Libri-Light: a Benchmark for ASR with Limited or no Supervision

J. Kahn*[1], M. Rivière*[1], W. Zheng*[1], E. Kharitonov*[1], Q. Xu*[1], P.E. Mazaré*[1], J. Karadayi*[2], V. Liptchinsky[1], R. Collobert[1], C. Fuegen[1], T. Likhomanenko[1], G. Synnaeve[1], A. Joulin[1], A. Mohamed[1], E. Dupoux[1,2]

ICASSP 2020

1 facebook AI Research 2 Inria

Motivation

- **Progress in ASR along two axes:**
  - usage of increasingly large deep neural networks
  - increasingly large amounts of annotated speech
- **Two challenges:**
  - annotating large amounts of speech is prohibitively expensive
  - annotation doesn’t scale beyond high resource languages
    - can’t address low-resources languages, accents, dialectical variants, etc.

Motivation

- **Progress in ASR along two axes:**
  - usage of increasingly large deep neural networks
  - increasingly large amounts of annotating speech
- **Two challenges:**
  - annotating large amounts of speech is prohibitively expensive
  - annotation doesn't scale beyond high resource languages
    - can't address low-resources languages, accents, dialectical variants, etc.
- **Research in weak supervision is growing:**
  - usage of datasets with fewer human annotations
  - labels from other languages
  - unsupervised objectives
  - zero-resource ASR

Motivation

We need a common benchmark across semi-supervised and unsupervised learning in speech.

Libri-Light defines:

- datasets
- evaluation metrics
- baselines

Existing Benchmarks and Datasets in ASR

Supervised:

- **Librispeech** (Panayotov et al. 2015)

  1000 hours of English audio books with textual annotations aligned at the sentence level

- **Mozilla's CommonVoice** (Ardila et al. 2019)

  2,900 hours of read speech in 37 languages

- **Wilderness** (Black et al. 2019)

  Text of the Bible read in 750 languages

Panayotov et al. *Librispeech: an ASR corpus based on public domain audio books*, ICASSP 2015
Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*, LREC 2020, to appear.
Black et al. *CMU Wilderness Multilingual Speech Dataset*, ICASSP 2019

Existing Benchmarks and Datasets in ASR

Supervised:

- **Librispeech** (Panayotov et al. 2015)
- **Mozilla's CommonVoice** (Ardila et al. 2019)
- **Wilderness** (Black et al. 2019)

Semi-Supervised:

- **Babel Project** (IARPA)

  Many languages; 10 hours of transcribed speech and large amounts of unlabeled audio, but no benchmark

  High Resource — English, German, French, Mandarin

  Low Resource (*2.5 — 50 hours*) — Xitsonga, Wolof, Indonesian, etc.

Panayotov et al. *Librispeech: an ASR corpus based on public domain audio books*, ICASSP 2015
Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*, LREC 2020, to appear.
Black et al. *CMU Wilderness Multilingual Speech Dataset*, ICASSP 2019
Roach et al. *BABEL: an Eastern European multi-language database*, ICSLP 1996

Existing Benchmarks and Datasets in ASR

Supervised:

- **Librispeech** (Panayotov et al. 2015)
- **Mozilla's CommonVoice** (Ardila et al. 2019)
- **Wilderness** (Black et al. 2019)

Semi-Supervised:

- **Babel Project** (IARPA)

Unsupervised:

- **Zero Resource Challenge** (Versteegh et al. 2015, Dunbar et al. 2017, Dunbar et al. 2019)

For unsupervised learning: 2.5 and 50 hours of speech

Panayotov et al. *Librispeech: an ASR corpus based on public domain audio books*, ICASSP 2015
Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*, LREC 2020, to appear.
Black et al. *CMU Wilderness Multilingual Speech Dataset*, ICASSP 2019
Roach et al. *BABEL: an Eastern European multi-language database*, ICSLP 1996
Versteegh et al. *The Zero Resource Speech Challenge 2015*, Interspeech 2015
Dunbar et al. *The zero resource speech challenge 2017*, ASRU 2017
Dunbar et al. *The Zero Resource Speech Challenge 2019: TTS without T*, Interspeech 2019

Why have Libri-Light?

- We need to **compare semi and unsupervised techniques on the same set of data**.
- Facilitate **scaling up the amount of unlabeled data** while **scaling down the amount of labeled data**.
- **Development and test sets are the same as LibriSpeech** — keeps evaluation consistent with other in-domain work.

Why have Libri-Light?

- We need to **compare semi and unsupervised techniques on the same set of data**.
- Facilitate **scaling up the amount of unlabeled data** while **scaling down the amount of labeled data**.
- **Development and test sets are the same as LibriSpeech** — keeps evaluation consistent with other in-domain work.
- Develop a **common set of metrics** to evaluate different settings:

Setting	Metric	Audio Only (hours)	Audio + Text (hours)	LM Data
zero-resources / unsupervised	ABX	600, 6k, 60k	-	-
semi-supervised	PER and CER	600, 6k, 60k	10 min; 1h; 10h	-
distantly-supervised	WER	600, 6k, 60k	10 min; 1h; 10h	800 million words +

Make everything open source!

Dataset

Train, dev, and test sets

Libri-Light — The Numbers

68.8k

Total hours of unlabeled audio.

7582

Distinct speakers represented in LibriVox.

7.96

Avg. hours of audio per speaker from LibriVox.

Dataset Details

Four components:

- A training set with unlabelled audio
- A training set with limited labeling
- Development/test sets
- A training set containing unaligned text

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489

subset	hours	per-spk minutes	female sprks	male spkrs	total spkrs
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets (from LibriSpeech)					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33

subset	tokens	vocab
Unaligned Text Training Set		
librispeech-LM (in-domain)	800M	200K

*Six different versions of the 10 min datasets have been constructed, the union of these small datasets make up the 1h dataset.

Dataset Details

Four components:

- **A training set with unlabelled audio**
- A training set with limited labeling
- Development/test sets
- A training set containing unaligned text

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489

subset	hours	per-spk minutes	female sprks	male spkrs	total spkrs
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets (from LibriSpeech)					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33

subset	tokens	vocab
Unaligned Text Training Set		
librispeech-LM (in-domain)	800M	200K

*Six different versions of the 10 min datasets have been constructed, the union of these small datasets make up the 1h dataset.

Dataset Details

Four components:

- A training set with unlabelled audio
- **A training set with limited labeling**
- Development/test sets
- A training set containing unaligned text

Training data is from half clean, half noisy subsets.

Provide phonetic transcriptions generated from a phonemizer.

*Six different versions of the 10 min datasets have been constructed, the union of these small datasets make up the 1h dataset.

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489
subset	**hours**	**per-spk minutes**	**female sprks**	**male spkrs**	**total spkrs**
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets (from LibriSpeech)					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33
subset			**tokens**	**vocab**	
Unaligned Text Training Set					
librispeech-LM (in-domain)			800M	200K	

Dataset Details

Four components:

- A training set with unlabelled audio
- A training set with limited labeling
- **Development/test sets**
- A training set containing unaligned text

All LibriSpeech dev/test set audio is removed from all training sets.

For ABX evaluation, force alignment is obtained with a model trained on LibriSpeech.

*Six different versions of the 10 min datasets have been constructed, the union of these small datasets make up the 1h dataset.

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489
subset	**hours**	**per-spk minutes**	**female sprks**	**male spkrs**	**total spkrs**
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets (from LibriSpeech)					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33
subset			**tokens**	**vocab**	
Unaligned Text Training Set					
librispeech-LM (in-domain)			800M	200K	

Dataset Details

Four components:

- A training set with unlabelled audio
- A training set with limited labeling
- Development/test sets
- **A training set containing unaligned text**

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489

subset	hours	per-spk minutes	female sprks	male spkrs	total spkrs
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets (from LibriSpeech)					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33

subset	tokens	vocab
Unaligned Text Training Set		
librispeech-LM (in-domain)	800M	200K

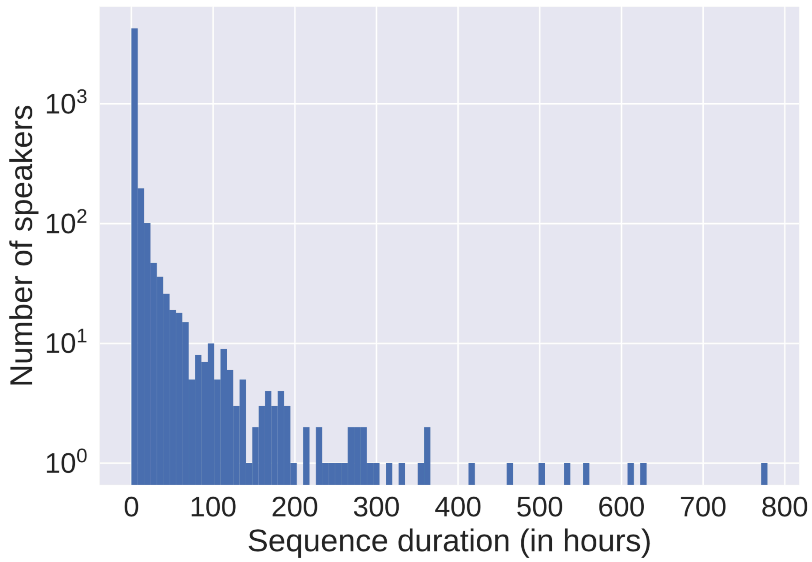
*Six different versions of the 10 min datasets have been constructed, the union of these small datasets make up the 1h dataset.

Dataset Preparation

The pipeline for audio preprocessing, voice activity detection, and segmentation.

Creating the Training Set — Unlabeled Audio

1. **Extract audio files** for English speech from LibriVox (public domain audio books)
2. **Filter files** with unknown or multiple speakers or speakers from LibriSpeech dev/test, or for duplications based on title

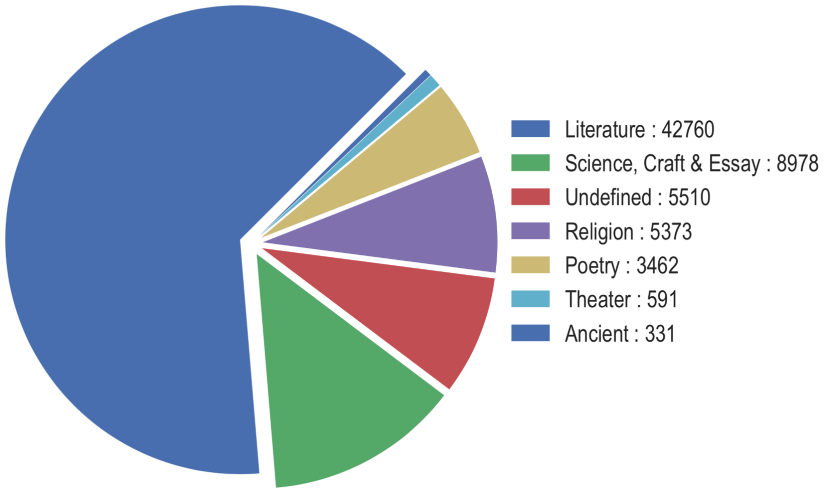


Creating the Training Set — Unlabeled Audio

1. **Extract audio files** for English speech from LibriVox (public domain audio books)
2. **Filter files** with unknown or multiple speakers or speakers from LibriSpeech dev/test, or for duplications based on title
3. Run **Voice Activity Detection** (VAD) using wav2letter++ models (Pratap et al. 2019) to tag onsets and offsets of speech segments; compute SNR for segments
4. **Prepare JSON metadata** containing title, a unique speaker ID, SNR, macro genre, and VAD data
    - Preserve genre distribution over different dataset splits

subset	hours	books	files	per-spk hours	total spkrs
Unlabelled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489

Genre's categories (in hours)

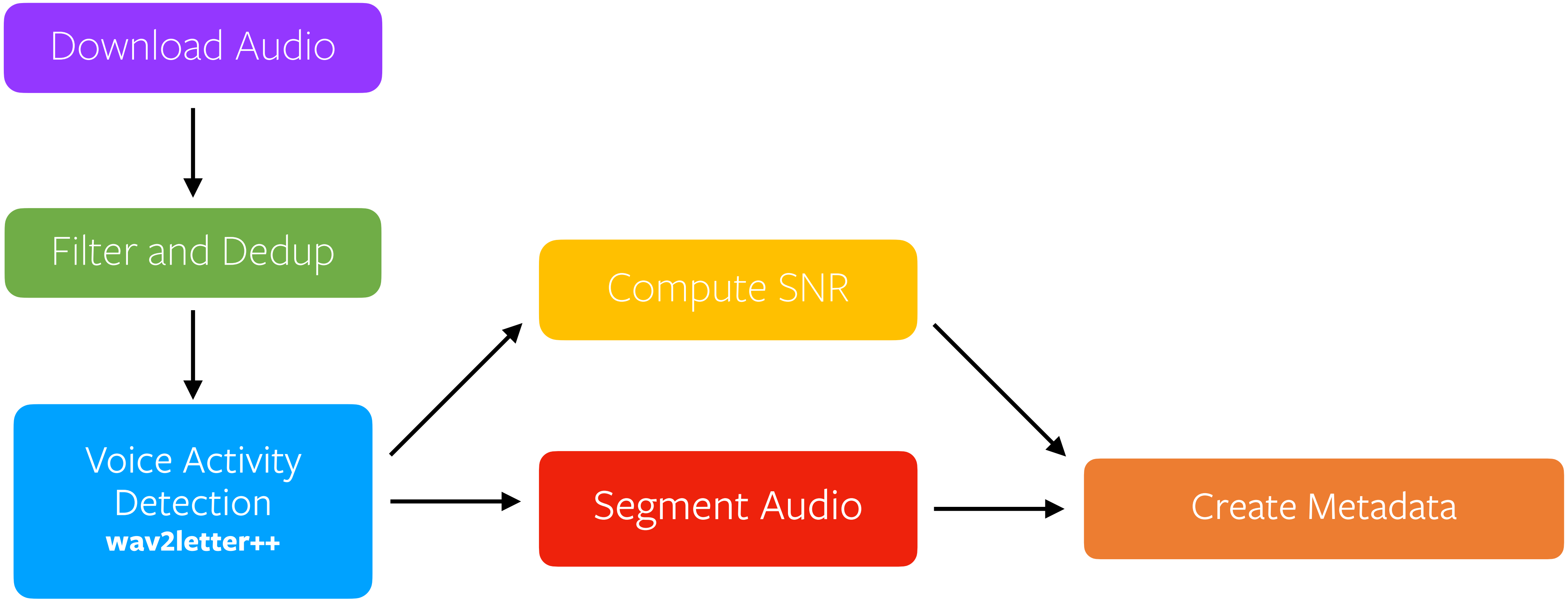


Pratap et al. *Wav2Letter++: A Fast Open-source Speech Recognition System*, ICASSP 2019

Creating the Training Set — Unlabeled Audio

A completely open source pipeline for preprocessing large amounts of unlabeled audio.

github.com/facebookresearch/libri-light



Metrics

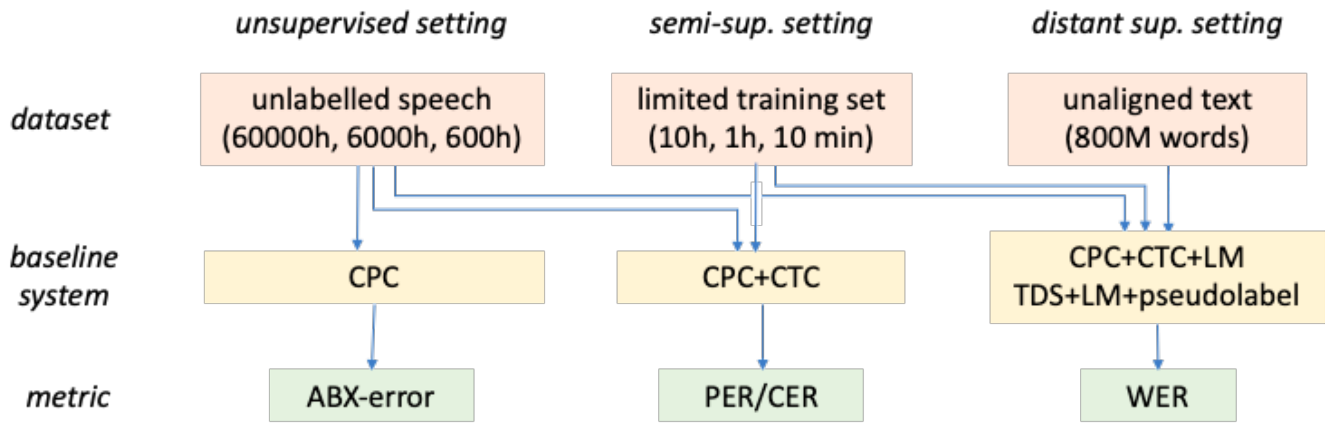
Evaluating dataset benchmarks with varying levels of supervision.

Baselines and Metrics

- **Unsupervised learning**
  - *Goal*: extract speech representations
  - Evaluate with ABX metrics (Schatz et al. 2013)
- **Semi-supervised learning**
  - *Goal:* evaluate learned speech representations learned with little annotated data
    - Train with character-based of phonemic targets
  - Evaluate with character and phoneme error rates
- **Distant Supervision**
  - *Goal*: evaluate how learned representations can decode speech at the word level in conjunction with a language model
  - Evaluate with word error rate (WER)

Schatz et al. *Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline*, 2013

Baselines and Metrics



Baseline Results

Unsupervised learning with Contrastive Predictive Coding (CPC) (Oord et al. 2018)

- CPC constructs embeddings with good ABX scores compared to an MFCC baseline
  - Results are in the same range as the best result from the Zero Resource Speech Challenge 2017 for English
- Increasing the amount of unlabeled data significantly improves ABX embedding quality.

System	ABX within speaker				ABX across speaker			
	dev-clean	dev-other	test-clean	test-other	dev-clean	dev-other	test-clean	test-other
MFCC Baseline	10.95	13.55	10.58	13.60	20.94	29.41	20.45	28.5
CPC unlab-600	7.36	9.39	6.90	9.59	9.58	14.67	9.00	15.1
CPC unlab-6k	6.51	8.42	6.22	8.55	8.48	13.39	8.05	13.81
CPC unlab-60k	**6.11**	**8.17**	**5.83**	**8.14**	**8.05**	**12.83**	**7.56**	**13.42**

Oord et al. *Representation Learning with Contrastive Predictive Coding*, 2018

Baseline Results

Semi-supervised learning with Contrastive Predictive Coding (CPC) (Oord et al. 2018)

- A pre-trained CPC system + a linear classifier trained on *just 10 hours of labeled audio* outperforms the same system trained only on labeled data from scratch.
- Pre-training is more effective even when only *10 minutes* of labeled audio is available.

System	dev-clean	dev-other	test-clean	test-other
no pretraining+train-10h	45.9	55.7	43.7	58.6
CPC unlab-60k+train-10m	40.1	51.5	39.4	53.3
CPC unlab-60k+train-1h	32.2	44.6	31.6	46.8
CPC unlab-60k+train-10h	**28.4**	**41.4**	**27.9**	**43.6**

Results given in phoneme-error rate (PER)

Oord et al. *Representation Learning with Contrastive Predictive Coding*, 2018

Baseline Results

Distantly-supervised learning with Contrastive Predictive Coding (CPC) (Oord et al. 2018)

- Use a model pre-trained on some labeled audio with a CPC model trained with unlabeled audio.
- Increasing the amount of unsupervised data helps pre-training.
  - Returns diminish with more unlabeled audio.

System	dev-clean	dev-other	test-clean	test-other
Supervised systems (LibriSpeech 1000 h)				
Gated Cnv+4gramLM[20]	4.6	13.8	4.8	14.5
Hybrid+seqdisc+4gramLM[21]	3.4	8.3	3.8	8.8
CPC pretrain + CTC fine-tuning + 4gram-LM				
CPC unlab-600+train-10m	97.3	97.6	97.1	97.7
CPC unlab-600+train-1h	72.2	84.5	70.1	86.3
CPC unlab-600+train-10h	52.5	71.6	49.3	74.1
CPC unlab-6k+train-10m	93.6	95.2	93.2	94.9
CPC unlab-6k+train-1h	67.5	81.3	65.4	82.0
CPC unlab-6k+train-10h	46.4	**66.7**	44.7	69.3
CPC unlab-60k+train-10m	92.5	94.2	92.5	94.4
CPC unlab-60k+train-1h	66.6	80.0	64.7	81.6
CPC unlab-60k+train-10h	**46.1**	**66.7**	**43.9**	**69.5**

Oord et al. *Representation Learning with Contrastive Predictive Coding*, 2018

Baseline Results

Distantly-supervised learning with pseudo-labeling

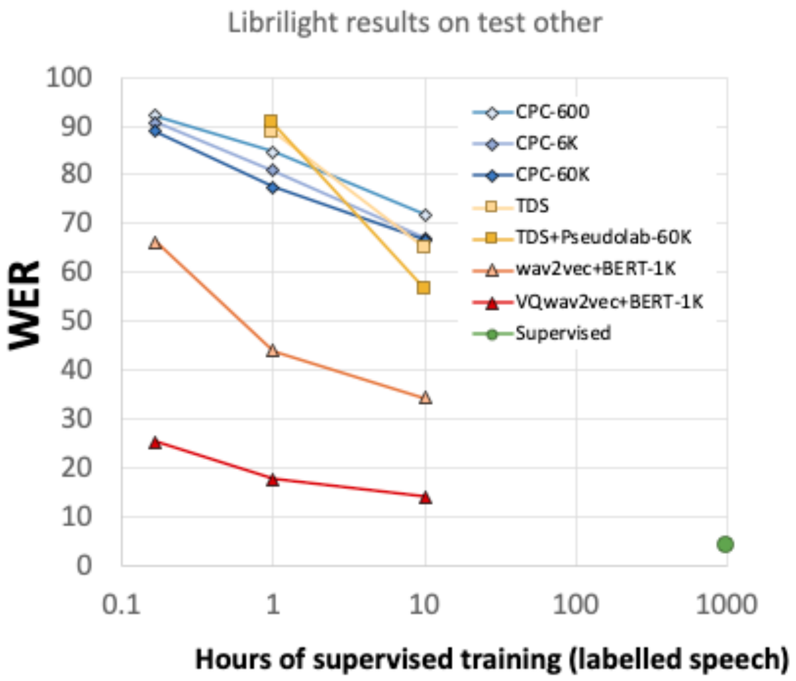
- Adding unlabeled audio helps in pretraining.
- Self-training is effective, but only if the pseudo-label-generating model is good.

System	dev-clean	dev-other	test-clean	test-other
Supervised systems (LibriSpeech 1000 h)				
Gated Cnv+4gramLM[20]	4.6	13.8	4.8	14.5
Hybrid+seqdisc+4gramLM[21]	3.4	8.3	3.8	8.8
MFSC + TDS + CTC + Grapheme + 4gram-LM				
train-1h	79.4	88.1	78.4	88.0
+ 60k pseudo-label	78.6	86.5	77.2	86.3
train-10h	34.0	60.9	33.5	62.1
+ 60k pseudo-label	30.5	55.8	30.1	57.2
MFSC + TDS + CTC + Phoneme + 4gram-LM				
train-1h	81.1	88.5	80.2	88.7
+ 60k pseudo-label	84.3	90.0	84.0	90.5
train-10h	44.1	64.2	43.8	65.1
+ 60k pseudo-label	**30.0**	**55.8**	**29.3**	**56.6**

Results given in word-error rate (WER)

Aggregated Results

Unlabeled audio pushes the low-resource setting forward.



Newer, Improved Results:

- wav2vec + BERT-1k (Baevski et al. 2019)
  - 34% WER with 10 hours of labeled audio
- vq-wav2vec + BERT-1k (Baevski et al. 2019)
  - 14% WER with 10 hours of labeled audio

Baevski et al. *vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations*, ICLR 2019

In Summary

- We introduce a **large new dataset** for benchmarking ASR systems trained with **limited or no supervision**.
- Unsupervised training with more unlabeled audio learns **better representations**.
- **Future work:**
  - Larger models
  - Speaker-adversarial losses
  - Fine-tuning systems end-to-end
  - Pseudo-label retraining

Download or Reproduce Libri-Light!

https://github.com/facebookresearch/libri-light