DEEP NEURAL NETWORKS FOR AUTOMATIC DETECTION OF SCREAMS AND SHOUTED SPEECH IN SUBWAY TRAINS

Pierre Laffitte

David Sodoyer

Charles Tatkeu

LEOST-IFSTTAR Lille, France

Laurent Girin

GIPSA lab, Univ. Grenoble Alpes

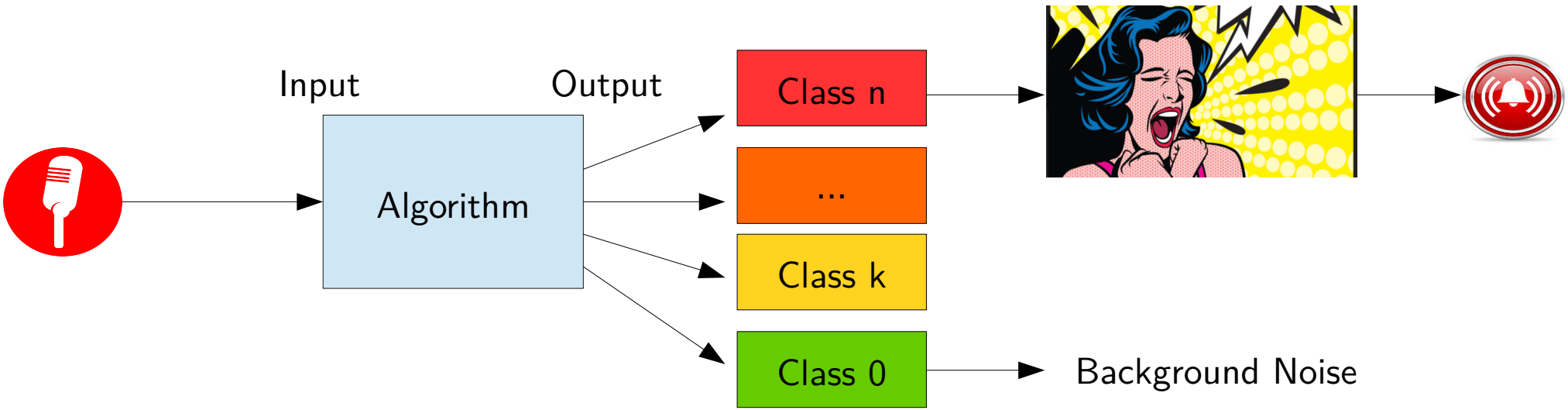
Grenoble, France

Index

1. Context

Scene Classification in a transportation environment

- Classification of the Acoustic Scene in terms of the situation of the passengers.
- Define the situation/classes as ranging from normal to critical from a security/surveillance perspective.

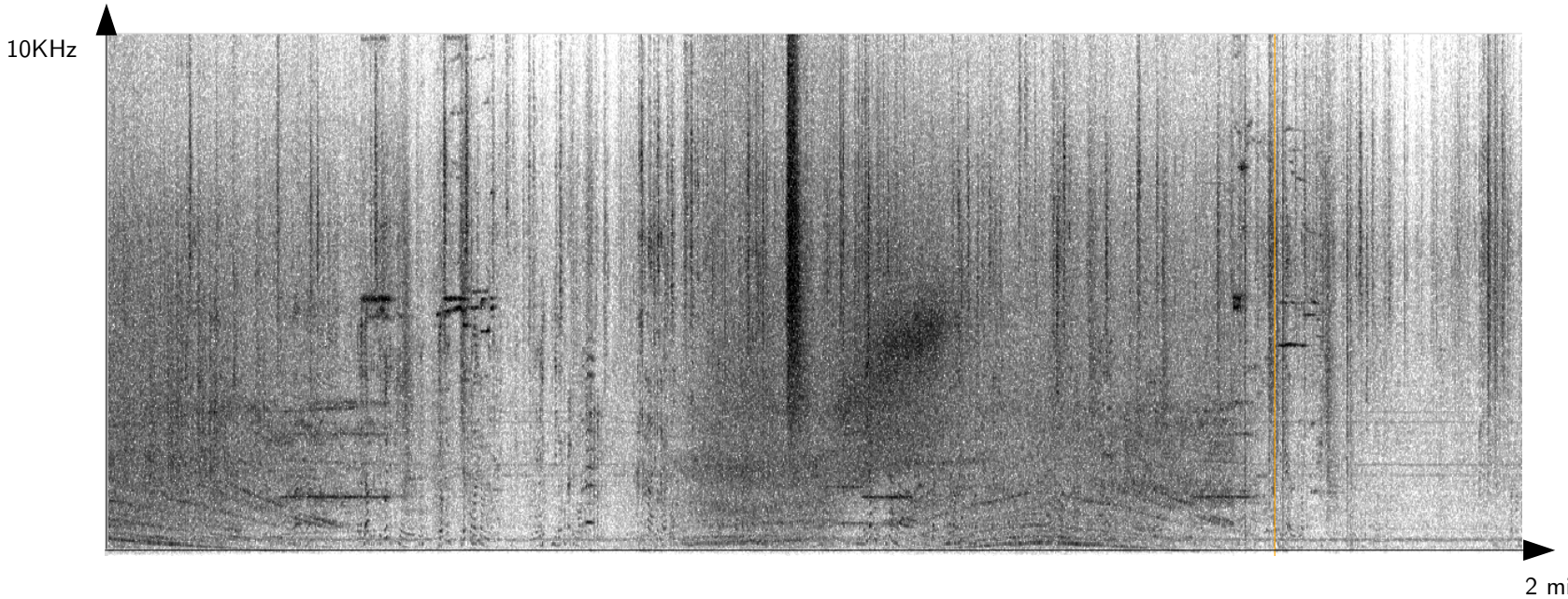


Embedded Transportation Acoustic Environment

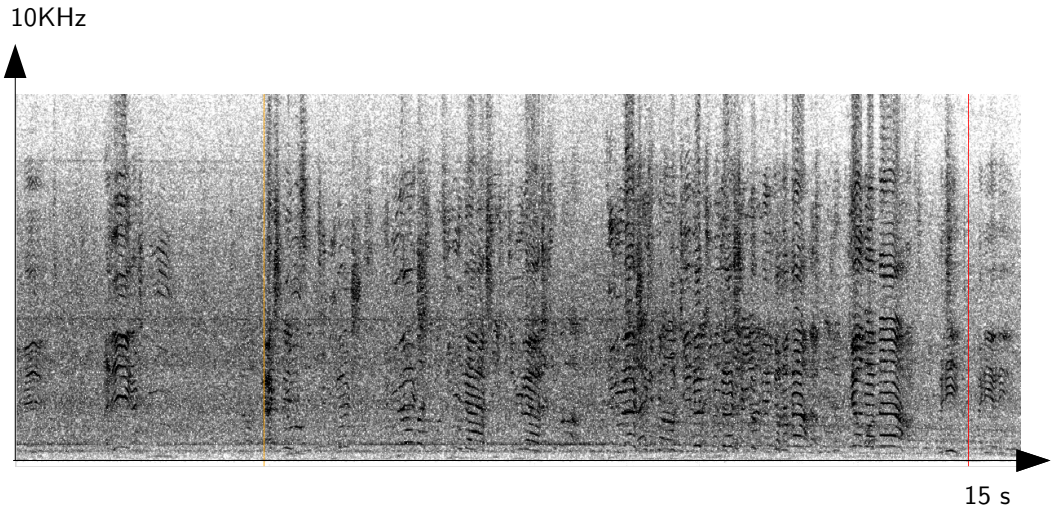
- Multi-source environment:
  - Vehicle (brake compressors, wheels screeching, etc..)
  - Passengers
  - Infrastructure and External environment (announcement signals, noises in the station)

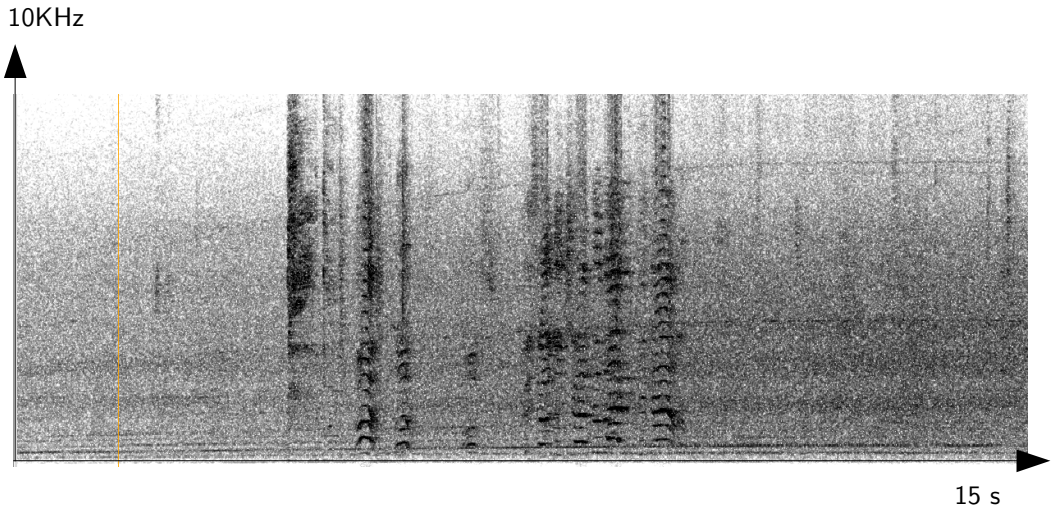
- Highly non stationary

- Varying number of sources

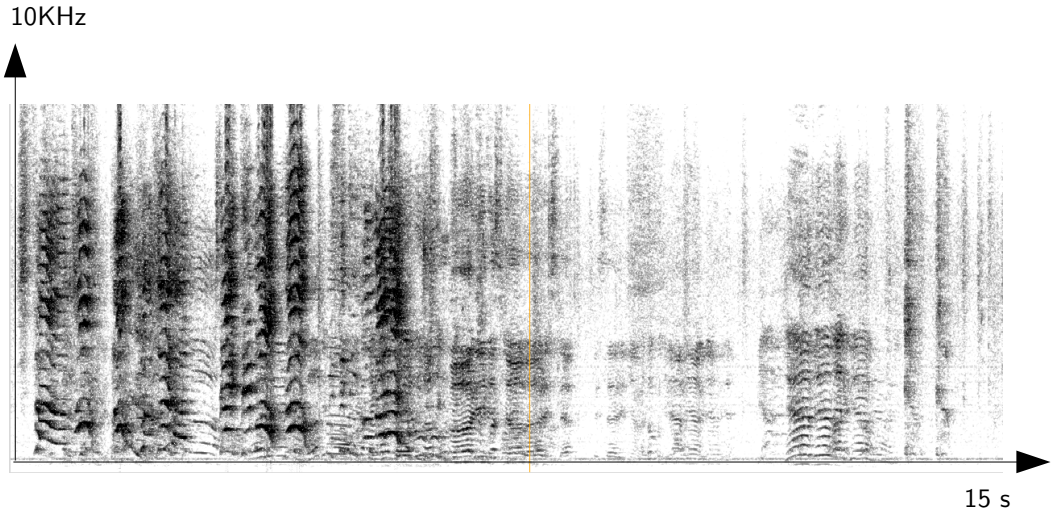


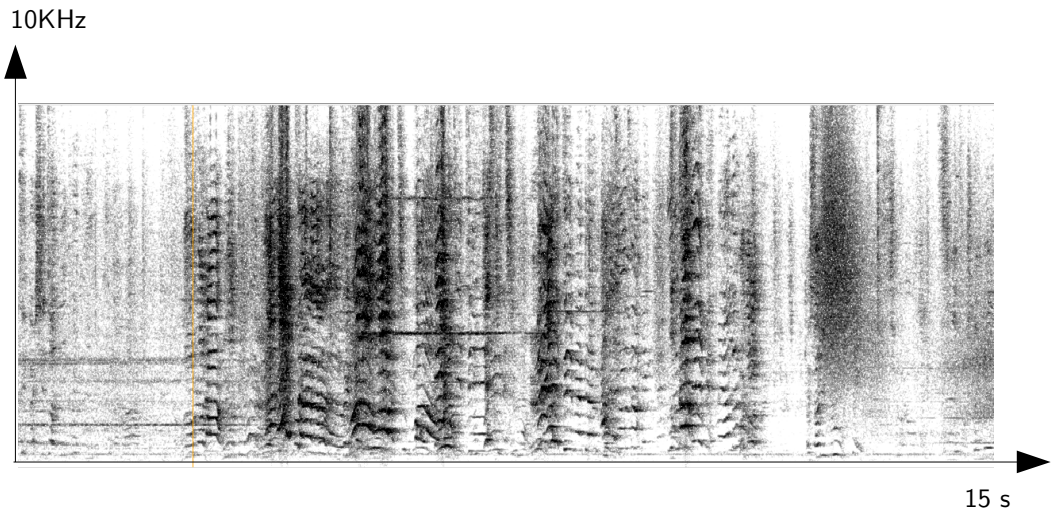
Voice/Speech signals are highly corrupted





Shouts/Screams in motion



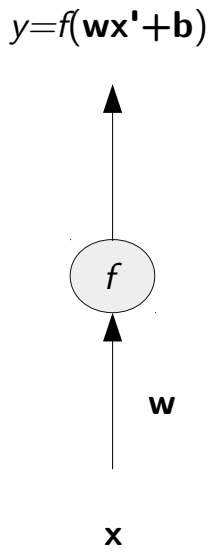


Shouts/Screams in standby

2. Model

Neural Networks

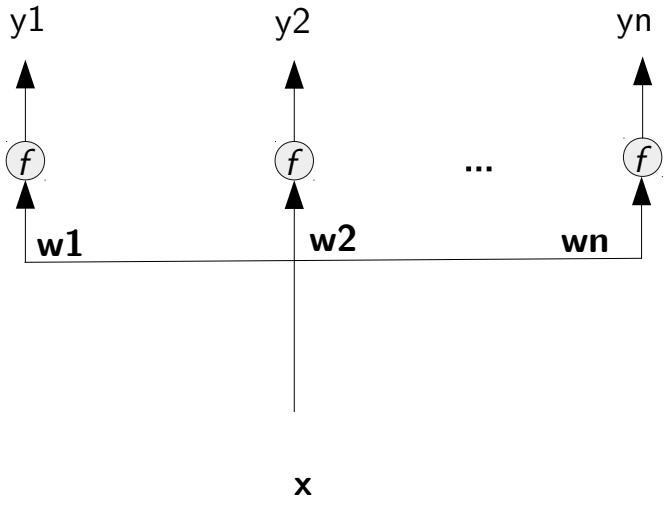
The simple neuron:



f: activation function : sigmoid,tanh, ...

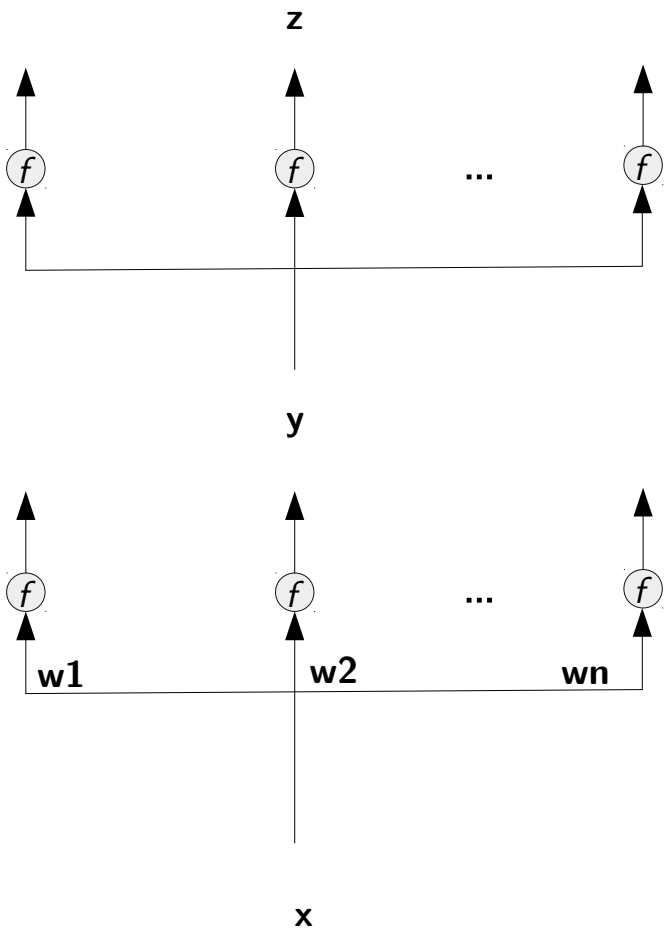
Neural Networks

Multiple neurons connected together:

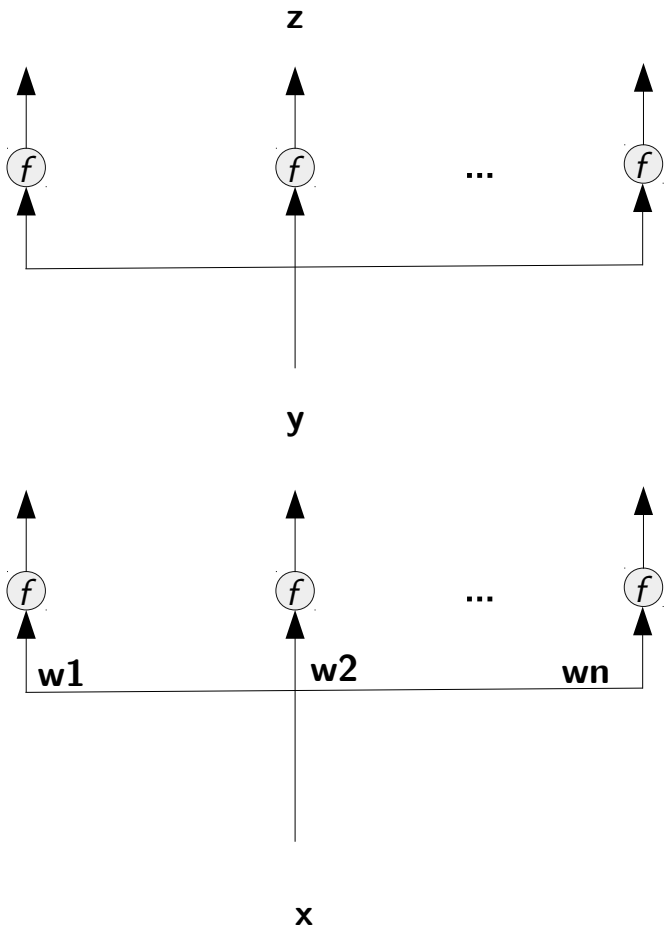


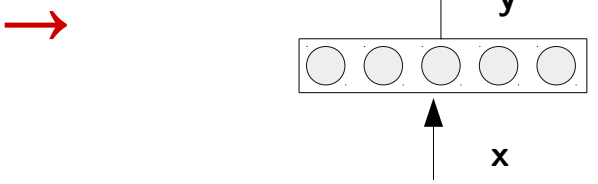
C.M. Bishop, *"Neural Networks for Pattern Recogni-tion"*

Neural Networks



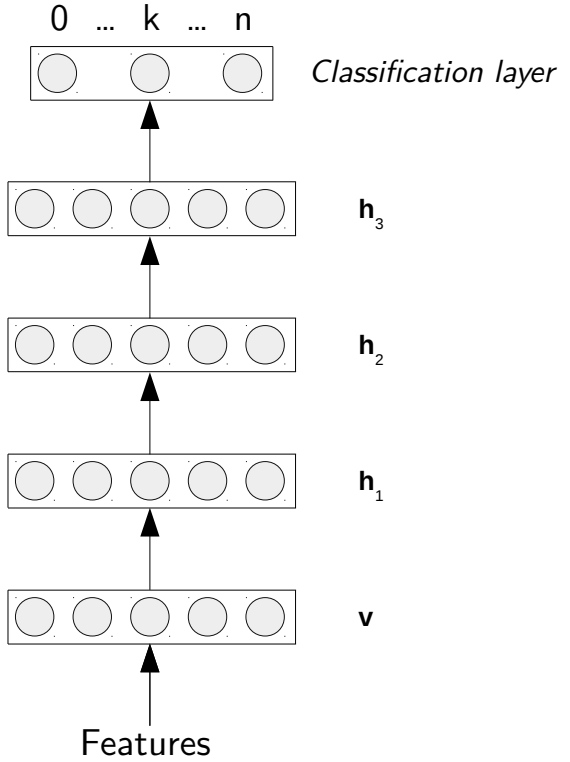
Neural Networks





Deep Neural Network

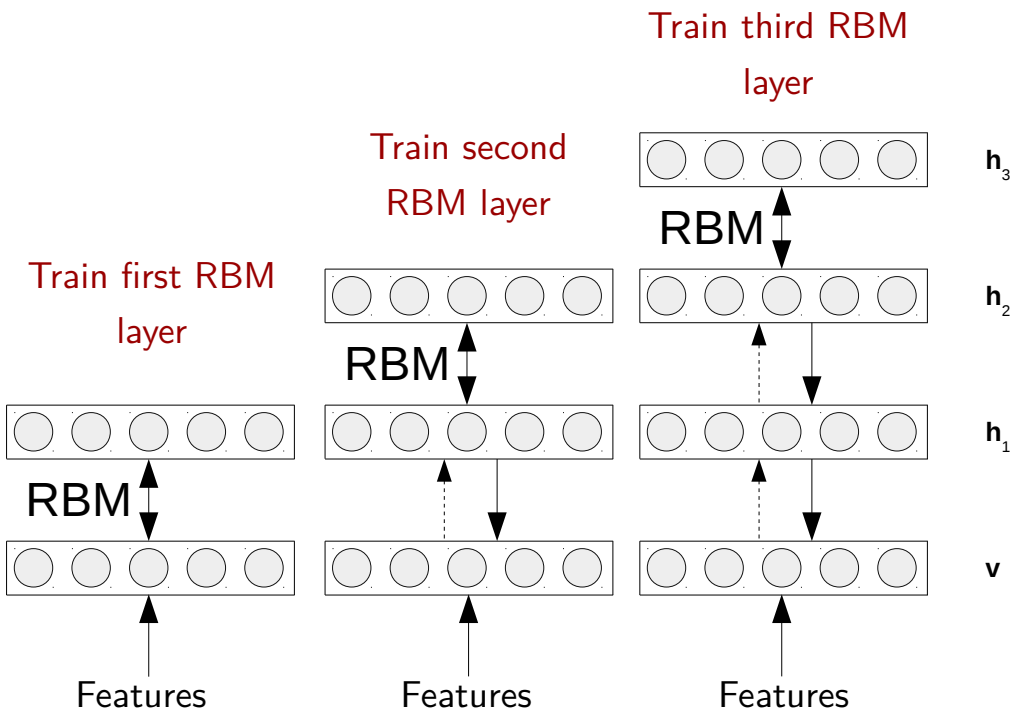
- Multi-Layer Perceptron (MLP)
- Discriminative model



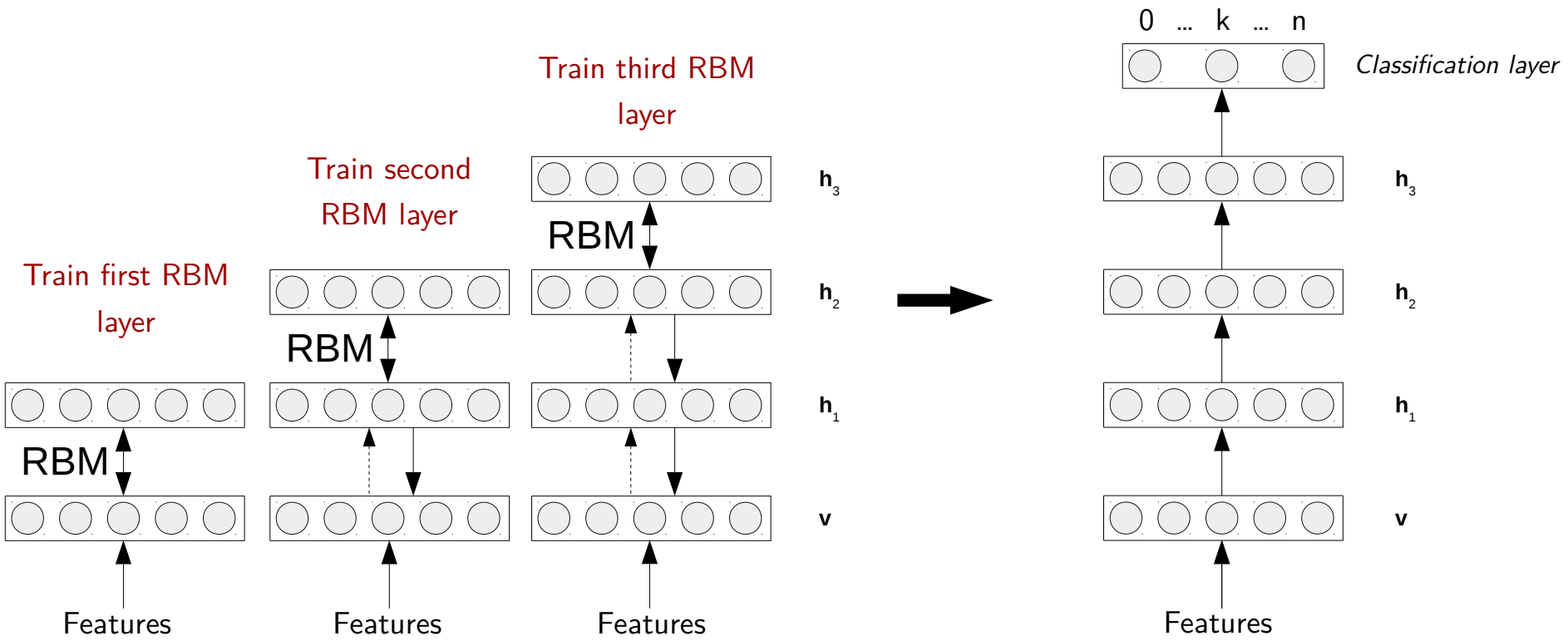
[Rumelhart, et al, 1986]

Deep Belief Network

- Stacked Restricted Boltzmann Machines (RBM)
- Generative model



DBN-DNN


Train third RBM layer
Train second RBM layer
Train first RBM layer
RBM
RBM
RBM
h3
h2
h1
v
Features
Features
Features
0 ... k ... n
Classification layer
h3
h2
h1
v
Features

Training the DBN-DNN

Unsupervised training of the Generative model DBN:

- Learn higher order features from the input data
- Tries to model the input distribution

Supervised training of the DNN (Fine-tuning):

- Discriminative learning
- Initialize DNN with weights from DBN

[Hinton, et al, 2006]

Python deep learning library by Yajie Miao:
https://github.com/yajiemiao/pdnn

3. Database description

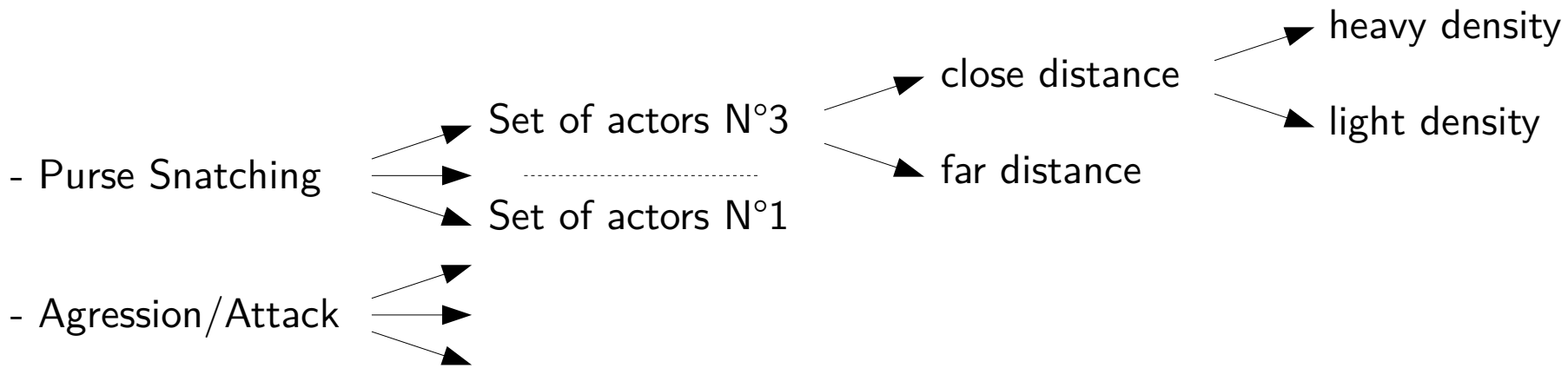
Recording the data

DéGIV project:

- subway car put at our disposal by RATP (Paris transportation authority)

Data captured from real-world environment:

- Line 14 in the Paris metro
- Train running amidst normal traffic
- Several days of recordings



4. Experiments

Experimental settings

Features:

- 12 MFCC coefficients + 1 energy term, over 25ms window, every 10ms
- Concatenation of 10 adjacent frames

Classes:

- Noise (~900s): everything not containing speech
- Background voice (~600s): unintelligible speech
- Conversation (~650s): intelligible speech
- Shout (~300s): shouted speech
- Scream (~100s): screams

All classes include very noisy occurrences (when train is moving)

3 different dataset:

- Train ~77%
- Valid ~77%
- Test ~23%

Classifier:

- 3 layers of 512 units
- 300 epochs for DBN training and 200 for DNN training (decided with validation result)

Results

2 classes

	Noise	*Shout*
Noise	**96.8**	20.8
Shout	3.20	**79.2**

Table 2. Confusion Matrix. Noise vs. Shout.

$\rightarrow$ Classes should be very different from each other by definition but separation surprisingly isn't perfect

Results

3 classes

	Noise	*Conversation*	*Shout*
Noise	**77.0**	21.6	7.20
Conversation	19.5	**66.1**	34.8
Shout	3.50	12.3	**58.0**

Table 3. Confusion Matrix. Noise vs. Conversation vs. Shout.

$\rightarrow$ A check on how classes Conversation and Shout separate:

They seem to have some similarities

Results

A practical case: abnormal event detection:

	Everything else	*Shout + Scream*
Everything else	**94.65**	34.96
Shout + Scream	5.35	**65.04**

Table 1. Confusion Matrix. Shout+Scream vs. Everyth. else.

$\rightarrow$ Semi-satisfying

Results

- Compared to a baseline GMM system, our results are slightly better on the whole
- Class Scream didn't have enough occurrence to provide significant results
- 5 classes experiment display big similarities between BG_voice and Conversation, and Shout and Scream, suggesting 3 bigger classes.

5. Conclusion

Conclusion

- Good recognition of abnormal sounds (Shout) versus Noise
- Some misclassification due to Shout containing some Noise intertwined
- Difficulties in distinguishing between Conversation and Shout $\rightarrow$ raise concerns about the definition of classes?
- Difficulty stemming from complexity of the classes: whether vehicle moving or not, within one class.
- Noise class contains too many smaller events (Brake compressor, door signal)

Perspective

- Add temporal information (Recurrent connections, LSTM?).
- Event detection instead of scene (sharpen class definition).
- Superimposition of the classes $\rightarrow$ multi-label learning.

Thanks for your attention