made by traversing each Cut in windows of duration seconds and creating new Cut out of them.

The last window might have a shorter duration if there was not enough audio, so you might want to use either .filter() or .pad() afterwards to obtain a uniform duration CutSet.

Parameters

- **duration** (float) – Desired duration of the new cuts in seconds.
- **keep_excessive_supervisions** (bool) – bool. When a cut is truncated in the middle of a supervision segment, should the supervision be kept.

Return type *CutSet*

Returns a new CutSet with cuts made from shorter duration windows.

compute_and_store_features(*extractor*, *storage*, *augment_fn=None*, *executor=None*, *mix_eagerly=True*)

Modify the current `CutSet` with by extracting features and attaching the feature manifests to the cuts.

Parameters

- **extractor** (*FeatureExtractor*) – A `FeatureExtractor` instance (either Lhotse's built-in or a custom implementation).
- **storage** (*FeaturesWriter*) – A `FeaturesWriter` instance used to store the features.
- **augment_fn** (`Optional[Callable[[ndarray, int], ndarray]]`) – an optional callable used for audio augmentation.
- **executor** (`Optional[Any]`) – when provided, will be used to parallelize the feature extraction process. Any executor satisfying the standard concurrent.futures interface will be suitable; e.g. ProcessPoolExecutor, ThreadPoolExecutor, or dask.Client for distributed task execution (see: https://docs.dask.org/en/latest/futures.html?highlight=Client#start-dask-client)
- **mix_eagerly** (`bool`) – Related to how the features are extracted for `MixedCut` instances, if any are present. When False, extract and store the features for each track separately, and mix them dynamically when loading the features. When True, mix the audio first and store the mixed features, returning a new `Cut` instance with the same ID. The returned `Cut` will not have a `Recording` attached.

Return type *CutSet*

Returns a new CutSet instance with the same `Cut``s, but with attached ``Features` objects

with_features_path_prefix(*path*)

Return type *CutSet*

with_recording_path_prefix(*path*)

Return type *CutSet*

map_supervisions(*transform_fn*)

Modify the SupervisionSegments by *transform_fn* in this CutSet.

Parameters **transform_fn** (`Callable[[SupervisionSegment], SupervisionSegment]`) – a function that modifies a supervision as an argument.

Return type *CutSet*

Returns a new, modified CutSet.

transform_text(*transform_fn*)

Return a copy of this `CutSet` with all `SupervisionSegments` text transformed with `transform_fn`. Useful for text normalization, phonetic transcription, etc.

Parameters **transform_fn** (`Callable[[str], str]`) – a function that accepts a string and returns a string.

Return type *CutSet*

Returns a new, modified CutSet.

__init__(*cuts*)
Initialize self. See help(type(self)) for accurate signature.

lhotse.cut.**make_windowed_cuts_from_features**(*feature_set*, *cut_duration*, *cut_shift=None*, *keep_shorter_windows=False*)
Converts a FeatureSet to a CutSet by traversing each Features object in - possibly overlapping - windows, and creating a Cut out of that area. By default, the last window in traversal will be discarded if it cannot satisfy the *cut_duration* requirement.

Parameters

- **feature_set** (*FeatureSet*) – a FeatureSet object.
- **cut_duration** (float) – float, duration of created Cuts in seconds.
- **cut_shift** (Optional[float]) – optional float, specifies how many seconds are in between the starts of consecutive windows. Equals *cut_duration* by default.
- **keep_shorter_windows** (bool) – bool, when True, the last window will be used to create a Cut even if its duration is shorter than *cut_duration*.

Return type *CutSet*

Returns a CutSet object.

lhotse.cut.**mix**(*reference_cut*, *mixed_in_cut*, *offset=0*, *snr=None*)
Overlay, or mix, two cuts. Optionally the *mixed_in_cut* may be shifted by *offset* seconds and scaled down (positive SNR) or scaled up (negative SNR). Returns a MixedCut, which contains both cuts and the mix information. The actual feature mixing is performed during the call to MixedCut.load_features().

Parameters

- **reference_cut** (Union[*Cut*, *MixedCut*, *PaddingCut*]) – The reference cut for the mix - offset and snr are specified w.r.t this cut.
- **mixed_in_cut** (Union[*Cut*, *MixedCut*, *PaddingCut*]) – The mixed-in cut - it will be offset and rescaled to match the offset and snr parameters.
- **offset** (float) – How many seconds to shift the mixed_in_cut w.r.t. the reference_cut.
- **snr** (Optional[float]) – Desired SNR of the *right_cut* w.r.t. the *left_cut* in the mix.

Return type *MixedCut*

Returns A MixedCut instance.

lhotse.cut.**append**(*left_cut*, *right_cut*, *snr=None*)
Helper method for functional-style appending of Cuts.

Return type *MixedCut*

lhotse.cut.**mix_cuts**(*cuts*)
Return a MixedCut that consists of the input Cuts mixed with each other as-is.

Return type *MixedCut*

lhotse.cut.**append_cuts**(*cuts*)
Return a MixedCut that consists of the input Cuts appended to each other as-is.

Return type Union[*Cut*, *MixedCut*, *PaddingCut*]

9.7 Recipes

Convenience methods used to prepare recording and supervision manifests for standard corpora.

9.8 Kaldi conversion

Convenience methods used to interact with Kaldi data directories.

lhotse.kaldi.**load_kaldi_data_dir**(*path*, *sampling_rate*)
Load a Kaldi data directory and convert it to a Lhotse RecordingSet and SupervisionSet manifests. For this to work, at least the wav.scp file must exist. SupervisionSet is created only when a segments file exists. All the other files (text, utt2spk, etc.) are optional, and some of them might not be handled yet. In particular, feats.scp files are ignored.

Return type Tuple[*RecordingSet*, Optional[*SupervisionSet*]]

lhotse.kaldi.**load_kaldi_text_mapping**(*path*, *must_exist=False*)
Load Kaldi files such as utt2spk, spk2gender, text, etc. as a dict.

Return type Dict[str, Optional[str]]

9.9 Others

Helper methods used throughout the codebase.

lhotse.manipulation.**combine**(**manifests*)
Combine multiple manifests of the same type into one.

Return type ~Manifest

lhotse.manipulation.**to_manifest**(*items*)
Take an iterable of data types in Lhotse such as Recording, SupervisonSegment or Cut, and create the manifest of the corresponding type. When the iterable is empty, returns None.

Return type Optional[~Manifest]

lhotse.manipulation.**load_manifest**(*path*)
Generic utility for reading an arbitrary manifest.

Return type ~Manifest

CHAPTER

TEN

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

l

INDEX

Symbols

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

R

S

T

U

V

W