DeepSpeech and Common Voice by Mozilla

Torino Coding Society, 6 Feb 2019

Agenda

- Hi, nice to meet you!
- Why did Mozilla focus on voice recognition?
- From a paper to DeepSpeech’s development
- The Common Voice dataset
- Your contribution and where we can meet again

Hi, I'm Stefania!

or astrastefania on those addictive things...

I'm working with M-Lab data in TOP-IX (used also by Mozilla for Internet Health Report)

I play Maths games at MathsJam and love being an Effective Altruist and a Mozillian

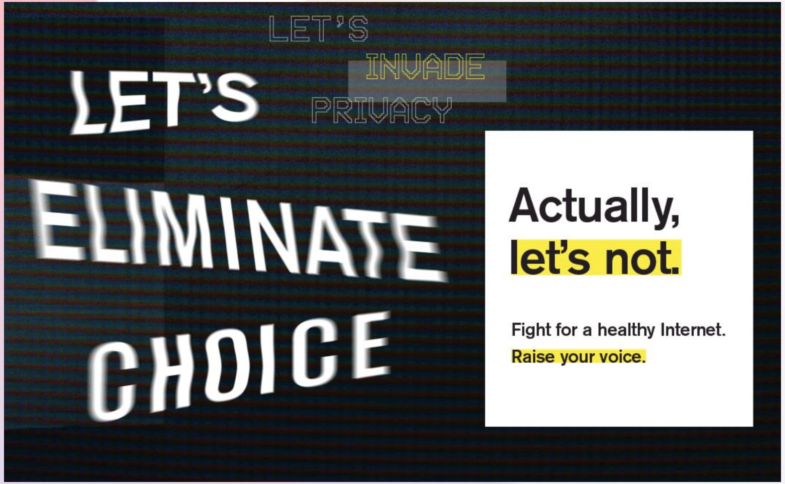
What is Mozilla?

Mozilla is a free software community supported by the not-for-profit organization Mozilla Foundation.

More than Firefox...

Here's others interesting projects:

- Thunderbird
- MDN web docs
- AV1 codec
- WebVR and A-Frame
- Rust and Servo
- Campaigns for the web



Mozilla’s focus on voice

Why did Mozilla focus on voice recognition?

Once upon a time...

- The project Vaani: a voice assistant for Firefox OS
- Connected devices, now iot.mozilla.org
- New team dedicated to Machine Learning in research.mozilla.org

Why did Mozilla focus on voice recognition?

- Need for a open source alternative
- Off-line accessibility and data storage
- State of the art Speech-to-Text
- End-to-end Machine Learning system

DeepSpeech by Mozilla

Baidu's Deep Speech paper v2

https://arxiv.org/abs/1412.5567

:1412.5567v2 [cs.CL] 19 Dec 2014

Deep Speech: Scaling up end-to-end speech recognition

Awni Hannun*, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab

Abstract

We present a state-of-the-art speech recognition system developed using end-to-end deep learning. Our architecture is significantly simpler than traditional speech systems, which rely on laboriously engineered processing pipelines; these traditional systems also tend to perform poorly when used in noisy environments. In contrast, our system does not need hand-designed components to model background noise, reverberation, or speaker variation, but instead directly learns a function that is robust to such effects. We do not need a phoneme dictionary, nor even the concept of a "phoneme." Key to our approach is a well-optimized RNN training system that uses multiple GPUs, as well as a set of novel data synthesis techniques that allow us to efficiently obtain a large amount of varied data for training. Our system, called Deep Speech, outperforms previously published results on the widely studied Switchboard Hub5'00, achieving 16.0% error on the full test set. Deep Speech also handles challenging noisy environments better than widely used, state-of-the-art commercial speech systems.

1 Introduction

Top speech recognition systems rely on sophisticated pipelines composed of multiple algorithms and hand-engineered processing stages. In this paper, we describe an end-to-end speech system, called "Deep Speech", where deep learning supersedes these processing stages. Combined with a

Further publications

Deep Speech v2 for
English and Mandarin
https://arxiv.org/abs/1512.02595

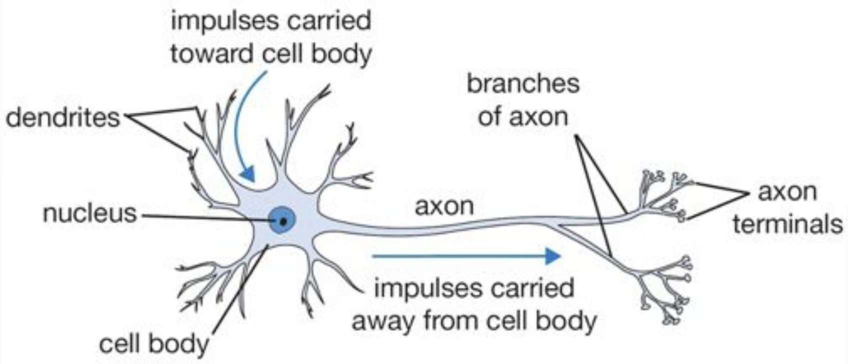
Deep Speech v3
http://research.baidu.com/Blog/index-view?id=90
https://arxiv.org/abs/1707.07413

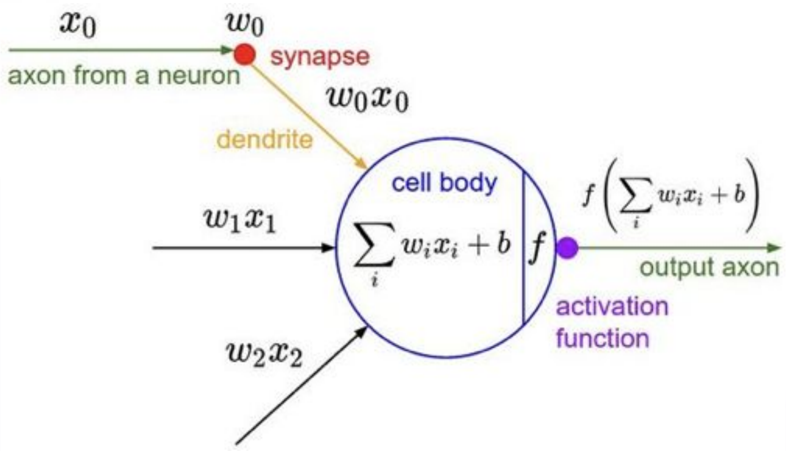
Baidu's Deep Speech paper

The main points to create a speech-to-text algorithm

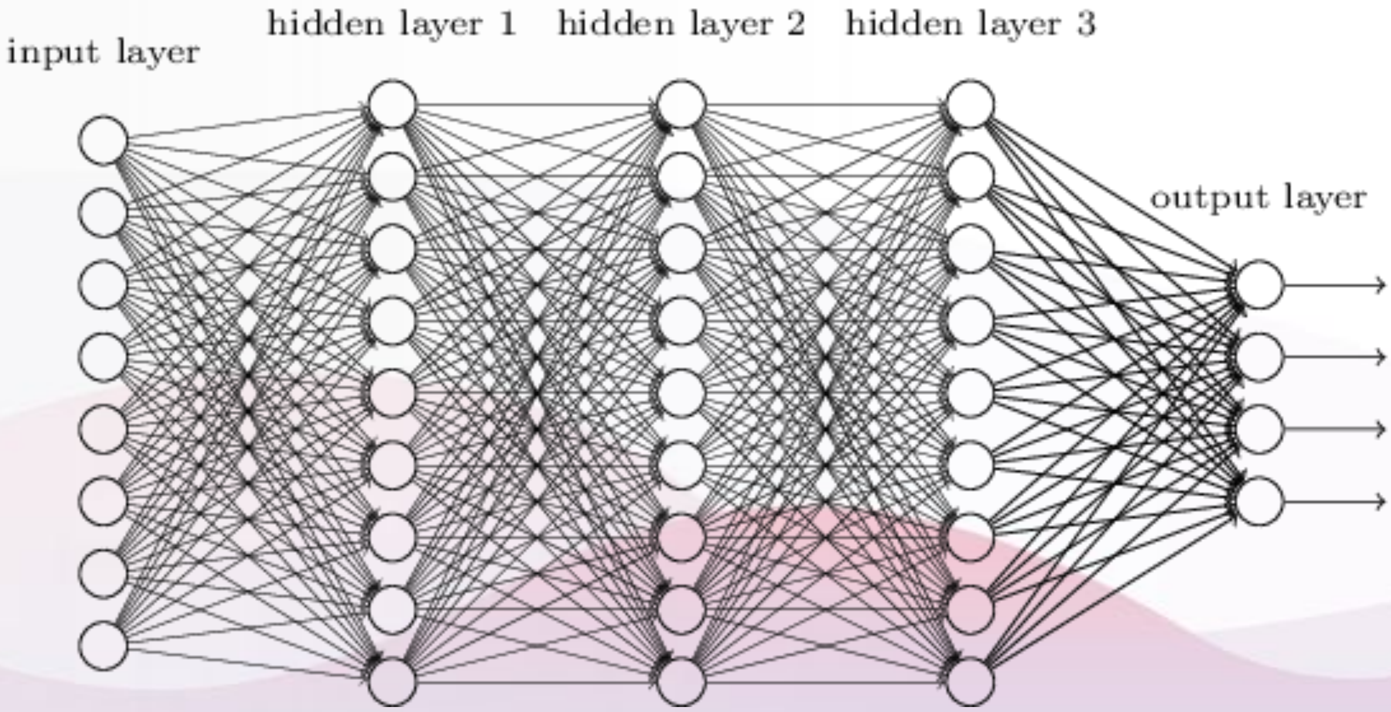
- No complex pipelines
- Using only Deep Learning through RNN
- Leverage GPUs and parallel processes
- A big dataset to learn even with background noises

ANN, Artificial Neural Network





DNN, Deep Neural Network



RNN, Recurrent Neural Network

"Recurrent Neural Networks in a class of ANN where connections between nodes form a **directed graph along a sequence**.

This allows it to exhibit temporal dynamic behavior for a time sequence.

Unlike feedforward neural networks, RNNs can **use their internal state** (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition."

https://en.wikipedia.org/wiki/Recurrent_neural_network

Baidu's Deep Speech paper, RNN

Recurrent Neural Networks in Baidu's paper

- 3 forward-layers
- 1 bi-directional recurrent layer
- 1 forward using forward and backward recurrent layers as input

$$h^{(6)}_{t,k} = \hat{y}_{t,k} \equiv \mathbb{P}(c_t = k|x) = \frac{\exp(W^{(6)}_k h^{(5)}_t + b^{(6)}_k)}{\sum_j \exp(W^{(6)}_j h^{(5)}_t + b^{(6)}_j)}$$

- h, inputs
- W, weight matrix for layers
- b, bias matrix for layers

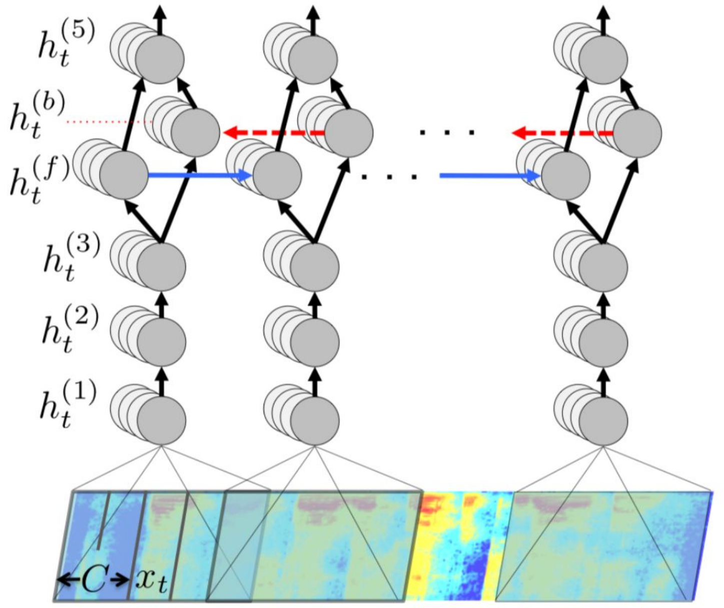


Figure 1: Structure of our RNN model and notation.

Problems

- Different background sounds
- "Lombard Effect" when speakers actively change their pitch or inflections of their voice to overcome noise around them
- Emotional tones and local accents

Solutions

- Have a dataset with a mix of tracks with a clean and a noisy background
- Analyse and test the divergence from average behaviour in a real speech
- Integrate different accents and tones

Developing DeepSpeech

Initial goals were...

- WER (Word Error Rate) below 10%
- Implement Neural Networks using TensorFlow
- Tuning of the Hyperparameters
- Found or create a decent dataset for training and testing

WER (Word Error Rate) below 10%

hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate

A Journey to <10% Word Error Rate

By Reuben Morais

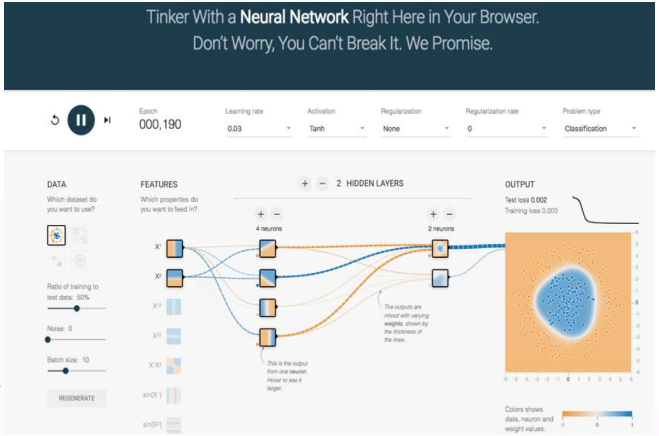
Posted on November 29, 2017 in Featured Article and Research ❤ Share This

At Mozilla, we believe speech interfaces will be a big part of how people interact with their devices in the future. Today we are excited to announce the initial release of our open source speech recognition model so that anyone can develop compelling speech experiences.

The Machine Learning team at Mozilla Research has been working on an open source Automatic Speech Recognition engine modeled after the Deep Speech papers (1, 2) published by Baidu. One of the major goals from the beginning was to achieve a Word Error Rate in the transcriptions of under 10%. We have made great progress: Our word error rate on LibriSpeech's test-clean set is 6.5%, which not only achieves our initial goal, but gets us close to human level performance.

Implement Neural Networks using TensorFlow

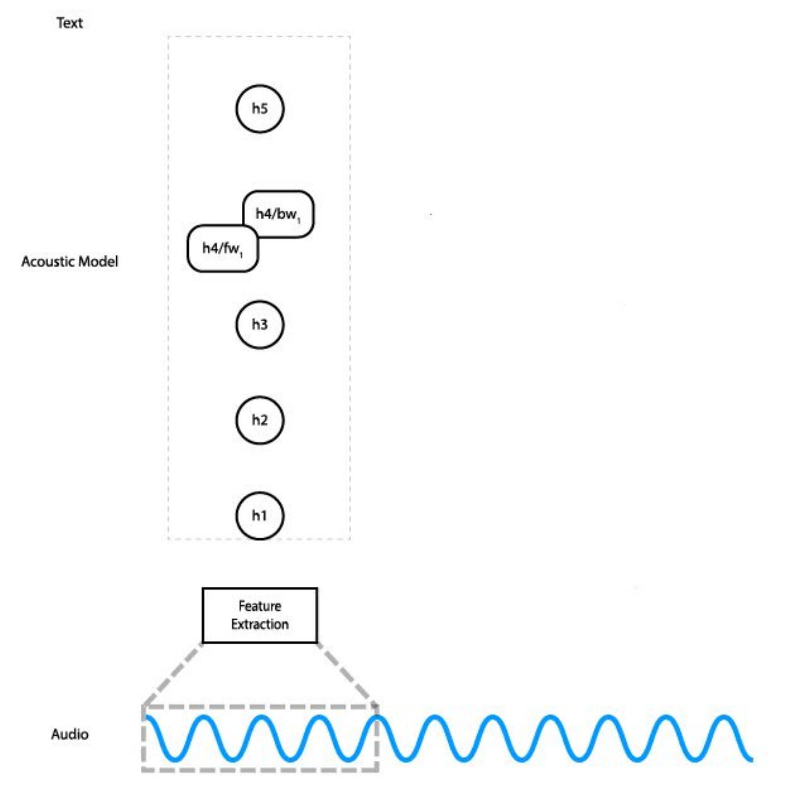
- To get an idea try online on
  playground.tensorflow.org



- Hidden fully connected layers use ReLU activation.
- RNN layer uses LSTM cells with tanh activation
- Additionally it has been used use an Adam optimizer
  arxiv.org/pdf/1412.6980.pdf

The initial bi-directional RNN

hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate

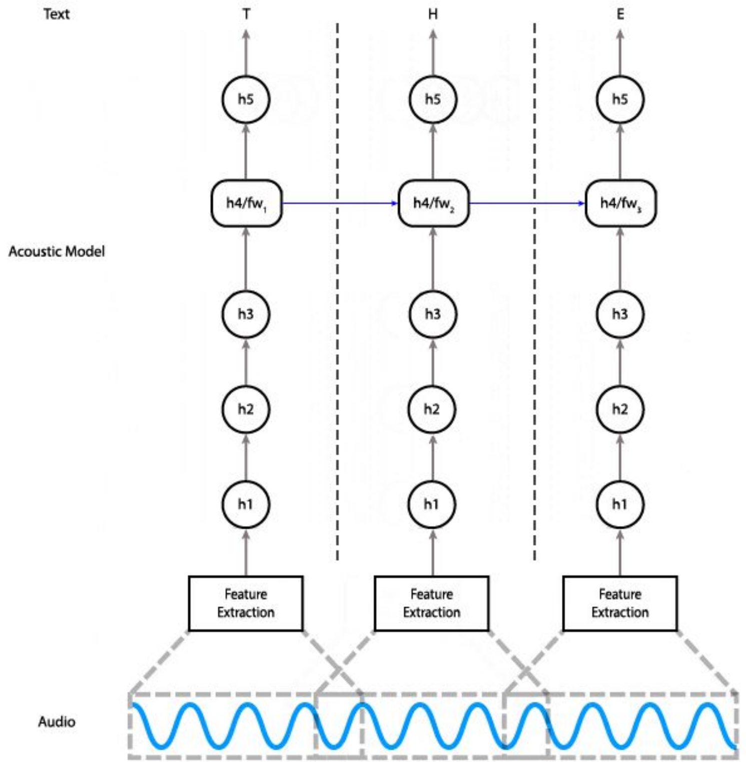


Moving to the unidirectional RNN for streaming

hacks.mozilla.org/2018/09/speech-recognition-deepspeech

"With a **unidirectional model**, instead of feeding the entire input in at once and getting the entire output, you can **feed the input piecewise**.

Meaning, you can input 100ms of audio at a time, get those outputs right away, and **save the final state** so you can use it as the initial state for the next 100ms of audio."



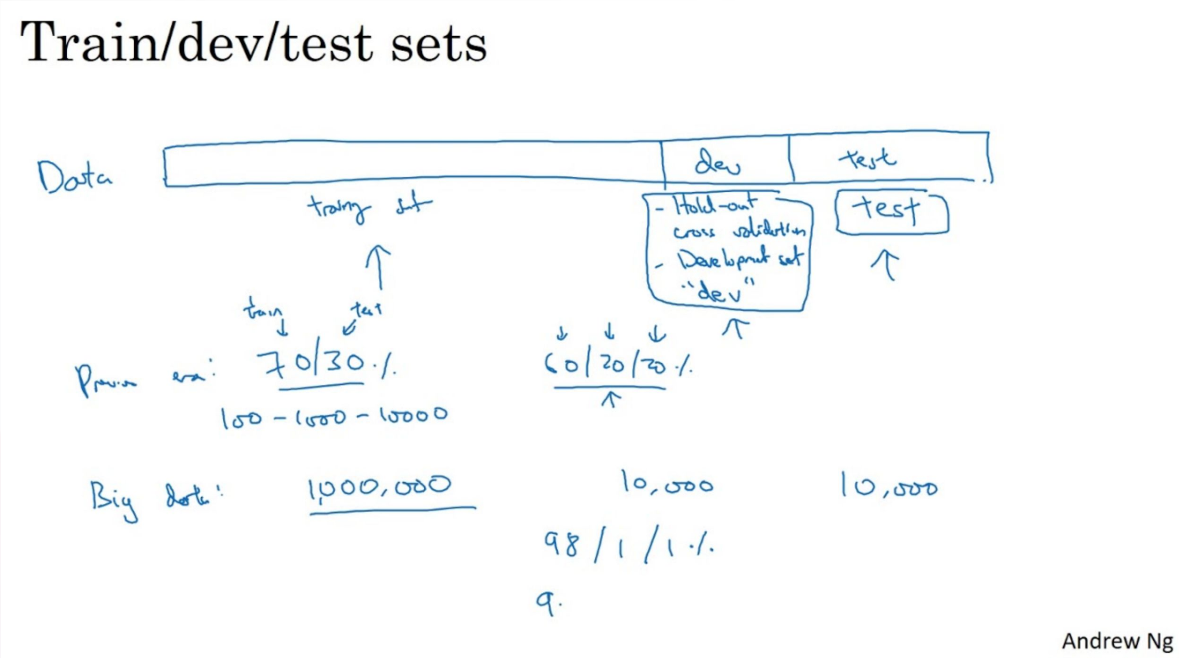
Tuning of the Hyperparameters

The hyperparameters used to train the model are useful for fine tuning.

Latest stable version v0.4.1

- `train_files` Fisher, LibriSpeech, Switchboard training corpora, as well as a pre-release snapshot of the English Common Voice training corpus.
- `dev_files` LibriSpeech clean and other dev corpora, as well as a pre-release snapshot of the English Common Voice validation corpus.
- `test_files` LibriSpeech clean test corpus
- `train_batch_size` 24
- `dev_batch_size` 48
- `test_batch_size` 48
- `epoch` 30
- `learning_rate` 0.0001
- `display_step` 0
- `validation_step` 1
- `dropout_rate` 0.15
- `checkpoint_step` 1
- `n_hidden` 2048
- `lm_alpha` 0.75
- `lm_beta` 1.85

Which dataset to train and test the algorithm?



https://www.coursera.org/specializations/deep-learning

Which dataset to train and test the algorithm?

voice.mozilla.org/en/datasets

LibriSpeech

LibriSpeech is a corpus of approximately 1000 hours of 16Khz read English speech derived from read audiobooks from the LibriVox project.

License: CC-BY-4.0

Go to LibriSpeech

TED-LIUM Corpus

The TED-LIUM corpus was made from audio talks and their transcriptions available on the TED website.

License: CC-BY-NC-ND 3.0

Download Data

VoxForge

VoxForge was set up to collect transcribed speech for use with Free and Open Source Speech Recognition Engines.

License: GNU-GPL

Download Data

Tatoeba

Tatoeba is a large database of sentences, translations, and spoken audio for use in language learning. This download contains spoken English recorded by their community.

License: Mixed

Download Data

Next releases

3 Open ✓ 3 Closed Sort

Deep Speech v0.6.0 Updated 15 days ago	Making the engine only require TFLite [Q1 (March)]
Deep Speech v0.5.0 Updated 14 days ago	Models robust to background noise [Q1 (February)]
Deep Speech v0.4.0 Updated on Jan 7	Reduction in language model size + integrate alternative CTC [Q1 (January)]

github.com/mozilla/DeepSpeech/projects

Stay in touch with DeepSpeech community

On GitHub you can find the common FAQ.

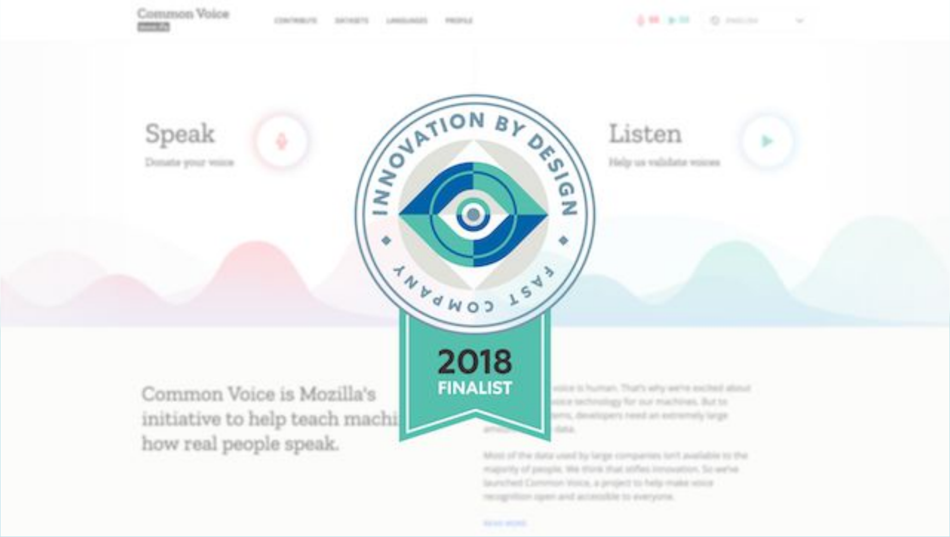
You can read the latest discussions, or add yours on the dedicated Discourse Forums or on the #machinelearning channel on Mozilla IRC for more direct help.

You can open contribute to issues on the repo or open one if not addressed anywhere else.

Common Voice by Mozilla

Creating a dataset of voices as a common

Launched in June 2017, it was named as one of eight finalists in Fast Company's 2018 Innovation by Design Awards Experimental category



Common Voice dataset

The English validated dataset can be downloaded at voice.mozilla.org/en/datasets

12 GB that includes train/dev/test mp3 files and cvs already divided and optimised for DeepSpeech training or your own voice recognition project!

Common Voice dataset

voice.mozilla.org/en/datasets

```
cv.shape
```

```
(3995, 8)
```

```
cv.head()
```

	filename	text	up_votes	down_votes	age	gender	accent	duration
0	cv-valid-test/sample-000000.mp3	without the dataset the article is useless	1	0	NaN	NaN	NaN	NaN
1	cv-valid-test/sample-000001.mp3	i've got to go to him	1	0	twenties	male	NaN	NaN
2	cv-valid-test/sample-000002.mp3	and you know it	1	0	NaN	NaN	NaN	NaN
3	cv-valid-test/sample-000003.mp3	down below in the darkness were hundreds of pe...	4	0	twenties	male	us	NaN
4	cv-valid-test/sample-000004.mp3	hold your nose to keep the smell from disablin...	2	0	NaN	NaN	NaN	NaN

Common Voice dataset

voice.mozilla.org/en/datasets

```
cv.accent.value_counts(dropna=False)
```

```
array([nan, 'us', 'england', 'scotland', 'african', 'indian', 'canada',
       'ireland', 'philippines', 'australia', 'newzealand', 'hongkong',
       'wales', 'southatlandtic', 'malaysia', 'singapore', 'bermuda'],
      dtype=object)
```

```
cv.age.value_counts(dropna=False)
```

```
NaN          2453
twenties      466
thirties      389
fourties      236
fifties       205
teens         117
sixties        88
seventies      36
eighties        5
Name: age, dtype: int64
```

What you can do?

How can you contribute?

Da luglio 2018 Common Voice è anche in italiano!

voice.mozilla.org/it

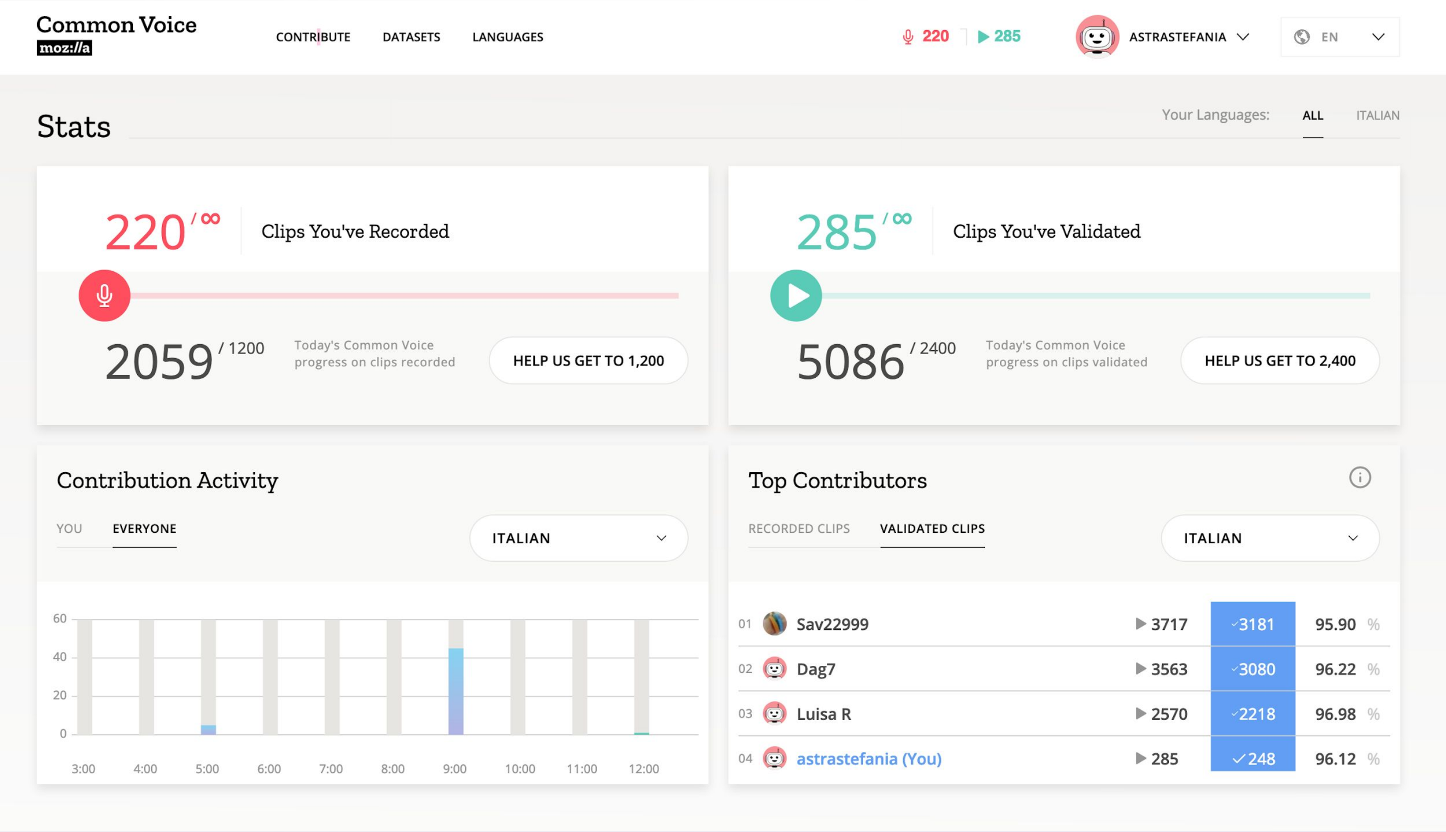
How can you contribute? Parliamoci chiaro!

voice.mozilla.org/it

How can you contribute? Validiamo ascoltando!

voice.mozilla.org/it

New dashboard and ranking



Who's using DeepSpeech and Common Voice?

Some example:

- Mozilla IoT: Experimental voice assistant for Web of Things Gateway
- Xaero: Grammar checker
- Mycroft AI: used Common Voice dataset to build an open source voice assistant
- You?

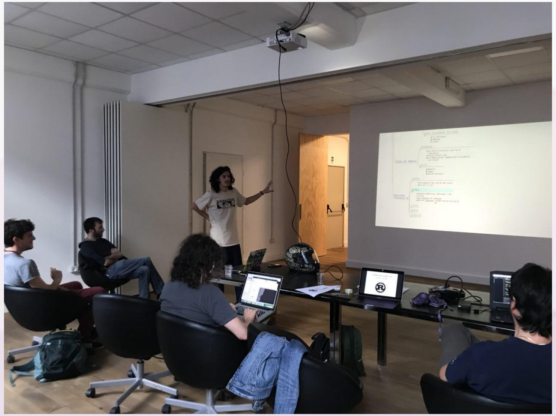
So to summarise...

- You can donate your voice in a bunch of different languages
- In italiano puoi contruibuire sia nelle traduzioni che alla frasi stesse
- You can use DeepSpeech and Common Voice dataset for your own project
- DeepSpeech is still evolving, you can get in touch with core developers and contribute via GitHub

Let's explore more together, rivediamoci!

- 28 Feb Gruppo di studio Rust
- 7 Mar Open Mozilla Night: proviamo DeepSpeech?

meetup.com/Mozilla-Torino



Other resources and credits

- Un grande ringraziamento a tutta la comunità italiana di Mozilla raggiungibile su Telegram **Mozilla Italia - HOME** o sul forum forum.mozillaitalia.org

- Slides background from voice.mozilla.org

Thank you!

Stefania Delprete

@astrastefania