



# Common Voice & Deepspeech

## Il riconoscimento vocale Open-Source nella lingua più bella del mondo

Seminario di Informatica Musicale

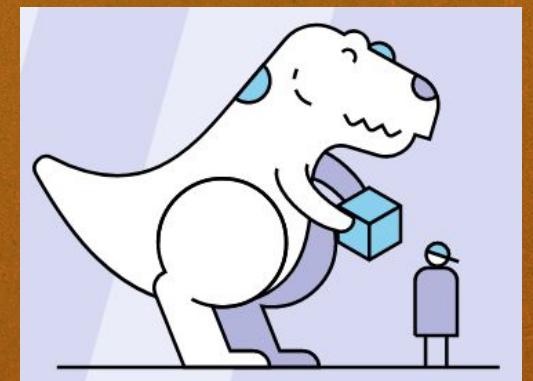
---

Dario Camonita

# Manifesto Mozilla

- Internet è una risorsa pubblica globale che deve rimanere aperta e accessibile.
- Internet deve arricchire la vita di ogni essere umano.
- L'efficacia di Internet come risorsa pubblica dipende dal suo carattere di interoperabilità (relativamente a protocolli, formati di dati, contenuto), dal suo grado di innovazione e dalla partecipazione decentralizzata a livello mondiale.
- La partecipazione commerciale allo sviluppo di Internet è in grado di apportare numerosi benefici, ma è fondamentale un equilibrio tra profitto commerciale e benefici pubblici.
- La valorizzazione degli aspetti di pubblica utilità di Internet rappresenta un obiettivo importante, che merita tempo, attenzione e impegno.

<https://www.mozilla.org/it/about/manifesto/>

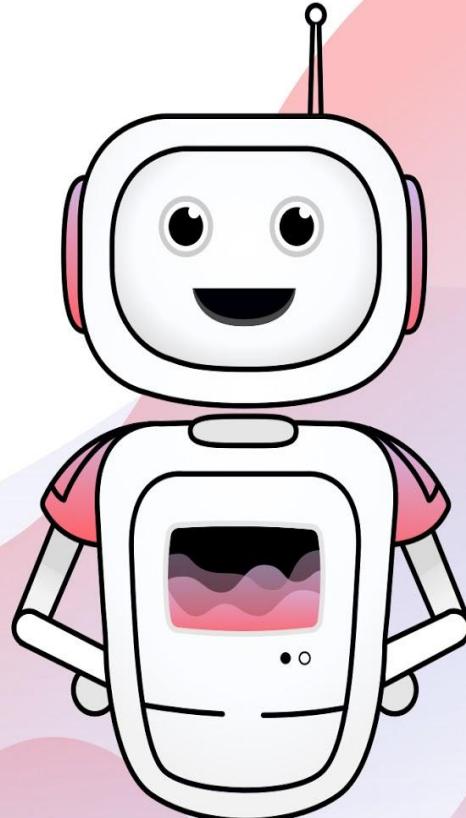


# Common Voice

[voice.mozilla.org](https://voice.mozilla.org)

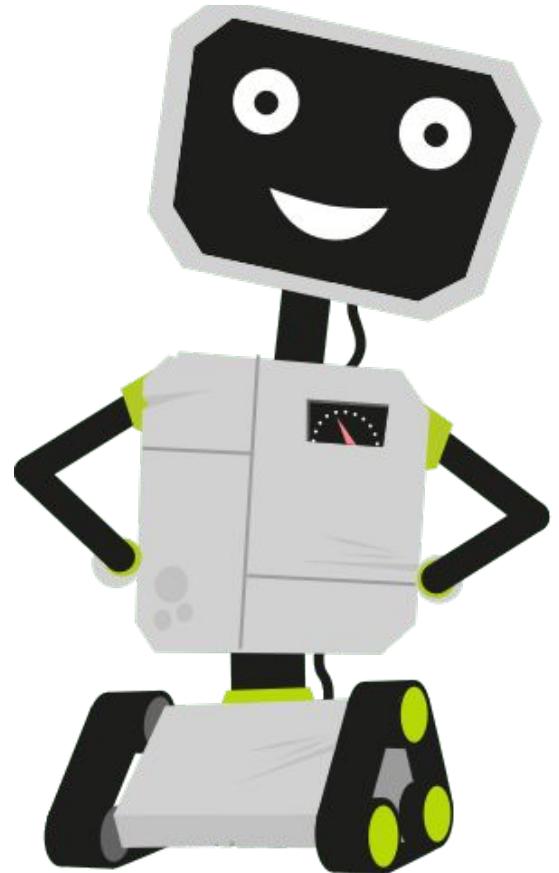
Progetto di Mozilla incentrato sul raccoglimento di registrazioni vocali, per poter creare un riconoscitore vocale basato su un dataset e modello di pubblico dominio per tutte le lingue che partecipano al progetto.

Il modello, basato sul deep learning, è implementato dal progetto DeepSpeech di Mozilla.



# Perché è nato il progetto?

- Contrastare la centralizzazione dei motori Speech-To-Text proprietari
- Contrastarli perché funzionano solo online con problemi legati alla privacy degli utenti
- Rendere il riconoscimento vocale disponibile per tutte le lingue in maniera open source
- Anche per le lingue con meno interesse commerciale
- Per non avere limitazioni sull'utilizzo di questi servizi

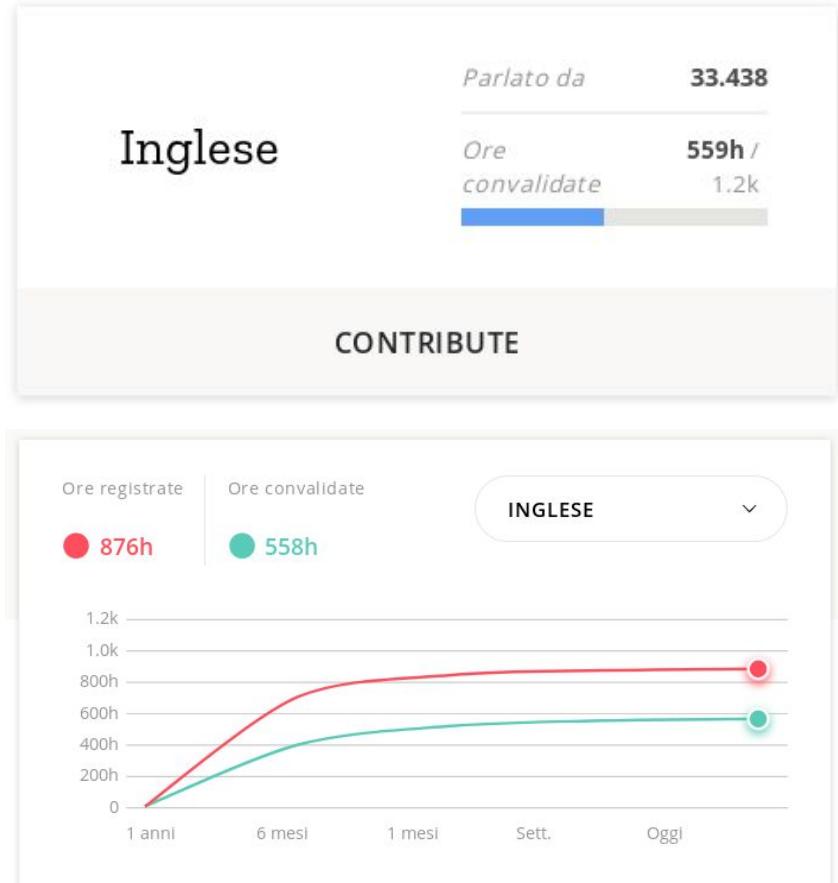


# Stato del progetto per la lingua inglese

- 20000 partecipanti nel primo anno
- 70 GB di dataset
- WER versione 0.1.0: **5.6%** su LibriSpeech ASR (Automatic Speech Recognition)
- WER versione 0.3.0: **11%**

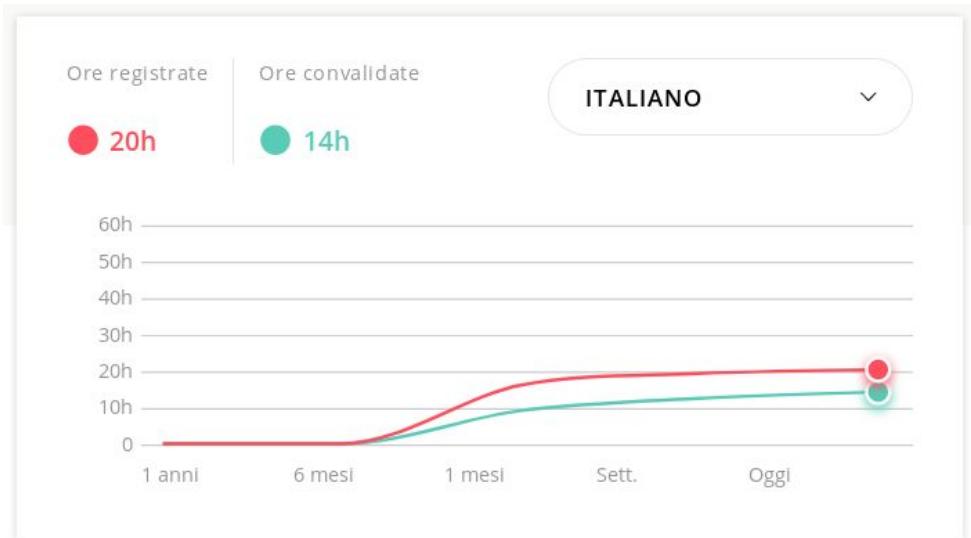
<https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/>

<https://hacks.mozilla.org/2018/09/speech-recognition-deepspeech/>



# Stato del progetto per la lingua italiana

- Oltre 10000+ frasi da diverse fonti
- Avviato il 3 Luglio 2017
- Promozione:
  - Social network
  - Linux Day
  - Italian Hacker Camp 2018
- TU!

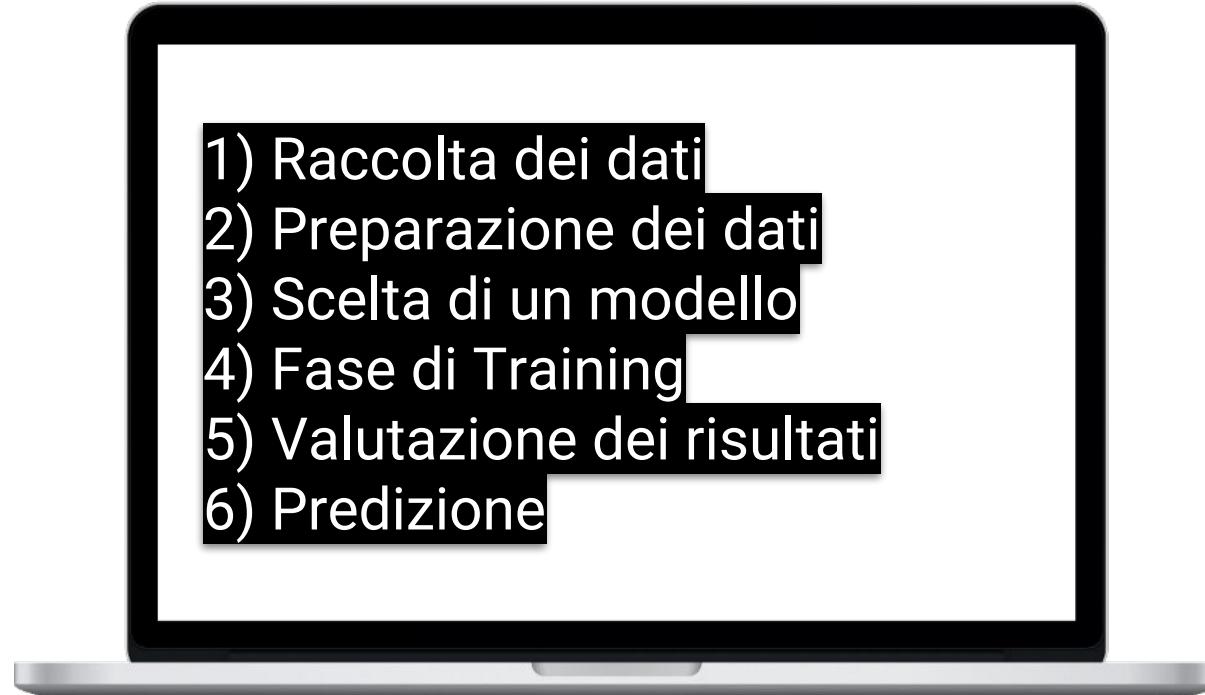


# Cos'è il Machine Learning?

È quel processo che riguarda lo studio, la costruzione e l'implementazione di algoritmi che permettono ad un qualsiasi calcolatore sul quale sono implementati di **imparare e fare previsioni in modo automatico** partendo da un insieme di dati in ingresso (files di testo, musicali, immagini, etc...).



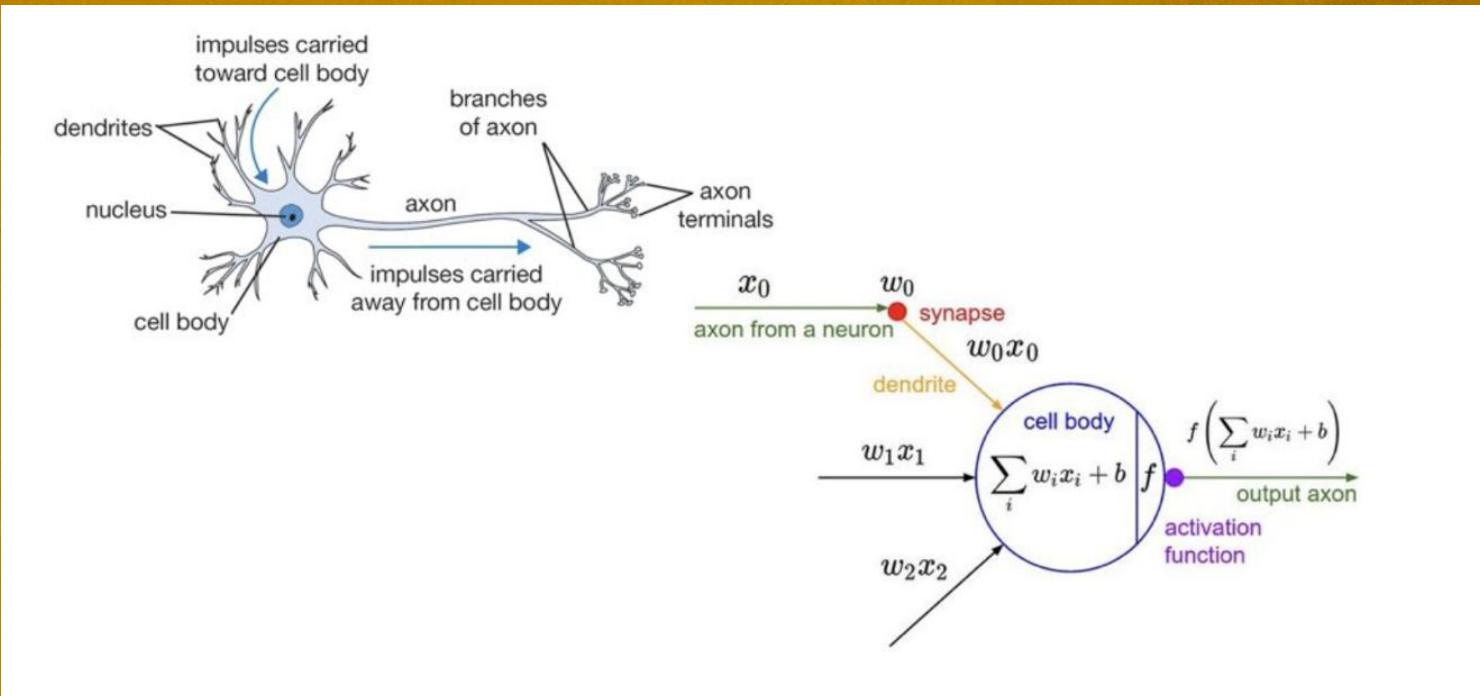
# Come funziona il Machine Learning?



# ANN, Rete Neurale Artificiale

"Una rete neurale è un modello di calcolo la cui struttura stratificata assomiglia alla struttura della rete di neuroni nel cervello, con strati di nodi connessi. Una rete neurale può apprendere dai dati, quindi può essere addestrata a riconoscere pattern, classificare i dati e prevedere eventi futuri."

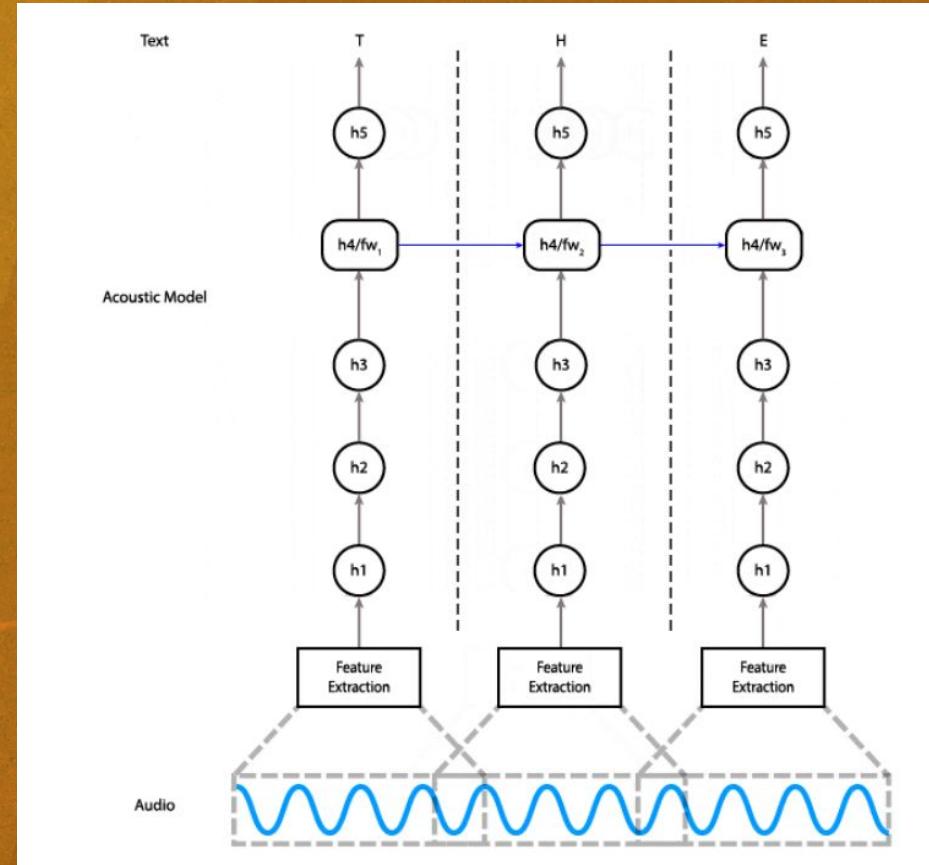
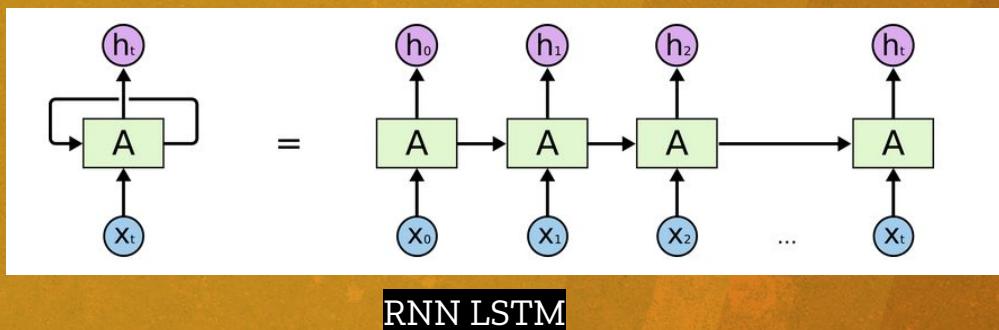
<https://it.mathworks.com/discovery/neural-network.html>



# RNN, Reti Neurali Ricorrenti

Le RNN sono una classe particolare delle ANN, le cui connessioni tra i nodi permettono di dare persistenza alle informazioni ricevute in input.

Ciò gli consente di mostrare un comportamento dinamico temporale per una determinata sequenza temporale.



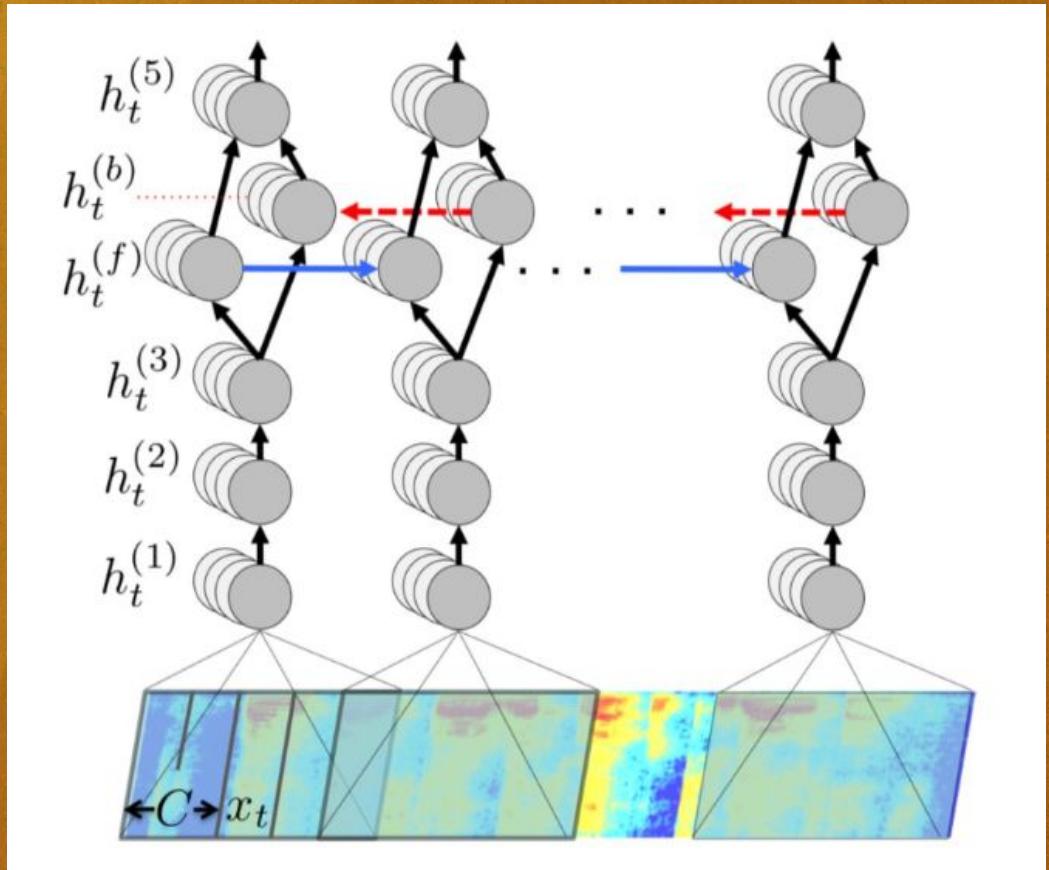
Struttura della RNN utilizzata in Deepspeech 0.3.0

# BRNN, Reti Neurali Ricorrenti Bidirezionali

- 3 forward-layers
- 1 bidirectional recurrent layer
- 1 forward-layer

Perché in DeepSpeech 0.3.0 sono state usate le RNN?

- Incremento del VAD (Voice Activity Detection)
- Incremento del WER (Word Error Rate)
- Far fronte ai problemi di disturbo/rumore sul dataset Common Voice, causati dai dispositivi di registrazione
- Test su dataset clean



Struttura della BRNN utilizzata in DeepSpeech 0.1.0

# Progetto DeepSpeech

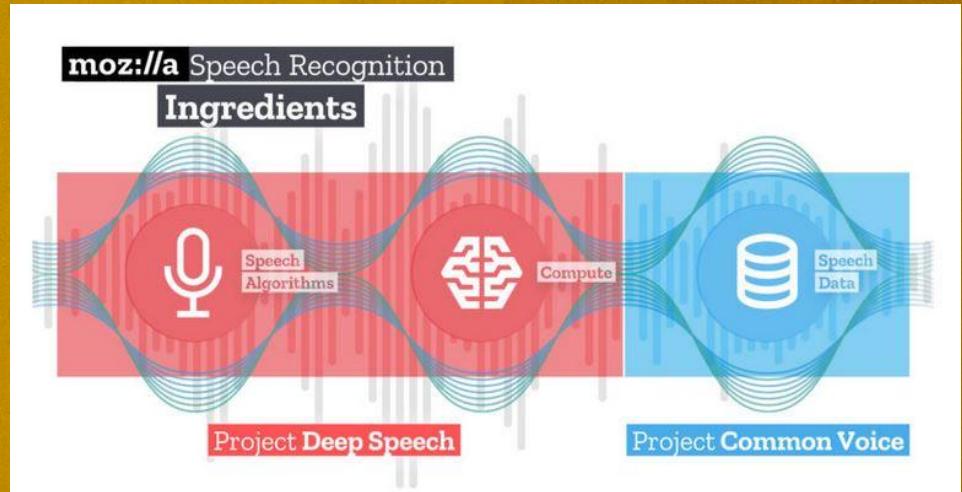
Motore di riconoscimento vocale, open source e sviluppato da Mozilla, siamo alla versione 0.3.0

Utilizza un modello TensorFlow, basandosi sul paper di Baidu:

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

i quattro punti fondamentali per un algoritmo speech-to-text:

- Deep Learning attraverso le RNN (reti neurali ricorrenti)
- Parallel processing
- Un grande dataset per il training
- Complessità delle pipelines nulla

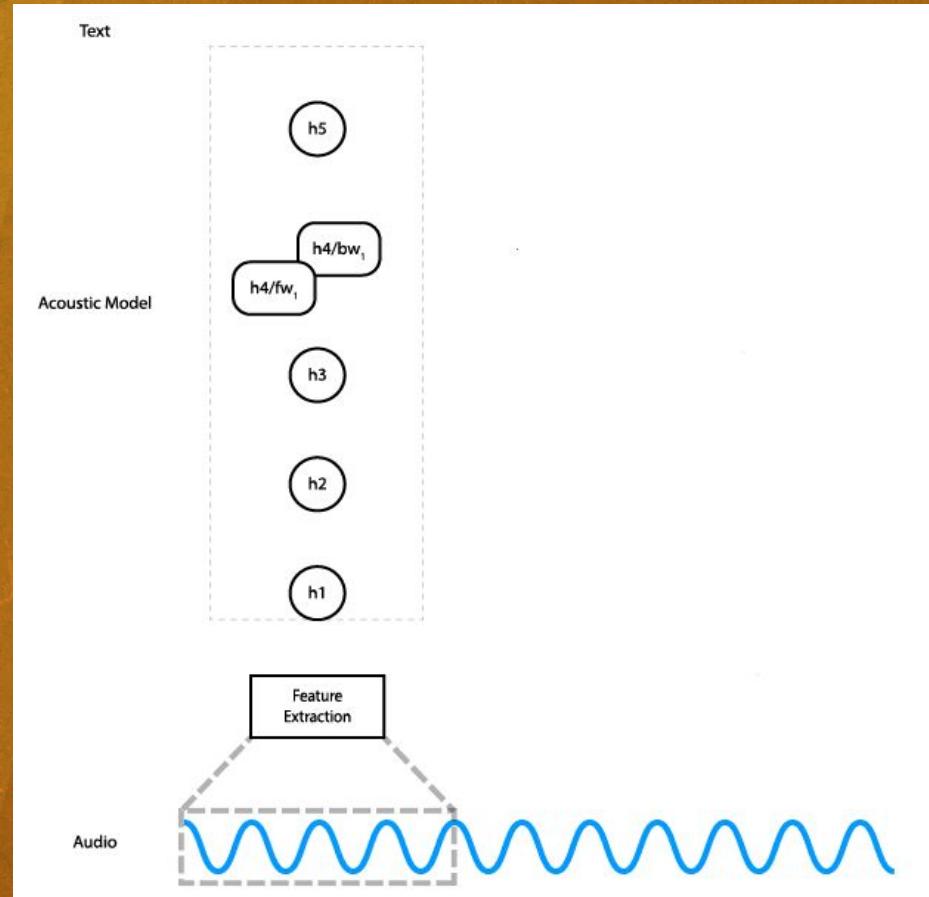


<https://github.com.mozilla/DeepSpeech>

# Sviluppo - DeepSpeech

Si tratta di un modello TensorFlow realizzato in Python, con "semplici" API che permettono di svolgere il riconoscimento vocale

- Utilizzo della *BRNN* (*bidirectional recurrent neural network*) nella versione 0.1.0, e della *RNN* dalla versione 0.2.0
- Ottimizzazione degli *iperparametri*
- Creazione di un *dataset* per il training



# Sviluppo - DeepSpeech

```
((demo) Kellys-MacBook-Pro:DeepSpeech kdavis$ pip install deepspeech
```

Obiettivi:  
WER < 10%

Implementazione delle  
reti neurali tramite  
TensorFlow

Personalizzazione degli  
iperparametri

Creazione di un dataset

# Sviluppo - DeepSpeech

Sono presenti bindings nativi per

- Node.js

<https://github.com/mozilla/DeepSpeech#using-the-nodejs-package>

- Python

<https://github.com/mozilla/DeepSpeech#using-the-python-package>

- Rust

<https://github.com/RustAudio/deepspeech-rs>

- Go Lang

<https://github.com/asticode/go-astideepspeech>

- GStreamer

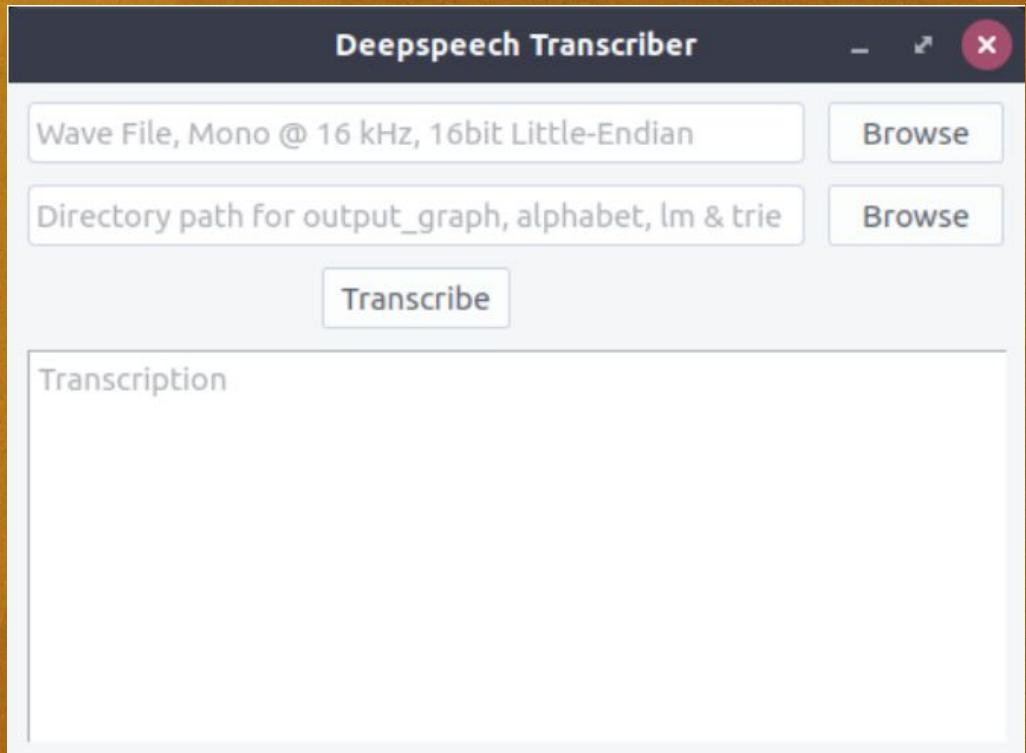
<https://github.com/Elleo/gst-deepspeech>

Documentazione:  
<https://deepspeech.readthedocs.io/en/latest/>



# Modello - DeepSpeech

- già scaricabile, versione 0.3.0  
stable, 0.4.0 alpha
- disponibile il pre-trained in  
inglese (~2 GB)
- Affidabilità maggiore del 94%  
(Novembre 2017)
- Interfaccia grafica in Qt 5



# DeepSpeech | Text-to-speech

Basato su Tacotron: A fully end-to-end text-to-speech Synthesis model

Utilizzo del framework di encoding-decoding seq2seq per Tensorflow

Discreti risultati col dataset LJ Speech contenente circa 24 ore di brevi registrazioni audio di un singolo speaker



Deepspeech TTS

★ Deepspeech Demo × +

localhost:3000

Cerca

# DEMO DeepSpeech

Provalanche tu  
DeepSpeech!

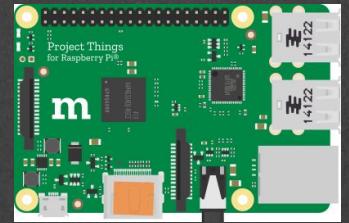
Listen

okay

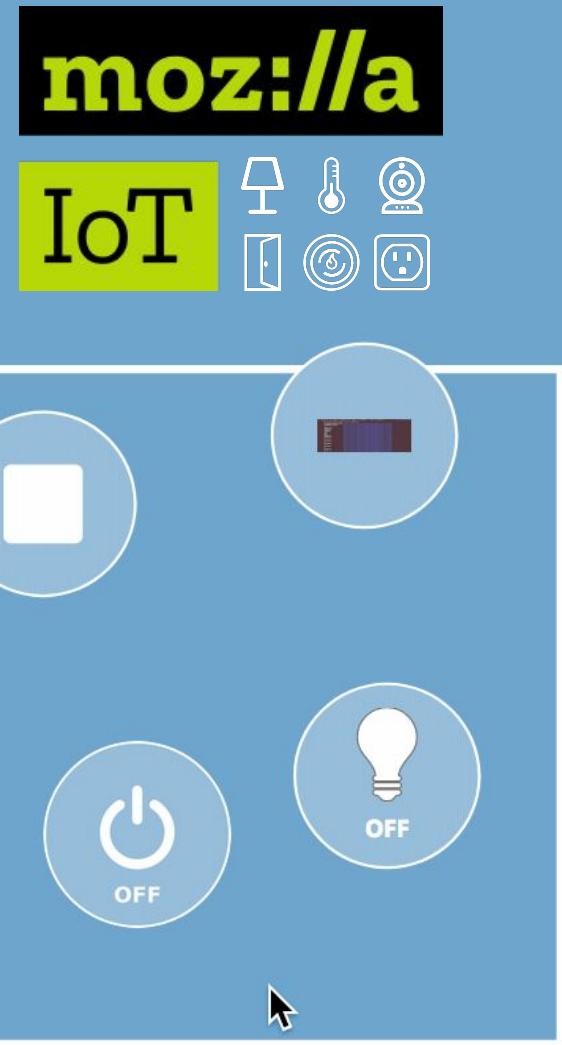
This screenshot shows a DeepSpeech demo interface running in Mozilla Firefox. The page title is "Deepspeech Demo". The URL in the address bar is "localhost:3000". The search bar contains the placeholder "Cerca". The main content features a large title "DEMO DeepSpeech" in a bold, sans-serif font. Below it is a black rectangular area containing a white waveform visualization of audio input. To the right, there is a call-to-action text "Provalanche tu DeepSpeech!" with an exclamation mark. At the bottom center is a red button labeled "Listen". A dark gray footer bar at the bottom contains the word "okay". The background is a complex, abstract pattern of overlapping white wireframe meshes forming spheres and cubes. A QR code is located on the right side of the page.

# Utilizzi - DeepSpeech/Common Voice

- Mozilla IoT:  
assistente sperimentale per il  
Web of Things Gateway
- xaero:  
correttore grammaticale
- Mycroft AI:  
uso del dataset di Common Voice  
per lo sviluppo di un assistente  
vocale completamente Open Source



- Piattaforma IoT Open Source
- Things Cloud per la registrazione del proprio WoT Gateway tramite il protocollo TLS
- Things Framework, snippet per la realizzazione di dispositivi direttamente connessi al Web of Things
- Things Gateway, implementazione Open Source per il collegamento dei dispositivi IoT nel web
- Things UI per il monitoraggio e il controllo di tutti i dispositivi smart tramite un'unica interfaccia web
- Ultima release 0.6



<https://iot.mozilla.org/gateway/>

# Come contribuire a Common Voice?

## Dona delle frasi

- Lunghezza max 125 caratteri a frase
- No linguaggio volgare e dialettale
- Grammatica e punteggiatura corretta
- Iniziale maiuscola in ogni frase



## Registra la tua voce

- Registrazioni massimo di 20 secondi
- Leggere i numeri correttamente
- Leggere tutto!
- Non preoccuparti del tuo accento!

## Revisiona le registrazioni

- Deve contenere tutto il testo
- Deve essere comprensibile/udibile



IHC 2018

# Come partecipare a Common Voice?

Vai su: [voice.mozilla.org/it](https://voice.mozilla.org/it) e inizia da subito  
**QUALUNQUE DISPOSITIVO TU ABBIA!**

Parla

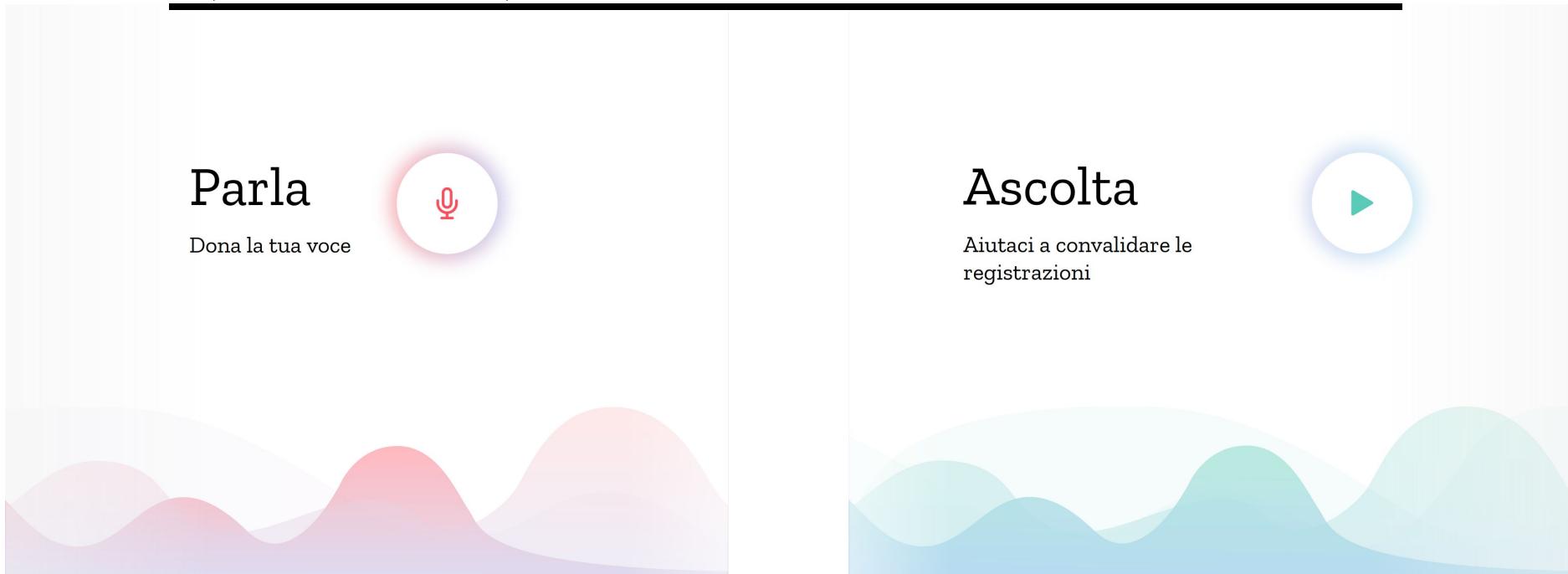


Dona la tua voce

Ascolta



Aiutaci a convalidare le  
registrazioni



# Grazie per l'attenzione!

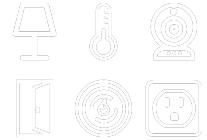
Contatti

Telegram: @dariocam

Email: dario.camonita@aol.com

GitHub:

<https://github.com/danterolle>



Informatica Musicale  
(6 CFU)