DL applied to analyzing speech and audio recordings

handling small amounts of labeled data

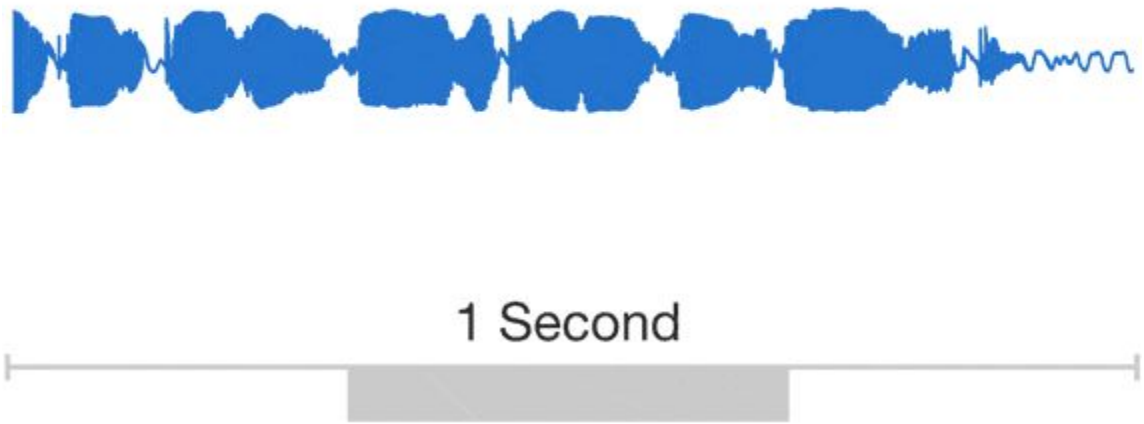
Deep Learning @ IRIT, 07/01/2021

Outline

- Introduction
  - How do we encode audio data?
  - What types of NNs do we use?
- Handling little labeled data
  - Audio data augmentation
  - Learning strategies

How to encode audio data with neural networks?

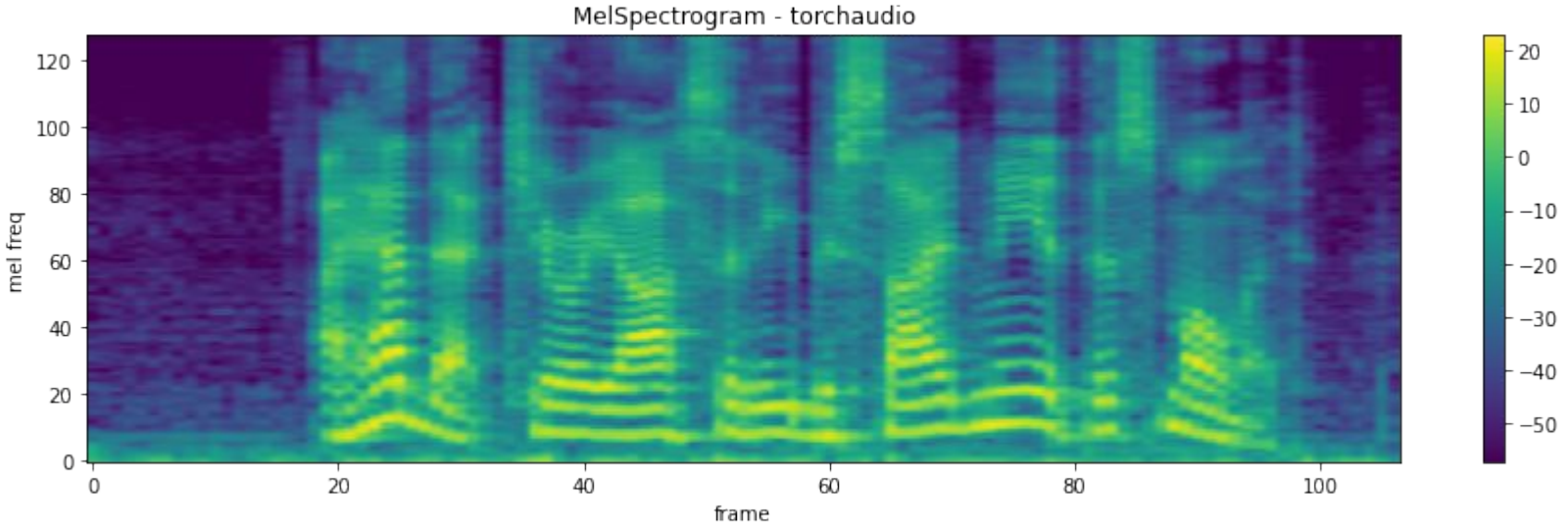
- Option 1: raw audio signal, usually between 8K to 44K points per second!



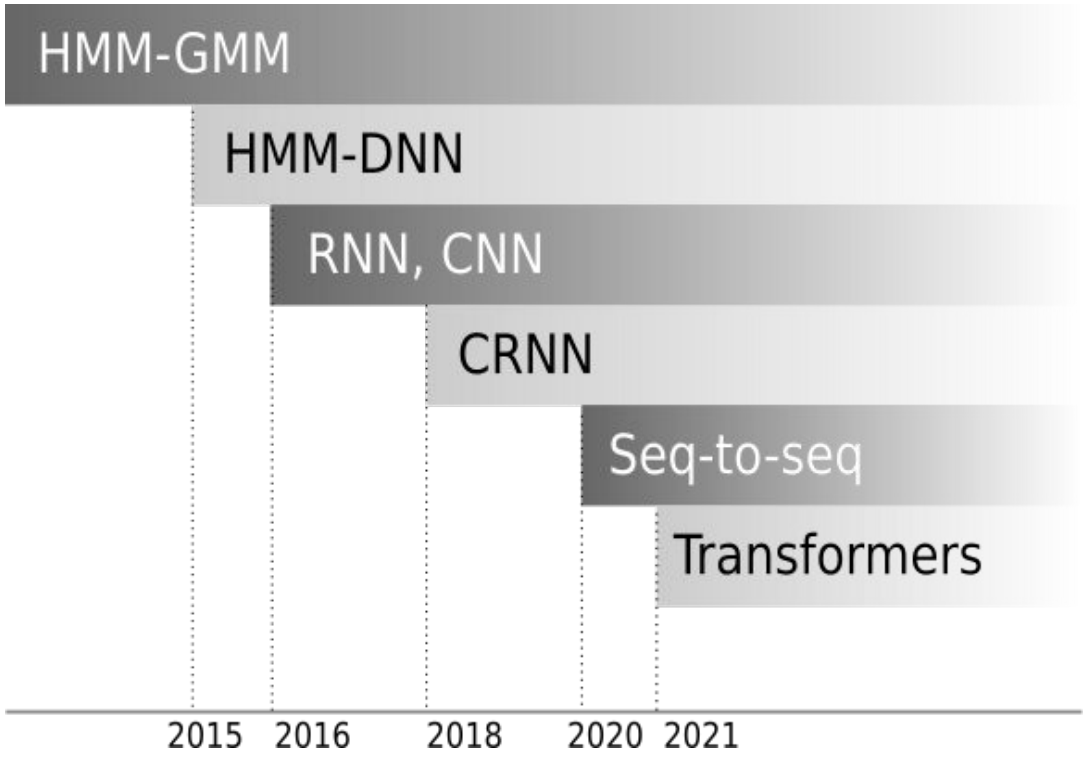
Source: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

How to encode audio data with neural networks?

- Option 2: time-frequency representation → audio data processed as images



What types of models do we use?



Handling little labeled data

Speech with manual transcriptions

French datasets sizes? ~1000 hours of standard adult speech

We work on child speech, L2 speech, pathological speech: from a few minutes to a few hours

Audio events

AudioSet: equivalent of ImageNet for audio events, 527 tag classes, 2.1M files, 5800 hours

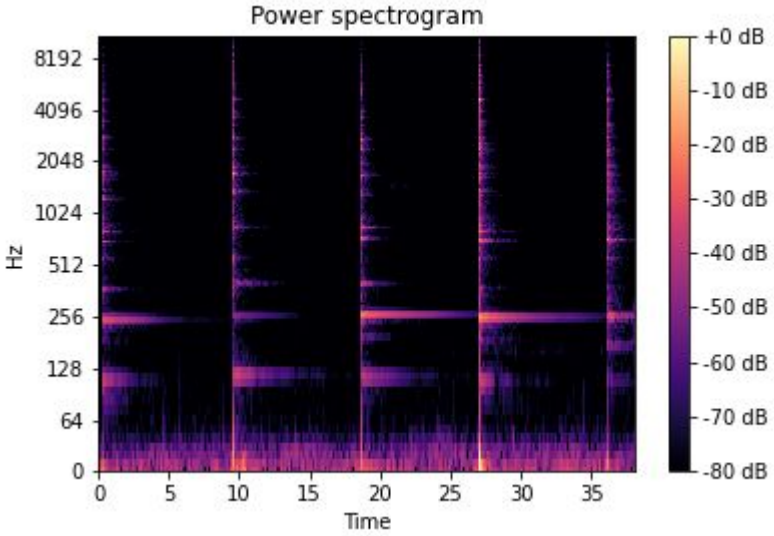
We work on sound event detection, strong labels: 5% of AudioSet

Handling little labeled data

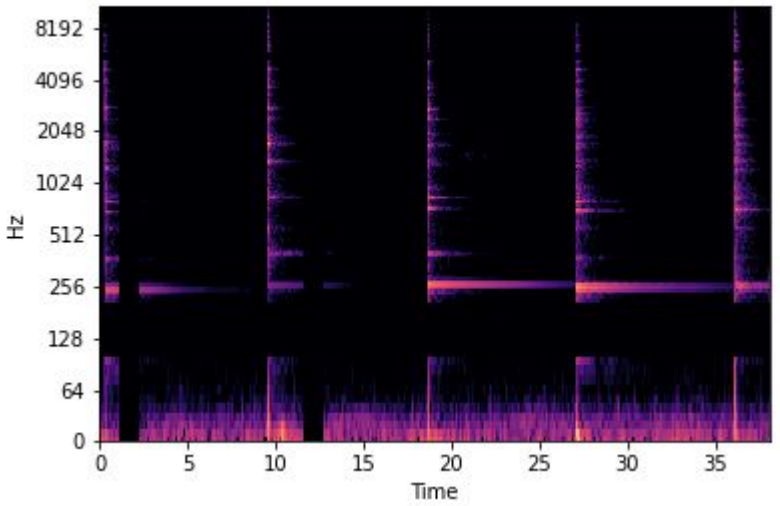
- Key component: data augmentation
  - SpecAugment, MixUp, etc.
  - Simulating errors: L2, child reading speech
- Learning approaches
  - Transfer learning
  - Semi-supervised learning
  - Few-shot learning

Handling little labeled data: audio data augmentations

SpecAugment



Original

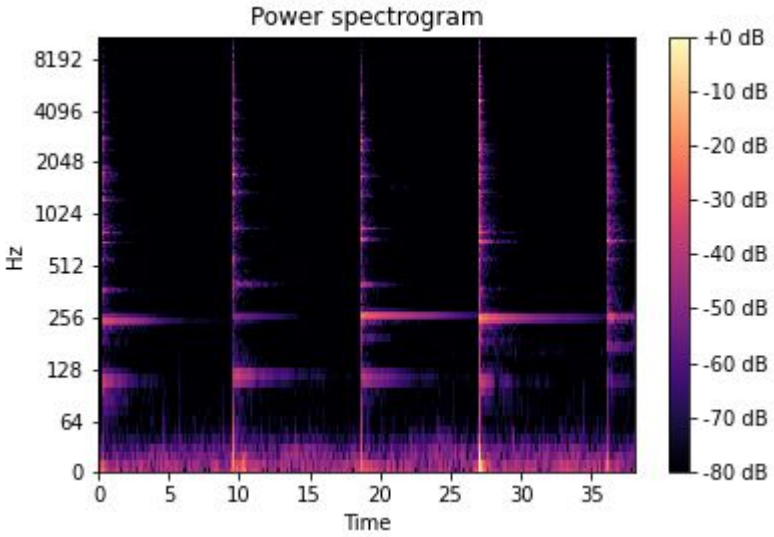


Transformed

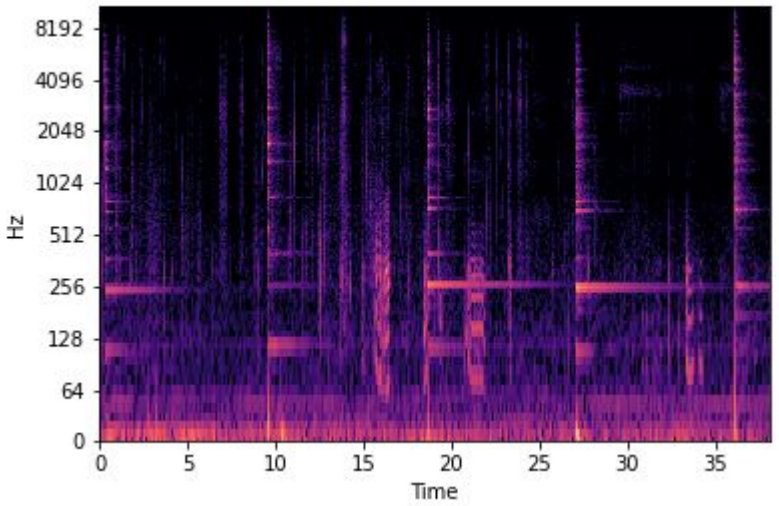
Handling little labeled data: audio data augmentations

MixUp

$$x^{mix} = \lambda x_1 + (1 - \lambda) x_2$$
$$y^{mix} = \lambda y_1 + (1 - \lambda) y_2$$



Original



Transformed

Handling little labeled data: learning approaches

Transfer learning: Adult speech → Child speech

Reading exercise: “Entre le pouce et le majeur, il y a l'index”

Manual ground truth:

A~ T L AE P u S A~ T R AE L AE P u S e L AE M a ZH AE R i L i a L E~ D e

Adult model: Error Rate = 48%

A~ R u T D AE P u S A~ T R AE L AE P u S K R i R E L AE M a ZH O R i L i a E~ D e L a E~ D e L a

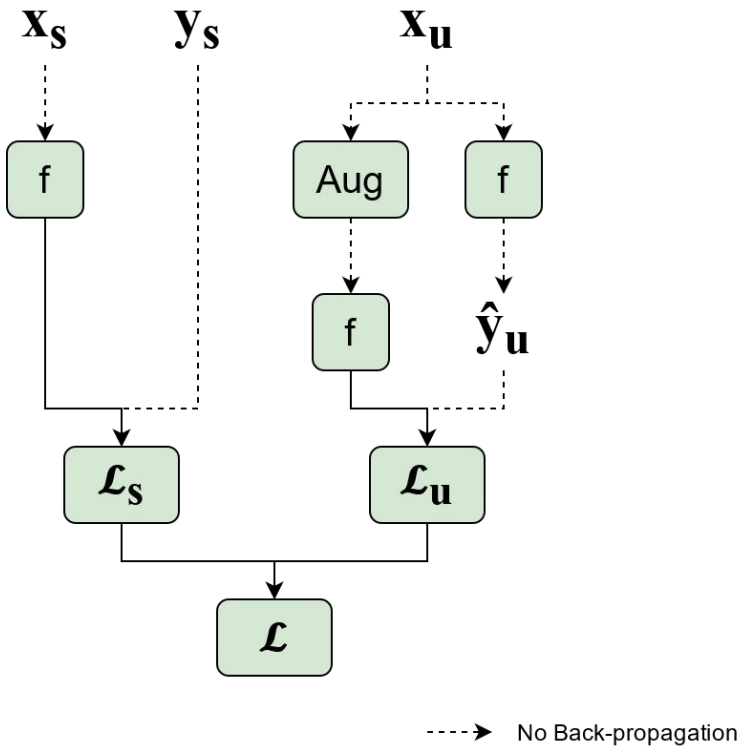
Child adapted model: Error Rate = 18%

A~ T R AE P u S A~ T R AE L AE P u S e L AE M a ZH AE R i L i a L E~ D E L J a

Handling little labeled data: learning approaches

Semi-supervised deep learning

- Labeled data : (x_S, y_S)
- Unlabeled data : x_U
- A single model : **f**



Handling little labeled data: learning approaches

Error Rates on three datasets

Method	Labeled fraction	ESC-10 (cross-val, %)	UBS8K (cross-val, %)	GSC (test, %)
Supervised	10%	32.00 ± 6.17	33.80 ± 4.82	10.01
Supervised+Aug	10%	22.67 ± 3.46	23.75 ± 4.73	6.58
Supervised	100%	8.00 ± 5.06	23.29 ± 5.80	4.94
Supervised+Aug	100%	**4.67** ± 1.39	**17.96** ± 3.64	**2.98**
Semi-supervised (MixMatch)	10%	15.33 ± 5.58	**18.02** ± 4.00	**3.25**

Thank you, questions?