THE DEVELOPER'S CONFERENCE

O que a IA pode extrair dos arquivos de **áudios**

(Audio insights)

Diego Augusto Silva • Data Scientist • diegosilva@ciandt.com

linkedin.com/in/diegoaugusto/

Julho • 2019

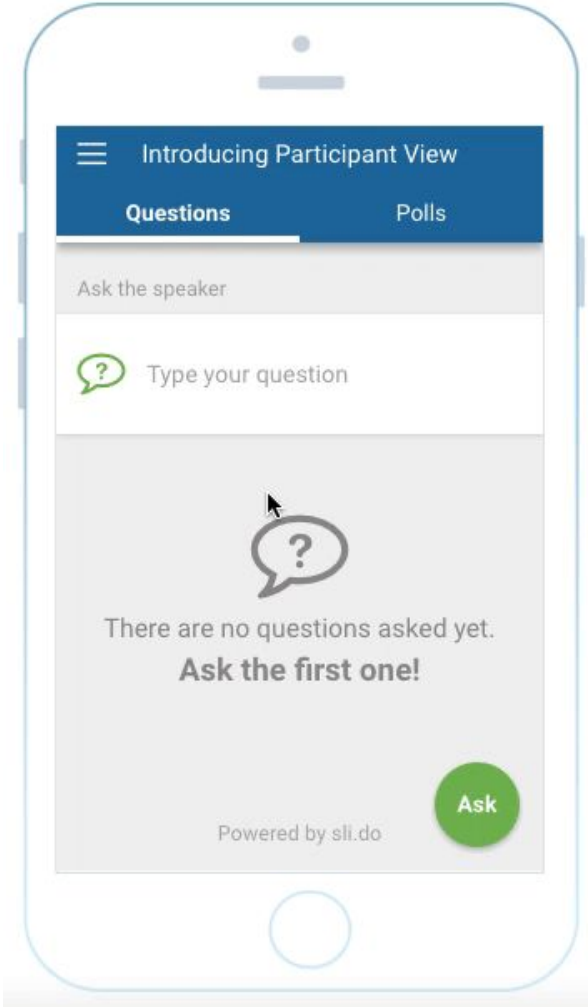
The next **20 minutes:**

- Challenges
- Databases
- Applications

The next **20 minutes:**

- Challenges
- Databases
- Applications

Deep neural nets' result in state of the art performance!



sli.do

Código: **iaudio**

Google play

Available on the App Store

Friendly reminder

sli.do
Código: iaudio

Audio: Refers to an electrical signal (recorded).
Sound: When electrical signal is converted into audible acoustical pressure.
Speech/Voice: Sound produced by a person's.

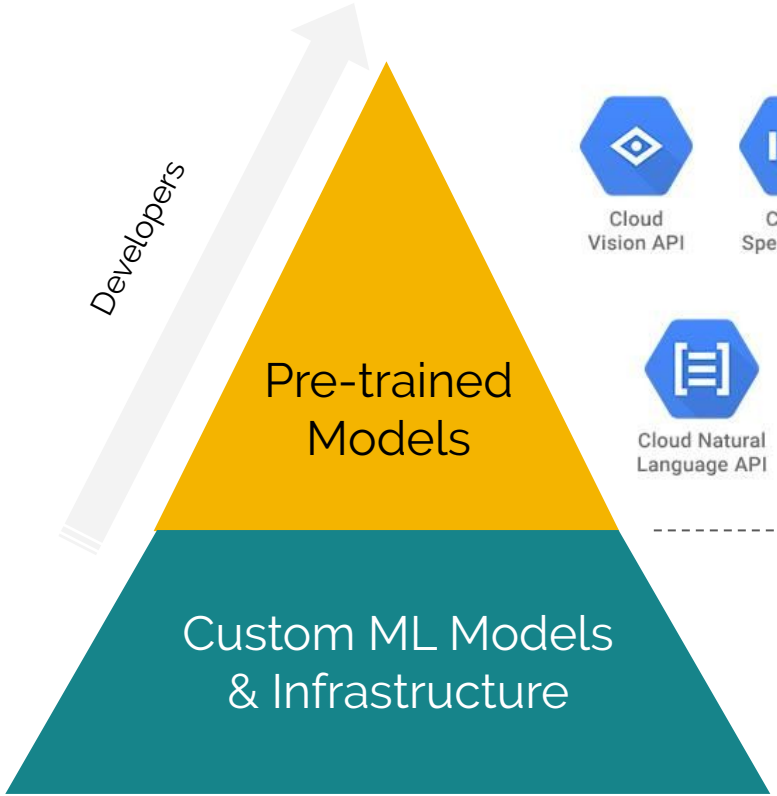
Speech/Voice AI:
-Recognition (ASR)
-Synthesizer (TTS)
-Biometric
-Translation

Music AI:
-Fingerprinting
-Genre
-Make

Sound AI:
-Scene Recog.
-Event Recog.

Framework and API

sli.do Código: iaudio



Google

Cloud Vision API

Cloud Speech API

Cloud Jobs API

Cloud Translation API

Cloud Natural Language API

Cloud Video Intelligence API

Amazon

Polly

Transcribe

IBM Watson

SPEECH TO TEXT

TEXT TO SPEECH

VISUAL RECOGNITION

Azure

Computer Vision

SpeechToText

Language Understanding

Google ML Engine

Azure ML Studio

Amazon SageMaker

NVIDIA RAPIDS

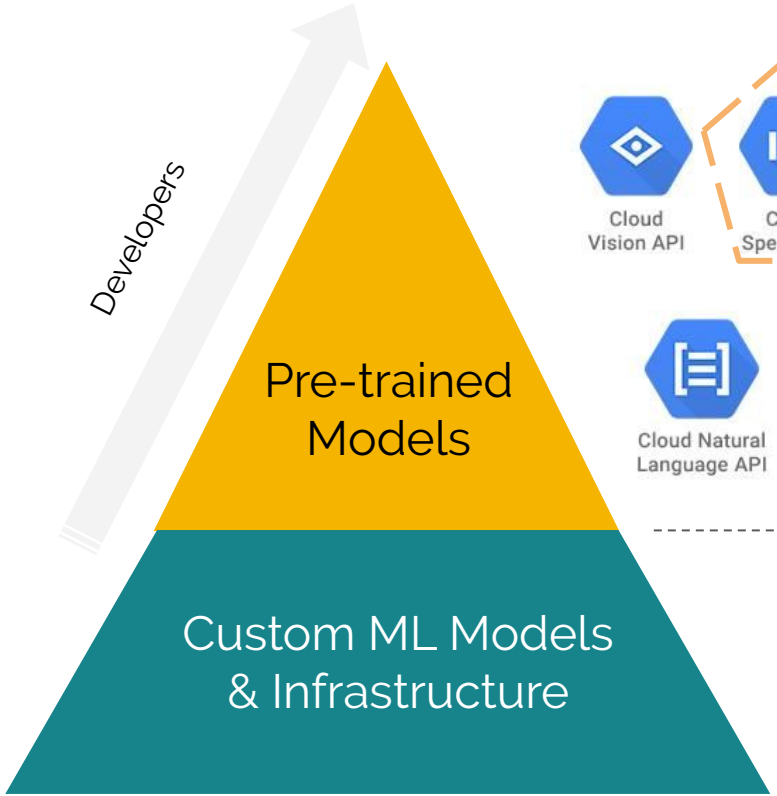
Spark ML

Frameworks

Languages

Framework and API

sli.do Código: iaudio



Google

Cloud Vision API

Cloud Speech API

Cloud Jobs API

Cloud Translation API

Amazon

Polly

Transcribe

IBM Watson

SPEECH TO TEXT

TEXT TO SPEECH

VISUAL RECOGNITION

Cloud Natural Language API

Cloud Video Intelligence API

Azure

Computer Vision

SpeechToText

Language Understanding

Google ML Engine

Azure ML Studio

Amazon SageMaker

NVIDIA RAPIDS

Frameworks

Languages

Why aren't sound events/scenes recognition done yet?

A successful **events/scenes recognition** system **requires**:

- The right representation.
- The right classification architecture.
- The right training data.

Why aren't sound events/scenes recognition done yet?

A successful **events/scenes recognition** system **requires**:

- The right representation.
- The right classification architecture.
- The right training data.

Discriminative model remains somewhat valid:

p(car alarm | beep)

Why aren't sound events/scenes recognition done yet?

A successful **events/scenes recognition** system **requires**:

- The right representation.
- The right classification architecture.
- The right training data.

Discriminative model remains somewhat valid:

p(car alarm | beep)

Oh no!!! It's a hard...

p(car alarm | beep and a lots of garbage)

Rocha et al: The open set recognition problem, ICIP 2016

sli.do

Código: iaudio

Audio Processing Challenges:

(1) Capture mode: Close-talking (headset) and Far-field (distant)

- IHM - Individual Headset Microphone(s)
- SDM - Single Distant Microphone
- MDM - Multiple Distant Microphones

(2) Distortion:

- Reverberation, Signal-to-Noise ratio, ...

(3) Events co-occurring:

- Music (jingle)
- Speech (overlap)

sli.do

Código: iaudio

Audio Processing Challenges:

(1) Capture mode: Close-talking (headset) and Far-field (distant)

- IHM - Individual Headset Microphone(s)
- SDM - Single Distant Microphone
- MDM - Multiple Distant Microphones

(2) Distortion:

- Reverberation, Signal-to-Noise ratio, ...

(3) Events co-occurring:

- Music (jingle)
- Speech (overlap)
- **Noise (background)**

sli.do

Código: iaudio

Audio Processing Challenges:

(1) Capture mode: Close-talking (headset) and **Far-field (distant)**

- IHM - Individual Headset Microphone(s)
- SDM - Single Distant Microphone
- **MDM - Multiple Distant Microphones**

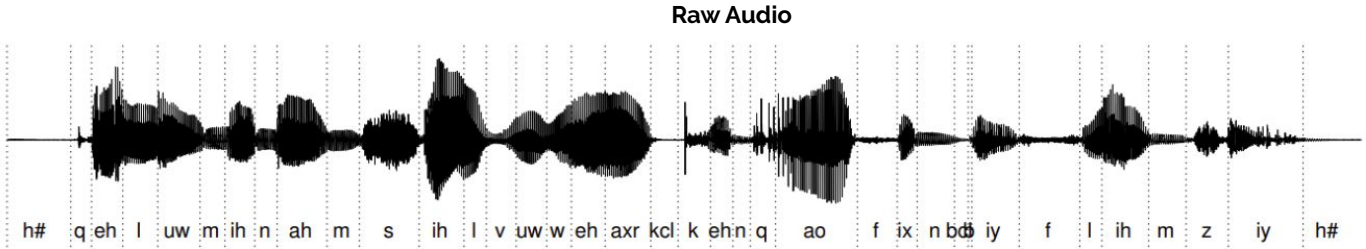
(2) Distortion:

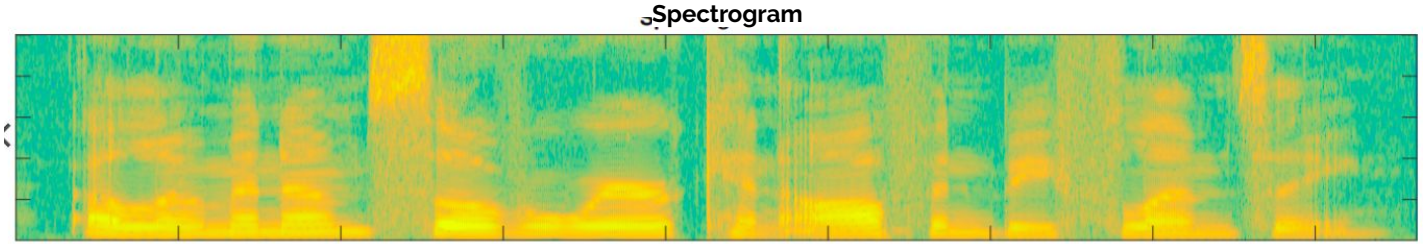
- Reverberation, Signal-to-Noise ratio, ...

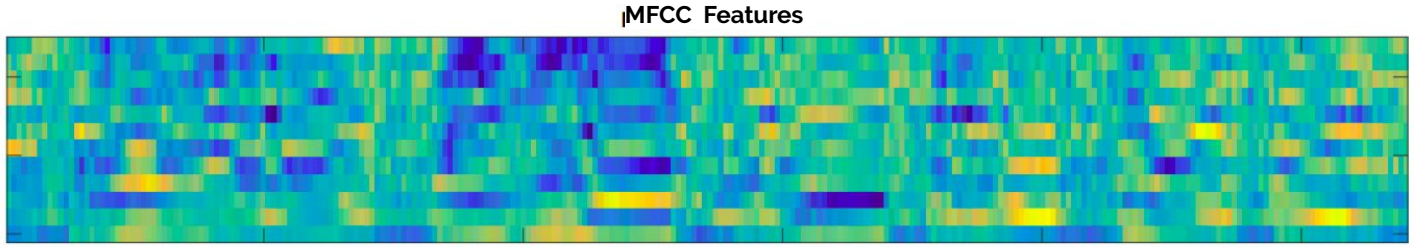
(3) Events co-occurring:

- Music (jingle)
- Speech (overlap)
- **Noise (background)**

Audio representation







Raw Audio:

16k feats/sec.

`shape:(1,16000)`

Spectrogram

~32k feats/sec.

`shape:(128,256)`

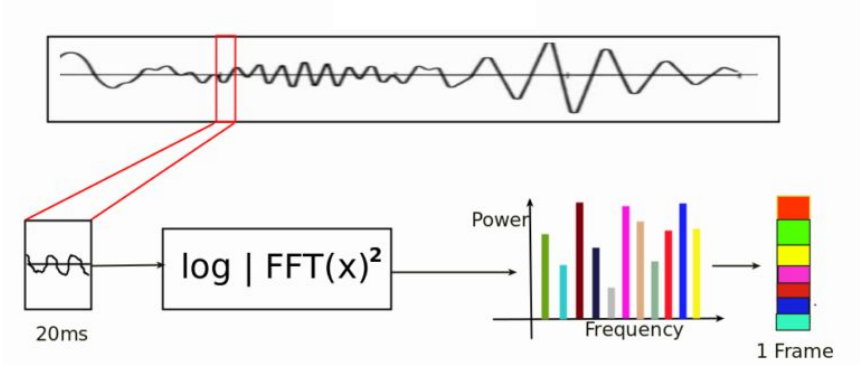
MFCC:

~3k feats/sec.

`shape:(13,256)`

Ravanelli et al: Speaker Recognition from raw waveform with SincNet, SLT 2018

Audio representation



Fourier Transform, Fourier Series, and Frequency Spectrum: https://www.youtube.com/watch?v=r18Gi8lSkfM

Databases: Labeled Acoustic Events

IEEE AASP Challenge on Detection/Classification of Acoustic Scenes/Events

- **DCASE** Challenges (since 2016): **Classification, Localization, Detec...**

Google Audioset:

- 527 sound events on YouTube videos (over 2 million recordings)

ESC-50:

- 50 classes (2k recordings)

sli.do Código: iaudio

Databases: Labeled Acoustic Events

IEEE AASP Challenge on Detection/Classification of Acoustic Scenes/Events

- **DCASE** Challenges (since 2016): **Classification, Localization, Detec...**

Google Audioset:

- 527 sound events on YouTube videos (over 2 million recordings)

ESC-50:

- 50 classes (2k recordings)

ESC-10:

- 10 classes - subset of ESC-50 (400 recordings)

sli.do Código: iaudio

ESC-50 Public Results - Accuracy:

86.50%: CNN based + Features Fusion

- https://pdfs.semanticscholar.org/f6fd/1be38a2d764d900b11b382a379efe88b3ed6.pdf

84.90%: EnvNet-v2 + Data Augmentation

- https://arxiv.org/pdf/1711.10282.pdf

83.50%: CNN pretrained on AudioSet

- https://arxiv.org/pdf/1711.01369.pdf

sli.do Código: iaudio

ESC-50 Public Results - Accuracy:

86.50%: CNN based + Features Fusion

- https://pdfs.semanticscholar.org/f6fd/1be38a2d764d900b11b382a379efe88b3ed6.pdf

84.90%: EnvNet-v2 + Data Augmentation

- https://arxiv.org/pdf/1711.10282.pdf

83.50%: CNN pretrained on AudioSet

- https://arxiv.org/pdf/1711.01369.pdf

81.30%: Human listeners: Crowdsourcing experiment classification

- http://karol.piczak.com/papers/Piczak2015-ESC-Dataset.pdf

https://paperswithcode.com/task/sound-event-detection

Applications:

Content-based Audio Search:

- Semantic search

Monitoring/Event Detection

- Secure: Gun shot, Window glass break, Door knock, Doorbell, Car alarm.
- Human: Shout, Laugh, Yawn, Cough, Snore.
- Nature: Rain, Wind.

Intelligent Content Processing (Recommendation)

- Dog bark, Cat meow, Bee buzz, Birds.
- Baby voice, Children playing.

Context Detection/Localization

- Home Indoor and Outdoor.
- Bus/Metro Station, Restaurant, Office.

sli.do Código: iaudio

Recap:

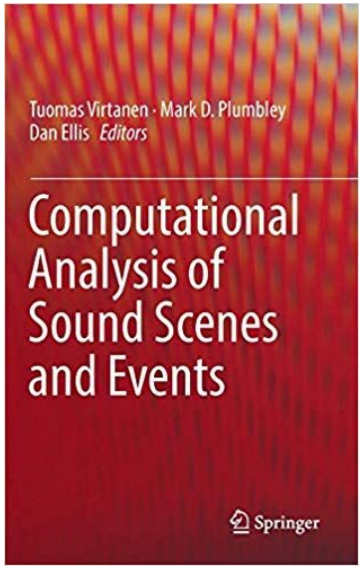
Sound is not speech!

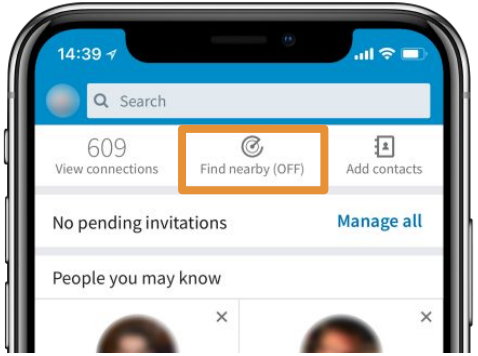
- Not exactly the same problem as speech/music
- Diversity of sound production
- Context sensitive behavior

Computation cost matters.

One against many.

Real world impact: Nonspeech Sound Event Captions





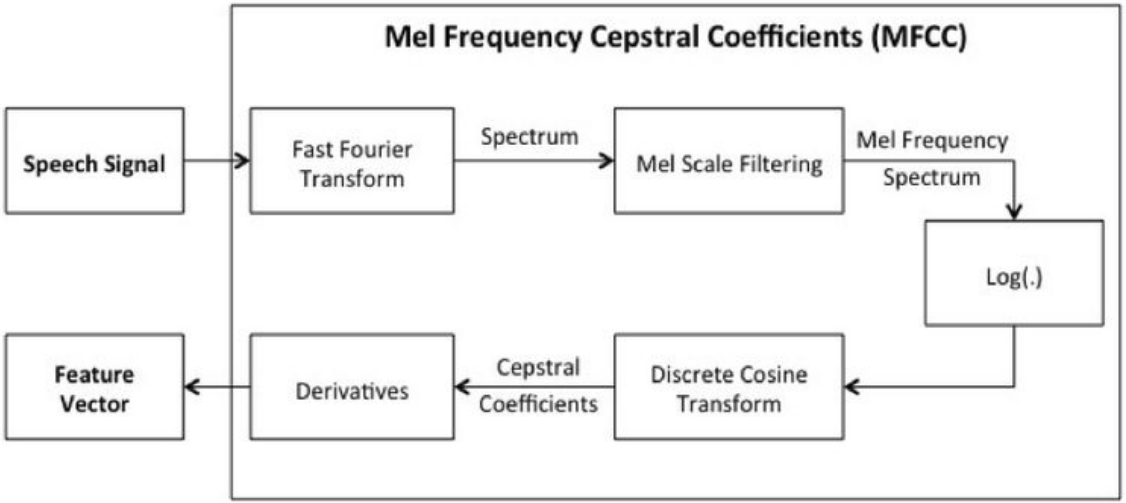
linkedin.com/in/diegoaugusto

CI&T

Diego Augusto Silva • Data Scientist • diegosilva@ciandt.com

Audio Features

- MFCC [1, 2]
- ...



[1] http://librosa.github.io/librosa/feature.html
[2] http://kaldi-asr.org/doc/group__feat.html