MULTICHANNEL RNN-BASED SEPARATION OF OVERLAPPING SPEECH

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:

- Ambisonics contents
- Multi-source localization
- Enhancement in overlapping speech conditions by estimating the parameters of a multichannel filter

1. HIGH ORDER AMBISONICS

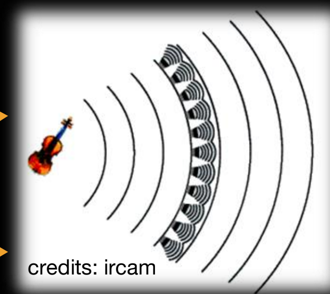
Capture

Ambeo

Eigenmike

HOA

Rendering



Wave Field Synthesis

binaural

5.1, ATMOS…

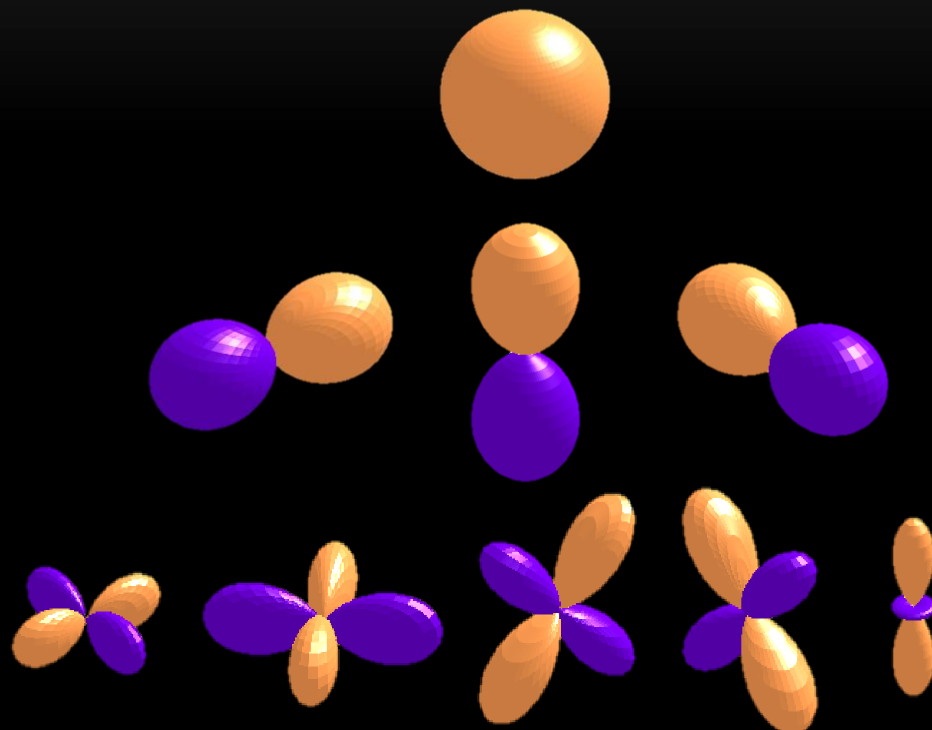
1. HIGH ORDER AMBISONICS

Order 0

Order 1

Order 2

...



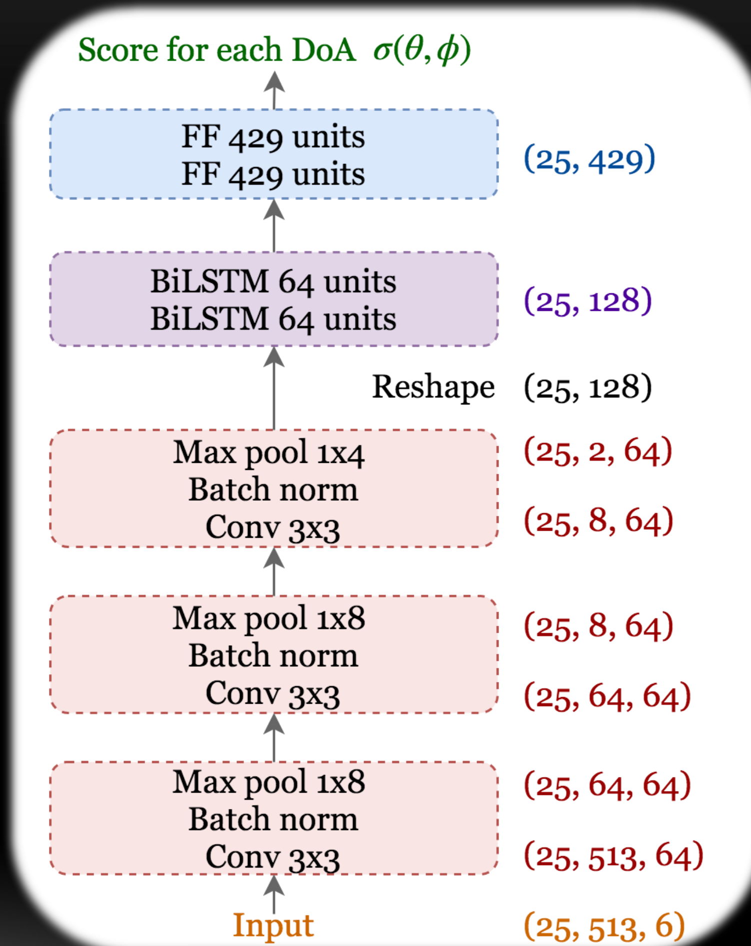
4 channels
W, X, Y, Z
≈ 4 virtual mics

2. PROPOSED SOLUTION FOR DIRECTION-OF-ARRIVAL ESTIMATION

Input feature based on acoustic intensity:

$$\mathbf{I}_{\mathrm{a}}(t,f) = \begin{bmatrix} \mathcal{R}\{W(t,f)^{*}X(t,f)\} \\ \mathcal{R}\{W(t,f)^{*}Y(t,f)\} \\ \mathcal{R}\{W(t,f)^{*}Z(t,f)\} \end{bmatrix}$$

$$\mathbf{I}_{\mathrm{r}}(t,f) = \begin{bmatrix} \mathcal{I}\{W(t,f)^{*}X(t,f)\} \\ \mathcal{I}\{W(t,f)^{*}Y(t,f)\} \\ \mathcal{I}\{W(t,f)^{*}Z(t,f)\} \end{bmatrix}$$



3. ENHANCEMENT: FULL-BAND BEAMFORMING

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

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

3. ENHANCEMENT: MULTICHANNEL WIENER FILTERING

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

Time-invariant multichannel Wiener filter:

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

→ Little distortion, but covariance matrices needed!

Mask – based covariance estimation:

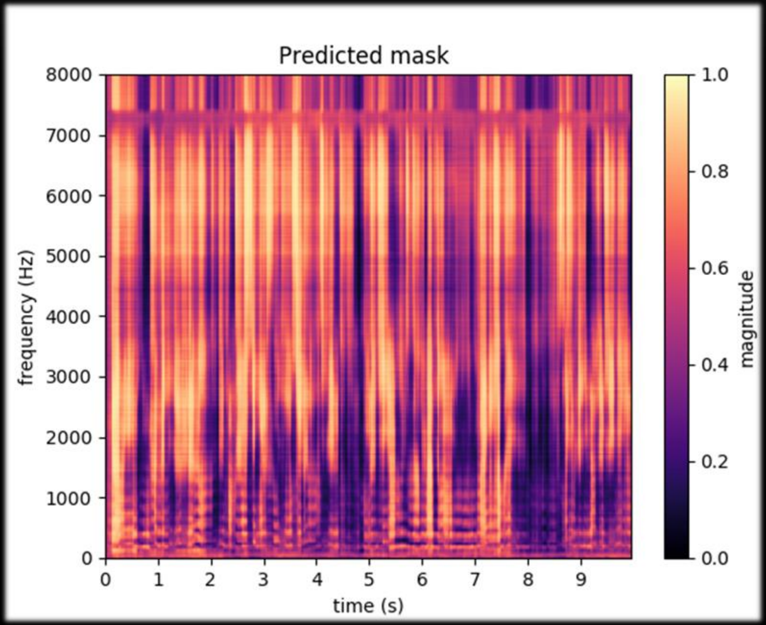
$$M_s(t,f) = \frac{|s(t,f)|}{|s(t,f)|+|n(t,f)|}$$

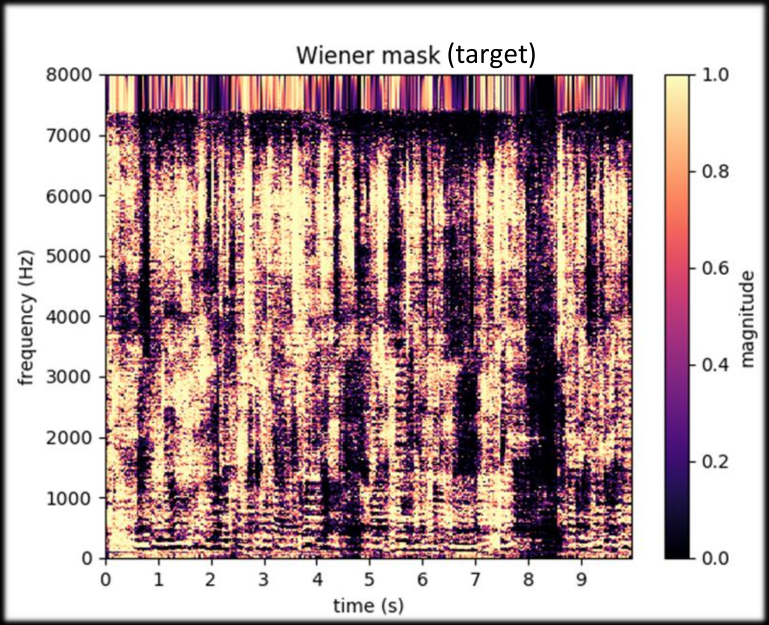
$$\tilde{\mathbf{s}}(t,f) = M_s(t,f)\mathbf{x}(t,f)$$

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

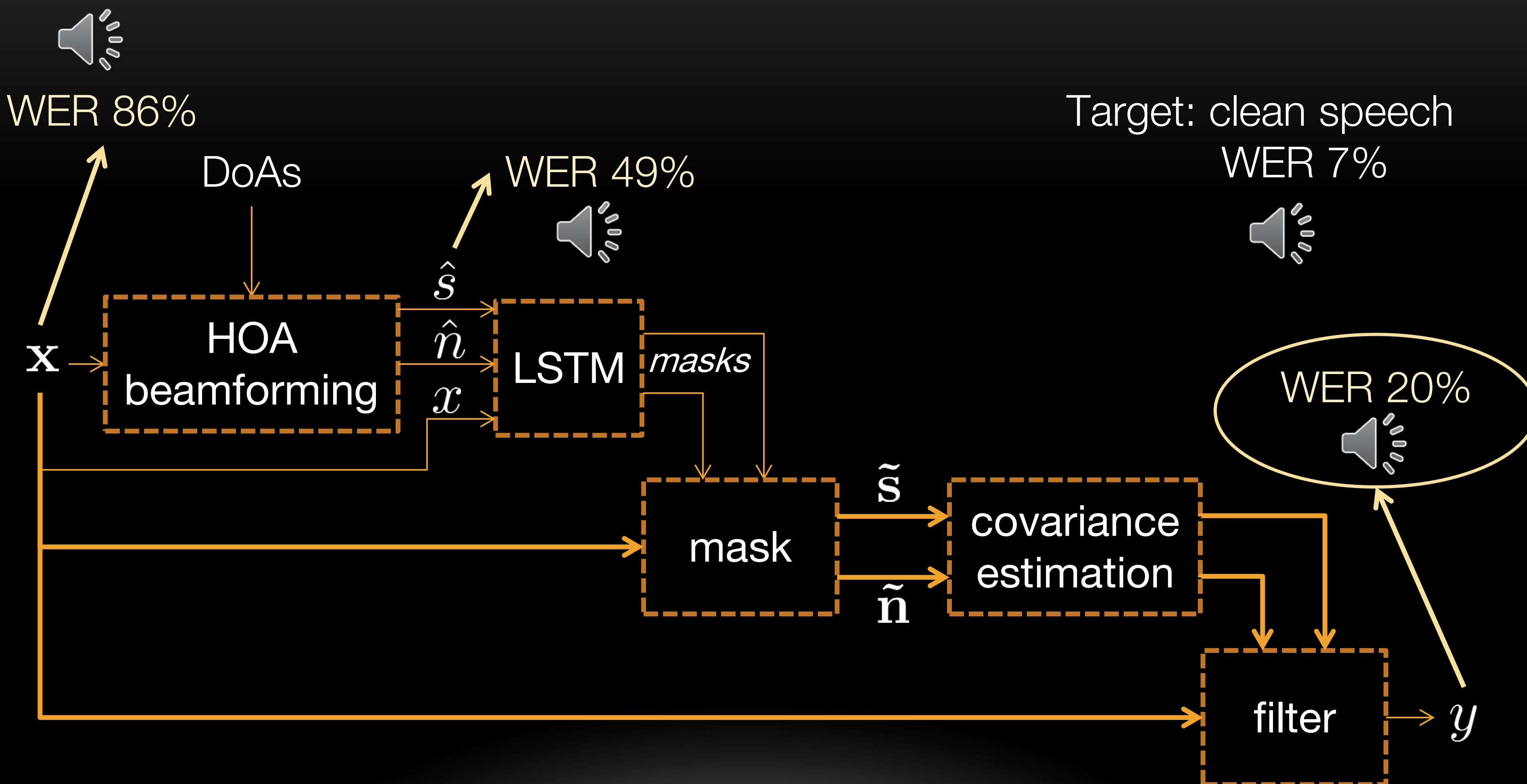
3. PROPOSED SOLUTION

Estimation of the mask via LSTM neural network:





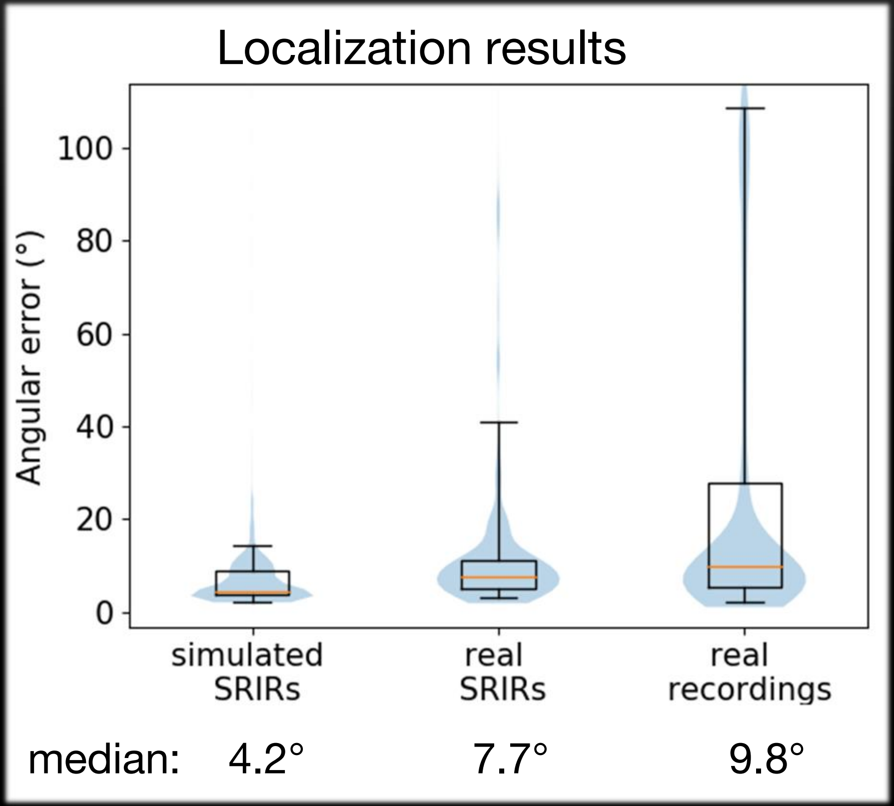
3. PROPOSED SOLUTION



3. RESULTS FOR LOCALIZATION

<u>Test data :</u>
2 overlapping speakers, static, no VAD
25 to 90° angular distance
SIR = 0 dB, SNR = 20 dB

- <u>Simulated SRIRs</u> (image method)
  RT60 = 0.2 to 0.8s
- <u>Real SRIRs</u>
  RT60 ≈ 0.5s
  random source/mic orientations
- <u>Real recordings</u>
  living-room, mic on coffee table



<u>Training data :</u>
SIR = 0 to 10 dB, SNR = 20 dB, different speakers and rooms
36 h of speech made from simulated SRIRs

3. RESULTS FOR ENHANCEMENT

Word Error Rate (%)	**Simulated SRIRs**	**Real SRIRs**	**Real recordings**
Clean speech	7.4	7.4	n/a
Mixture	81.9	86.0	89.5
Oracle DoA + beamformer	32.3	49.2	53.8
Oracle DoA + proposed filter	12.3	19.7	20.3
Estimated DoA + beamformer	33.1	53.1	57.9
Estimated DoA + proposed filter	13.4	25.3	26.5

<u>Training data</u>

10h of mixed speech at SIR = 0 dB

44 different speakers

16 positions in a single room at RT_{60} = 270ms

CONCLUSION

order 1 Ambisonics
2 speakers + noise

Directions of arrival by CRNN

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

Largely outperforms beamforming or sole masking, including with close speakers in real conditions