Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

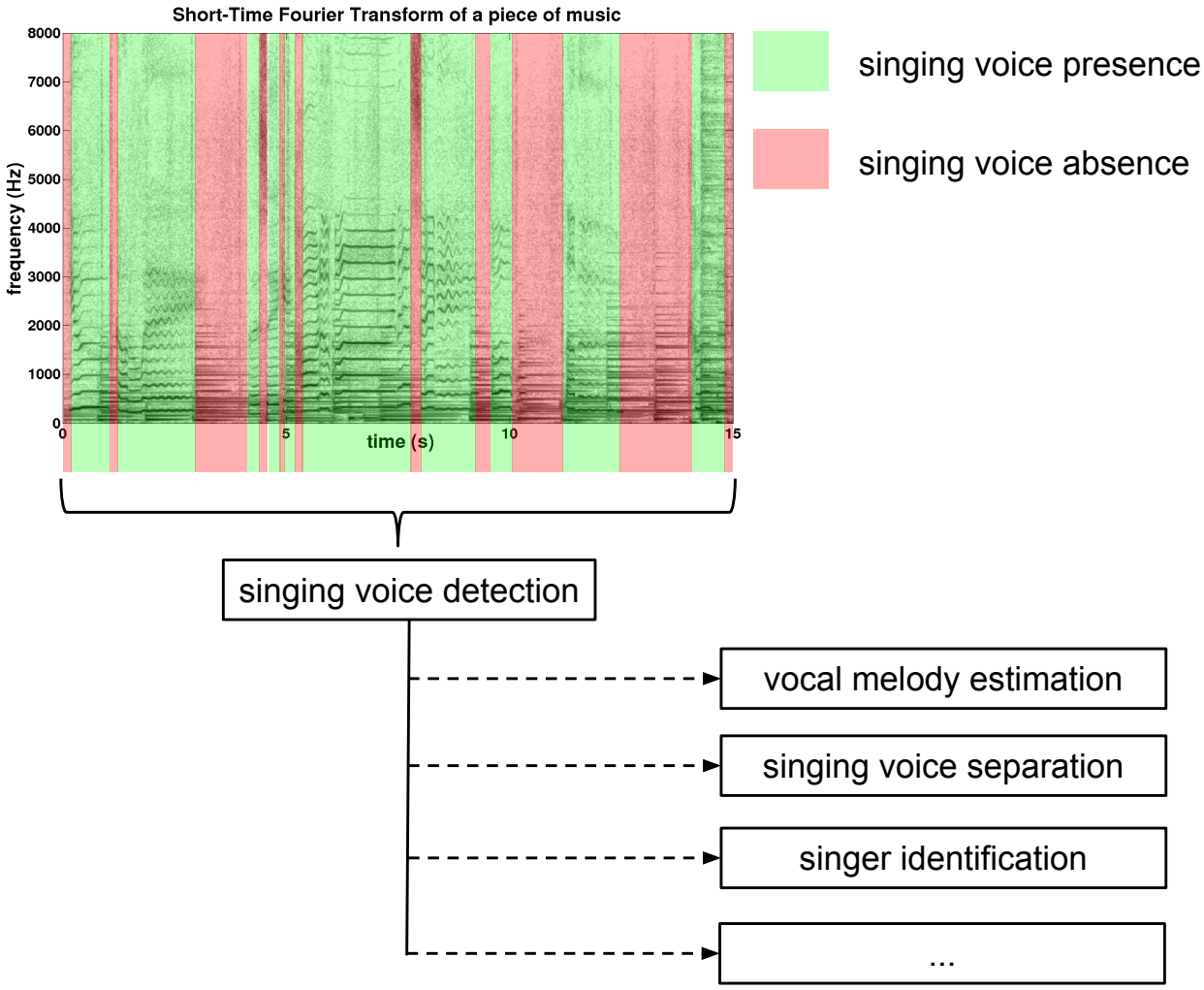
Singing Voice Detection with Deep Recurrent Neural Networks

Simon Leglaive, Romain Hennequin and Roland Badeau

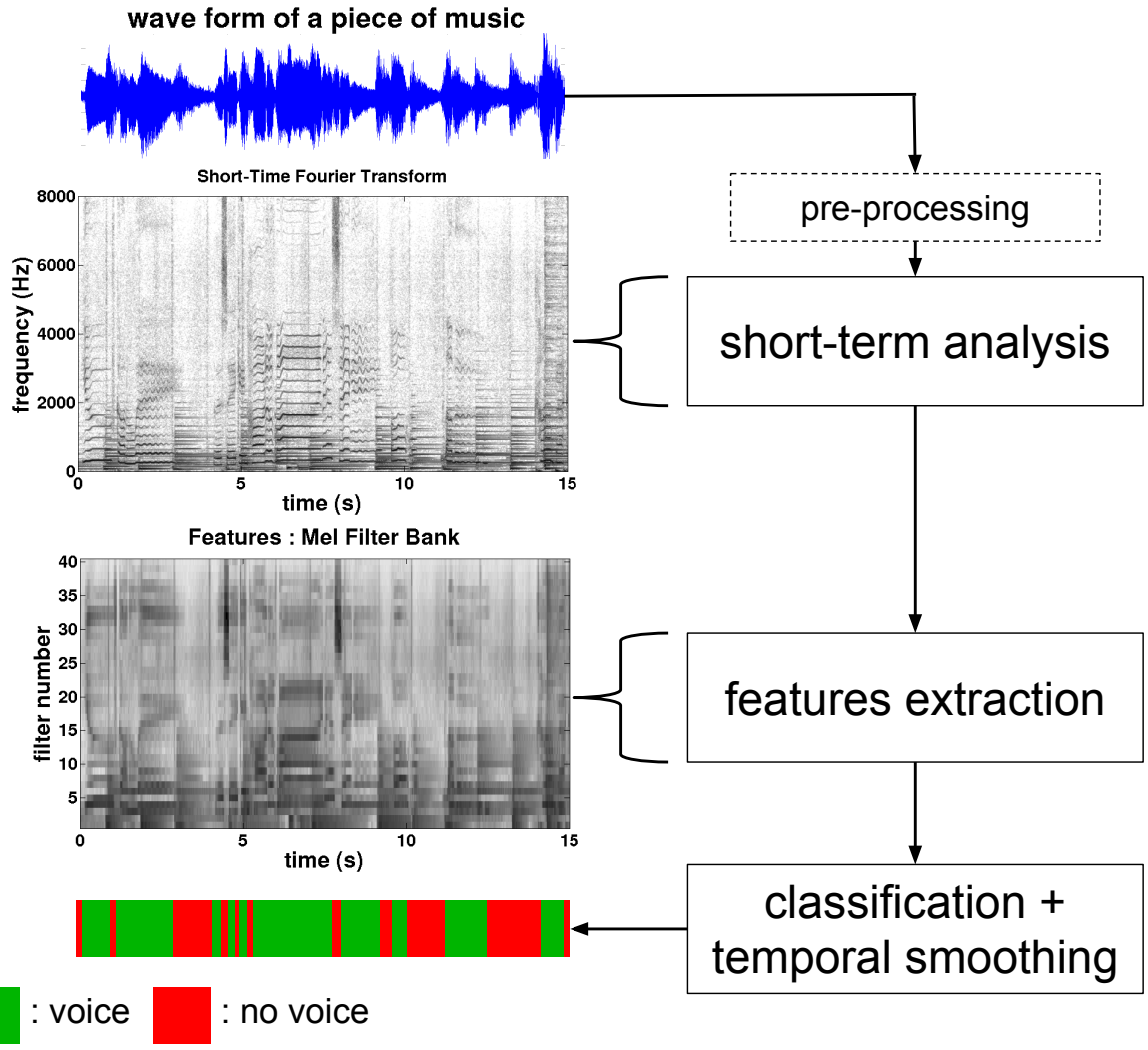
April 24, 2015

40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, April 19-24 2015, Brisbane, Australia

Introduction



Introduction



Introduction

Our method :

- Bidirectional Long-Short Term Memory (BLSTM) Recurrent Neural Network (RNN)

  $\longrightarrow$ long past and future **temporal context**

- With several hidden-layers

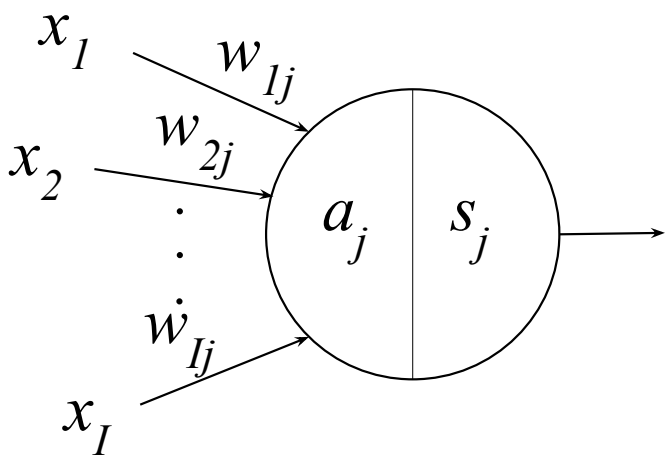
  $\longrightarrow$ extract simple useful information from **low level features**

Outline

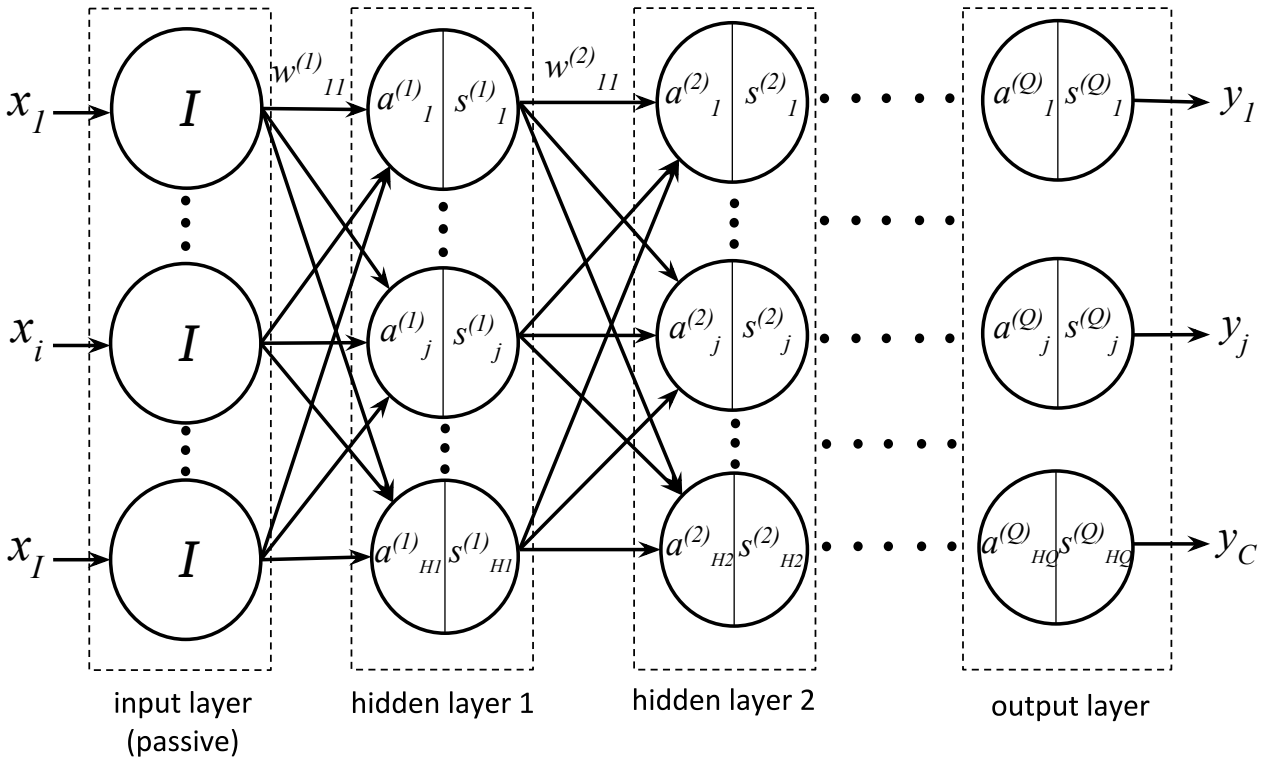
Outline

Formal Neuron

- **Activation** : $a = \sum_{i=1}^{I} w_i x_i$
- **Output** : $s = f(a)$ with f the nonlinear activation function (e.g. step function, sigmoid, hyperbolic tangent, ...)

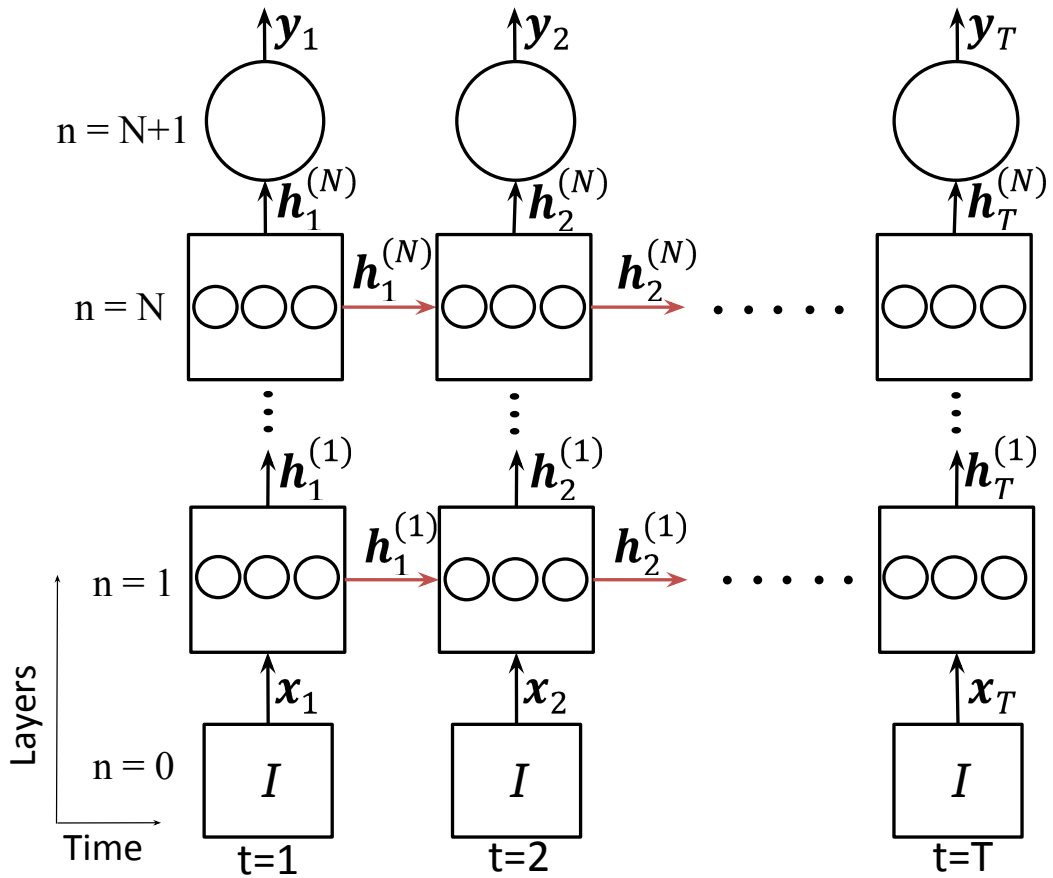


Multi-Layer Perceptron



- Feedforward Artificial Neural Network
- Maps inputs to outputs by propagating data through the layers
- Training : Gradient descent using backpropagation

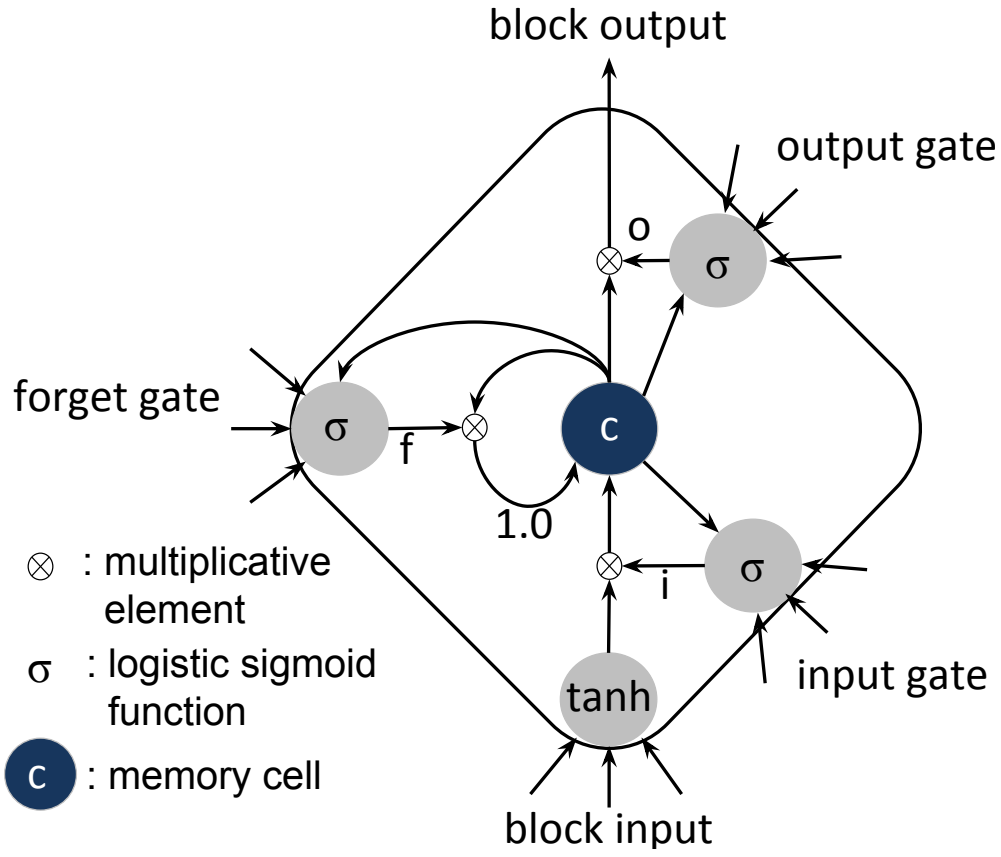
Recurrent Neural Network



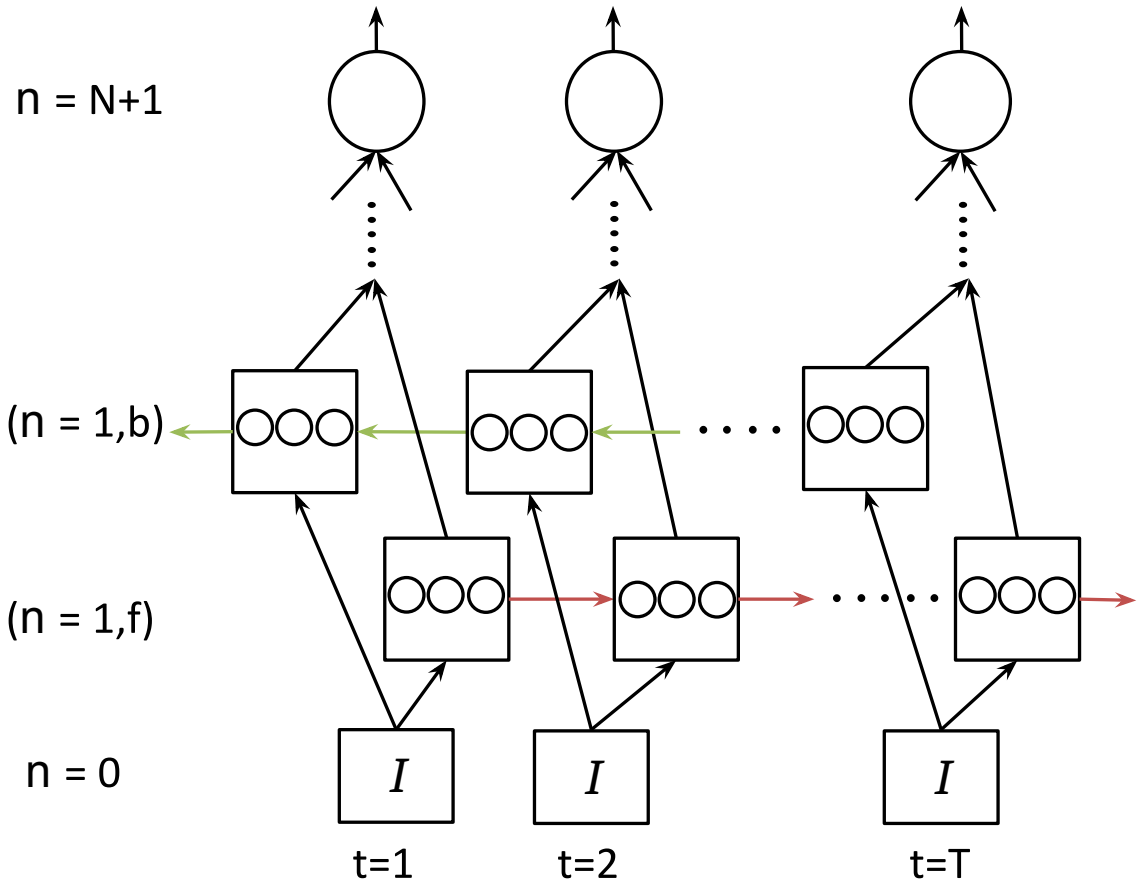
Recurrent Neural Network unfolded in time

Long Short-Term Memory

- Memory cell
- Input Gate, Output Gate, Forget Gate = Write, Read and Reset



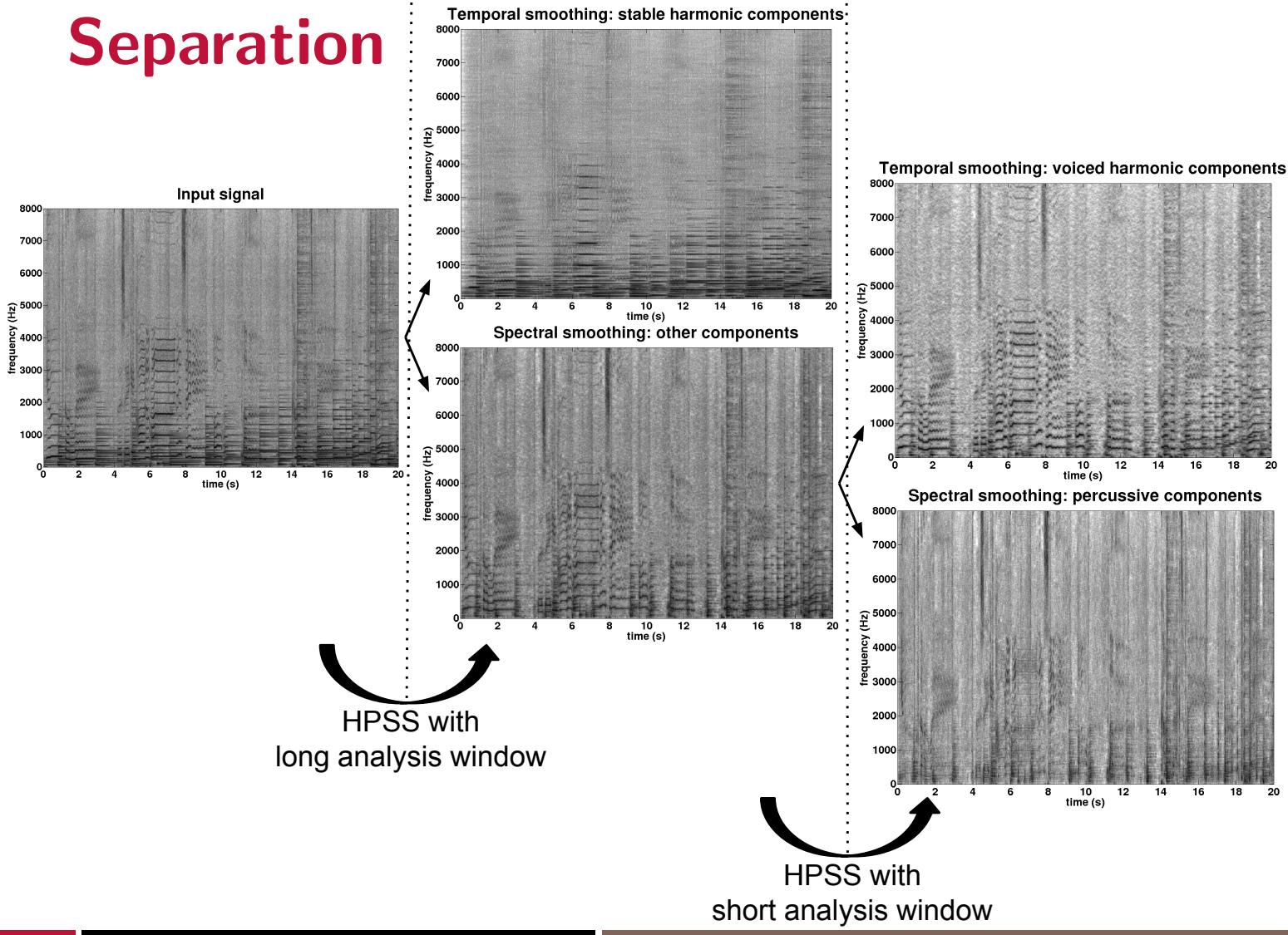
Bidirectional Recurrent Neural Network



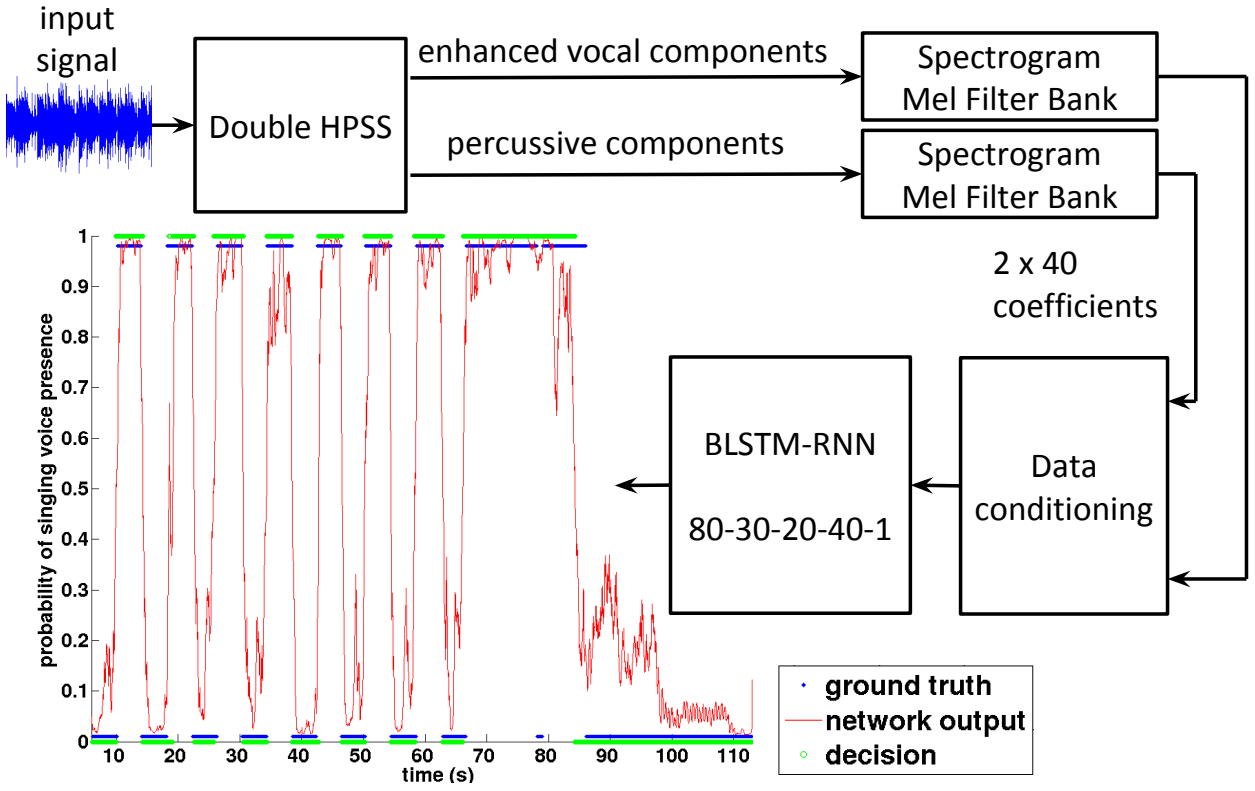
Bidirectional Recurrent Neural Network unfolded in time

Outline

Double Harmonic/Percussive Source Separation

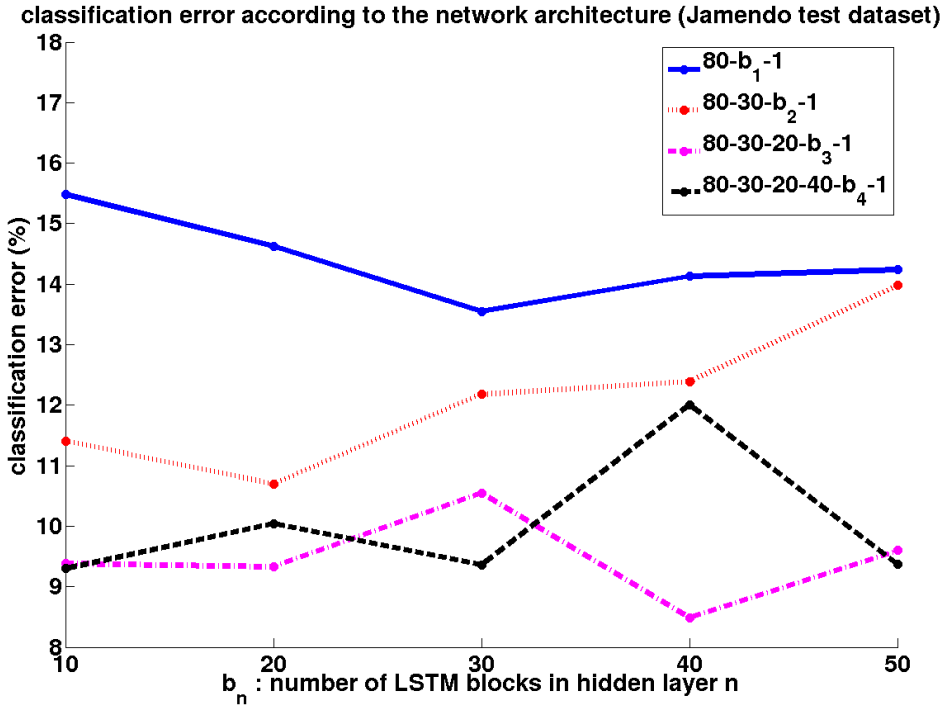


Global System



Building the Network

- No theoretical evidence $\longrightarrow$ empirical approach, not much discussed in papers
- Incremental procedure : depth increased by progressively adding hidden layers

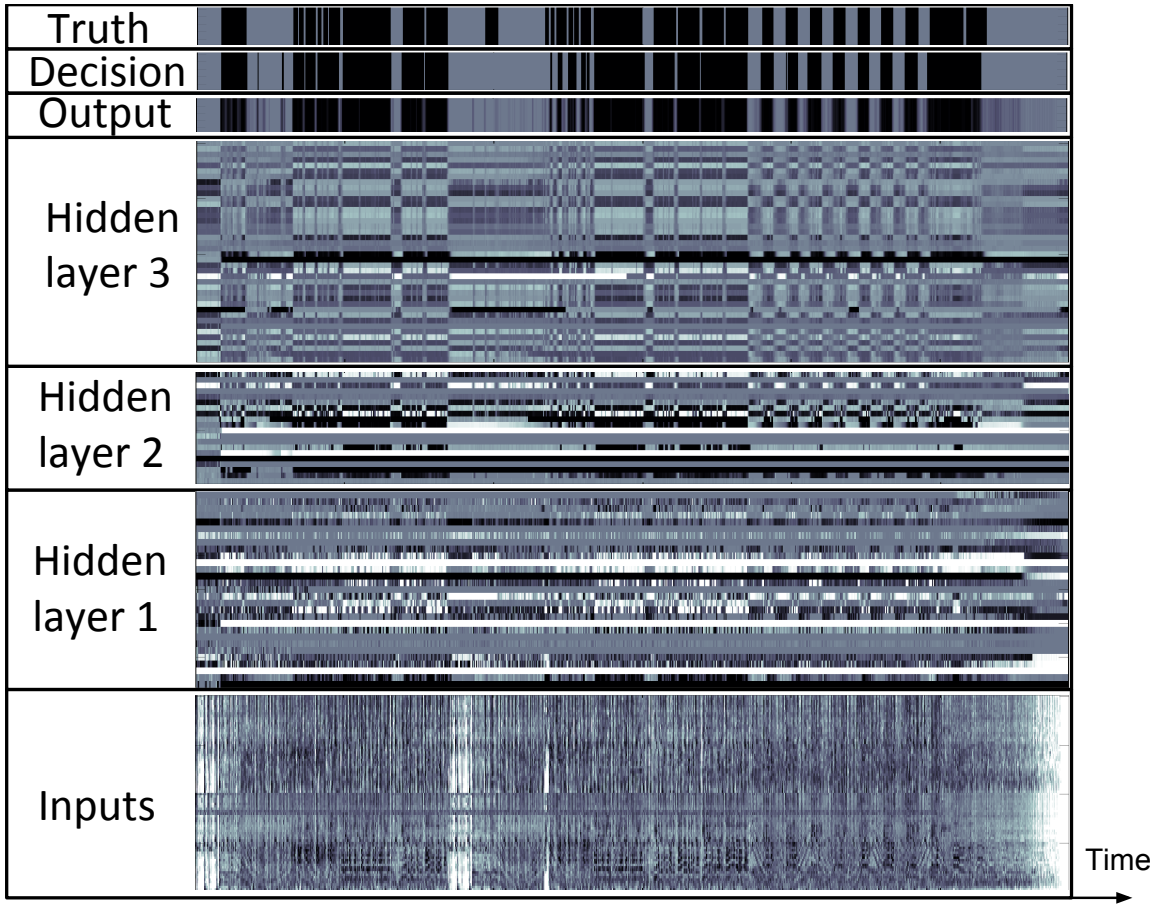


Outline

Jamendo : A Common Benchmark Dataset

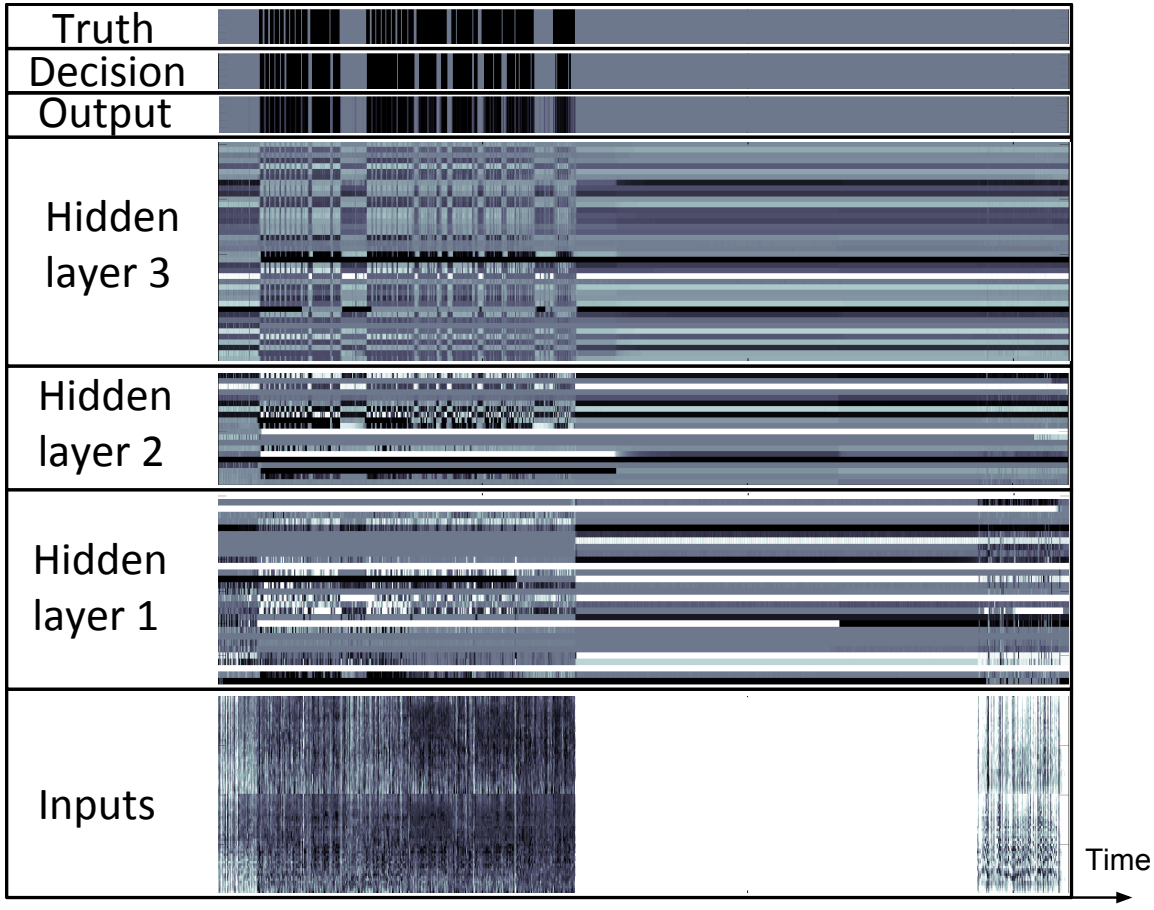
- Publicly available dataset
- Singing voice activity annotations
- Training set : 61 files
- Validation and Test sets : 16 files each
- Common database $\longrightarrow$ fair comparison of our approach

Internal Netwok Functioning

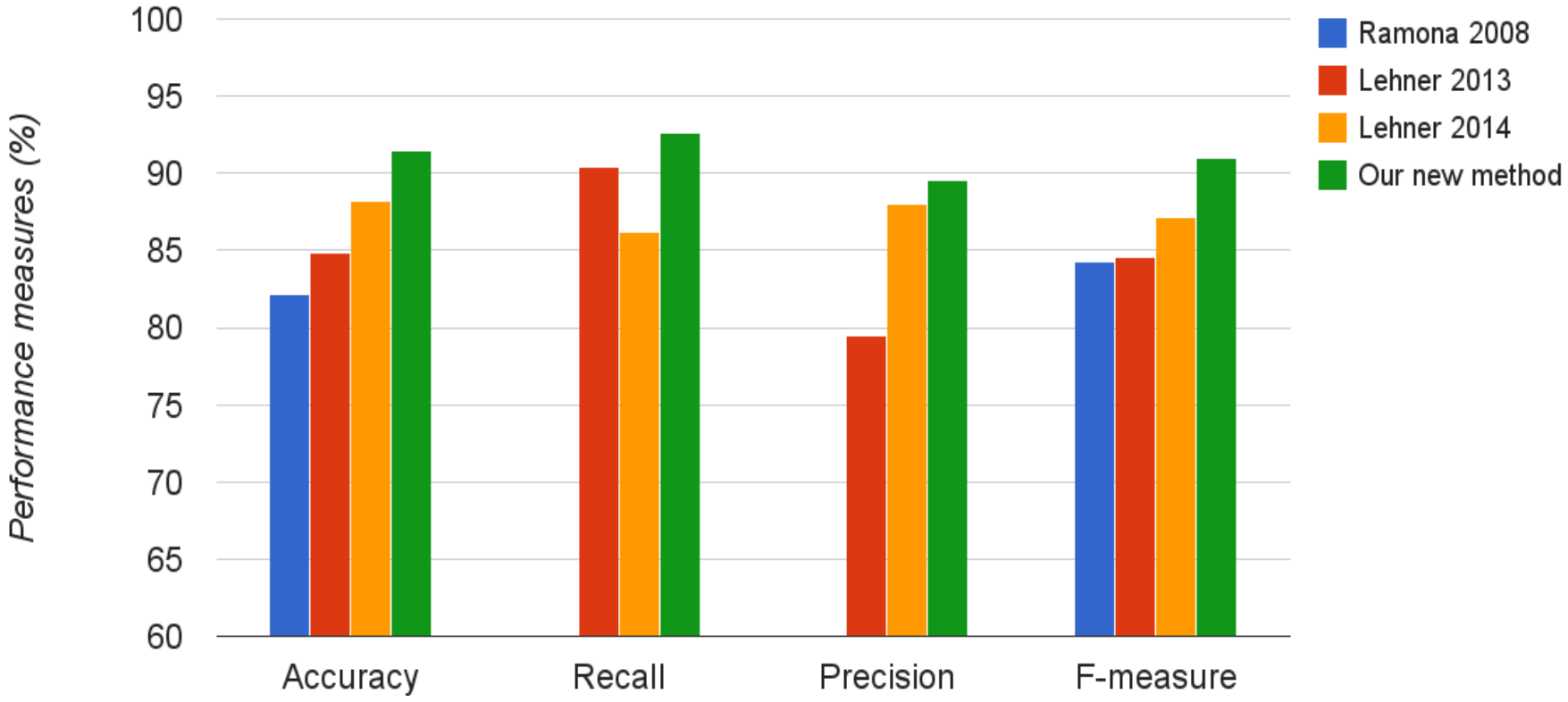


Color scale between -1 (white) and 1 (black)

Consideration of a Temporal Context



Results on Jamendo Dataset



Conclusion

- New approach for singing voice detection
- We do not focus on defining a complex feature set
  $\longrightarrow$ may be suboptimal
- We make use of neural networks to extract a simple representation, fitted to our task
- A past and future temporal context is considered by the classifier
  $\longrightarrow$ no need for temporal smoothing
- The results we obtain encourage further work with BLSTM-RNN in MIR for sequence classification tasks, e.g. melody estimation

Thank you