A brief introduction to multichannel noise reduction with deep neural networks

Romain Serizel

Loria

Laboratoire lorrain de recherche en informatique et ses applications

A brief introduction to multichannel noise reduction with deep neural network.

Romain Serizel

LORIA, Univeristé de Lorraine, Inria, CNRS (France)

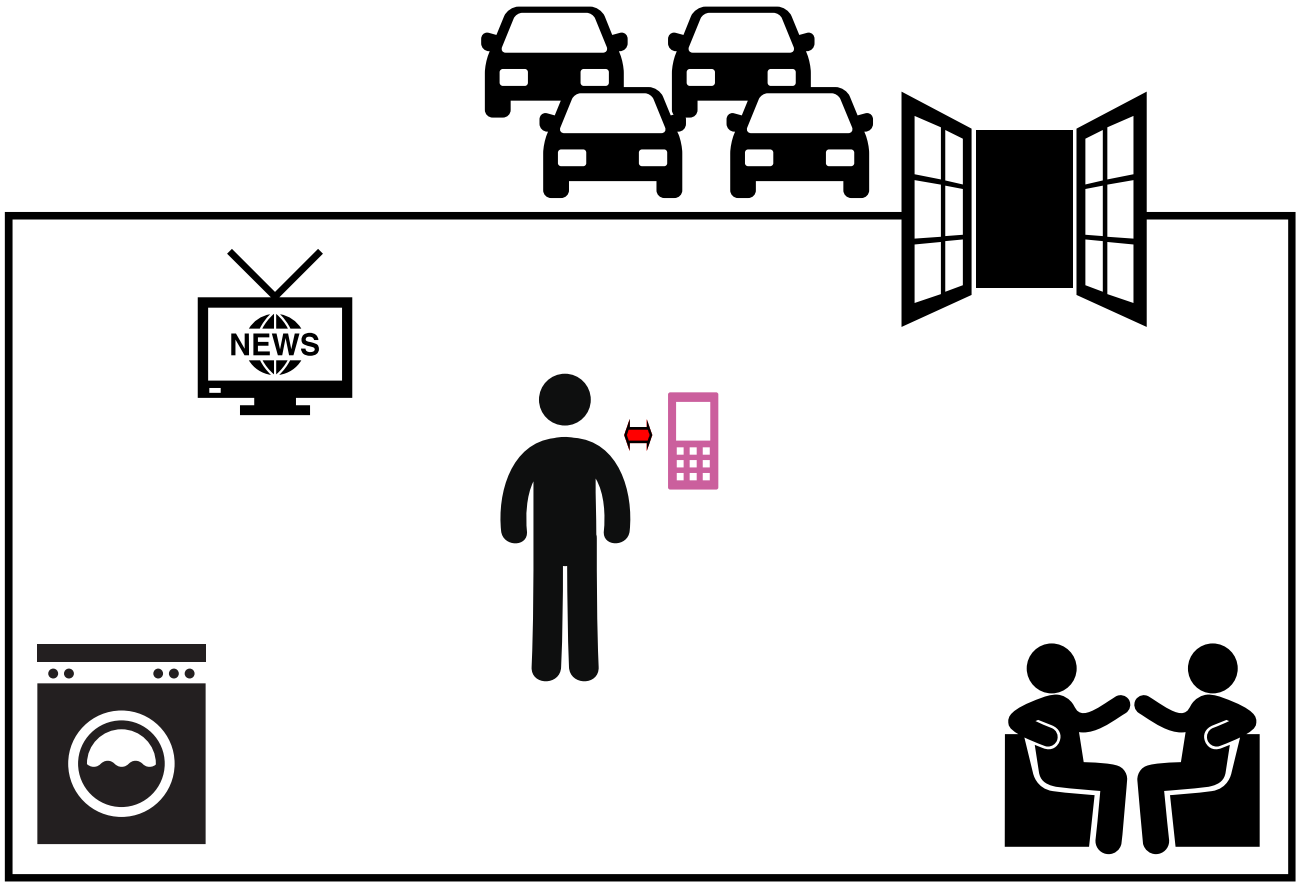
Thursday 9th, January 2020

Outline

Speech in real conditions

Problem “solved”

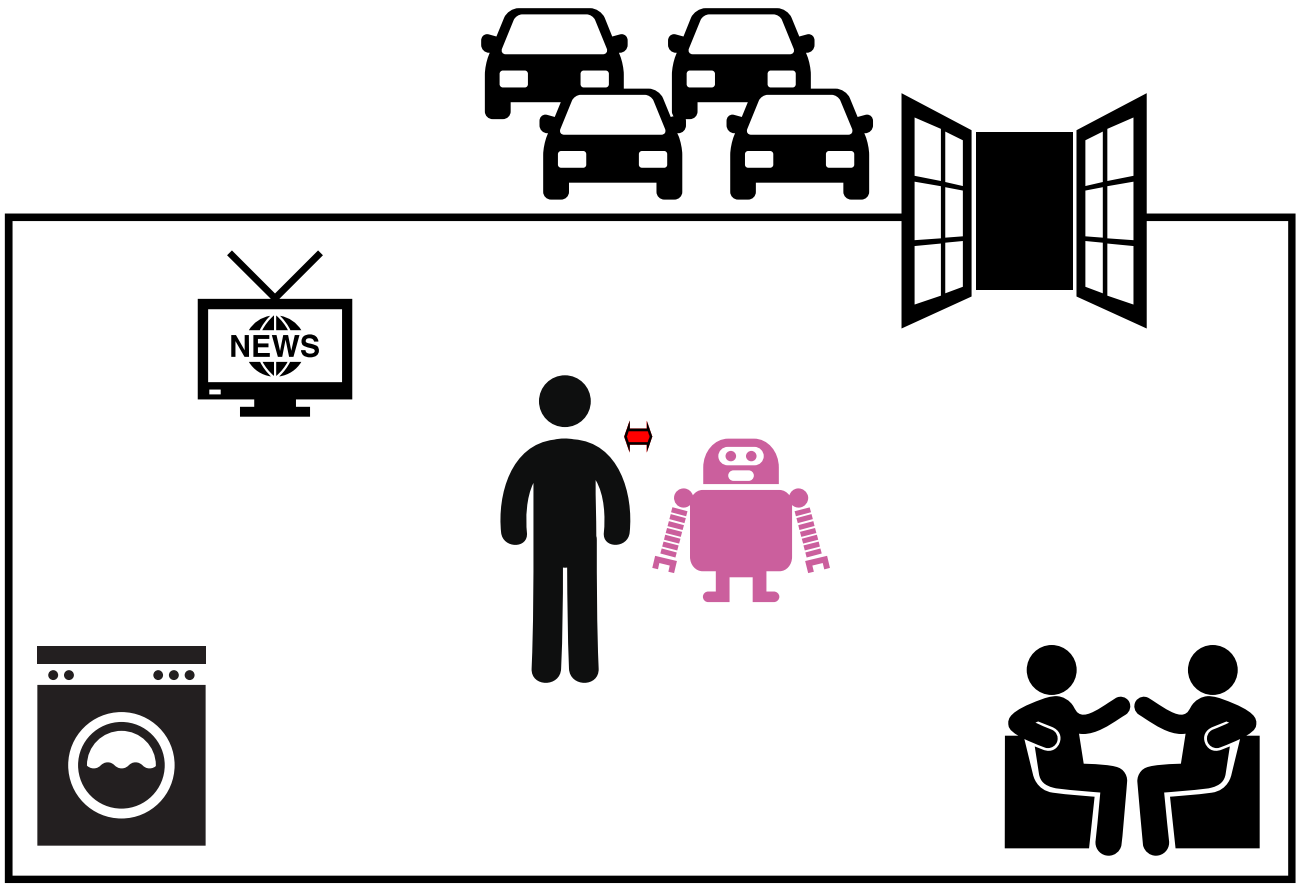
Close-up microphones, clean speech



Speech in real conditions

There is is some work to do :

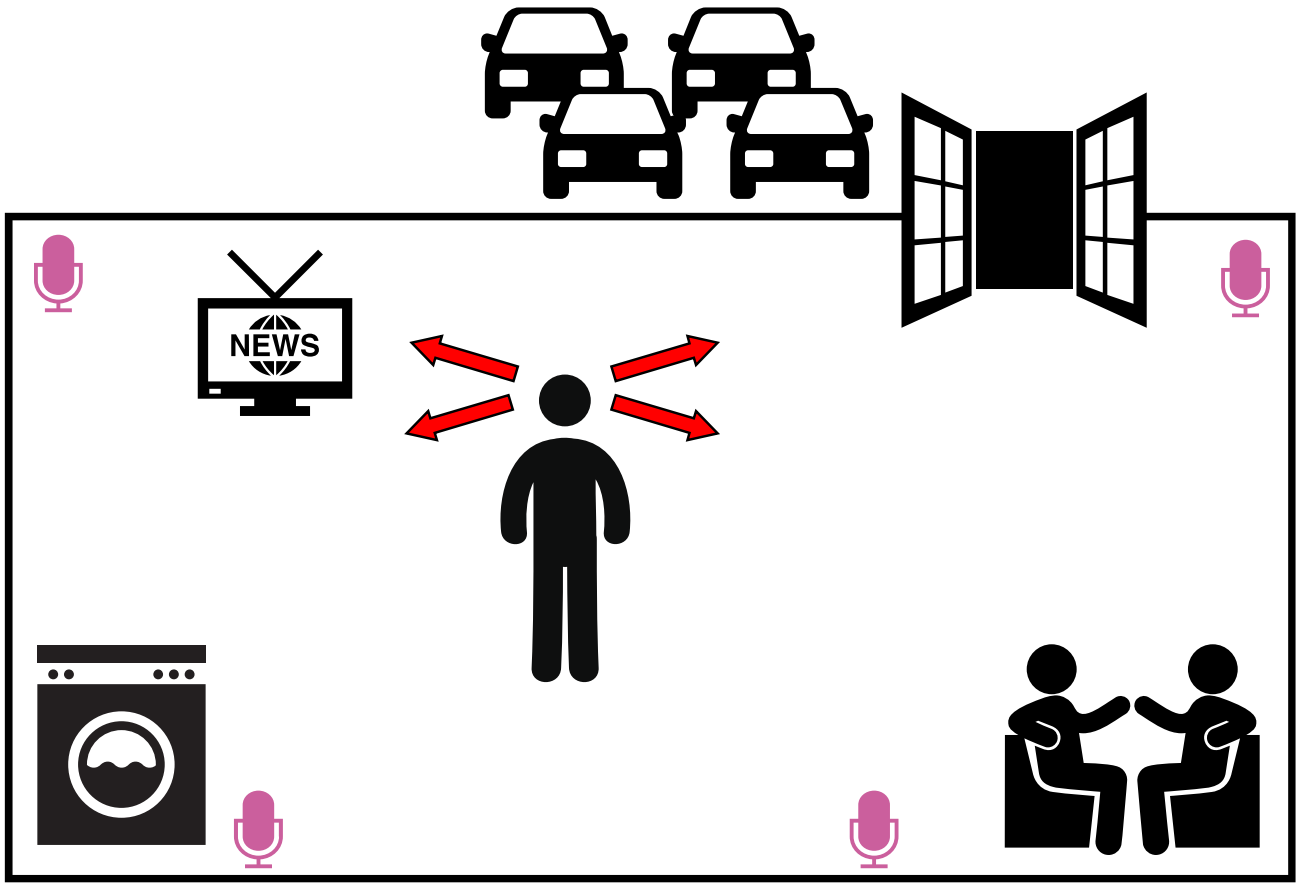
Moving microphones



Speech in real conditions

There is is some work to do :

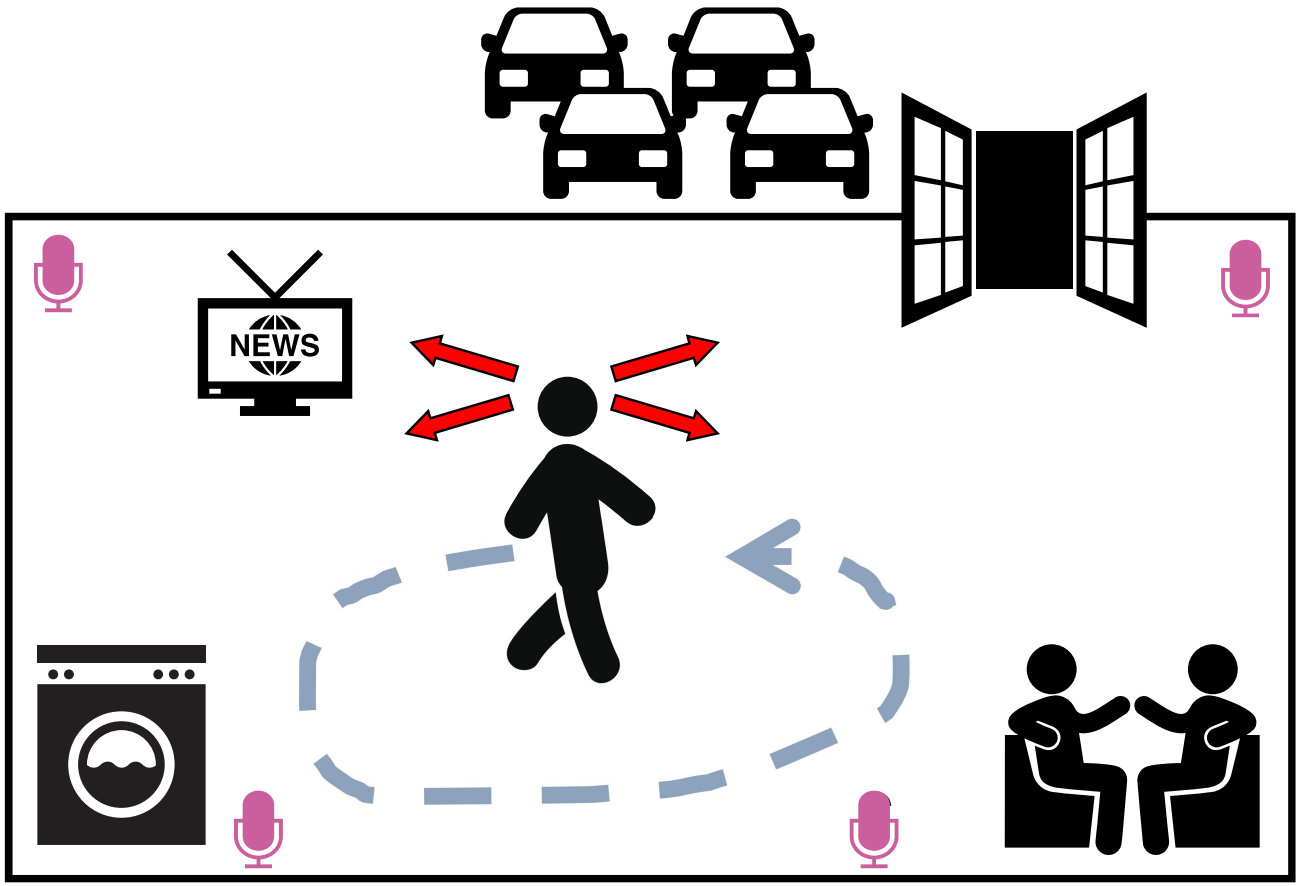
Distant microphones



Speech in real conditions

There is is some work to do :

Distant microphones, mobile sources. . .



Multichannel speech enhancement ?

Context

- Reduce the impact of the acoustic perturbations :
  - Cleaning the signal

In this talk

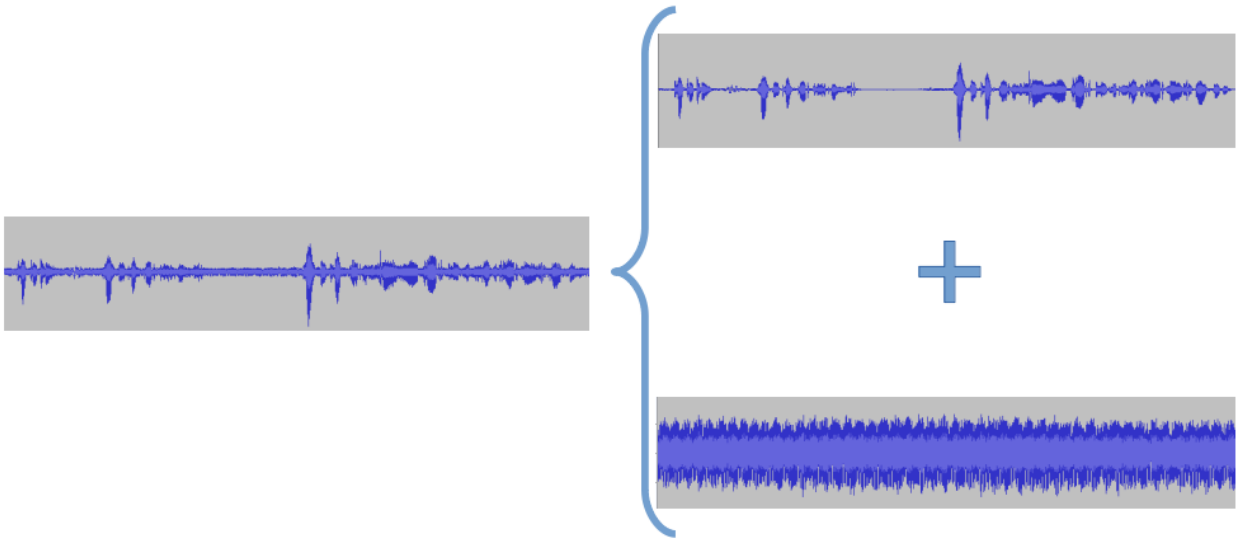
- Focus on noise removal/attenuation
  - Noise reduction
  - Noise suppression
  - De-noising
  - ...

Outline

Signal model

Signal observed

$$\mathbf{X} = \mathbf{X}_s - \mathbf{X}_n$$



Filtering

Single-channel filter

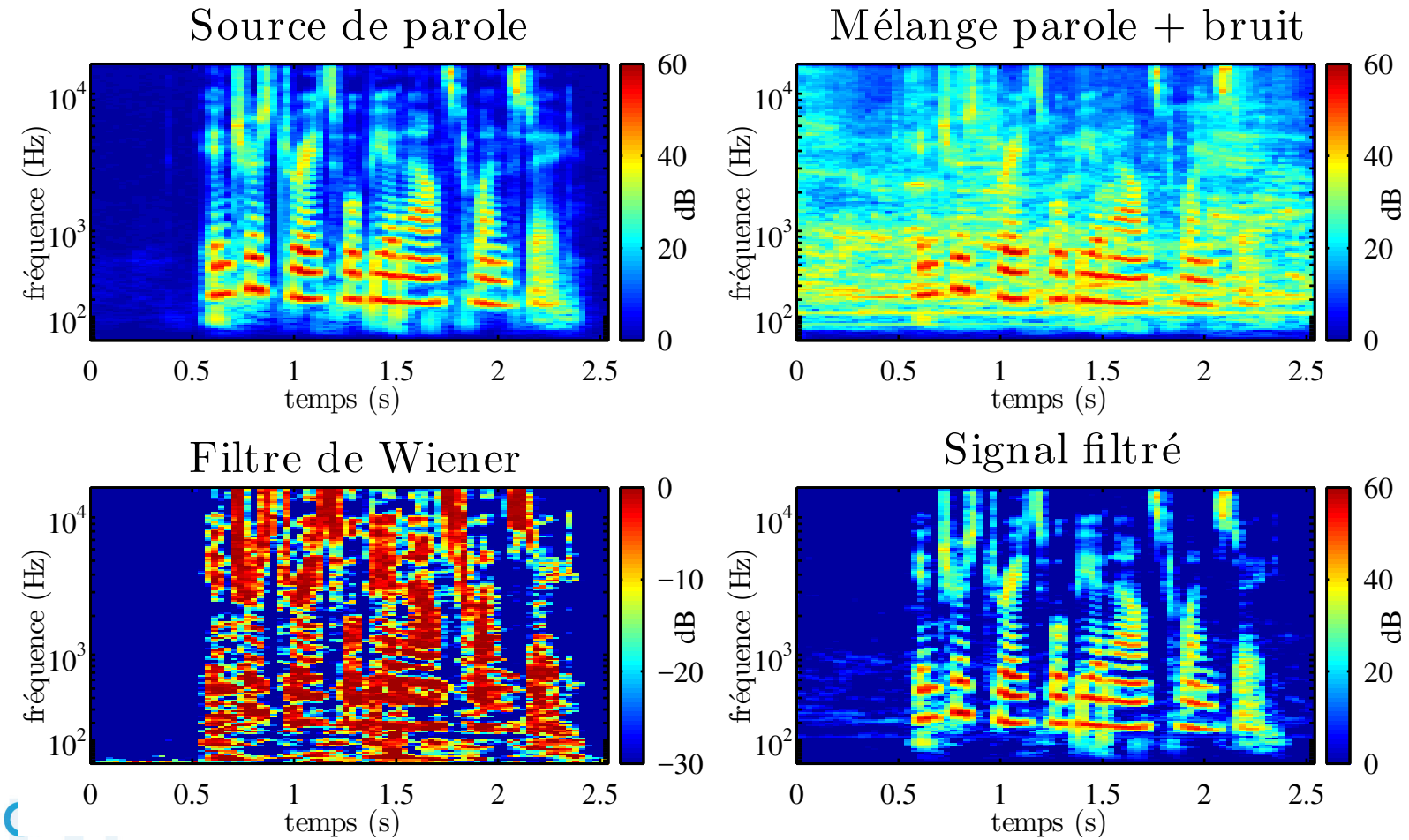
Filtering

Single-channel filter

Multichannel filter

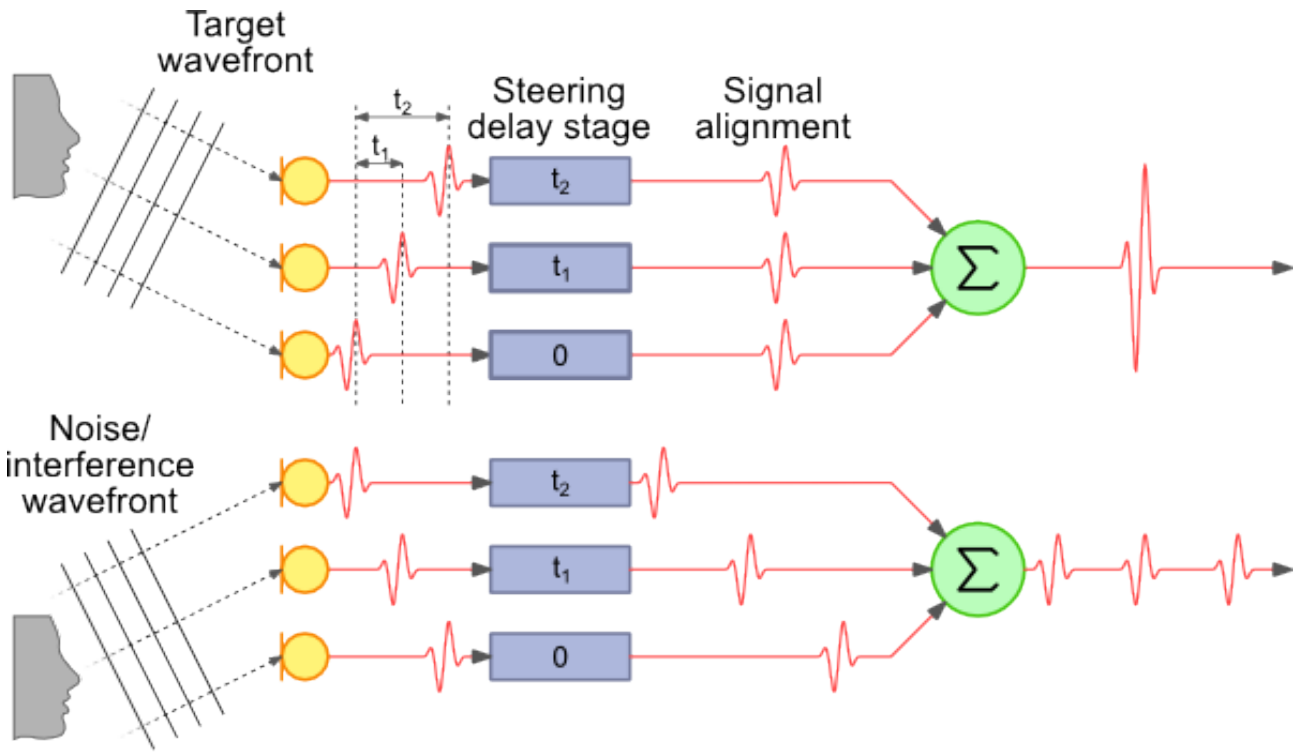
Single-channel enhacement

On single-mic data, perform spectral filtering/masking.



Multichannel enhancement (1)

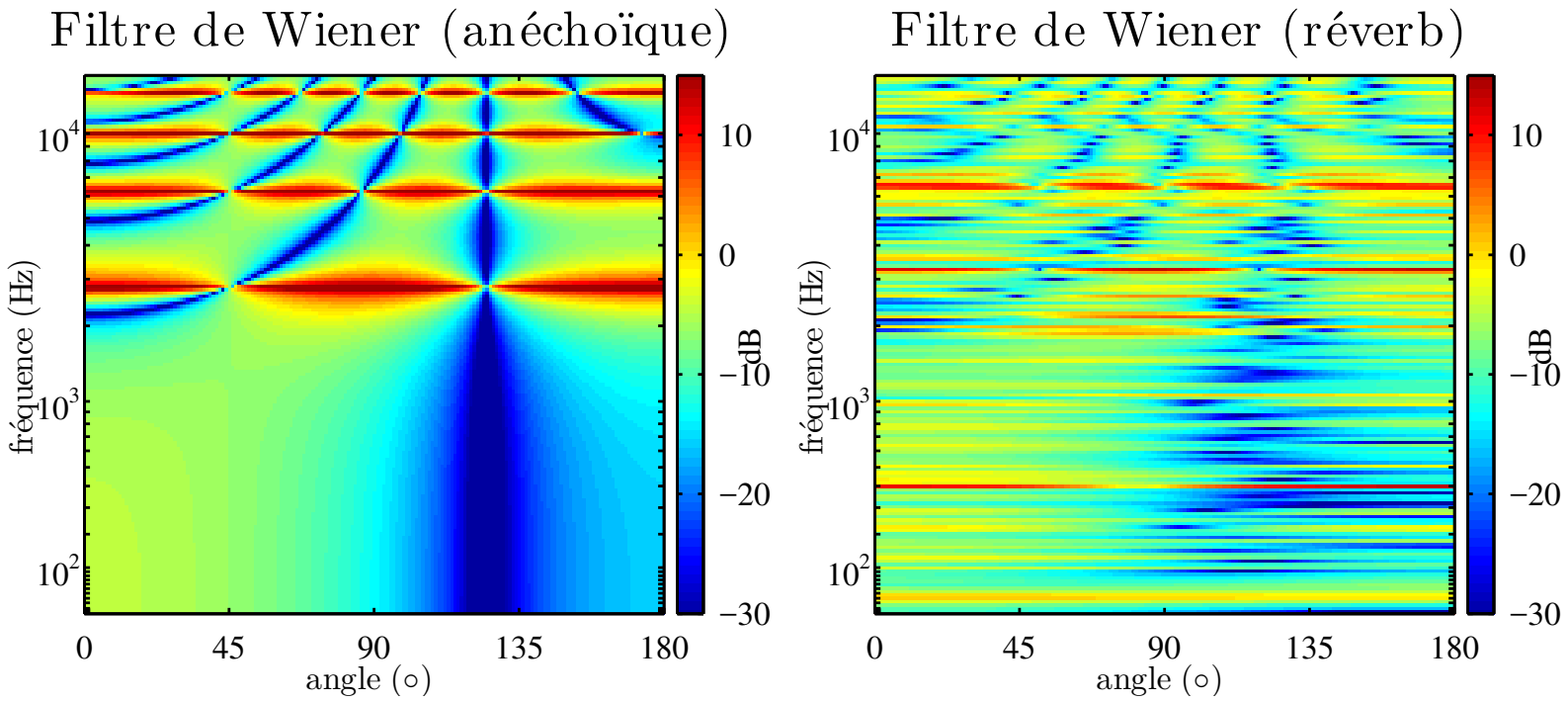
On multi-mic data, exploit spatial localization.



. Image source : the lab book page

Multichannel enhancement (2)

On multi-mic data, perform joint spatial and spectral masking.



MWF-based speech enhancement (1)

Speech distortion weighted MWF (MMSE criterion)

MMSE estimate of the (unknown) speech component :

$$\min_{\mathbf{W}} \; E\{|X_{\text{Ref}} - \mathbf{W}^H \mathbf{X}_s|^2\} + \mu\{\mathbf{W}^H \mathbf{X}_n|^2\}$$

SDW-MWF

$$\mathbf{W} = (\mathbf{R}_s + \mu \mathbf{R}_n)^{-1} \mathbf{R}_s \mathbf{e}_{\text{Ref}}$$

- $\mathbf{X}_s$, $\mathbf{X}_n$: speech, noise components of the microphone signals
- $\mathbf{W}^H\mathbf{X}$: output signal of the filter $\mathbf{W}$
- $\mathbf{R}_s$ and $\mathbf{R}_n$: speech and noise correlation matrices

MWF-based speech enhancement (2)

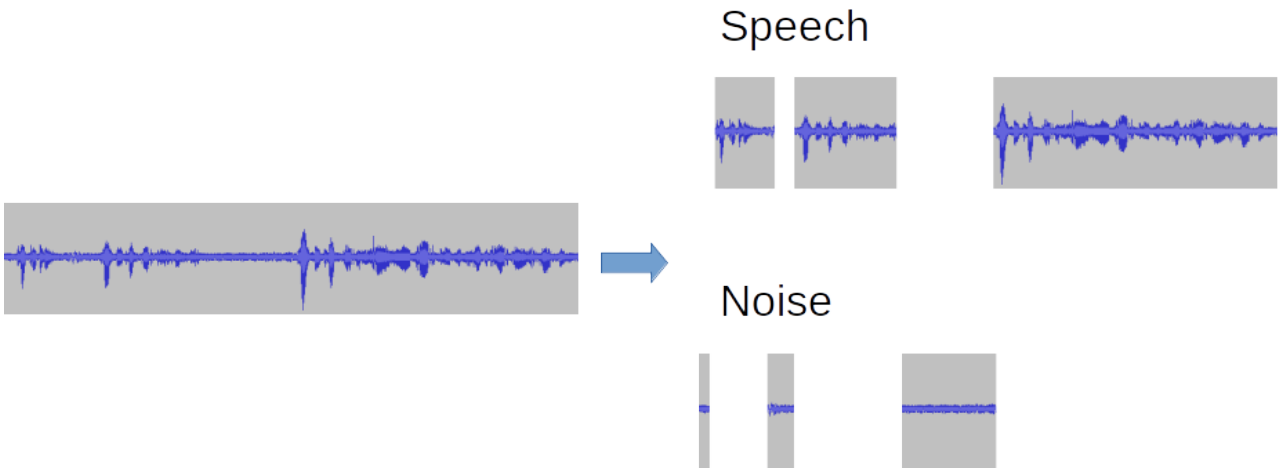
Estimation of the speech correlation matrix

$\mathbf{R}_s$ is estimated from speech+noise correlation matrix $\mathbf{R}_x$
and noise correlation matrix $\mathbf{R}_n$:

$$\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}_n$$

➔ Requires a voice activity detector !

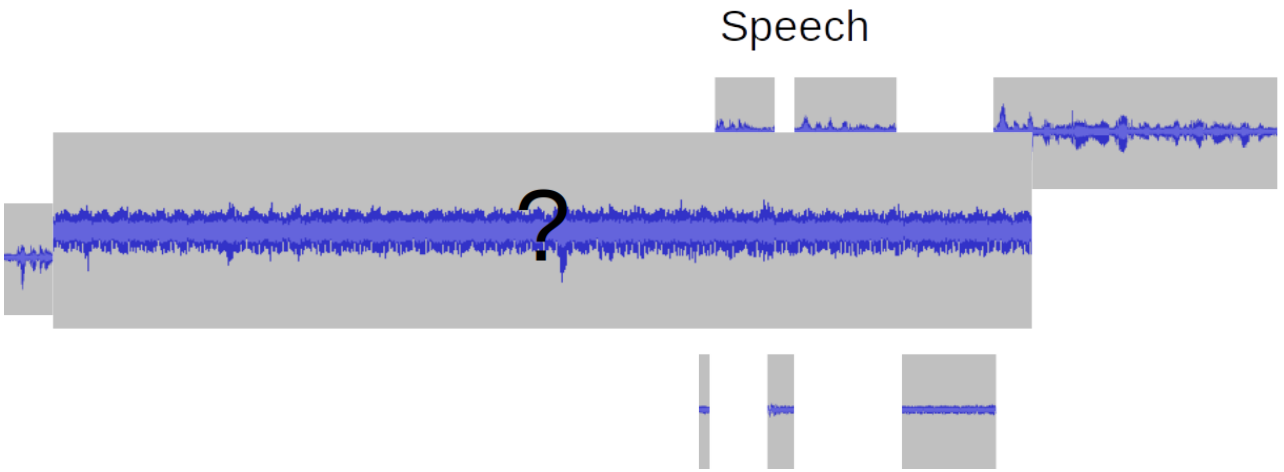
Voice activity detector



Limitations

- Limited to full frequency band
- Energy-based VAD do not work well at low SNR

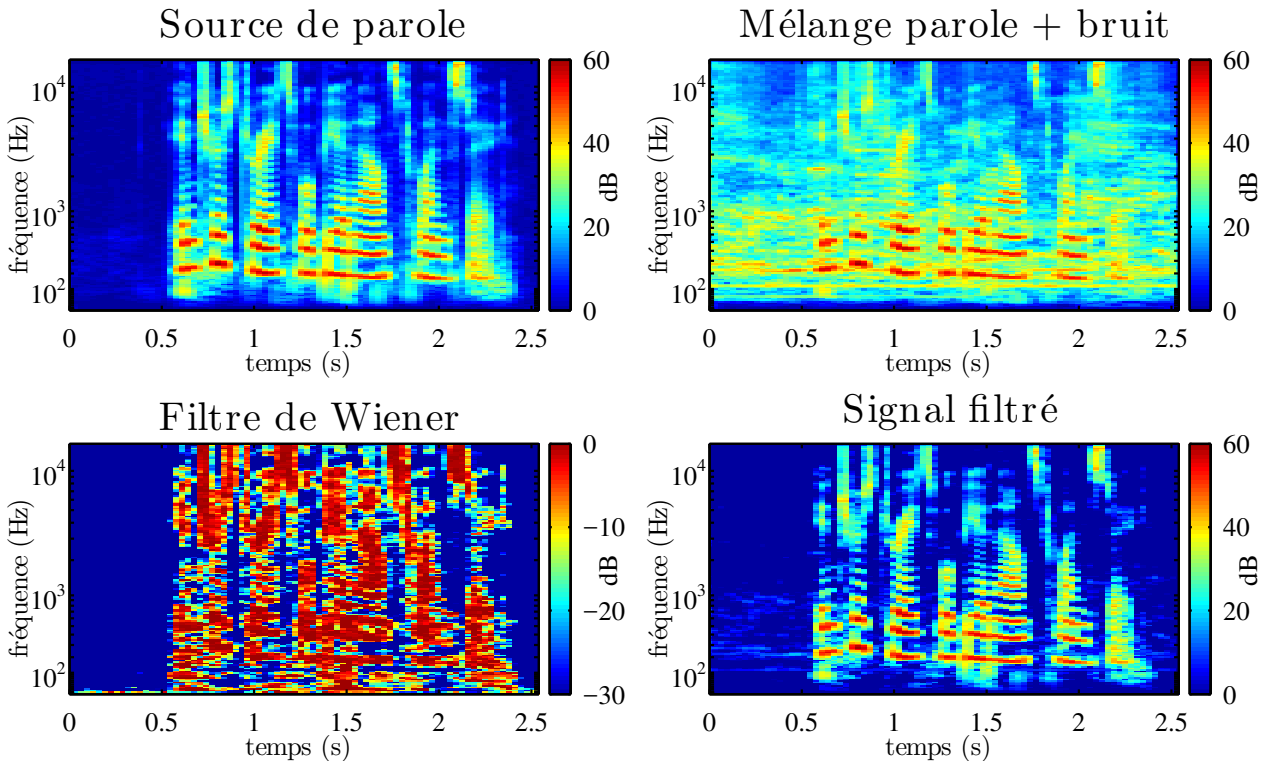
Voice activity detector



Limitations

- Limited to full frequency band
- Energy-based VAD do not work well at low SNR

Time-frequency masks

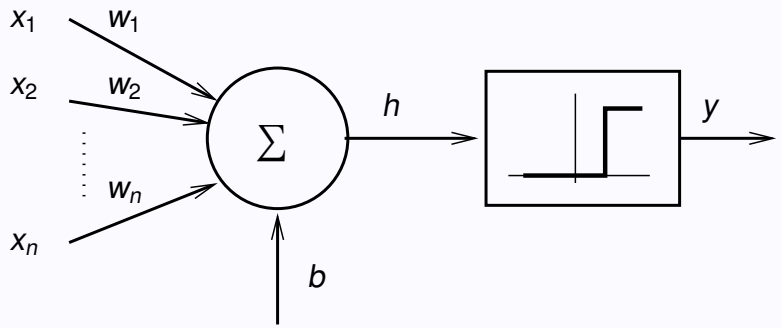


How can we obtain reliable masks ?

Outline

The perceptron 1

Perceptron model



A perceptron is a linear classifier :

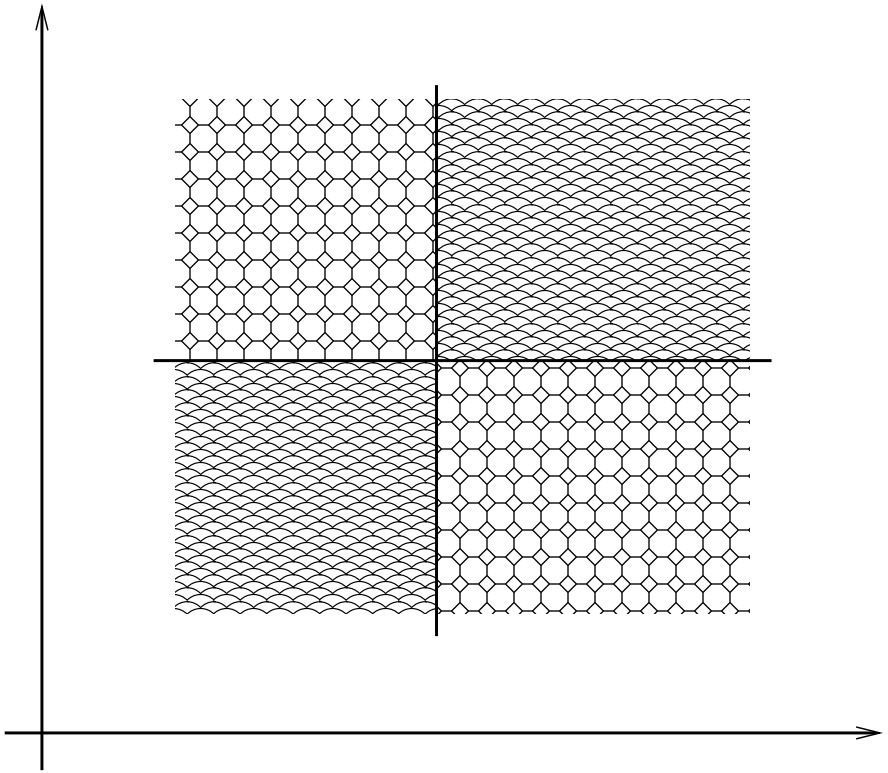
$$h = \mathbf{w}.\mathbf{x} + b$$

$$y = f(h) \; e.g. \begin{cases} 1 & if \quad \mathbf{w}.\mathbf{x} + b > 0 \\ 0 & else \end{cases}$$

The perceptron 2

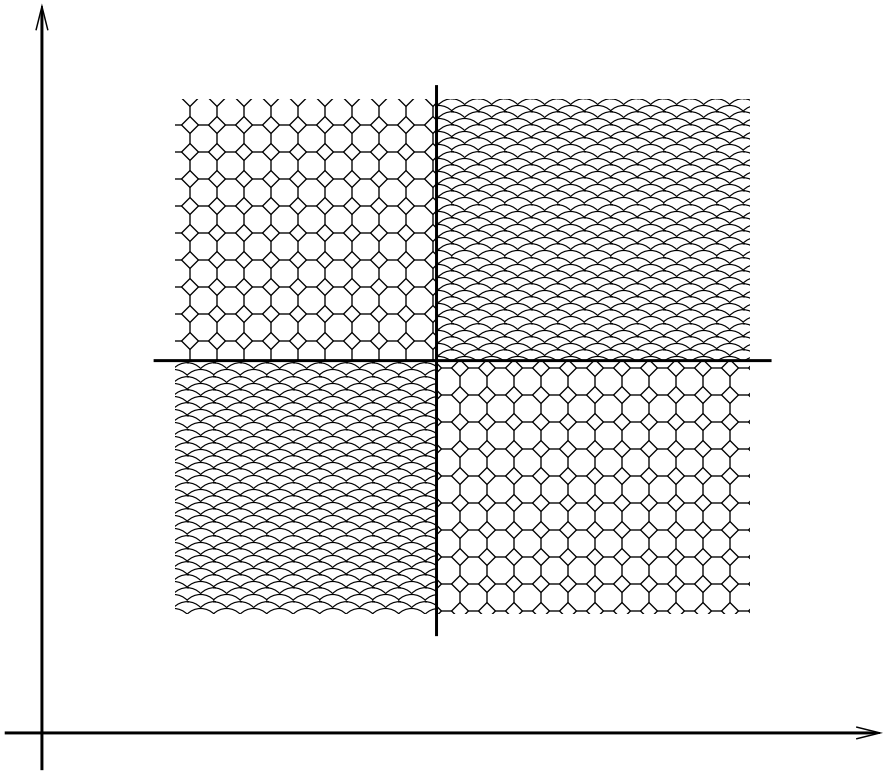
Perceptron cannot model a XOR [Minsky *et al.*, 1969]

⇒ Work on perceptron stopped



The perceptron 2

Perceptron cannot model a XOR [Minsky *et al.*, 1969]
⇒ Work on perceptron stopped

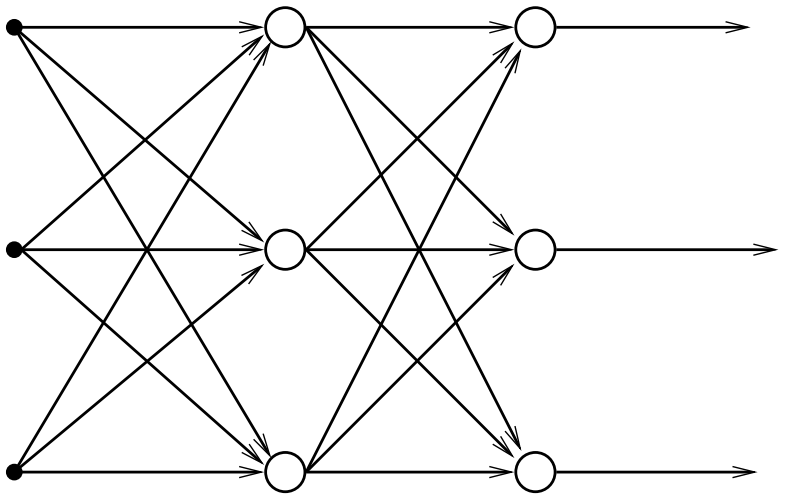


At least single perceptron cannot !

Multilayer perceptrons

Multilayer perceptrons (MLP) :

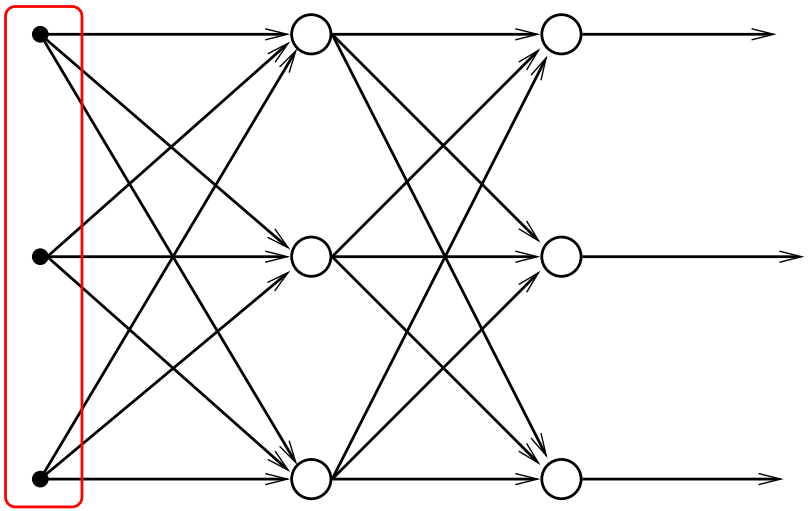
- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification



Multilayer perceptrons

Multilayer perceptrons (MLP) :

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

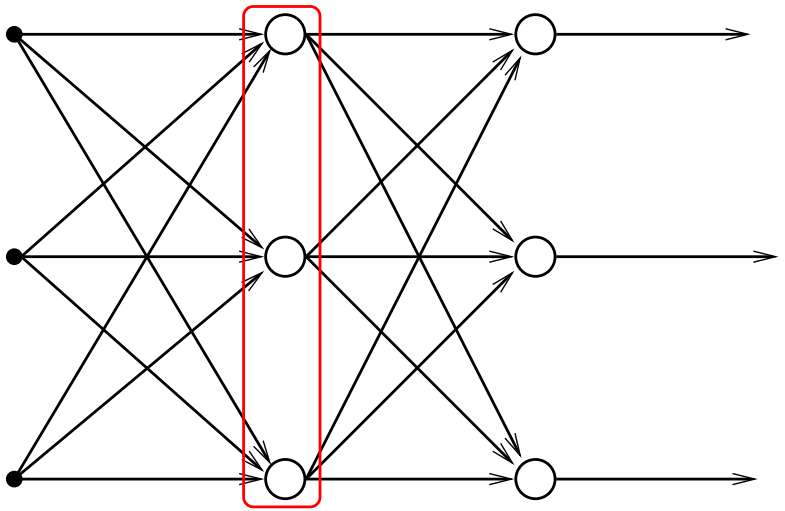


Input layer

Multilayer perceptrons

Multilayer perceptrons (MLP) :

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

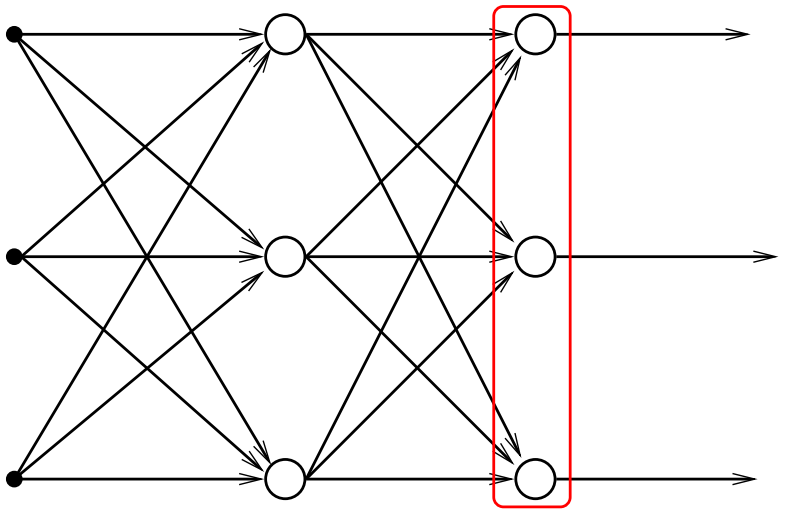


Hidden layer(s)

Multilayer perceptrons

Multilayer perceptrons (MLP) :

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

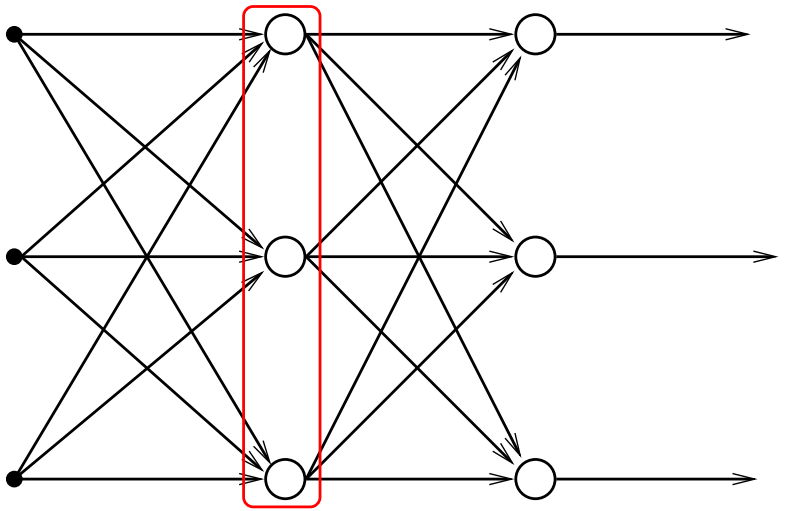


Output layer

Multilayer perceptrons

Multilayer perceptrons (MLP) :

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

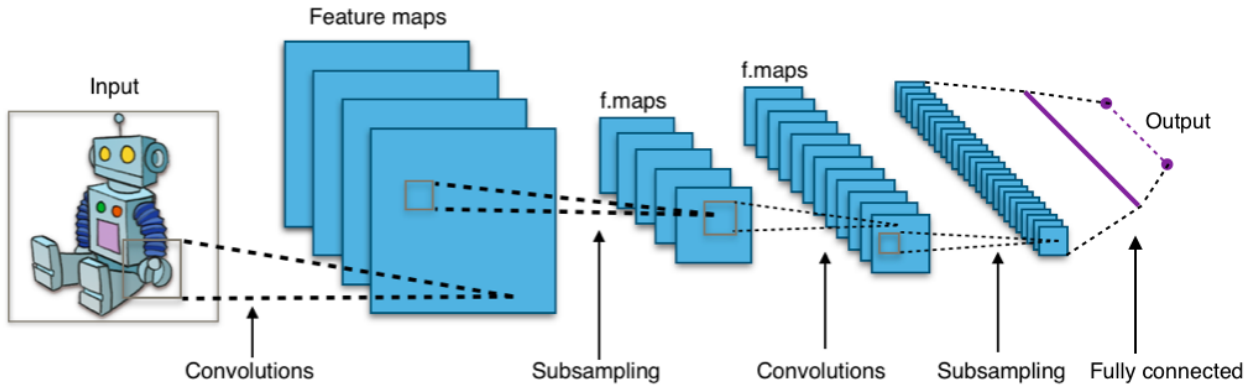


Deep neural networks have more than 1 hidden layers

Convolutional neural networks (CNN)

Keys ideas

- Detect patterns in images/spectrograms
- Apply sets of filters to obtain feature maps
- Subsample to control dimensionality

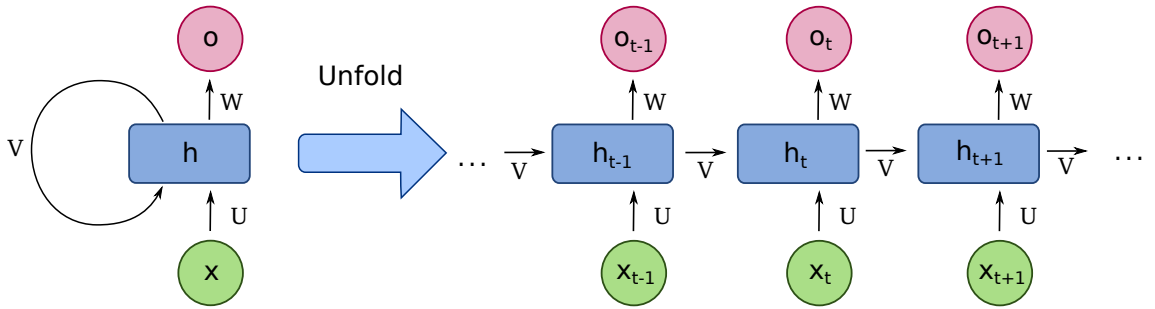


. Image source : wikimedia

Recurrent neural networks (RNN)

Keys idea

- Model context

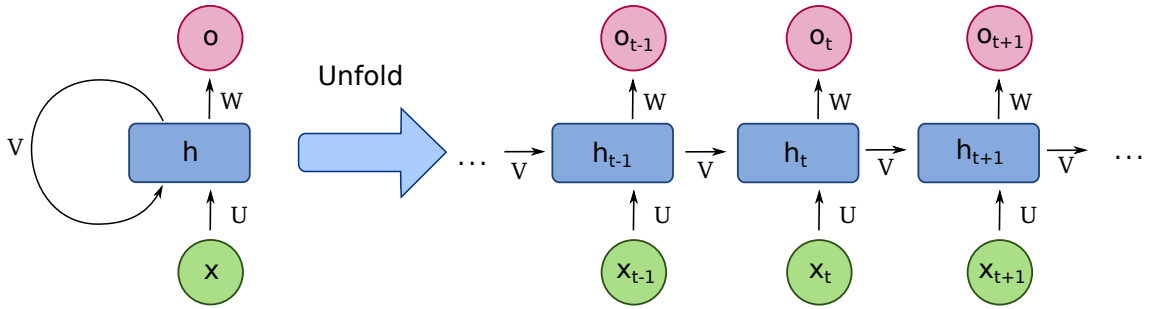


. Image source : wikimedia

Recurrent neural networks (RNN)

Keys idea

- Model context



Problems

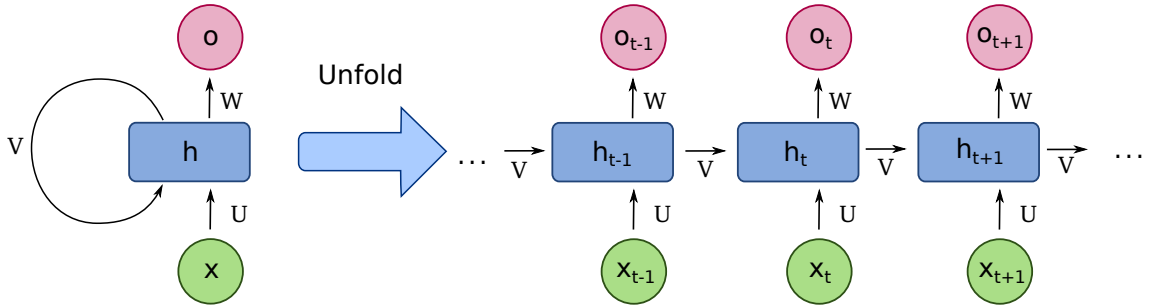
- Hard to train
- Exploding/vanishing gradient

. Image source : wikimedia

Recurrent neural networks (RNN)

Keys idea

- Model context



Solutions

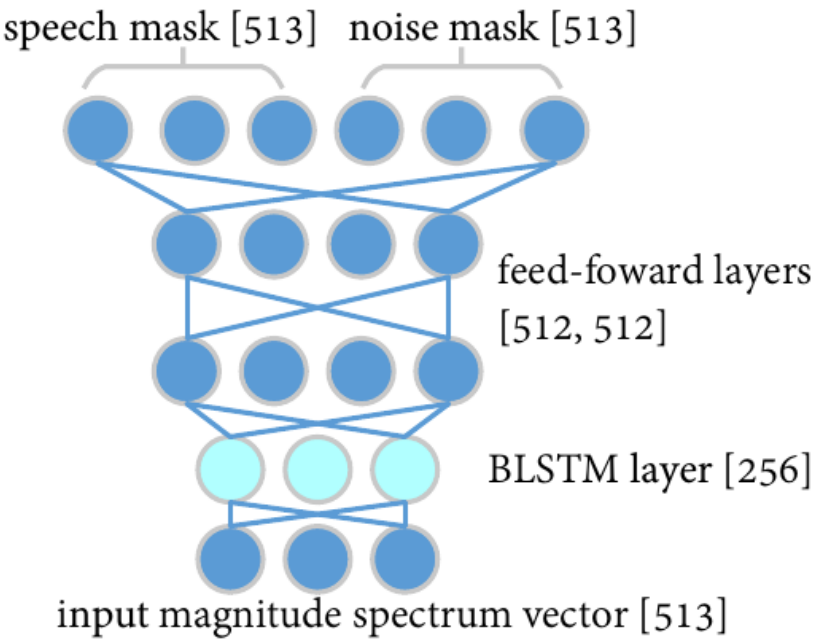
- LSTM, GRU. . .

. Image source : wikimedia

Outline

DNN-base mask estimation for multichannel filtering

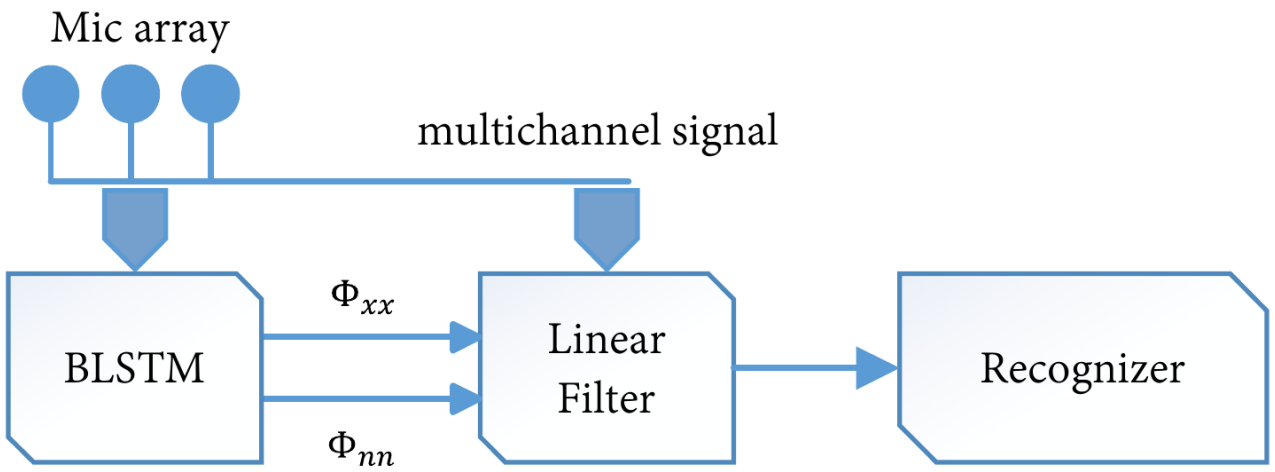
Use LSTM to estimate speech/noise masks [Heymann et al.].



. J. Heymann, L. Drude, R. Haeb-Umbach. *Neural network based spectral mask estimation for acoustic beamforming*. Proc ICASSP 2016

DNN-base mask estimation for MWF [Z. Wang]

Use the masks to estime the correlation matrices and MWF filters.



. Z. Wang, E. Vincent, R. Serizel, Y. Yan. *Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments*. Computer Speech and Language

Exemple performance

Evaluation in terms of word error rate (WER).

Noisy		WER baseline
Single-channel DNN		no WER reduc.
Delay-and-sum		21% rel. WER reduc.
DNN post-filter		20% rel. WER reduc.
Multichannel DNN		39% rel. WER reduc.

Greatly outperforms multichannel NMF (25% rel. WER reduc.) and DNN used to directly predict multichannel filters.

CHiME-3 : speech recorded in a bus/café. Single DNN iteration, no post-processing.

Loria

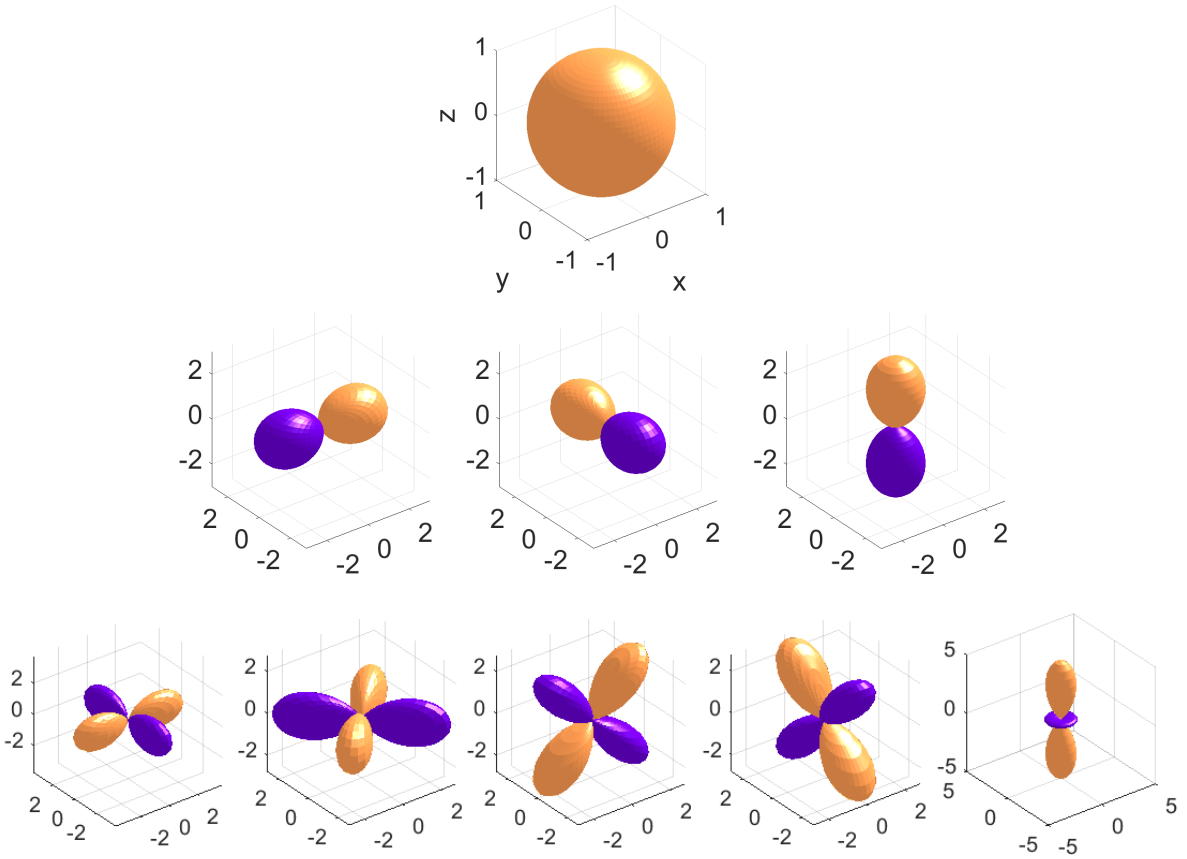
. A. A. Nugraha, A. Liutkus, E. Vincent. *Multichannel audio source separation with deep neural networks*. IEEE/ACM TASLP, 24 (9), pp. 1652 - 1664, 2016.

What input should we give to the network ?
(Part of Phd work from L. Perotin)

[With E. Vincent and A. Guérin]

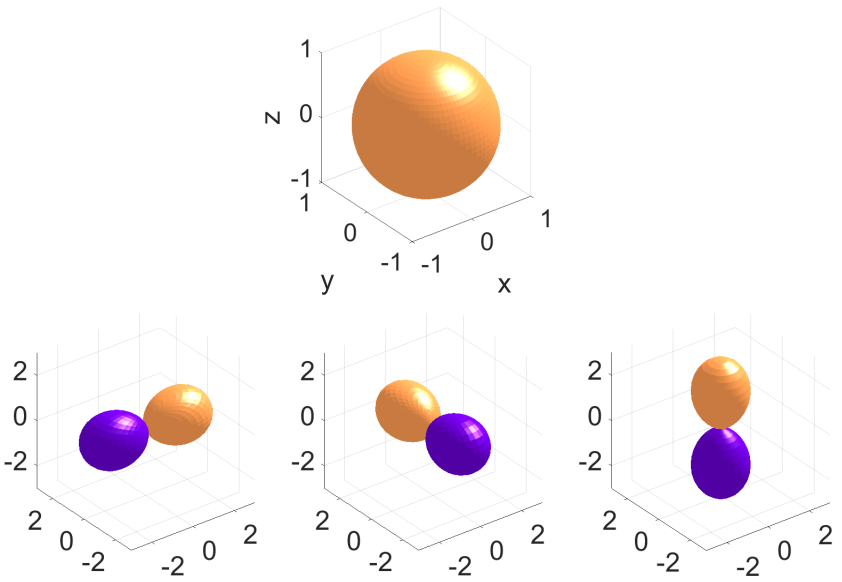
Multichannel speech enhancement for HOA

Represent the sound field in a basis of spherical harmonics



Multichannel speech enhancement for HOA

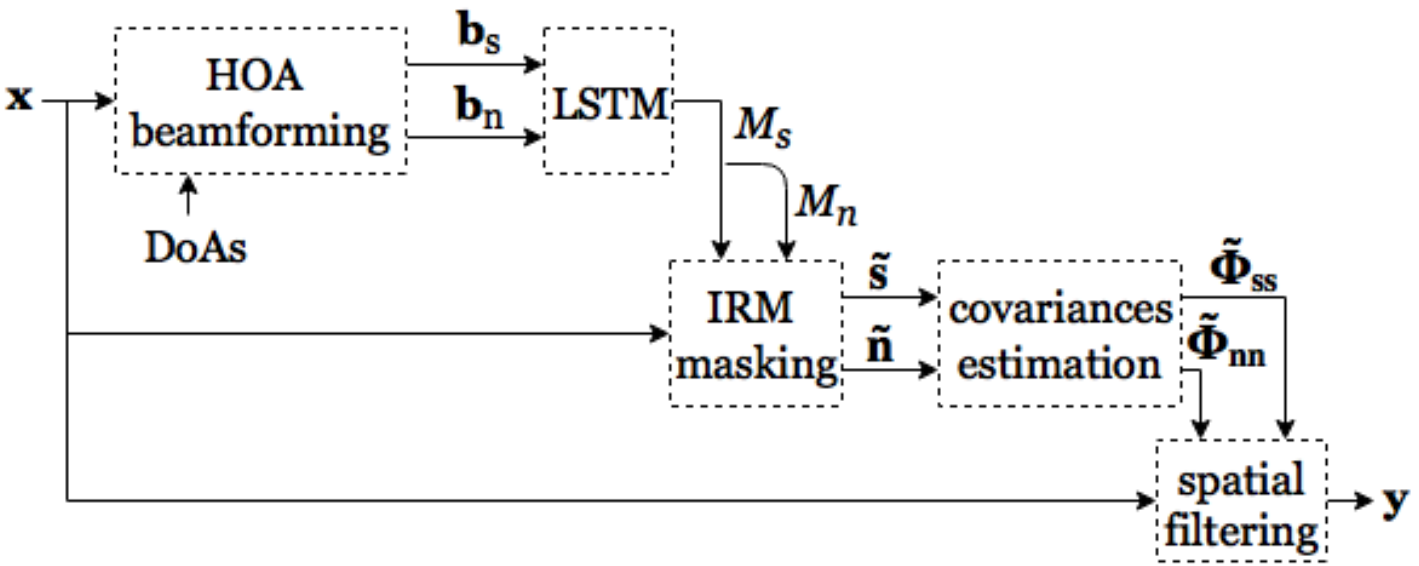
Use only the 4 channels from the first 2 orders



. L. Perotin, R. Serizel, E. Vincent, A. Guérin. *Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings*. In proc. of ICASSP, 2018,

Multichannel speech enhancement for HOA [L. Perotin]

- HOA beamformer
- LSTM to estimate speech/noise masks
- GEVD-MWF filter



. L. Perotin, R. Serizel, E. Vincent, A. Guérin. *Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings*. In proc. of ICASSP, 2018,

Results

Setup

- Tested on a subset of ESTER
- 20 sentences (4043 words)
- Acoustic conditions : TR_{60}=350ms, room-mic = 1.65m

Results

Interference	**noise**	**speaker**		
		25°	**45°**	**90°**
Reverberant source	9.80	10.80	10.82	10.76
Mixture	68.25	91.96	88.56	89.00
Simple beamformer	40.80	75.42	39.06	20.54
GEVD (w/o noise est.)	21.69	78.43	18.55	12.02
GEVD (w noise est.)	21.72	23.1	16.05	12.47

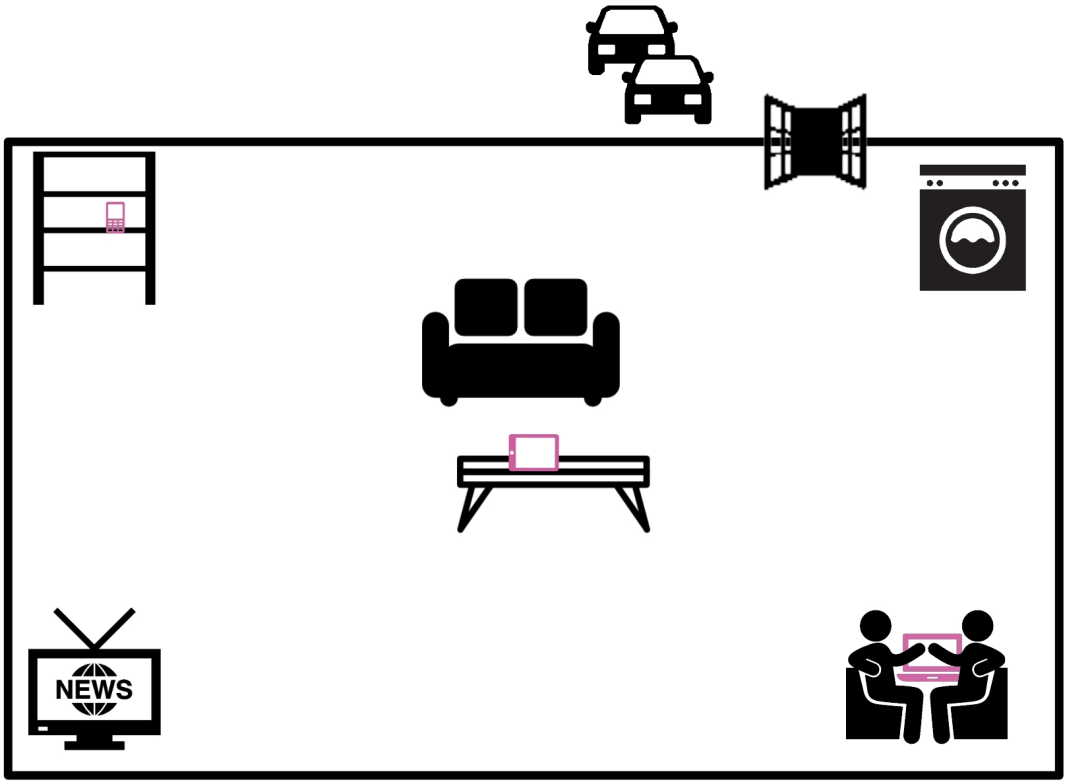
How can we deal with ad-hoc microphone arrays ? (Part of Phd work from N. Furnon)

[With I. Illina and S. Essid]

Heterogeneous unconstrained microphone arrays

Target :

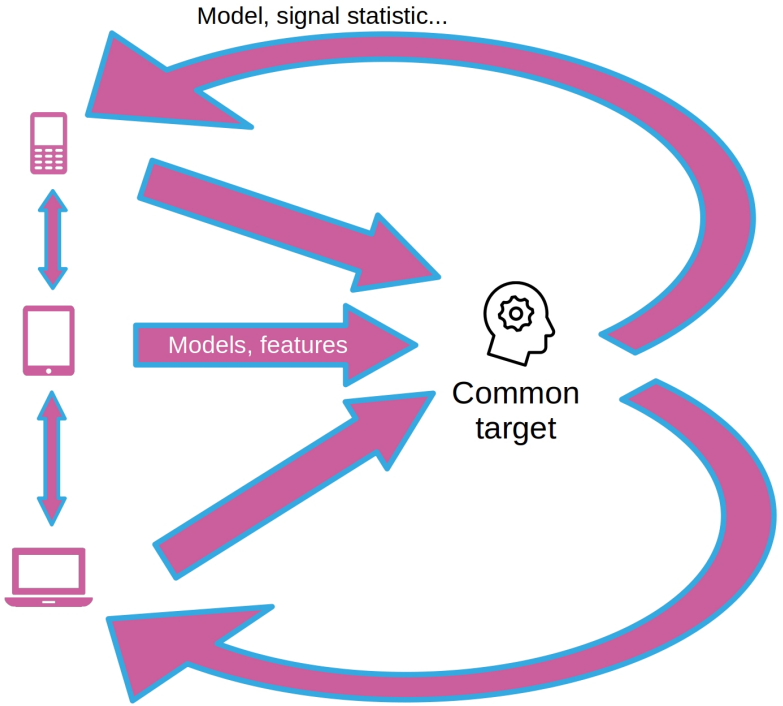
- Use the microphones already available



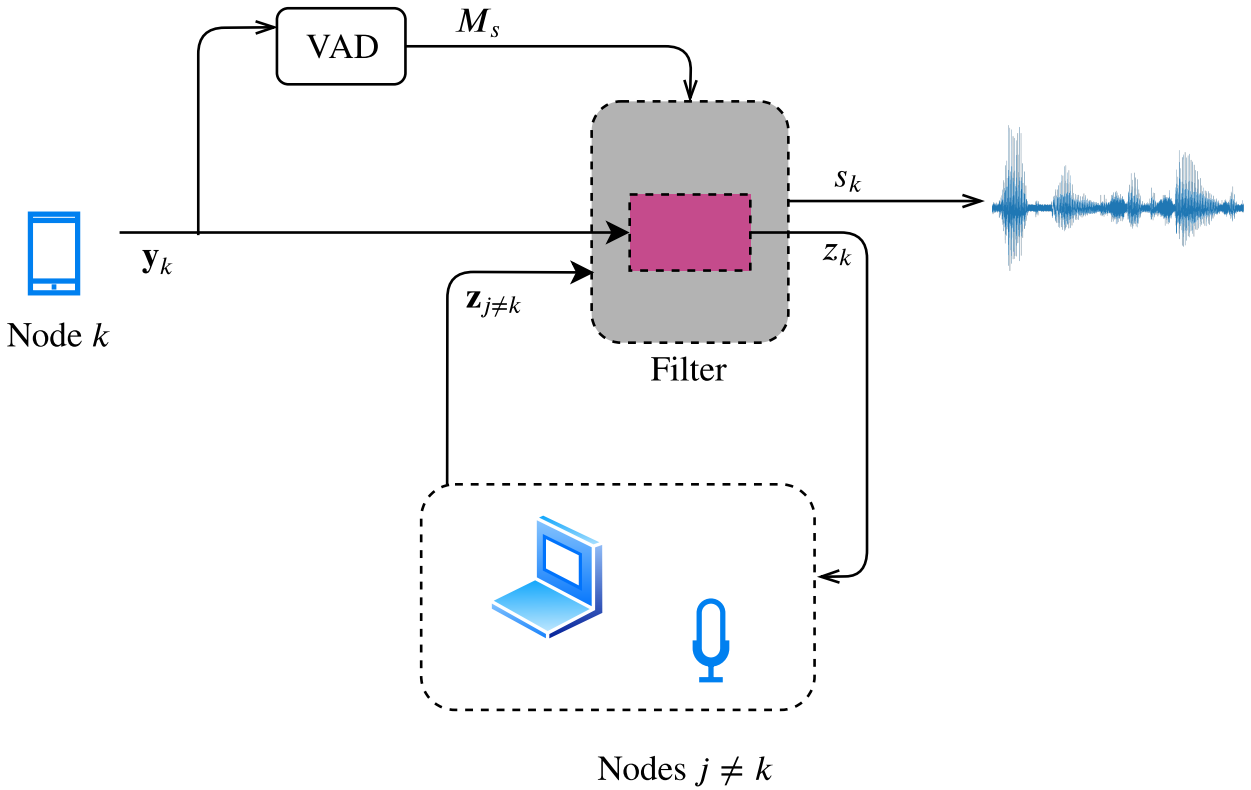
Collaborative approach

Targets :

- Collaboration between the nodes (common target)
- Feedback from the common target to the node level

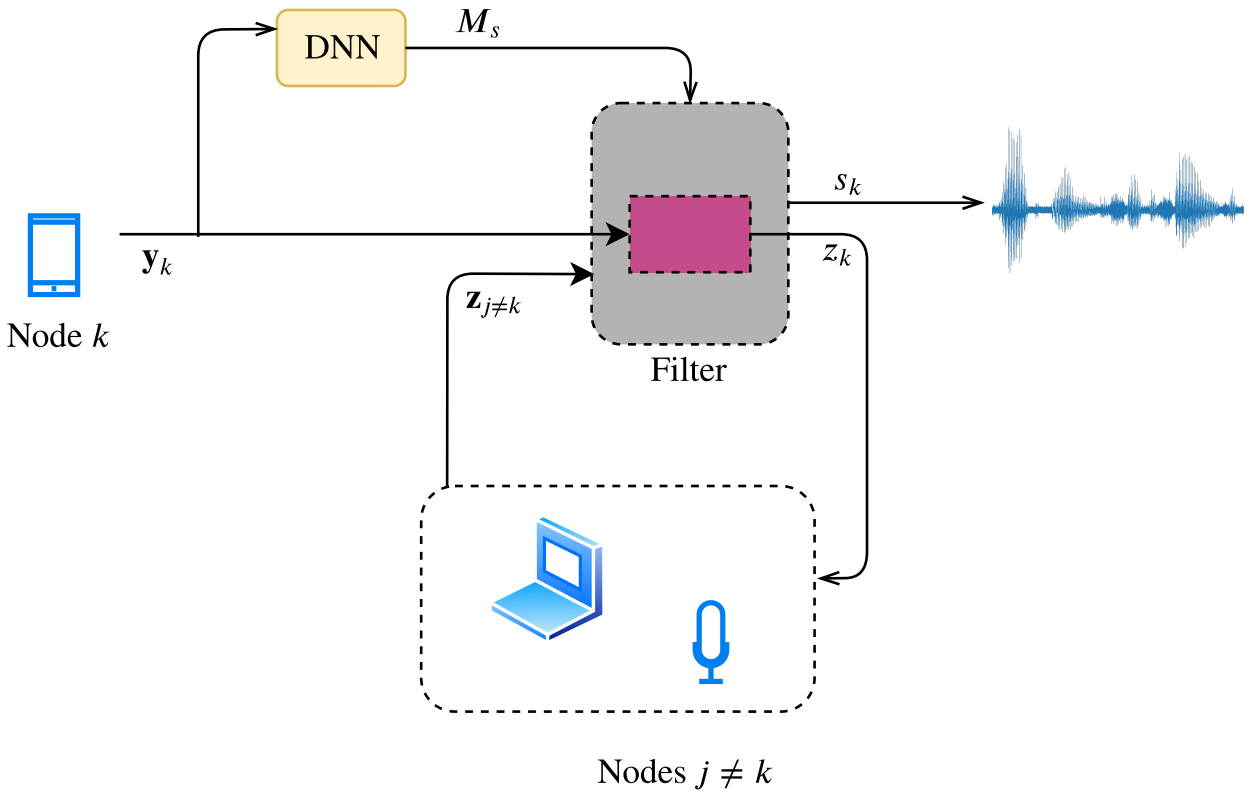


Distributed node specific MWF



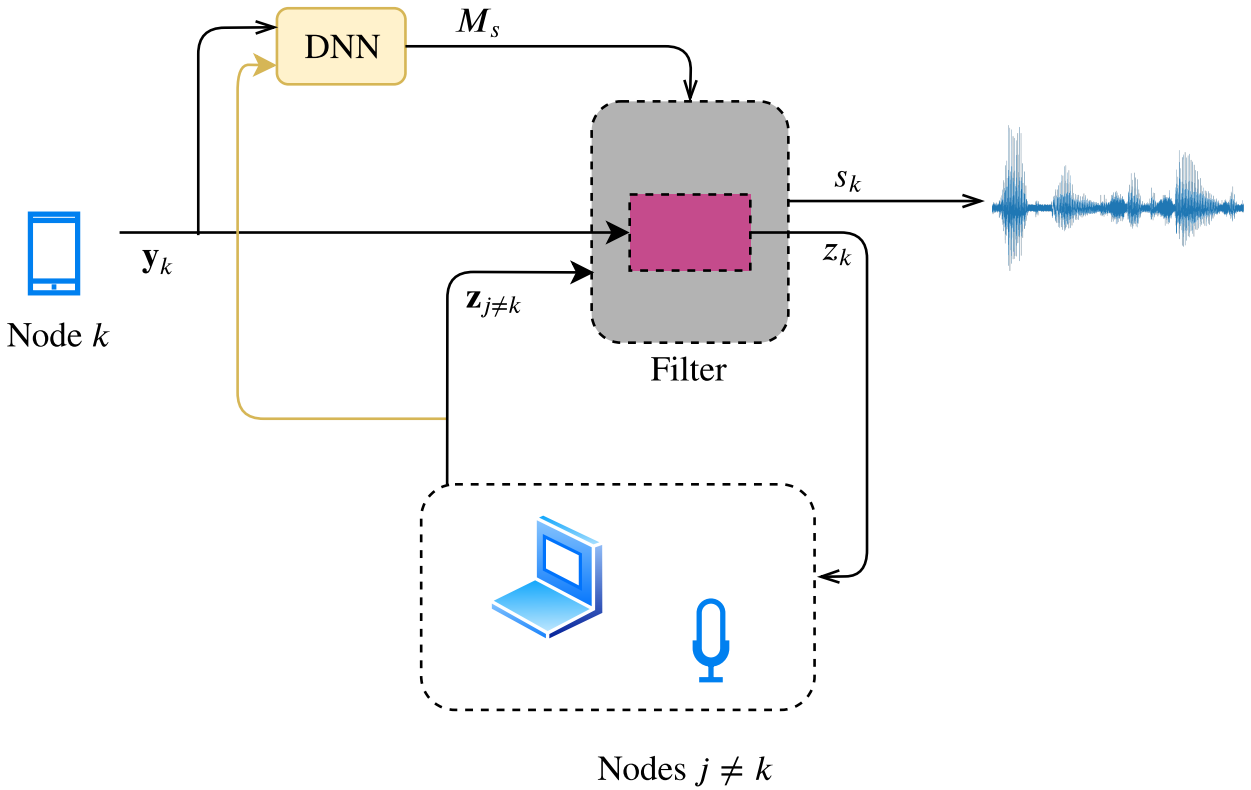
. A. Bertrand, M. Moonen. *Distributed adaptive node-specific signal estimation in fully connected sensor networks—Part I : Sequential node updating*. IEEE Transactions on Signal Processing.

Distributed MWF with DNN [N. Furnon]



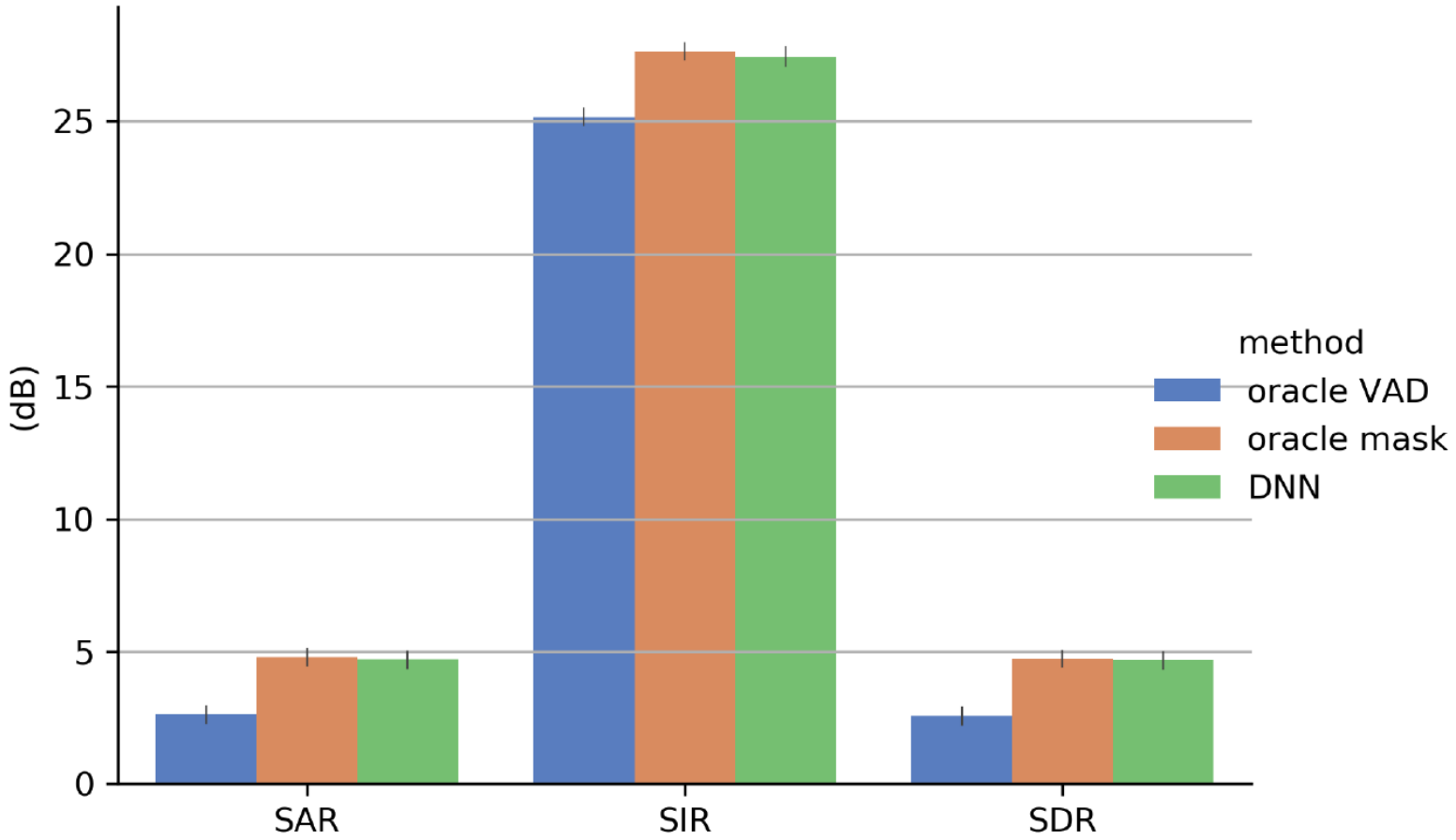
. N. Furnon, R. Serizel, I. Illina, S. Essid. *DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays.* Submitted to ICASSP 2020.

Distributed MWF with DNN [N. Furnon]



. N. Furnon, R. Serizel, I. Illina, S. Essid. *DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays.* Submitted to ICASSP 2020.

Distributed MWF with DNN (Results)



. N. Furnon, R. Serizel, I. Illina, S. Essid. *DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays.* Submitted to ICASSP 2020.

Outline

Conclusions and perspectives

Conclusions

- Deep learning is now the state-of-the-art in multichannel speech enhancement
- Important improvements compared to previous approaches

Perspectives

- Unconstrained microphone array processing,
- Distributed learning,
- Application to real-world conditions,
- Extension to other related domains. . .