MULTICHANNEL SPEECH SEPARATION WITH RNN FROM HOA RECORDINGS

Lauréline Perotin[†‡], Romain Serizel[‡], Emmanuel Vincent[‡], Alexandre Guérin[†]

[†]Orange Labs
[‡]Université de Lorraine, Inria, LORIA

PROBLEM STATEMENT

Distant-microphone voice command for personal digital assistant

- Real room conditions
- Competing speakers
- Ambient babble noise

→ Enhance the target speaker

PROBLEM STATEMENT

State of the art:

Neural networks to estimate time-frequency masks or multichannel filter parameters

Current challenges:

Overlapping speech

Contributions:

- New location-based method to estimate the parameters of a multichannel filter in overlapping speech conditions
- Ambisonics contents

1. HIGH ORDER AMBISONICS

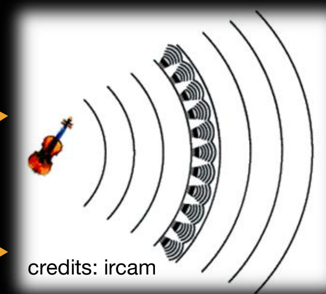
Capture

Eigenmike

Ambeo

HOA

Rendering



Wave Field Synthesis

binaural

5.1, ATMOS…

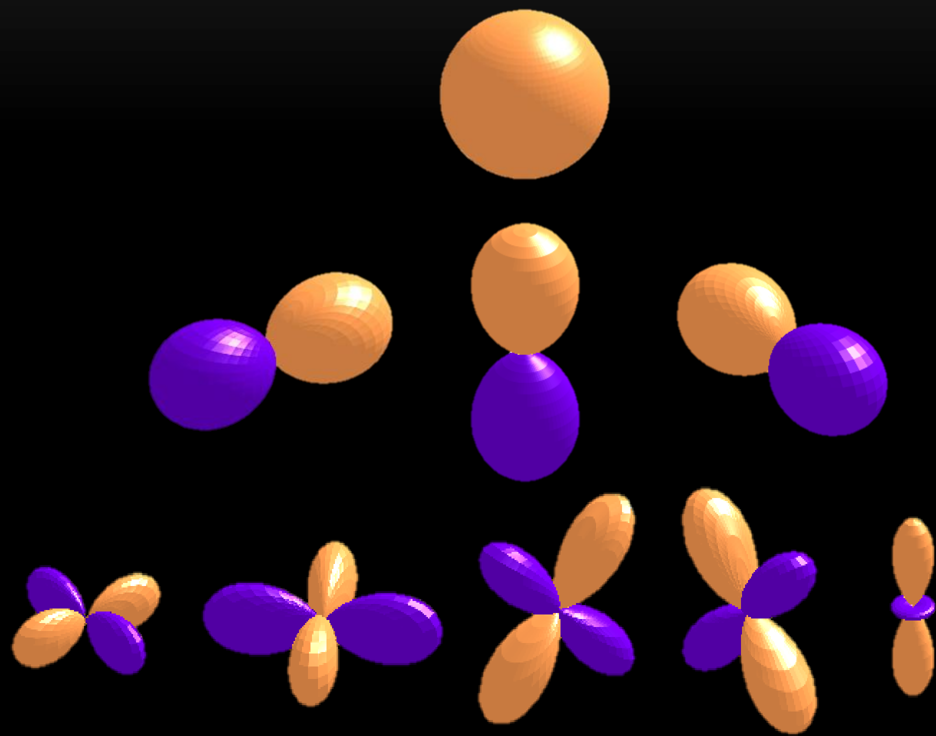
1. HIGH ORDER AMBISONICS

Order 0

Order 1

Order 2

...



4 channels
≈ 4 virtual mics

2. FULL-BAND BEAMFORMING

Mixture: $\mathbf{x}(t,f) = \mathbf{s}(t,f) + \mathbf{n}(t,f)$

Assumption: known direction of arrival of the target (at least)

HOA anechoic mixing matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & \dots & 1 \\ \sqrt{3}\cos\theta_0\cos\phi_0 & \dots & \sqrt{3}\cos\theta_J\cos\phi_J \\ \sqrt{3}\sin\theta_0\cos\phi_0 & \dots & \sqrt{3}\sin\theta_J\cos\phi_J \\ \sqrt{3}\sin\phi_0 & \dots & \sqrt{3}\sin\phi_J \end{bmatrix}$$

HOA beamformer: $\hat{s}(t,f) = \mathbf{u}_1^T \mathbf{A}^{\dagger}\mathbf{x}(t,f)$

→ not robust to reverberation and close speakers

2. MULTICHANNEL WIENER FILTERING

Mixture: $\mathbf{x}(t,f) = \mathbf{s}(t,f) + \mathbf{n}(t,f)$

Optimization criterion:

$$\min \mathbb{E}\{|y(t,f) - \mathbf{u}_1^H \mathbf{s}(t,f)|^2\}$$

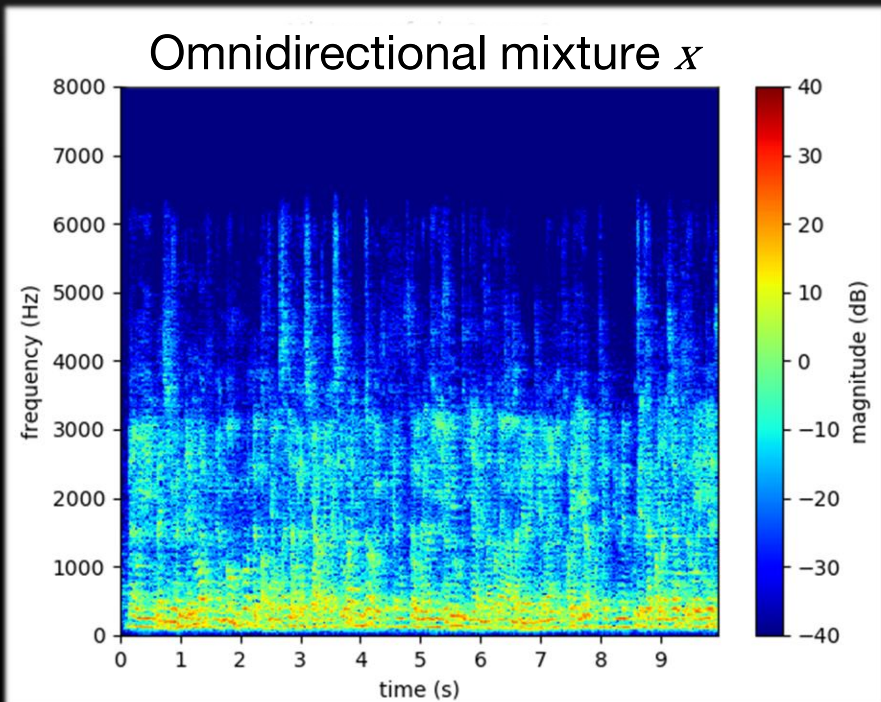
$$\text{with } y(t,f) = \mathbf{w}(f)^H \mathbf{x}(t,f)$$

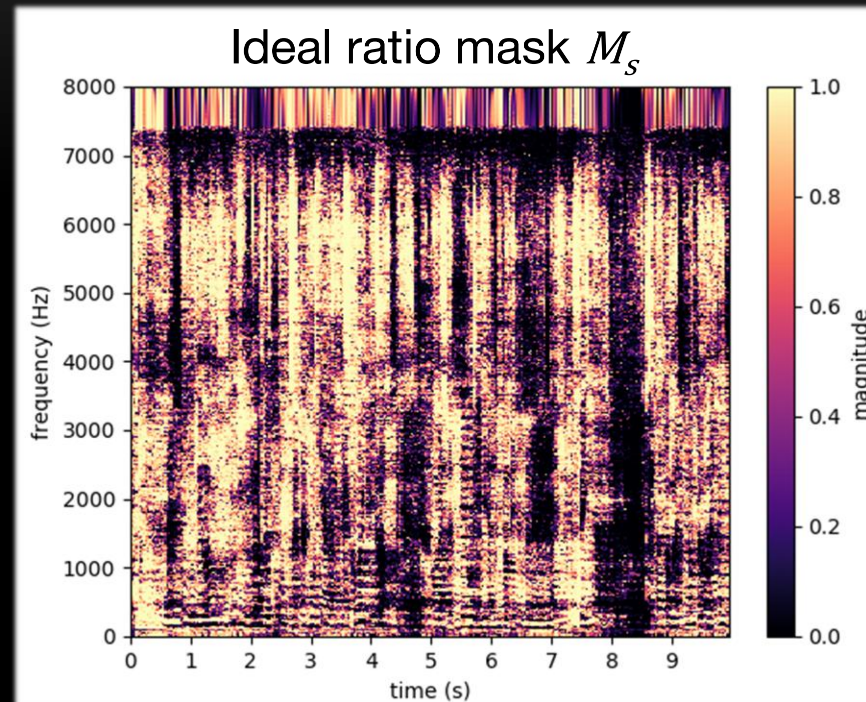
Time-invariant multichannel Wiener filter:

$$\mathbf{w}(f) = [\mathbf{\Phi}_{\mathbf{ss}}(f) + \mathbf{\Phi}_{\mathbf{nn}}(f)]^{-1} \mathbf{\Phi}_{\mathbf{ss}}(f) \mathbf{u}_1$$

→ Little distortion, but we need the covariance matrices!

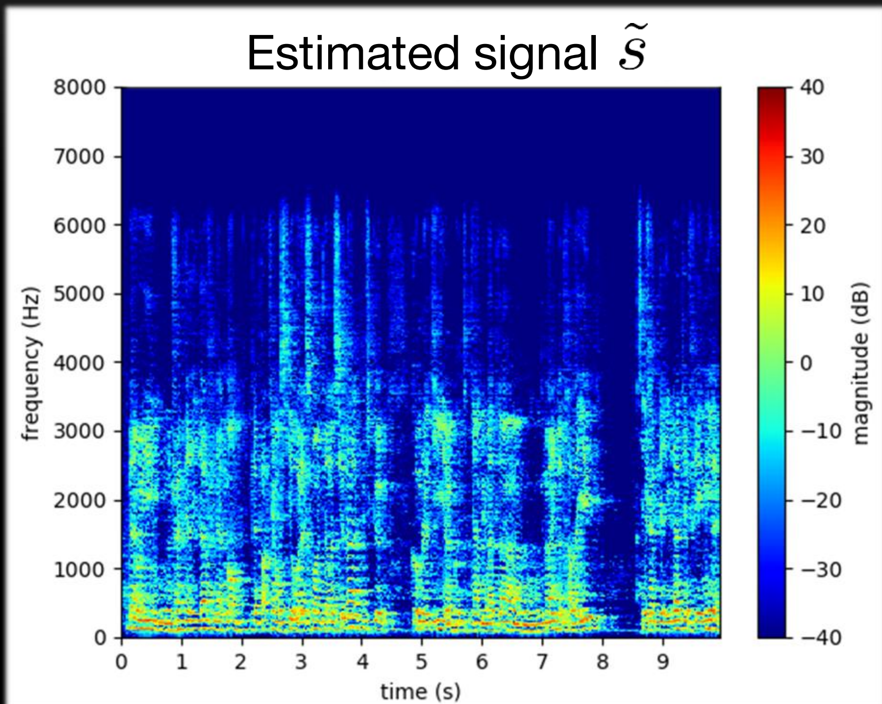
2. MASKING-BASED COVARIANCE ESTIMATION

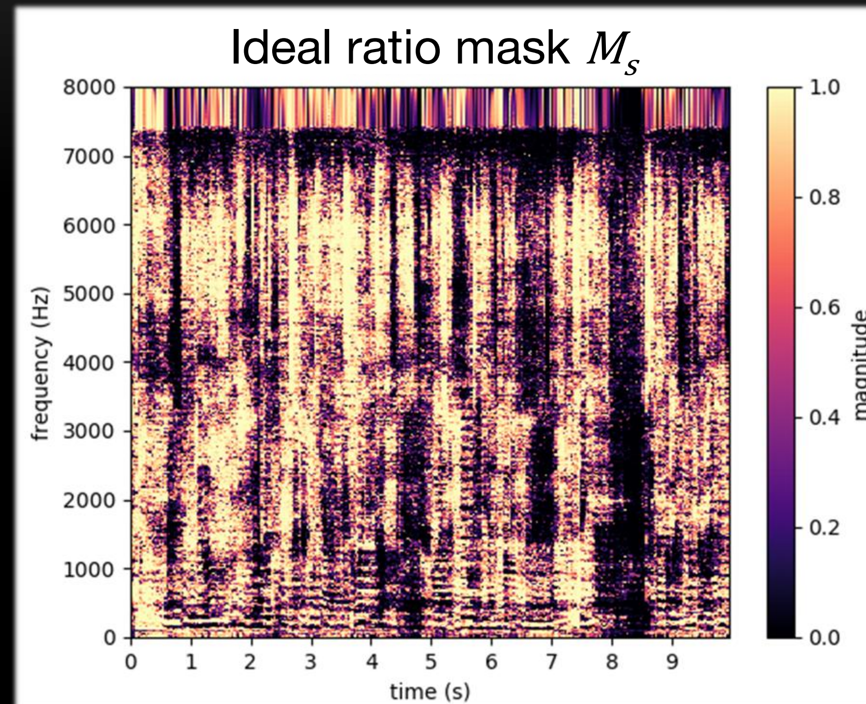




$$M_s(t, f) = \frac{|s(t,f)|}{|s(t,f)|+|n(t,f)|}$$

2. MASKING-BASED COVARIANCE ESTIMATION



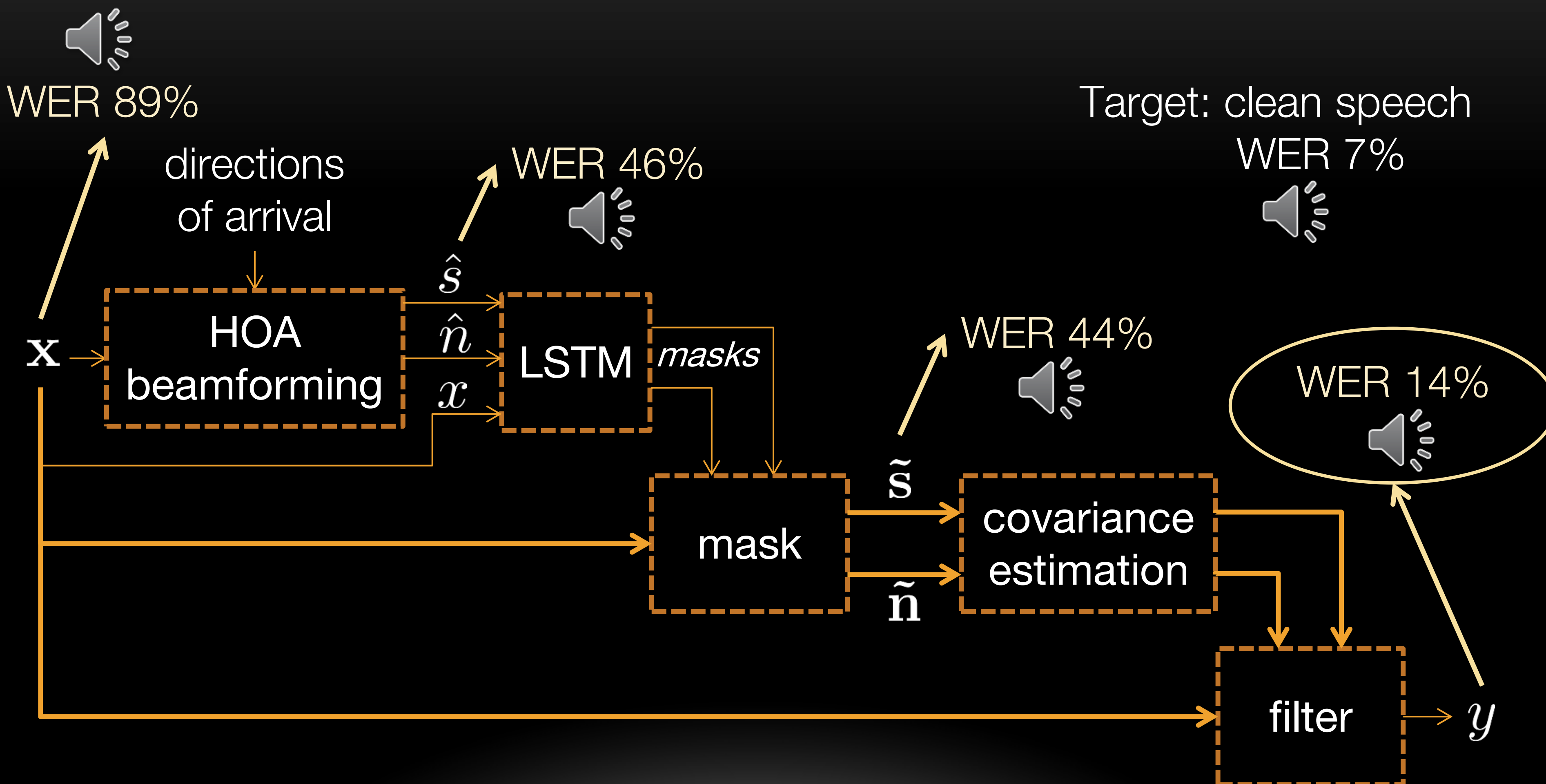


$$\tilde{\mathbf{s}}(t,f) = M_s(t,f)\mathbf{x}(t,f)$$

$$M_s(t,f) = \frac{|s(t,f)|}{|s(t,f)|+|n(t,f)|}$$

$$\rightarrow \tilde{\mathbf{\Phi}}_{\mathbf{ss}}(f) = \frac{1}{T}\sum_{t=0}^{T-1} \tilde{\mathbf{s}}(t,f)\tilde{\mathbf{s}}^H(t,f)$$

3. PROPOSED SOLUTION



3. RESULTS

<u>Training data :</u>

10h of mixed speech

SIR = 0 dB

44 different speakers

Room 1

16 positions, RT_{60} = 270ms

<u>Test data :</u>

20 min of mixed speech

SIR = 0 dB

20 different speakers

4000 words

Room 2

42 positions, RT_{60} = 350ms

Word Error Rate (%)			1 spk	2 spk, angle diff. 25°	45°	90°
Clean speech			7.4	7.4	7.4	7.4
Mixture			68.5	91.7	88.9	85.4
Beamformer			24.3	76.0	45.9	20.6
Ideal mask			18.3	16.3	15.0	16.3
Filter from ideal mask			13.1	23.0	16.5	11.1
Network inputs	x	mask	68.6	91.8	84.5	85.7
		filter	25.0	91.6	87.1	86.6
	$\hat{s}$	mask	61.2	90.8	84.8	78.3
		filter	19.6	67.2	27.1	12.9
	x, $\hat{s}$	mask	55.9	86.4	61.6	45.0
		filter	**17.1**	80.9	21.0	**10.5**
	x, $\hat{s}$, $\hat{n}$	mask	n/s	60.9	43.9	37.2
		filter		**22.3**	**14.5**	**11.0**

CONCLUSION

order 1 ambisonics
2 speakers + noise

↓

Directions of arrival ⇢

LSTM-based multichannel Wiener filter
Inputs: omnidirectional mixture
+ beamformer toward target speech
+ beamformer toward competing speech

↓

Performs as good as the filter computed from the ideal mask including with 25° apart speakers