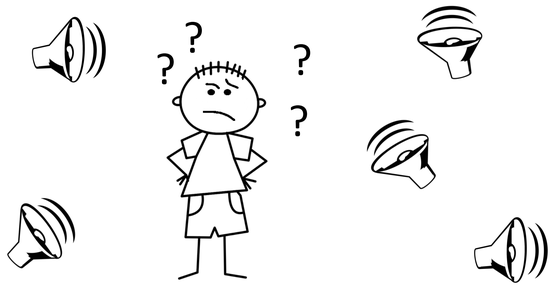
Regression versus classification for neural network based audio source localization

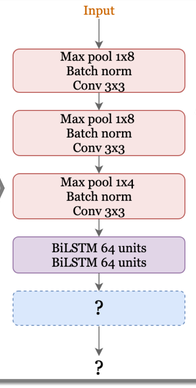
Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain Serizel, Alexandre Guérin

facebook Artificial Intelligence Research

Your mission: localize a source



Your tools



A CRNN

A multichannel recording (e. g. first-order Ambisonics)

Your environments

Recorded SRIRs

Various source-microphone positions and orientations

Real life situation

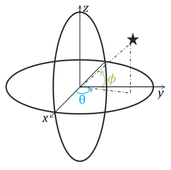
In a living room, recordings of people talking around a coffee table, with the microphone on the table and a TV at 10 dB SNR

How do you formulate the problem?

by regression?

FF 429 units
FF 2 or 3 units

Estimated coordinates



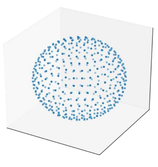
Target	Loss	Output layer	Accuracy (%) ±5° | 10° |15° (SRIRs) ±5° | 10° |15° (real)	Conclusion
spherical coordinates $\begin{cases}\theta \in (-180°, 180°] \\ \phi \in [-90°, 90°]\end{cases}$	MSE $(\theta - \hat{\theta})^2 + (\phi - \hat{\phi})^2$	sigmoid	28.5 | 60.2 | 81.2 6.0 | 23.2 | 37.1	You loose! - worst accuracy - many outliers
spherical coordinates $\begin{cases}\theta \in (-180°, 180°] \\ \phi \in [-90°, 90°]\end{cases}$	angular error $\delta[(\theta,\phi),(\hat{\theta},\hat{\phi})] = \arccos\{\sin(\hat{\phi})\sin(\phi) + \cos(\hat{\phi})\cos(\phi)\cos(\hat{\theta}-\theta)\}$	sigmoid	29.7 | 64.3 | 80.0 **12.5** | 35.6 | 47.9	You loose! - not so good accuracy - many outliers
Cartesian coordinates (x, y, z) with $x^2 + y^2 + z^2 = 1$	MSE $(x - \hat{x})^2 + (y - \hat{y})^2 + (z - \hat{z})^2$	sigmoid	**37.1** | **72.9** | **88.3** **12.0** | 32.3 | 45.3	You win! + best accuracy with SRIRs + good 5° accuracy in real life - yet many outliers in real life

+ faster to train than classification methods
+ precise in controlled conditions
- how to deal with several sources?

by classification?

FF 429 units
FF 429 units

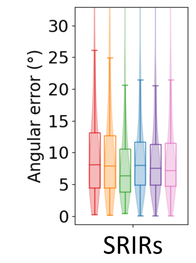
Score for each DoA σ_{ij}

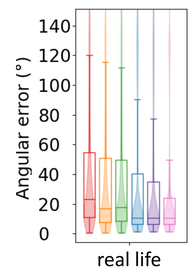


one-hot encoding 0 0 0 1 0	cross entropy $-\log(\sigma_{ij}) - \sum_{\substack{(i',j') \\ \neq(i,j)}} \log(1 - \sigma_{i'j'})$ (output for ψ_{ij}, the closest DoA on the grid)	softmax	26.6 | 66.8 | **85.3** 9.8 | **44.6** | **63.2**	Not bad + pretty good accuracy
Gibbs distribution $\mathcal{G}(\psi_{ij}) = e^{-\delta[\psi_{ij},\psi]^2/\beta^2}$ (neighborhood size; DoA on grid; true DoA)	MSE $\sum_{(i,j)} \mathcal{G}(\psi_{ij})^2$	sigmoid	26.0 | 67.7 | **86.4** **11.7** | 43.1 | **62.0**	You win! + best accuracy in real life - slowest to train
one-hot encoding 0 0 0 1 0	Gibbs cross entropy $-\log(\sigma_{ij}) - \sum_{\substack{(i',j') \\ \neq(i,j)}} [(1 - \mathcal{G}(\psi_{i'j'})) . \log(1 - \sigma_{i'j'})]$	sigmoid	27.6 | 67.7 | **84.5** **12.5** | 38.8 | 56.0	You win! + pretty good accuracy + least outliers in real life

+ robust to real life conditions with interference
+ can handle several sources
- high precision accuracy limited by the discretization of the sphere

Further results:





reg. sph.
reg. sph. ang. err.
reg. cart.
class. one-hot
class. Gibbs distrib.
class. Gibbs loss