Advanced user interfaces

AY 2020/2021

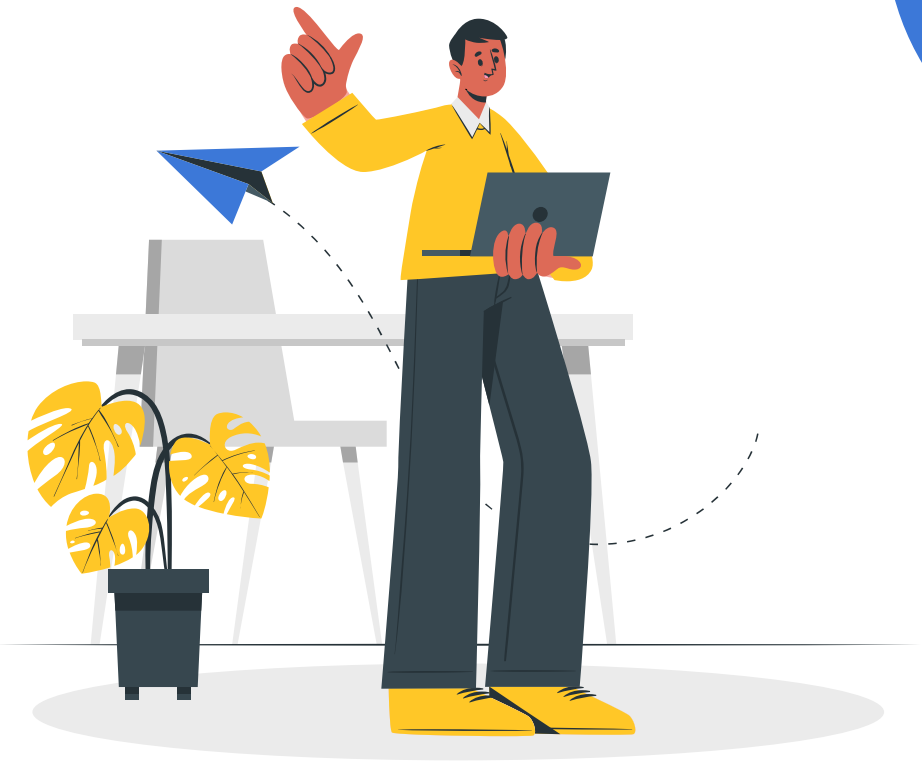
Speech-To-Text module

Aida Gasanova aida.gasanova@mail.polimi.it Computer Science and Engineering
Max Griesmayer max.griesmayer@mail.polimi.it,Erasmus - Computer Science and Engineering
Victor Mouradian victor.mouradian@gmail.com, Erasmus - Computer Science and Engineering

The problem

Main Target Group(s), Context, Needs, Constraints and Goals, State of the art

01



DLD - Developmental Language Disorder

DLD increases the risk of a range of negative impacts on education, employment, and social and emotional problems

DLD affects 8% of children

Speech and language therapists (SLTs) teach strategies to children with DLD and those around them, which aim to reduce the impact of their difficulties

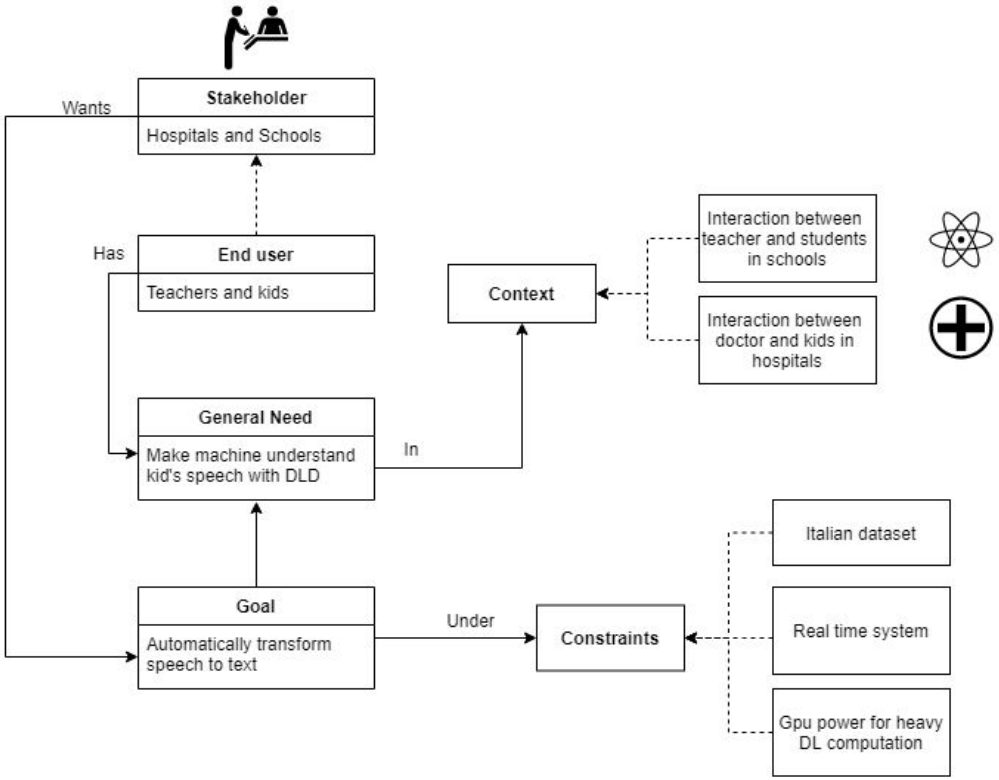
The main target groups

Children with DLD

The therapists working with the children

The parents or other relatives of the child

Context, needs, constraints and goals



State of the Art- Proprietary

Comparison of the APIs for Speech Processing					
	API	Tasks supported	Main details	Languages supported	Results quality
amazon	Transcribe	Speech to text converting	Punctuation and formatting, telephony audio, customization and multiple speakers recognition	English Spanish	GOOD
	Polly	Text to speech converting	Real-time mode, pronunciation, volume, pitch, speed rate, etc customization	27 + dialects	EXCELLENT
Google Cloud	Speech API	Speech to text converting	Customization, batch and real-time modes, noise robustness, filters for wrong words relative to the context, flexibility in the source files storage	120	INTERMEDIATE
IBM Watson	Speech to Text	Speech to text converting	Real-time mode, custom models, keywords spotting, speaker labels (in beta), word confidence, word timestamps, profanity filtering, word alternatives, smart formatting (in beta)	11	GOOD
	Text to Speech	Text to speech converting	Pronunciation customization, custom words, expressiveness, word timings	8 + dialects	EXCELLENT
Microsoft Azure	Bing Speech API	Speech to text converting	Real-time mode, customization, formatting, profanity filtering, text normalization, integration with Azure LUIS, speech scenarios	10 conversational mode 29 + dialects interactive and dictation modes	GOOD
		Text to speech converting	Pronunciation, volume, pitch etc customization	78 + dialects	EXCELLENT

Open Source - Data Privacy

Transparency

- full visibility into the code
- allows discussions about the development
- protected against lock-in risks

Independency

- Ability to make changes & add additional features
- Deploy model wherever you want

Privacy

- Be the owner of data, model and results

State of the Art - Open source

DeepSpeech is an open-source speech to text engine which can run in real-time using a model trained by machine learning techniques based on Baidu's Deep Speech research paper and is implemented using Tensorflow.

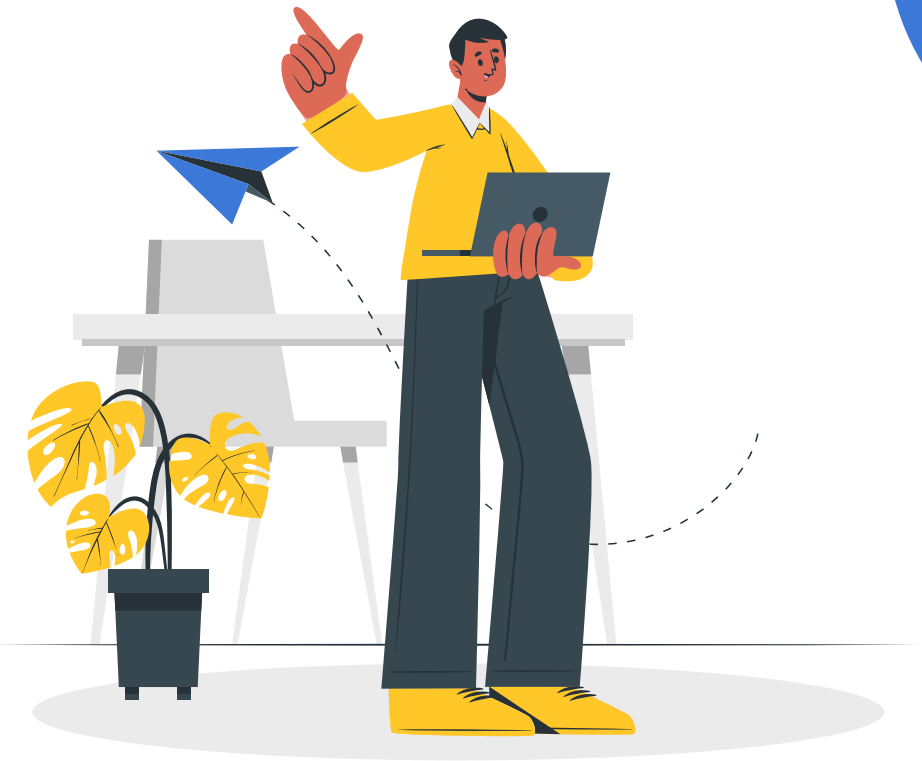
Leon is an open-source personal assistant who can live on your server and is able to perform task when you ask him to.

Wav2letter++ is Facebook AI Research's end-to-end Automatic Speech Recognition Toolkit written entirely in C++, supporting a wide range of models and learning techniques.

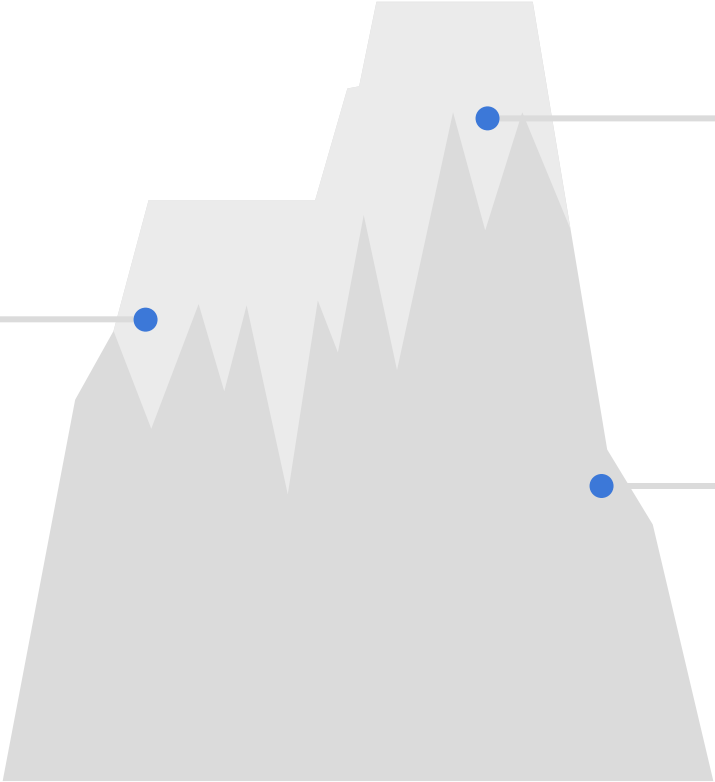
The solution

Design process, «Concept», Interaction paradigms, Details of interaction and interfaces, Scenarios, Technology

02



Concept



What?

Development of an Italian STT module which is able to understand the speech of children with DLD

How?

Achieve this by using deep learning using available Italian speech data from the internet

Interaction

Model that can be used locally offline as online.

Concept & User Scenarios

Giovanna 45 y.o is (teacher)

is helping some students with DLD. She wants to understand the typical mistakes in an automatic way. She uses our module to achieve her goal

Marco 30 y.o. is at the library

receives a voice message and she doesn’t have headphones. Then he is using our Speech to Text module to transcribe the voice message into the text

The solution

We ended up choosing deepspeech as the base layer of our project

We used colab and jupyter notebooks to implement our solution

Constraints

- Easy to use
- Real time results

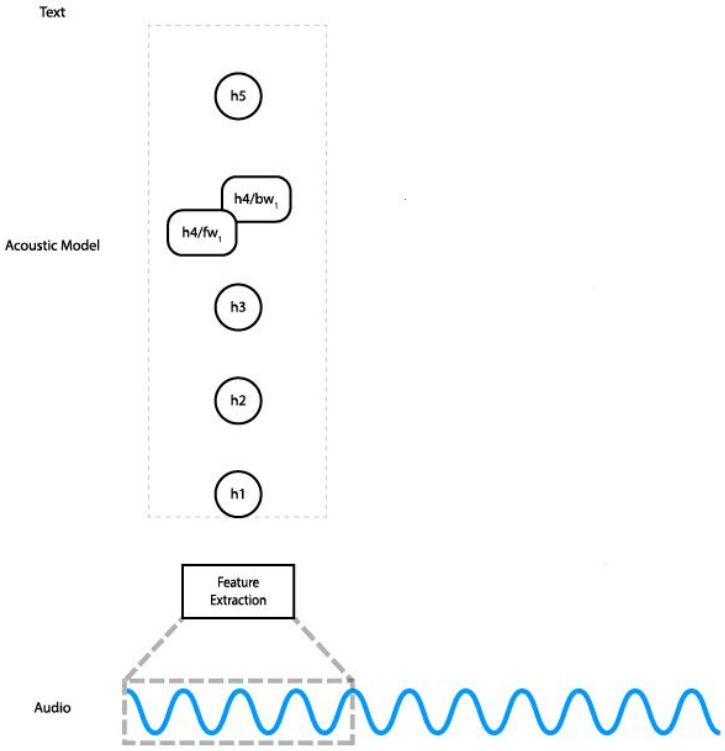
Why DeepSpeech?

- Well documented Open Source project
- Easy to implement examples (python, js, c#, android…)

Deepspeech

Underlying technology

- The deepspeech model follows the Baidu Deep Speech Research paper
- Implemented using Tensorflow



DeepSpeech implements a Recurrent Neural Network

DLD Data

How did we transform the DLD data into something usable by DeepSpeech

```
@UTF8
@Begin
@Languages:     ita
@Participants:  CHI Y_2_18 Child, OBS Cerutti Observer
@ID:    ita|Narrative|CHI|6;11.|female|bilingue_1||Child|||
@ID:    ita|change_corpus_later|OBS||female|||Observer|||
@Time Duration:  00:05-01:11
*CHI:   che c' era un topo e un cane.
*CHI:   un cane voleva prendere il topo.
*CHI:   ma il topo si a +//.
*CHI:   il cane non er [//] era stato to abbastanza veloce e così andò
        contro un albero infatti si fece molto male.
*CHI:   il bambino si_spaventò_col [//] si spaventò e la [//] scivolò il
        palloncino.
*CHI:   il palloncino era sull' albero.
*CHI:   il cane aveva pensato di mangiare le salsicce.
*CHI:   poi il bambino saltò in_al [//] in alto e poi era felice di aver
        recuperato il suo amato palloncino.
@End
```

```
extracted/M_11_17/1-m_11_17.wav, 709676,  giorno un cane vede un topo seduto vicino a un albero e decise di mangiarlo.
extracted/M_11_17/2-m_11_17.wav, 709676,  bambino tornato dalla spesa che ha comprato +//.
extracted/M_11_17/3-m_11_17.wav, 709676,  bambino allegro prese dalla spesa un pacco di salsicce e un palloncino.
extracted/M_11_17/4-m_11_17.wav, 709676,  vide il cane cercando [//] cercava [//] che cercava di prendere il topo.
extracted/M_11_17/5-m_11_17.wav, 709676,  non_f [//] non [//] non era troppo veloce e il cane si fa male.
extracted/M_11_17/6-m_11_17.wav, 709676,  bambino s' è spaventato e ha lasciato andare il suo palloncino.
extracted/M_11_17/7-m_11_17.wav, 709676,  poi decise il bambino di riprenderlo.
extracted/M_11_17/8-m_11_17.wav, 709676,  lora faccio un salto per riprenderselo.
extracted/M_11_17/9-m_11_17.wav, 709676,  ntre il cane si_man [//] mangia le suo salsicce che ha lasciato.
extracted/M_12_17/01-m_12_17.wav, 661292,  era il cane un po' dispettoso che voleva prendere un topo.
extracted/M_12_17/02-m_12_17.wav, 661292,  era il cane che voleva prendere un topo che è molto dispettoso.
extracted/M_12_17/03-m_12_17.wav, 661292,  [//] dopo [//] lo stava per prendere ma però ha sbattuto nell' albero e il topo era
troppo veloce.
extracted/M_12_17/04-m_12_17.wav, 661292,  quindi passò un bambino con le salsicce in mano e un palloncino.
```

Initial Data

LD_audio_ITA (806mo)

- 94 audio recordings
- 1411 Kbit/s.

Target Data

Extracted_audio_wav (554 mo)

- 848 audio recordings
- 16000 Kbit/s
- Mono-Channel

Data cleaning

How did we transform the DLD data into something usable by DeepSpeech

Two python scripts

- **extractaudio.py:** Extracts audio from the conversations, Extract the phrases for each conversation
- **convertwav.py:** Converts the extracted file into the right format

Python library used

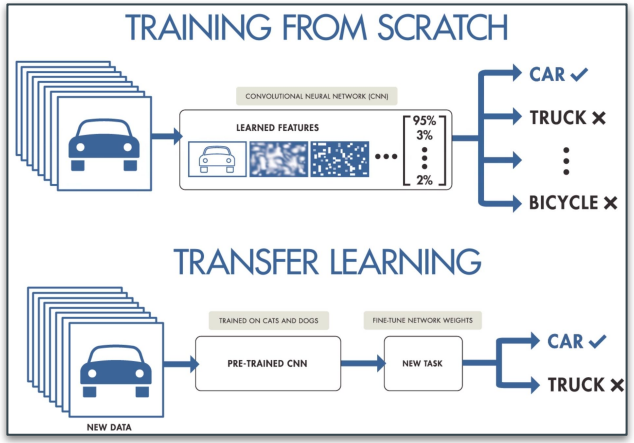
- **AudioClipExtractor:** This utility allows one to cut multiple clips from a single or multiple audio files
- **PyDub:** Manipulate audio with a simple and easy high level interface

Training

How did we train a DeepSpeech model

English to Italian (Transfer Learning)

- Used a jupyter notebook on colab
- Two datasets:
  - Mozilla Common Voice (130h - 4 GB)
  - The M-AILABS Speech Dataset (127h 40m - 14 GiB)
- The size of the data allows for convincing results



Regular Italian to DLD (Fine tuning)

- Allows for fast results for a small size dataset
- Heavily dependent on data quality
- Due to the quality of the data the overall performance of the model went down

Demo

Demo time !

chrome://flags/#unsafely-treat-insecure-origin-as-secure

http://35.181.8.180:3000/

Value proposition

Best as we could for an open source Speech To Text module

A good groundwork solution

Easily iterable and upgradable with the colab notebooks

Deepspeech has a lot of easily implementable examples

Never going to be better than paid services (<< data)

Limitations and Challenges

Data

- Data Quality: noisy data..
- Data size:: not enough data

Hardware

- Computational Power
- Memory
- File space

Language

Data Gathering Tool

Build a tool that makes it easier to gather DLD Data for therapists

To reach the goal of developing a functional DLD STT model, it is necessary to gather a lot of high quality data.

What are the benefits:

- Easy to use
- Getting the data in the right format
- Gamification
- Use different sentences
- Use data to automatically retrain the model
- Build dataset for research