Distributed learning for speech recognition in the context of privacy protection

Laboratoire Informatique d'Avignon (LIA)

Institut des sciences du Digital Management et Cognition (IDMC)

Université de Lorraine

Author:
Eunice AKANI

Supervisor:
Yannick ESTEVE

February 24, 2020 to August 16, 2020

Table of Contents

- **Introduction**
- **Problematic**
- **Experiments and Results**
  - Experimental setup
  - Generic acoustic model and speaker based model
  - Clustering of weight matrix
  - Experiment 1: Speaker clustering
  - Experiment 3: Speaker identification
  - Experiment 3: Gender grouping
- **Discussion**
- **Conclusion**

Introduction

Introduction

- LIA: Computer Sciences Labs located in Avignon
- LIA thematic:
  - Natural Language Processing
  - Networks
  - Operational research
- Neural acoustic model for speech recognition
- Distributed learning with communication between the nodes
- Analyse the shared information

Problematic

Problematic

- How to improve ASR System by considering the privacy of user?

- DEEP-PRIVACY Project:
  - Learning privacy-preserving representations of the speech signal
  - Distributed algorithms and personalization for speech recognition

- Measure the extent to which the weight matrix of a neural network adapted to a speaker could identify that speaker

Automatic Speech Recognition System

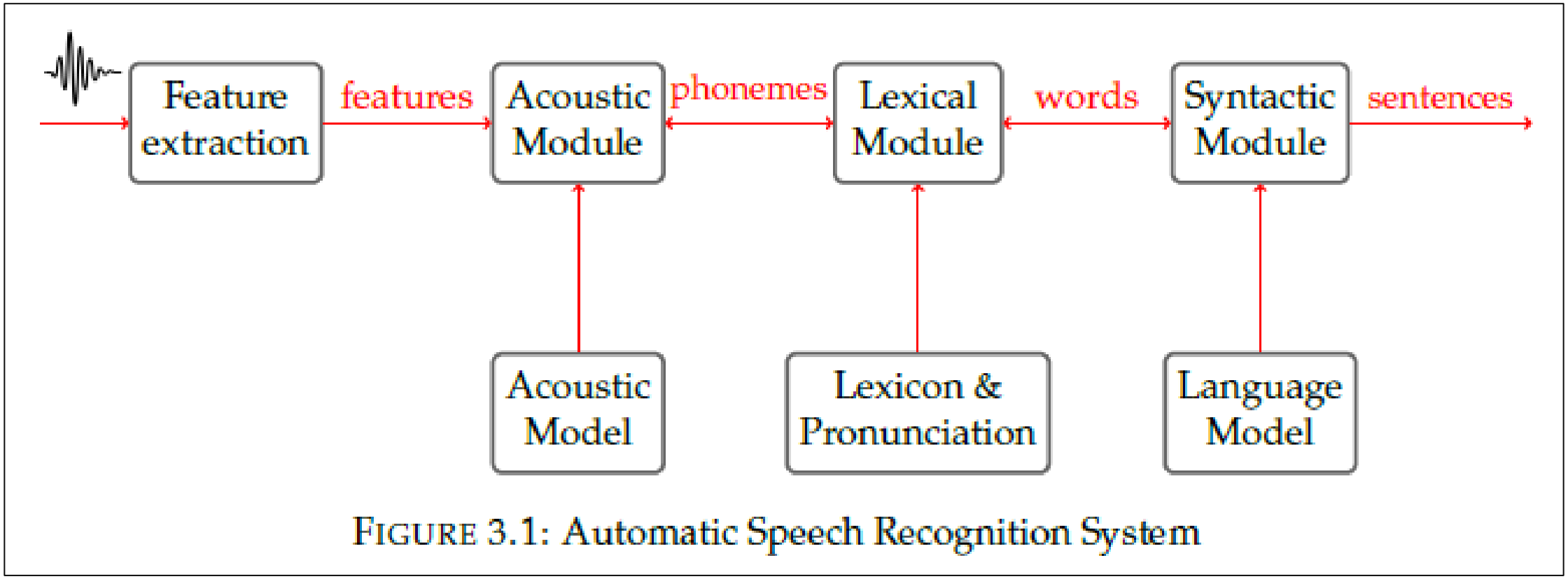


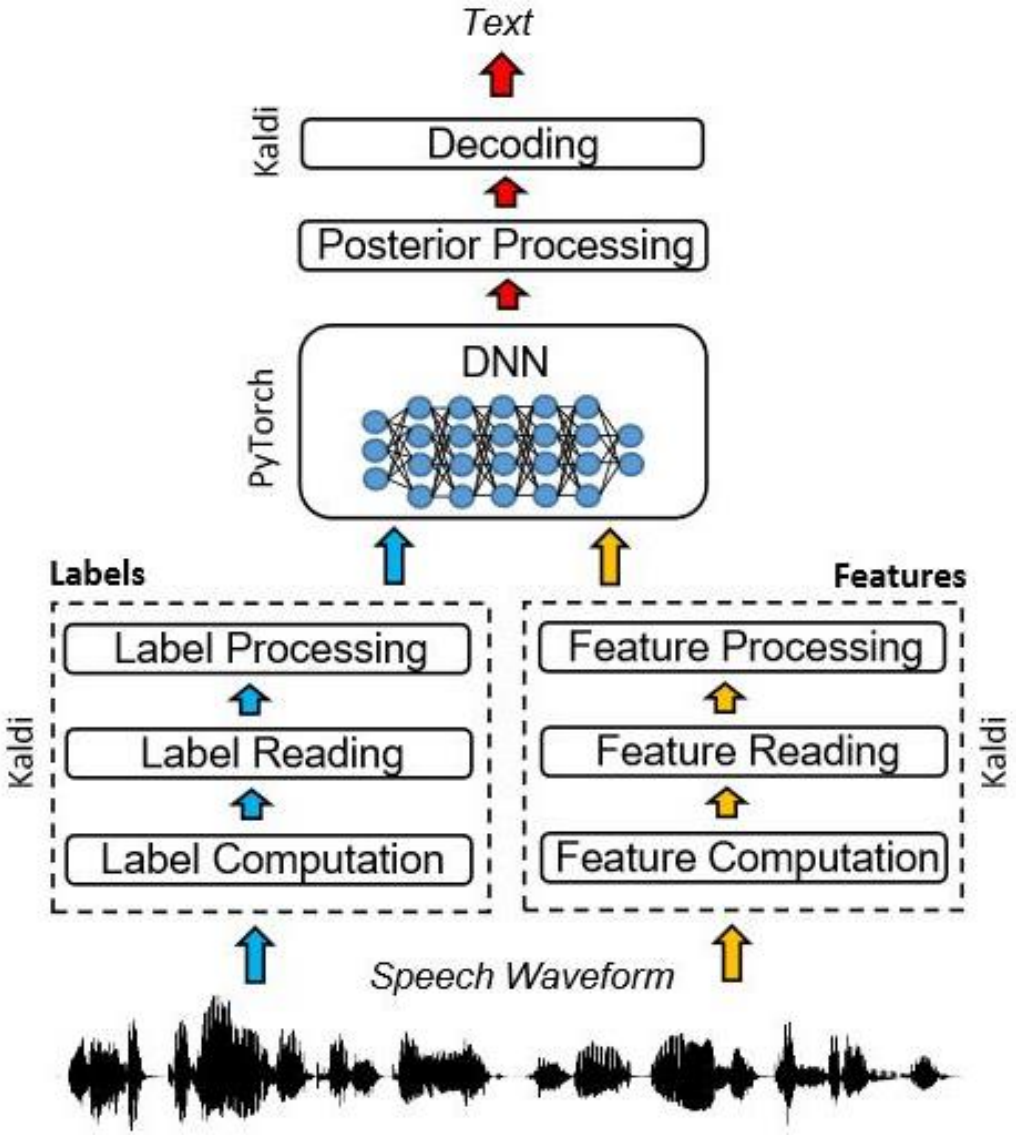
FIGURE 3.1: Automatic Speech Recognition System

Experiments and results

Experimental setup: Pytorch-Kaldi toolkit

[Ravanelli et al., 2019]

- ASR toolkit for bridging the gap between Kaldi and PyTorch
- Kaldi: ASR toolkit
- PyTorch: Machine learning framework



Experimental setup: dataset

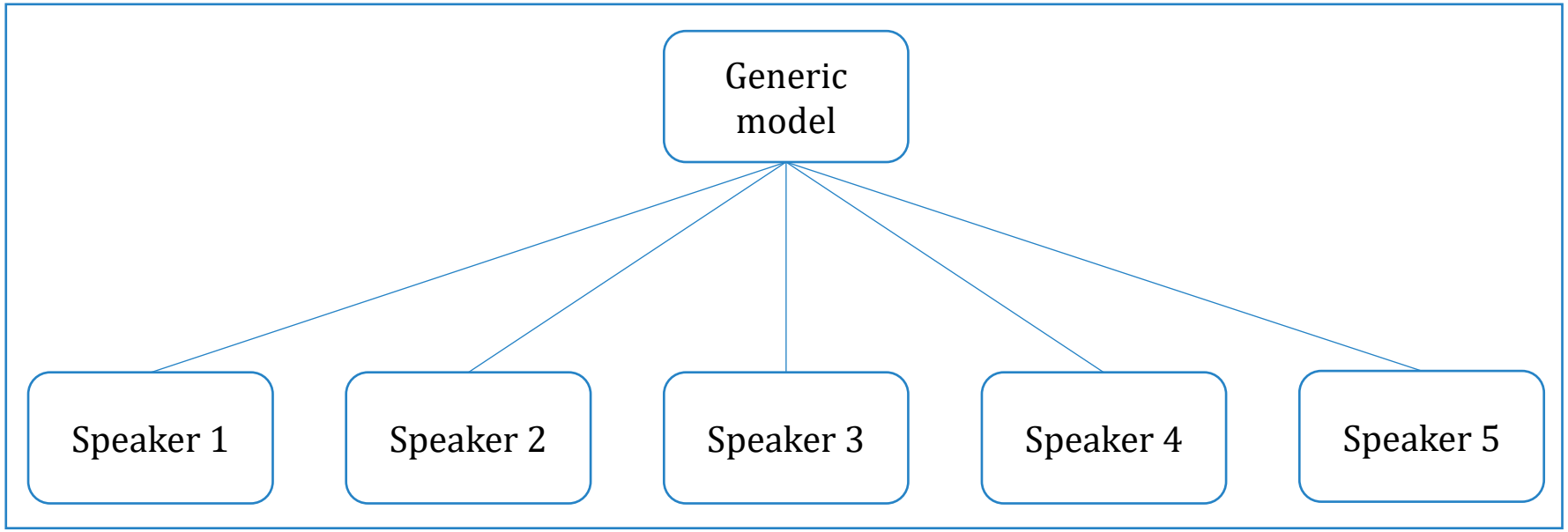
Dataset TEDLIUM Release 3 [Hernandez et al., 2018]

- Based on the TED Talks
- Corpus for Automatic Speech Recognition in English

- Two distributions:
  - Legacy version
  - Speaker adaptation version

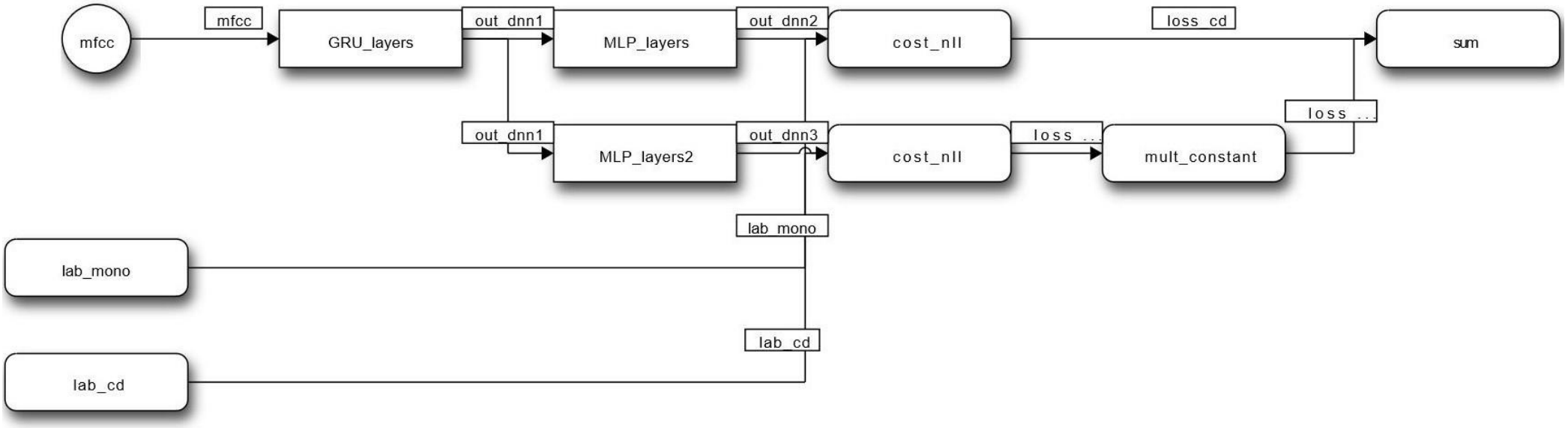
- Speaker adaptation distribution content:
  - 1938 speakers
  - 2281 talks
  - 346,17 hours
  - Without silence, noise, …

Experimental setup: dataset partitioning



Part	Number of speaker	Number of talks
2/3	1288	1514
1/3	644	765

Generic acoustic model [Hernandez et al., 2018]



- 5 GRU Layers and 2 MLP Classifiers
- Trained on 2/3 of the datasets for 23 epochs (14,5 days)
- %WER = 14,45 on non-training data (test set)

Speaker based model

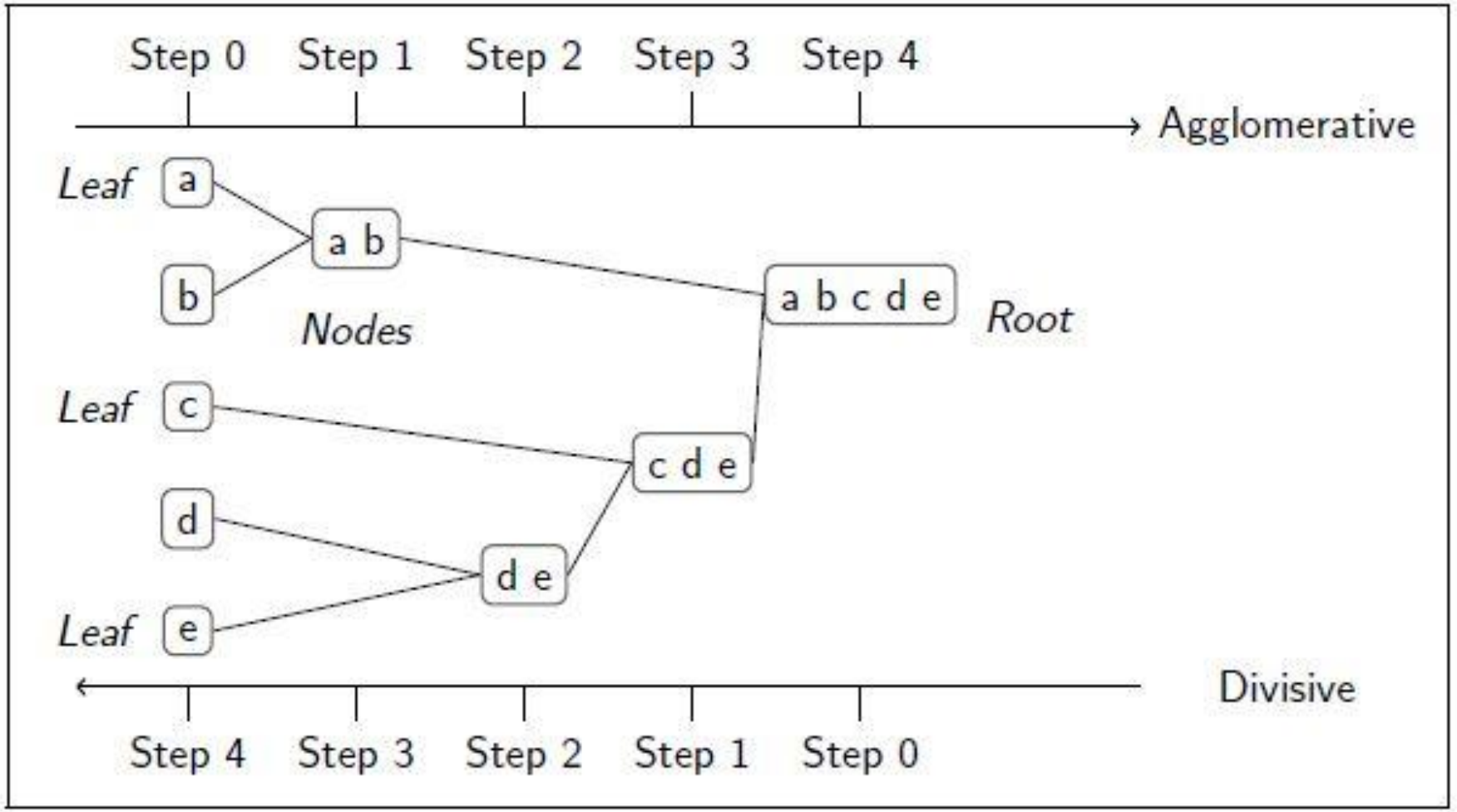
- Create new dataset based on the 1/3 datasets

Number of segments	Number of subsets
$x \geq 80$	4
$60 \leq x < 80$	3
$60 < x$	2

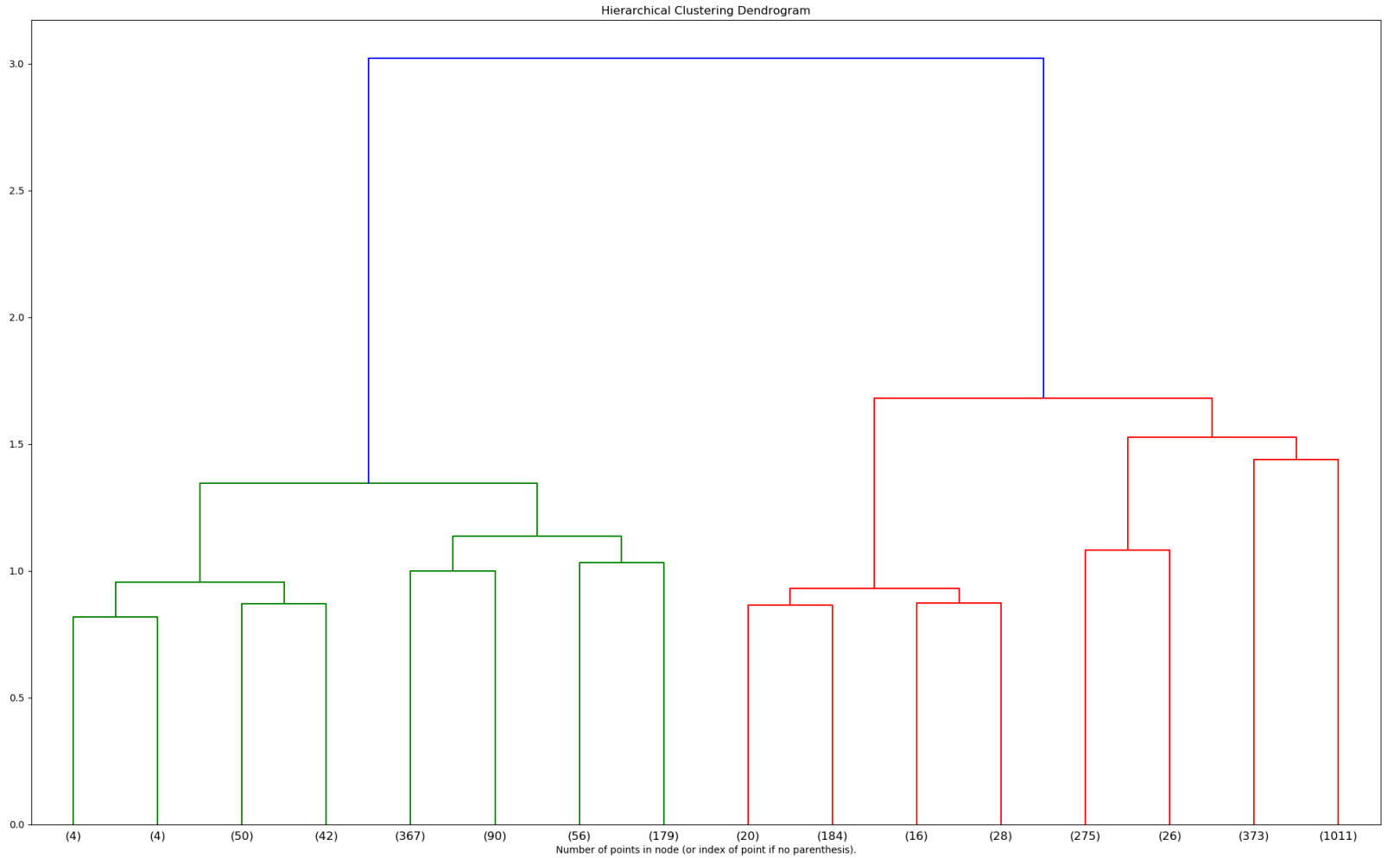
- From 765 speeches to 2745 sessions of speeches (average session duration = 200 s)

- Finetuning: initialisation of all the 2745 models with the weights of the generic model.
- % WER = 13.94 on average for the 2745 models after 4 epochs
- Extraction of the weights of each model every 2 epochs

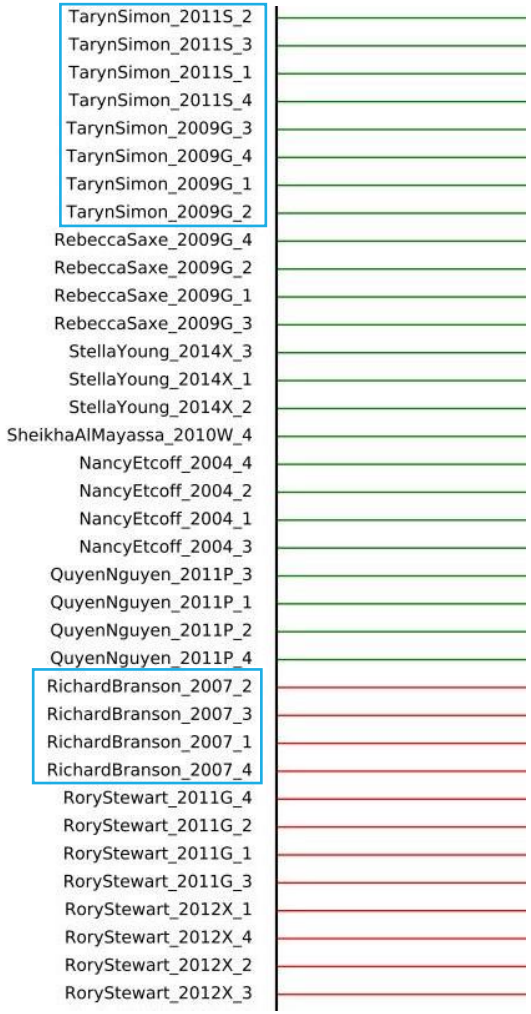
Clustering of weight matrix: Hierarchical clustering



Results: dendrogram



Results : Unwrapped dendrogram



- Speech subsets from the same speaker are usually clustered together
- The main split (green vs. red here) separates women and men
- This seems to confirm that update values of neural weights during AM speaker adaptation is useful to measure speaker similarity

Speaker clustering: experiments

- Random clustering:
  - Form randomly clusters of 4 elements

- Method 1:
  - Based on divisive hierarchical clustering
  - Restriction on the number of elements per cluster
  - Maximum: $4 \times \varepsilon$ elements per cluster

- Method 2 :
  - Based on agglomerative hierarchical clustering
  - Limit number of elements in each cluster
  - Minimum: 4 elements per cluster

Speaker clustering: results

- Case 1: Each talk forms a class

		Pure cluster	Purity	Entropy
Random clustering		0	0,253	1,99
Method 1	Epoch 0 Epoch 2	494/656 528/611	0.878 0.942	0,36 0.18
Method 2	Epoch 0 Epoch 2	344/548 466/564	0.822 0.919	0.49 0.22

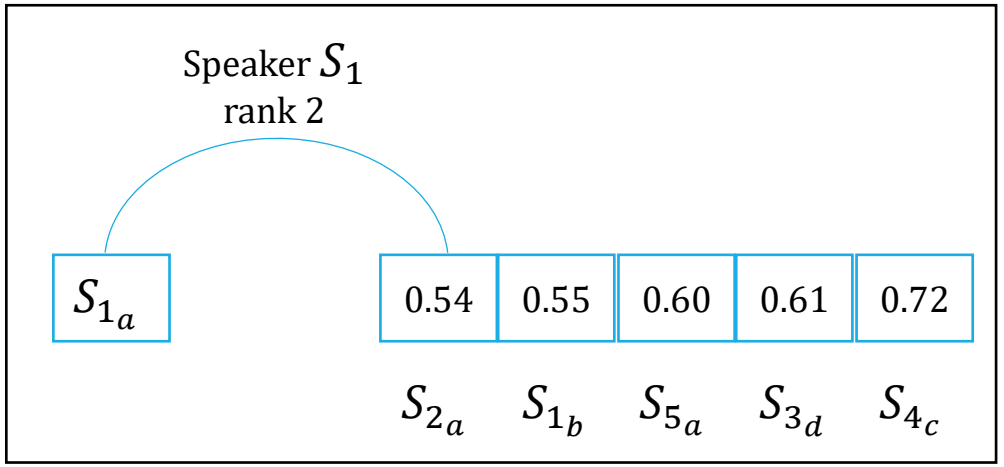
- Case 2: The talks of the same speaker form a class

		Pure cluster	Purity	Entropy
Random clustering		0	0.254	1.99
Method 1	Epoch 0 Epoch 2	509/656 545/611	0.893 0.955	0.32 0.14
Method 2	Epoch 0 Epoch 2	346/548 480/564	0.84 0.929	0.45 0.19

- Percentage of pure cluster for all layers considering case 2 (method 2)

Epochs	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4
Epoch 0 Epoch 2	67.77 82.62	62.48 64.70	64.26 67.26	62.64 65.42	63.00 62.86

Speaker identification: experiment



- Take two sessions of the same speaker and 49 sessions (randomly) from other speakers
- Compute Euclidean distance between one session of the speaker and the other 50 (one from the speaker and 49 from others)
- See how closed the two session are (rank)

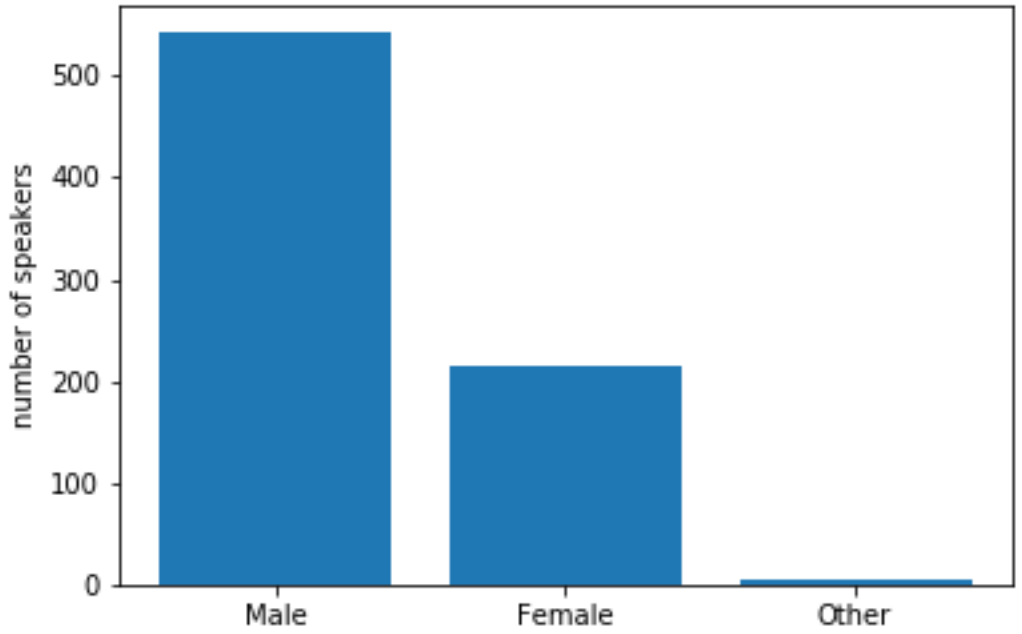
Speaker identification: results

Layers	Epochs	Average rank
Layer 0	Epoch 0 Epoch 2	19.47 17.69
Layer 1	Epoch 0 Epoch 2	23.33 22.03
Layer 2	Epoch 0 Epoch 2	22.55 20.53
Layer 3	Epoch 0 Epoch 2	22.14 21.36
Layer 4	Epoch 0 Epoch 2	24.04 23.96

Average of speaker rank (per layer and epoch)

Gender grouping: experiment

- Manually labeled dataset



- Clustering using Agglomerative clustering and KMeans

Gender grouping: results

Layer 0	KMeans	Cluster	Male	Female
		0	195	707
		1	1746	65
	Agglomerative	Cluster	Male	Female
		0	1868	91
		1	73	681

Layer 1	KMeans	Cluster	Male	Female
		0	169	732
		1	1772	40
	Agglomerative	Cluster	Male	Female
		0	1862	86
		1	79	686

Layer 2	KMeans	Cluster	Male	Female
		0	822	209
		1	119	686
	Agglomerative	Cluster	Male	Female
		0	596	193
		1	1045	579

Layers	Kmeans		Agglomerative	
	Purity	Entropy	Purity	Entropy
Layer 0	0.904	0.40	0.939	0.32
Layer 1	0.92	0.33	0.94	0.32
Layer 2	0.72	0.85	0.72	0.83

Distribution of each data into each cluster; Entropy and Purity

(Epoch 2)

Discussion

Discussion

- Speaker clustering and Speaker identification
  - Best result on epoch 2 layer 0
  - When will we stop the learning ?

- Gender grouping
  - Good results for layer 0 and layer 1
  - Pour results for the other layers
  - Gender can be identified at the low level

Conclusion

Conclusion

- The weight matrix:
  - contain speaker-specific information (gender)
  - contain information not enough for speaker identification task

- Future work:
  - Try speaker identification for more epoch
  - Learn the trajectory of the weight matrix
  - Find information that the weight matrix may contain: Age, Mood, Speaker accent, . . .

Thank you for your attention

References

- Bell, Peter, & Renals, Steve. 2015 (Apr.).
  Regularization of context-dependent deep neural networks with context-independent multi-task training.
  Pages 4290{4294 of: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

- Hernandez, Francois, Nguyen, Vincent, Ghannay, Sahar, Tomashenko, Natalia A., & Esteve, Yannick. 2018.
  TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation.
  CoRR, abs/1805.04699.

- Ravanelli, M., Parcollet, T., & Bengio, Y. 2019.
  The PyTorch-Kaldi Speech Recognition Toolkit.
  In: In Proc. of ICASSP.