Machine Learning @ ARI

An Overview

Acoustics is all around us





@ ARI we do fundamental research in various sciences related to acoustics

Clusters at the Acoustics Research Institute

Biology

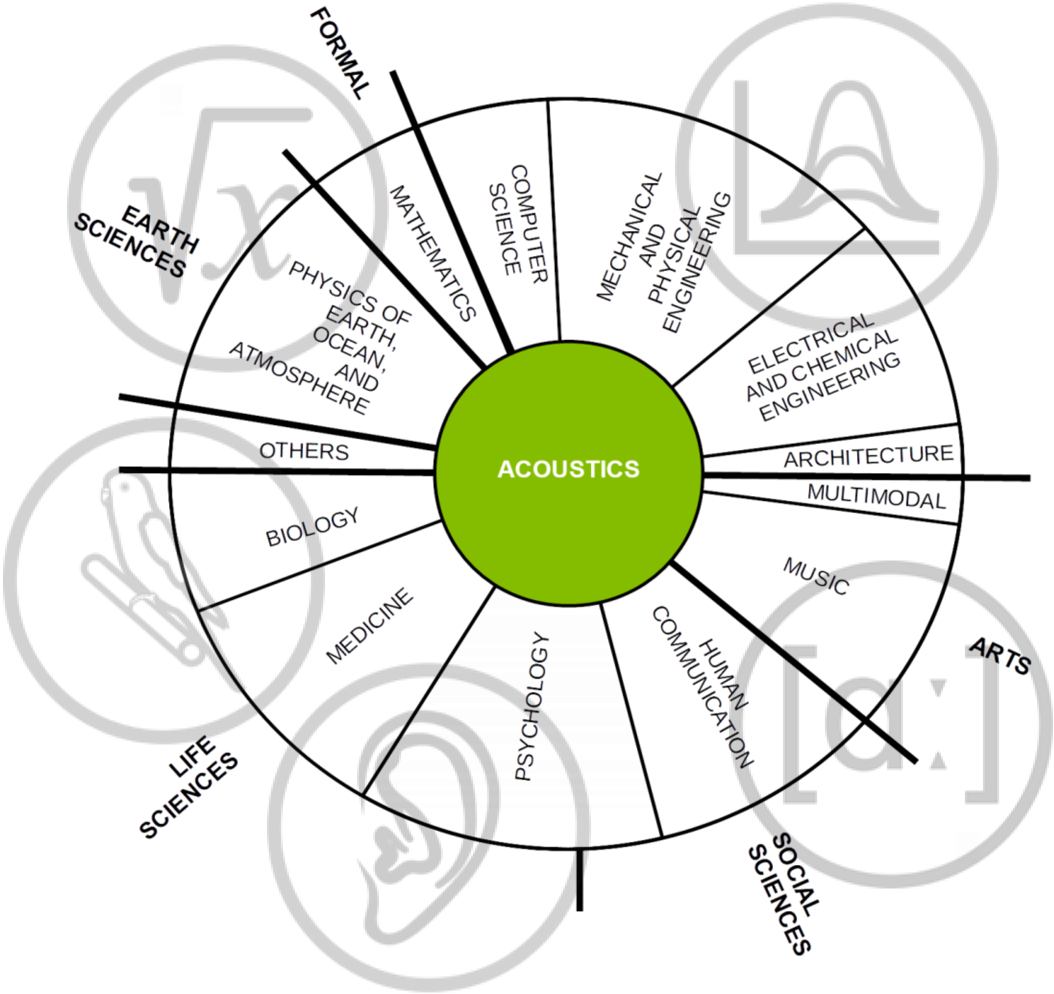
Phonetics

Hearing

Numerics

Mathematics

The Wheel of Acoustics and a curious gap



Machine Learning @ ARI

- No dedicated machine learning workgroup
- Usually project-based and therefore fluctuating
- Different interests in different clusters
  (Use vs. Study and Development)

ARI's Machine Learning Team

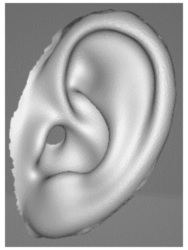
Platform for inter-cluster exchange, discussion, and collaboration on research opportunities concerning machine learning and computational statistics

Some Projects
(Recent and Ongoing)

AUSTRIAN ACADEMY OF SCIENCES

Mesh2PPM: Estimation of Parametric Pinna Model Parameters from a Pinna-Mesh Representation

F. Pausch, F. Perfler, N. Holighaus, P. Majdak



Pinna-mesh representation (e.g., photo) → Trained DNN → PPM parameters (personalised pinna mesh) —Mesh2HRTF*→ Personalised HRTFs

[*] www.mesh2hrtf.org

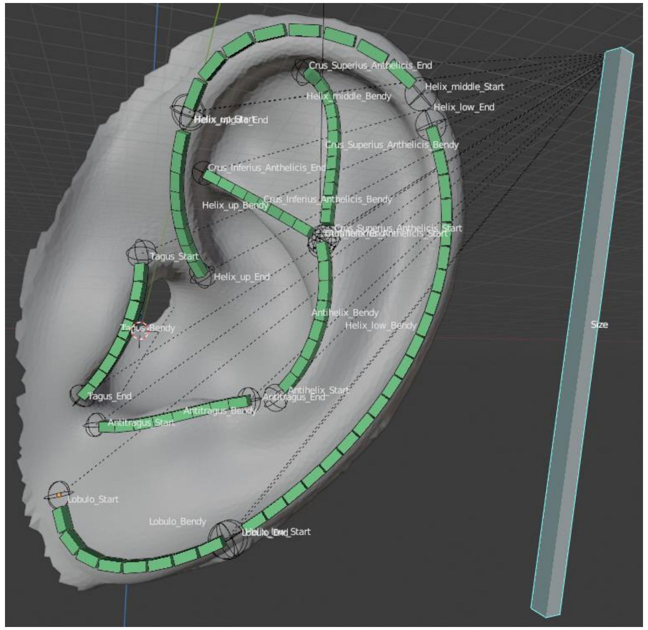
- Deep neural network (DNN) for parameter prediction from from ear images
- Synthesis of a personalised pinna mesh
- Numerical calculation of head-related transfer functions (HRTFs)

SONICOM

The SONICOM project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no.101017743.

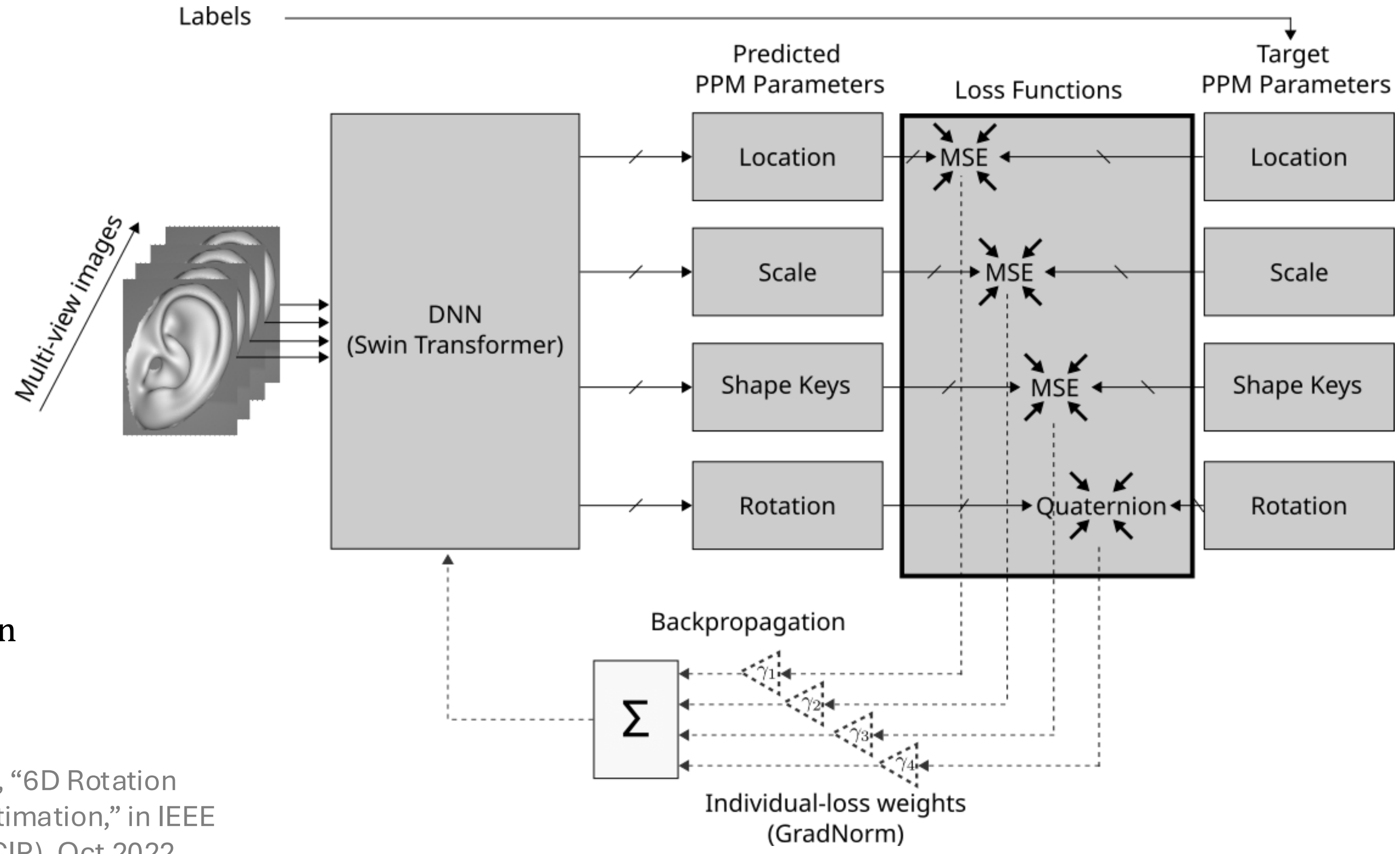
The parametric model: BezierPPM*

- Default model mesh: obtained via principal component analysis of 119 individual ear meshes (WiDESPREaD**)
- Armature definitions in BLENDER***
- 144 parameter dimensions:
  - Global parameters (parent bone)
  - Local shape curves (bendy bones)
  - Local shape weights (shape keys)
  - Four parameter types:
    Location, rotation, scale, shape keys



[*] F. Perfler, F. Pausch, K. Pollack, N. Holighaus, and P. Majdak, "Accurate Parametric Modeling of the Human Pinna Inspired by Nature Using Bézier Curves," 2024, Unpublished manuscript (in review). Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria.
[**] Guezenoc, C.; Renaud, S. (2020), "A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings", JASA 147: 4087-4096 https://doi.org/10.1121/10.0001461
[***] www.blender.org

Parameter Estimation Framework

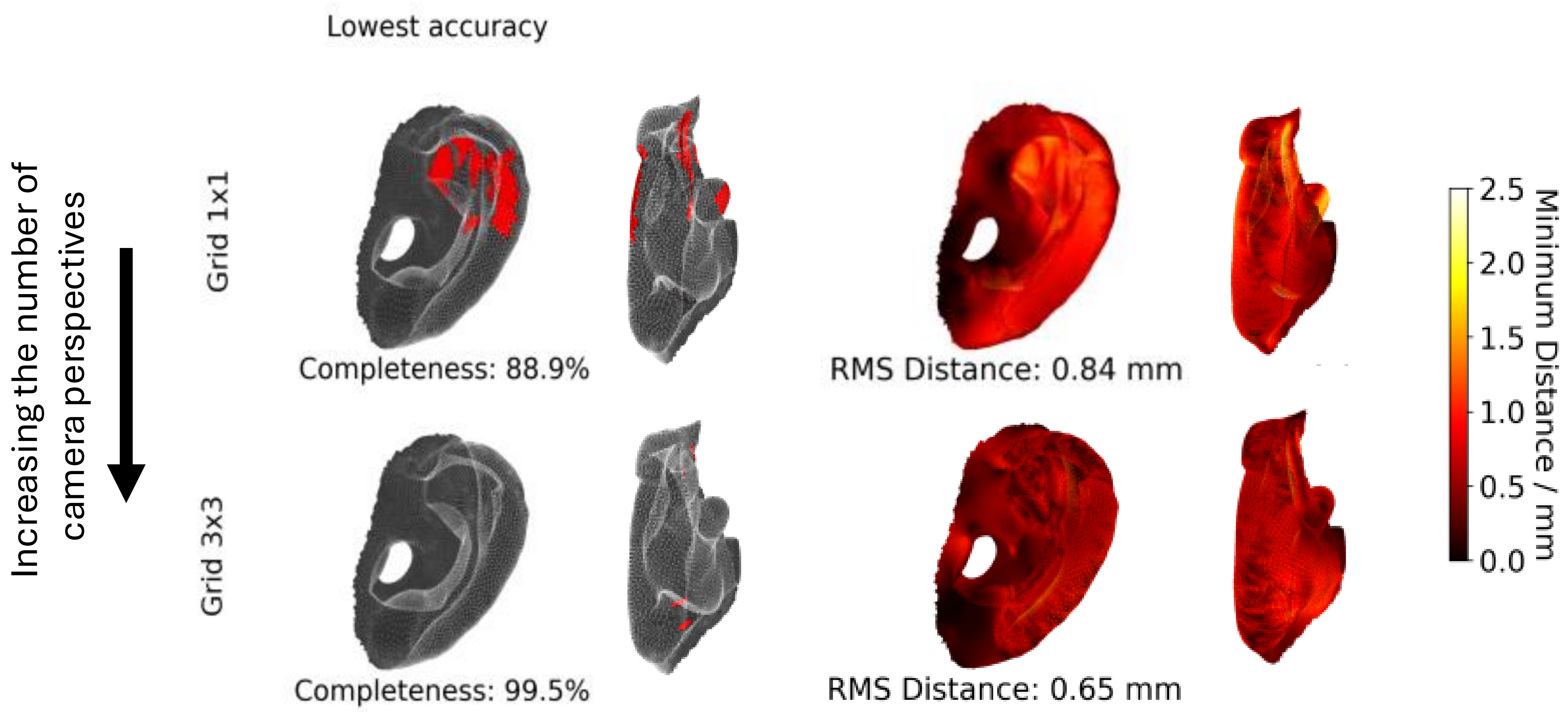


$$\mathcal{L} = \gamma_1 \mathcal{L}_{\text{Location}} + \gamma_2 \mathcal{L}_{\text{Scale}} + \gamma_3 \mathcal{L}_{\text{Shape Keys}} + \gamma_4 \mathcal{L}_{\text{Quaternion}}$$

Geodesic quaternion loss

T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, “6D Rotation Representation For Unconstrained Head Pose Estimation,” in IEEE International Conference on Image Processing (ICIP). Oct 2022.

Example Results

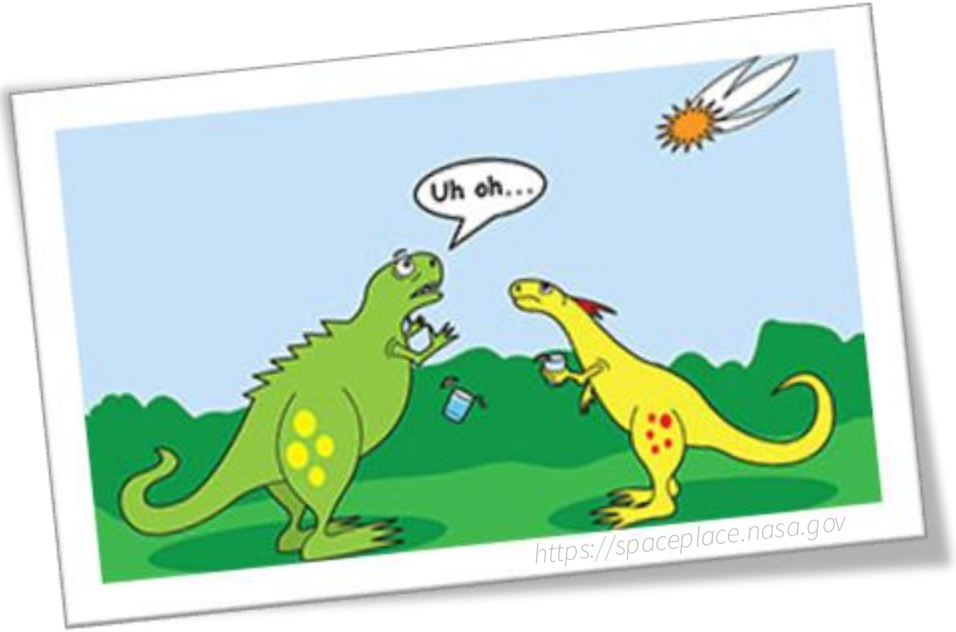


Parameter estimation for a Linear Ballistic Accumulator model of auditory change detection with Markov-Chain Monte Carlo

R. Barumerli, K. Ignatiadis, D. Baier, B. Tóth, R. Baumgartner

Auditory looming bias

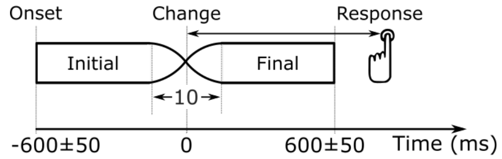
- Approaching sounds perceptually more salient than receding sounds
- Potential reason: more hazardous, evolutionary advantage

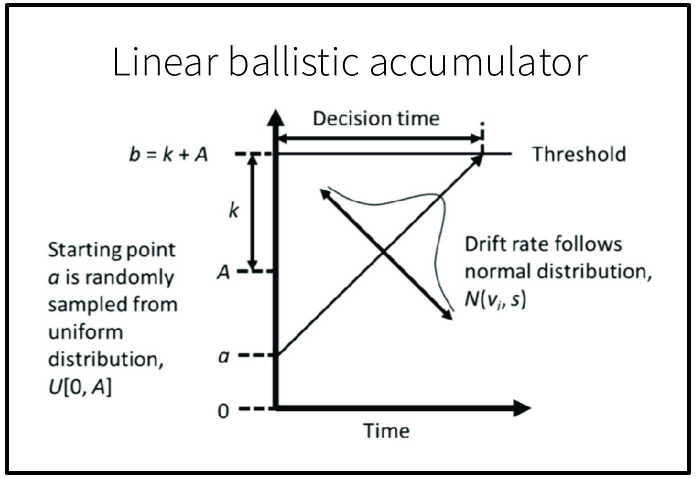


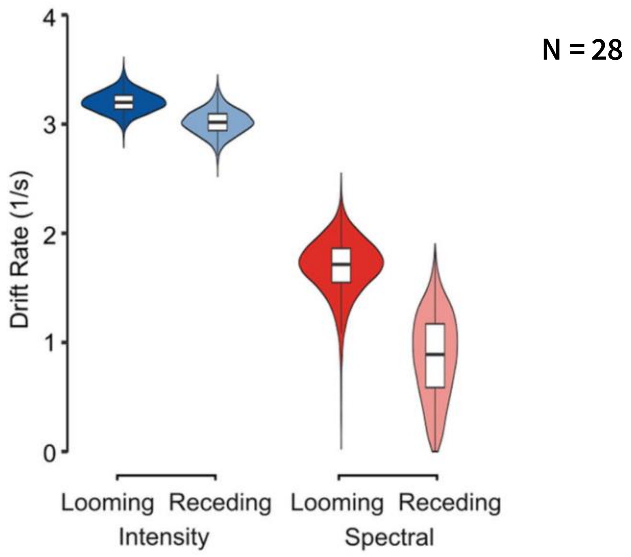
Neuhoff, J. G. (1998). Perceptual bias for rising tones. *Nature*, *395*(6698), 123–124.

Sensory evidence is accumulated faster for looming sounds

- Discrimination task: looming vs. receding (Human experiment)
- Prediction of human responses by computational model (parameter estimation via MCMC)





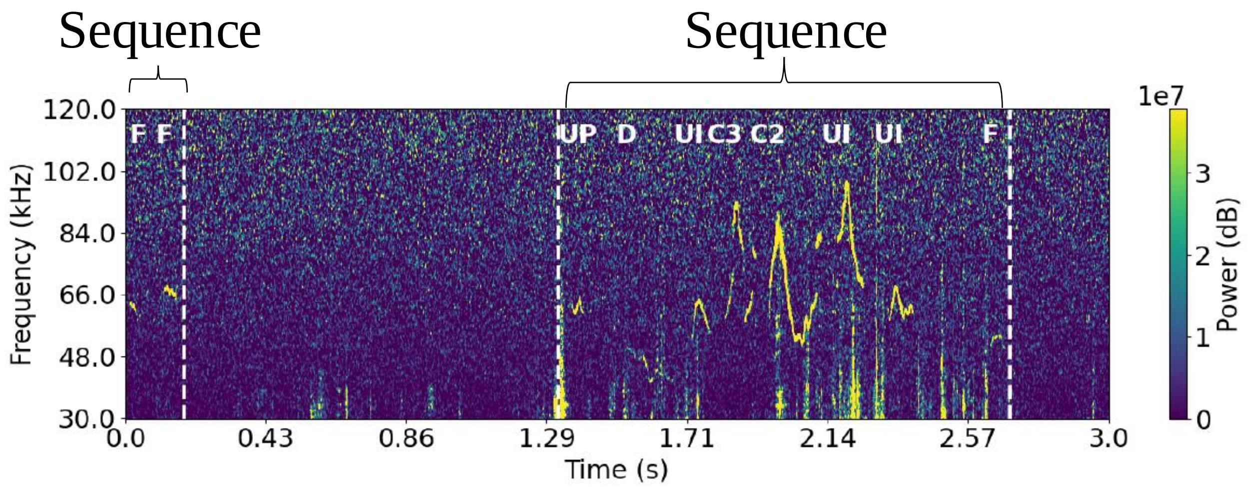


Ignatiadis, K., Baier, D., Barumerli, R. et al. Cortical signatures of auditory looming bias show cue-specific adaptation between newborns and young adults. Commun Psychol 2, 56 (2024).

Classification of Sequential Data

R. Abbasi, P. Balazs, F. Hlawatsch, S.M. Zala, G. Koliander

- **Applications**: bioinformatics, machine translation, speech recognition, animal vocalizations
- **Classifiers**: RNNs & LSTMs which capture temporal dependencies
- **Challenges**: High data demands lead to ignoring sequential patterns, thus reducing accuracy (e.g., animal vocalization studies)
- **Aim**: Develop more explainable method without relying on extensive data.

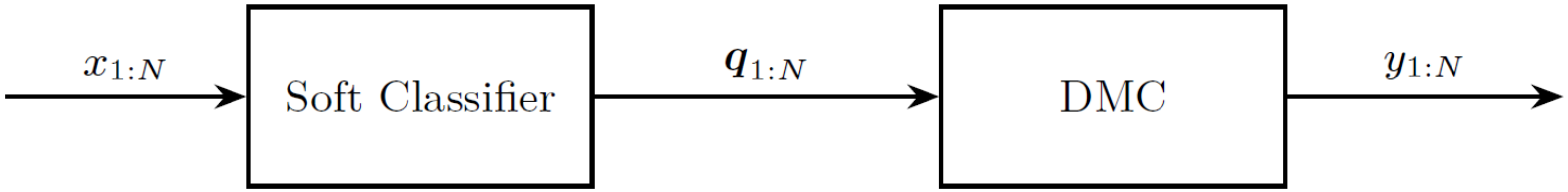


Abbasi, R., Balazs, P., Hlawatsch, F., Zala, S. M., & Koliander, G. (in preparation). "Classifying sequential data: A Bayesian framework integrating a soft classifier and a Markov model".

Dirichlet-Markov Classifier (DMC)

Proposed classification architecture

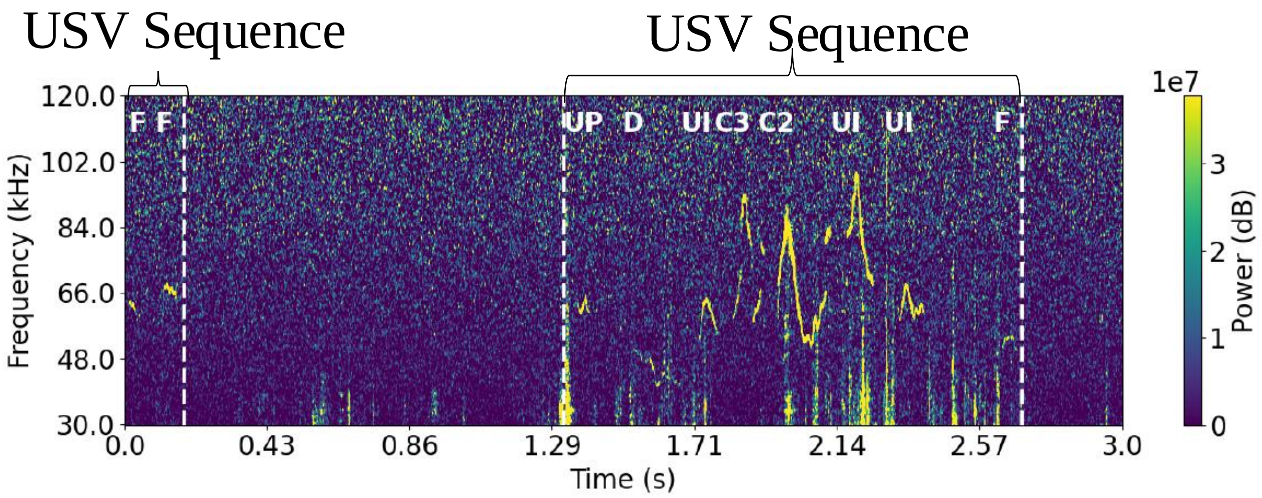
- a soft classifier generates intermediate outputs $\boldsymbol{q}_{1:N}$ for an input sequence $x_{1:N}$
- DMC integrates a Markov sequence model with $\boldsymbol{q}_{1:N}$ through Bayesian inference and assigns labels $y_{1:N}$ to the input sequence $x_{1:N}$



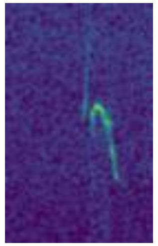
$$p(y_{1:N} \mid \boldsymbol{q}_{1:N}) \propto \prod_{n=1}^{N} p(\boldsymbol{q}_n \mid y_n) p(y_n \mid y_{n-1}).$$

Soft Classifier

Soft Classifier (BootSnap)



Soft Classifier



Neural Network layers

q_n: SoftMax output of BootSnap

C	C2	C3	D	F	FP	H	S	U	UI	UP	US
0.01	**0.6**	0.05	0.22	0.01	0.01	0.05	0.01	0.01	0.01	0.01	0.01

Max → C2

Results

- DMC compared with Dirichlet-based model, Markov-based model, and CNN
- DMC outperformed all methods on both synthetic and real data

synthetic data

		Class-wise F1 (%)									
Model	$F1$	C	$C2$	$C3$	D	F	H	S	U	UI	UP
CNN	74.6	75.7	31.1	26.7	89.8	92.0	71.0	87.8	85.9	90.1	95.7
Markov-based approach	75.1	75.1	31.3	28.5	89.5	92.7	72.8	89.5	85.8	89.9	95.6
Dirichlet-based approach	80.1	79.9	38.9	47.5	92.1	94.4	77.0	93.3	89.0	92.0	96.7
DMC	**82.6**	**82.1**	**46.7**	**53.9**	**93.1**	**95.3**	**81.7**	**93.3**	**90.3**	**92.9**	**96.7**

real data

		Class-wise F1 (%)									
Model	$F1$	C	$C2$	$C3$	D	F	H	S	U	UI	UP
CNN	68.3	70.2	25.5	33.3	80.0	81.5	**67.0**	95.5	67.3	77.0	85.5
Markov-based approach	67.9	70.2	23.4	31.7	79.4	81.5	66.8	95.5	67.9	76.8	**85.7**
Dirichlet-based approach	69.8	70.5	**31.5**	41.8	**80.6**	81.8	63.4	95.3	**69.8**	77.6	85.5
DMC	**70**	**71.7**	29.6	**44.4**	80.3	**81.9**	64.5	95.3	68.8	**78.0**	85.6

Machine Learning @ ARI

WWTF

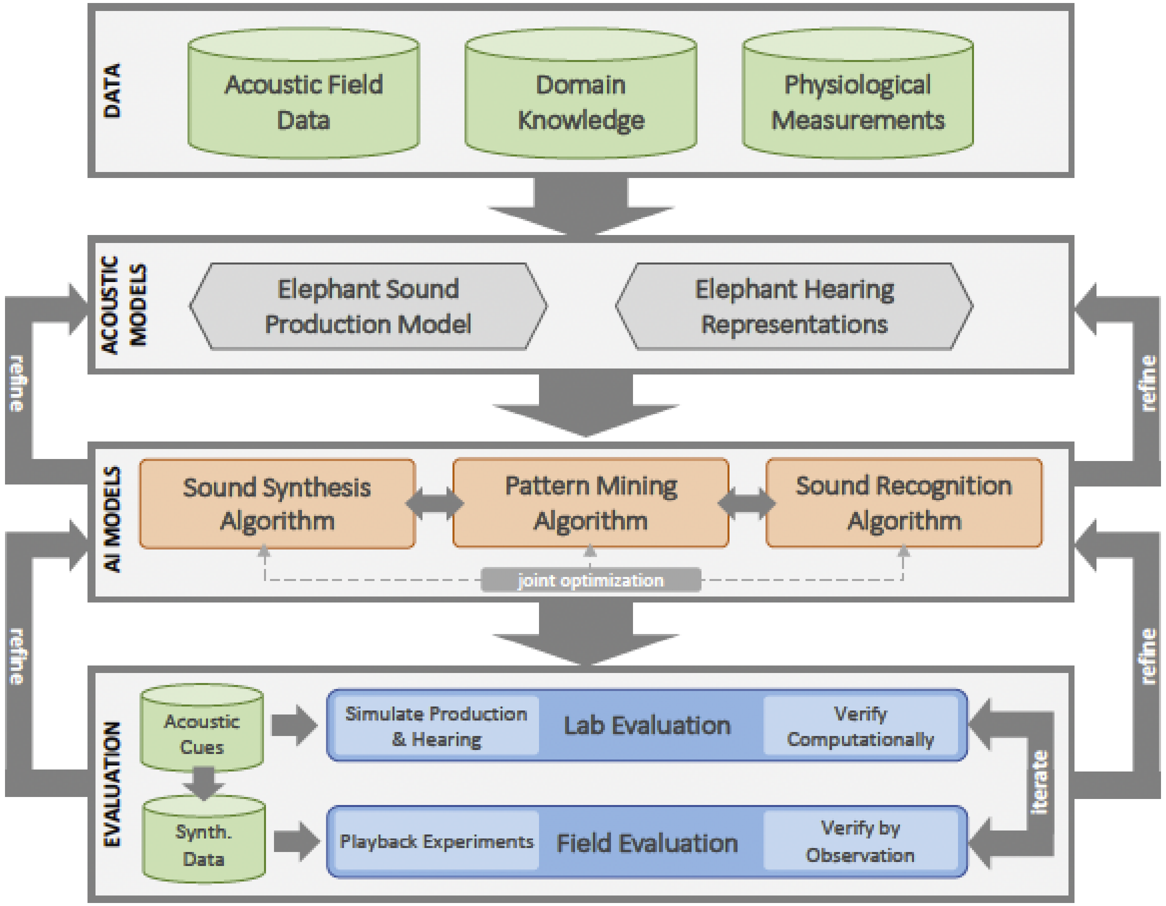
WWTF-funded focus on using AI to study animal communication (2024-2028)

Decoding Elephant Communication with AI

Principal Investigators: Angela Stöger-Horwath, Peter Balazs

- Wildlife preservation in increasingly human-dominated environment requires deeper understanding of animal behaviour, cognition, perception, and *communication*
- Develop models to identify acoustic cues relevant for elephant communication
- Create/work with largest dataset of annotated/curated African savannah elephant vocalizations

Planned project pipeline



- Combine advanced acoustic models with machine learning
- Computational models for elephant sound production and hearing + Evaluation in the wild
- Data- and knowledge-driven

ÖAW AUSTRIAN ACADEMY OF SCIENCES

ARI

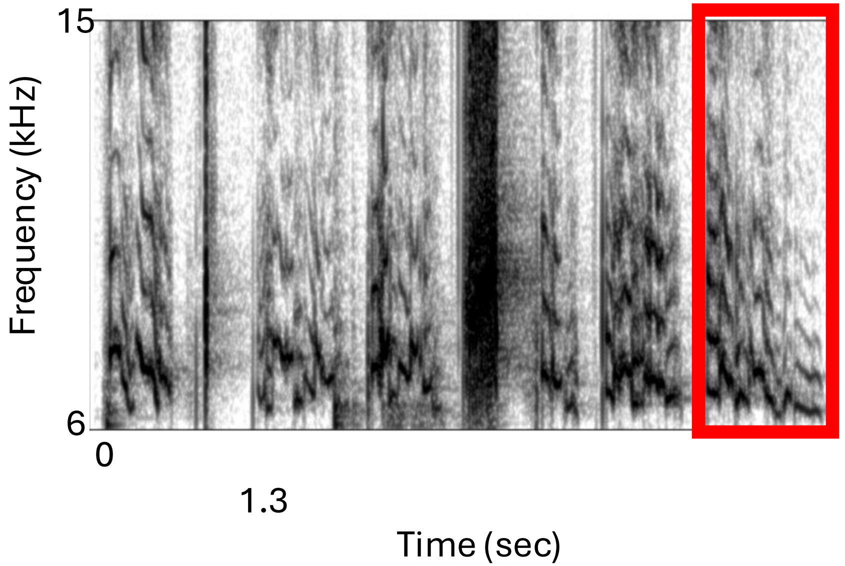
ANIML – Understanding Animal Communication with Machine Learning

Core Team: M. Hoeschele (PI), N. Holighaus (CoPI), G. Koliander (CoPI), J. Oh (PD), Z. Katona (PhD)

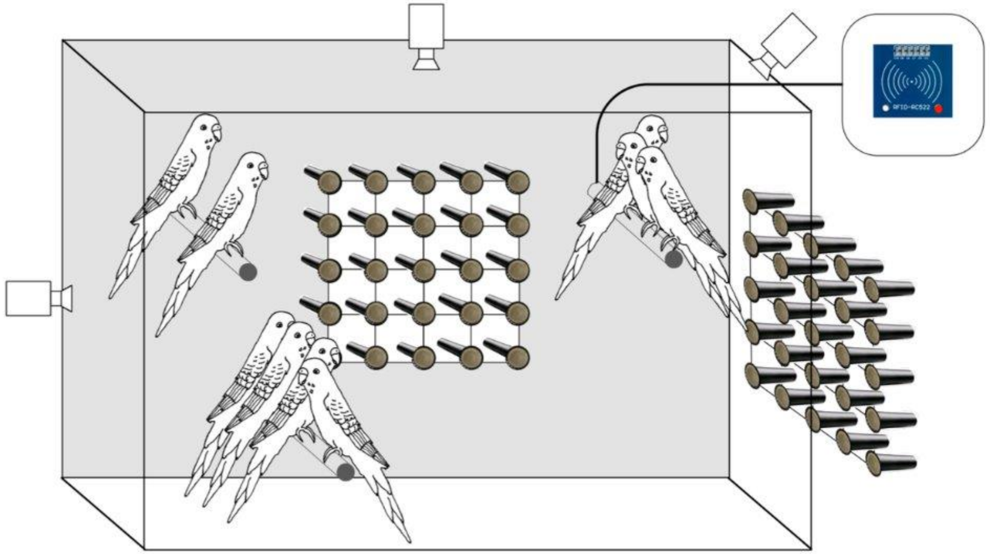
Understanding communication (human or animal) requires knowledge about **context** and **structure**.

Context: When is a vocalization performed? Who or what else is present?

Structure: What pieces are us to build vocalizations, how are they ordered?



Recording in context



- Humans and many animals mostly vocalize in social contexts and in groups
- Obtaining clean individual recordings in natural(-istic) situations is difficult

The approach of ANIML:

- Obtain a large dataset of multi-microphone recordings of animals (budgies) in a group
- Retrieve auxilliary position information via additional recording modalities (e.g., video)
- Separate into individual sources using physical models, state-of-the-art audio processing and ML

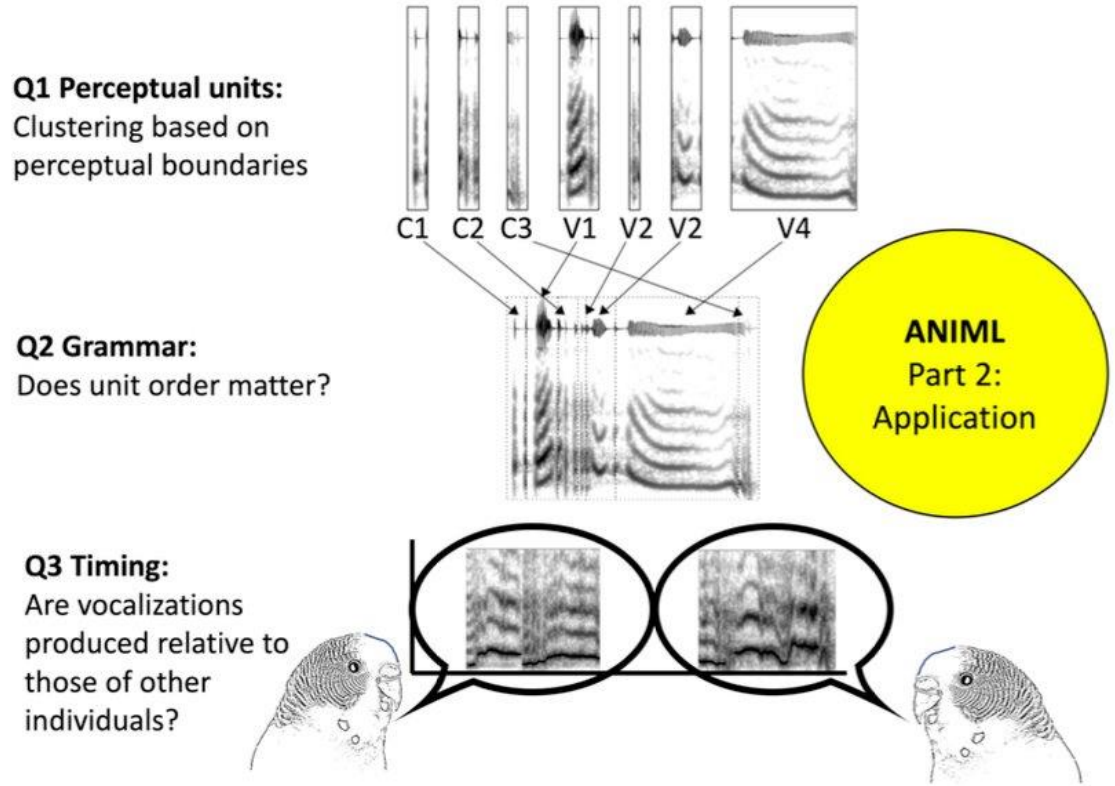
Making sense of complex vocalizations

- Segmenting complex animal (or human) vocalizations at silence is not sufficient
- How can meaningful segmentation be achieved?

The approach of ANIML [Q1]:

- Expanding prior work on applying a **universal speech segmenter** to **budgie vocalizations**

- Verification of results using recombined, synthetic budgie vocalizations in behavioral tests



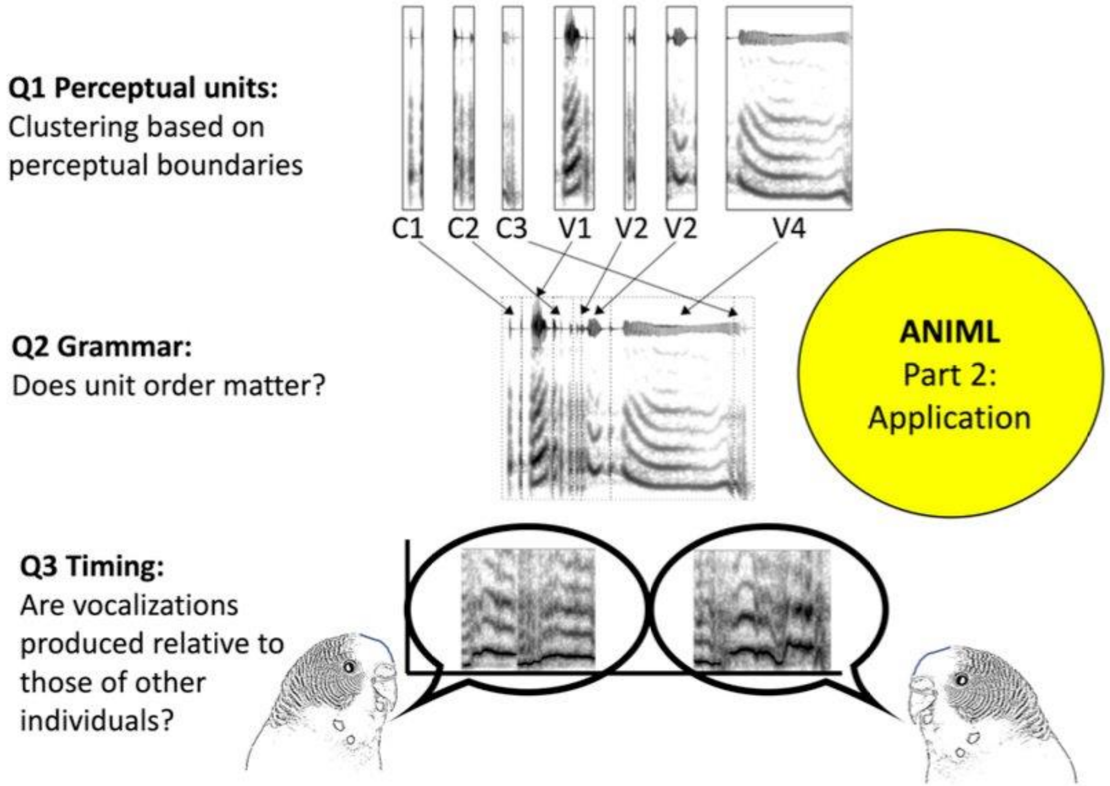
Making sense of complex vocalizations - II

- Semantics are (probably) important
- Search and test for meaningful patterns

The approach of ANIML [Q2 & Q3]:

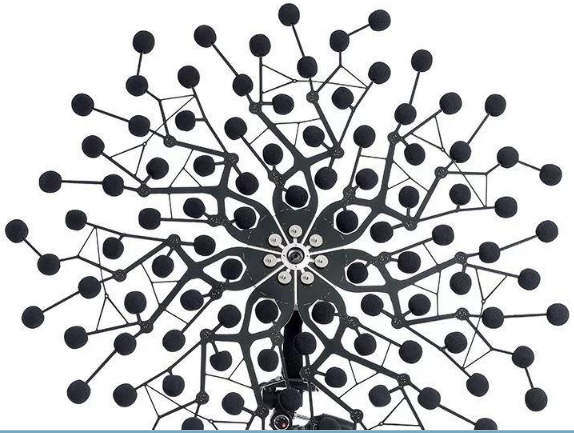
- Analyze recordings for repeating patterns on the scales of phrases and segments

- Analyze recordings with respect to timing between vocalizations in a 'conversation'



Goals for the dataset

- Large group size (8-16 individuals in an aviary)
- Consistent, reproducible recording conditions
- Raw audio (~110 channels) and video (~20 channels) data
- Preprocessed (separated, denoised) streams per individual
- Extended duration (100h+)
- Meta-data: Recording conditions, individuals, time-stamps, pre-segmentation (by silence), etc.
- Publicly available

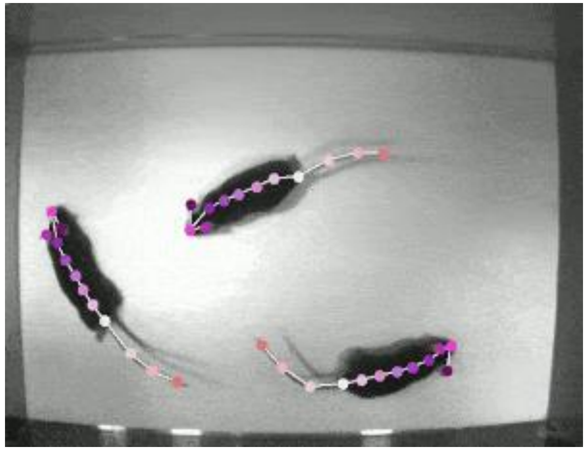


ÖAW AUSTRIAN ACADEMY OF SCIENCES

ARI

Technical Challenges

- Synchronize recordings between channels and modalities
- Track (many) individuals in video for position information
- Physics-based beamforming not good enough for separation
- Prior separation techniques (ML or otherwise) use only few channels (usually <10)
- Large number of sources
- Little training data (Synthesize?, Augment?, Fine-Tune?)



?

Machine Learning @ ARI

Invertibility and Stability of Neural Networks

[Peter Balazs and Team(s)]

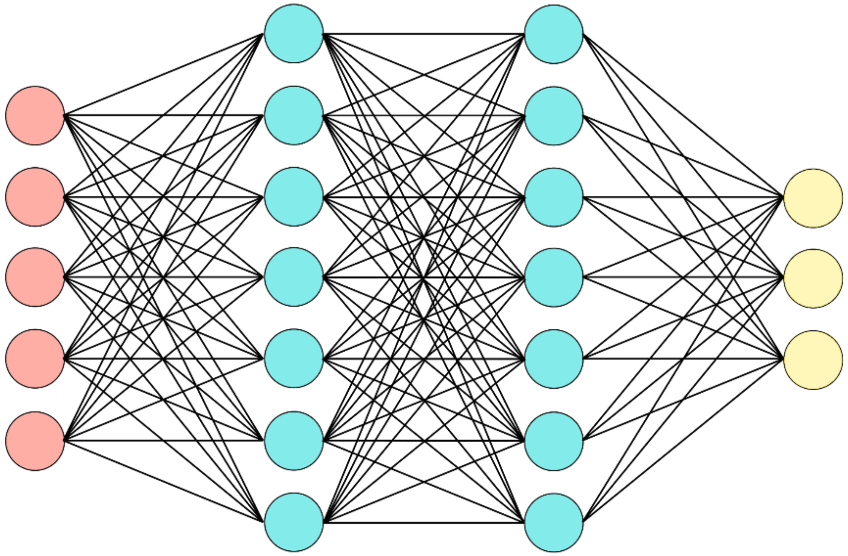
Mathematics for Machine Learning

A solid mathematical foundation is crucial for ML. While the mathematical understanding of what a neural networks can do – e.g. approximation properties – has impressively progressed recently, we set the focus on understanding why and how neural networks produce their output given their input.

Deep neural networks (DNNs) comprised of (affine) linear operators and (usually non-expanding, pointwise) non-linearities

$$x^{(i)} = \sigma^{(i)}\left(A^{(i)}x^{(i-1)} + b^{(i)}\right) =$$

$$= \left(\sigma^{(i)}\left[\left\langle x^{(i-1)}, \psi_n^{(i)}\right\rangle + b_n^{(i)}\right]\right)_{n\in N^{(i)}}$$



Long-term Goal

Understand (invertible) neural networks by expanding frame theory to include non-linear activation functions and developing new interpretable ML approaches in acoustics.

Frame theory:

$$A_\Psi \|f\|_{\mathcal{H}}^2 \leqslant \sum_{n \in N} |\langle f, \psi_n \rangle_{\mathcal{H}}|^2 \leqslant B_\Psi \|f\|_{\mathcal{H}}^2, \quad \forall f \in \mathcal{H}.$$

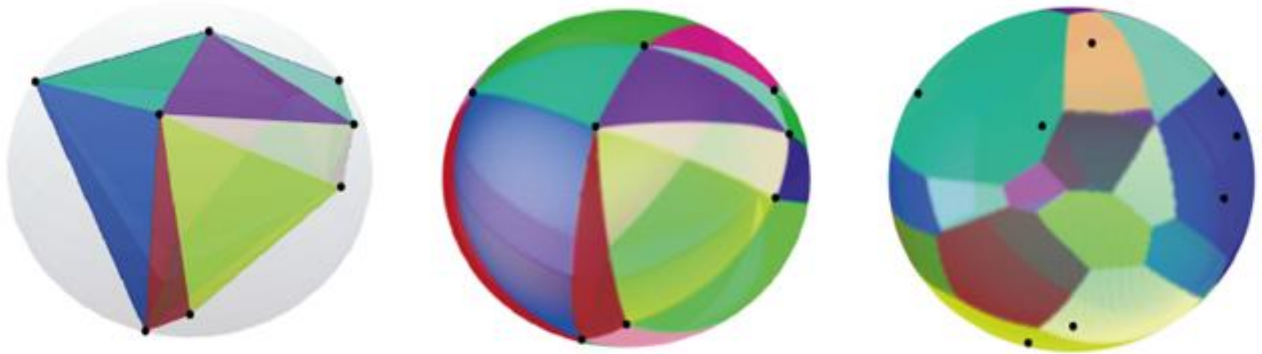
Non-linear Frame theory:

$$A_\Psi^\sigma \|f\|_{\mathcal{H}}^2 \leqslant \sum_{n \in N} \left|\sigma\left(\langle f, \psi_n \rangle_{\mathcal{H}}\right)\right|^2 \leqslant B_\Psi^\sigma \|f\|_{\mathcal{H}}^2 \quad \forall f \in W.$$

Injectivity and stability of ReLU-layers

D. Haider, M. Ehler, H. Eckert, D. Freeman, P. Balazs

- Characterization of invertible ReLU-layers using frame theory
- Estimates for lower Lipschitz-bound of ReLU-layers
- Algorithms for verification

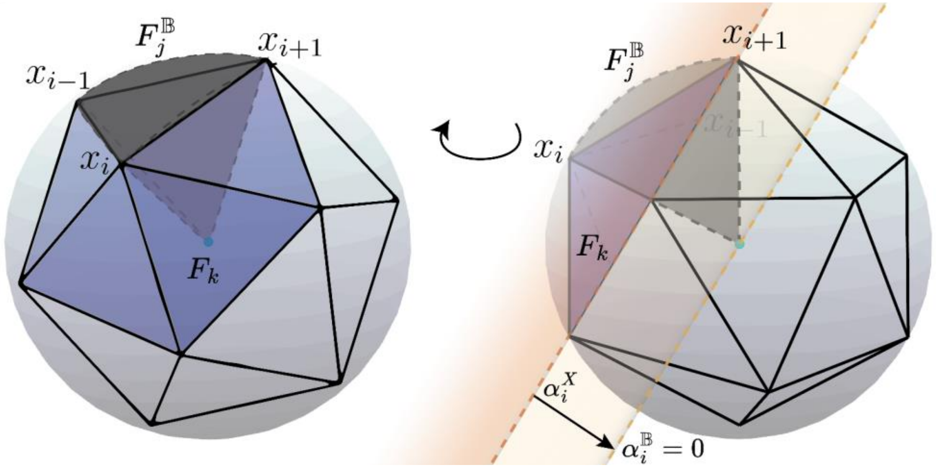


Domain decompositions for computing maximal bias ensuring invertibility

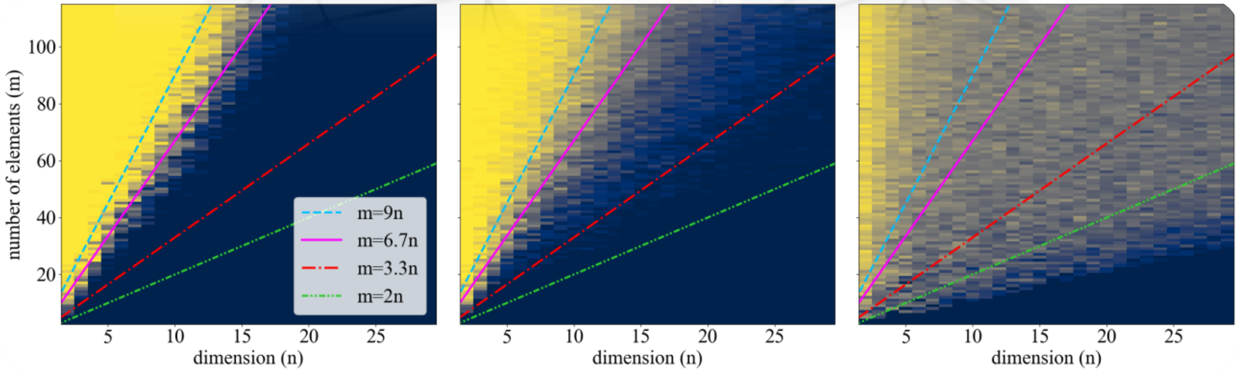
Haider, D., Balazs, P., Ehler, M., "Convex Geometry of ReLU-layers, Injectivity on the Ball and Local Reconstruction" ICML 2023

Verification algorithms

- **Deterministic:** Polytope bias estimation

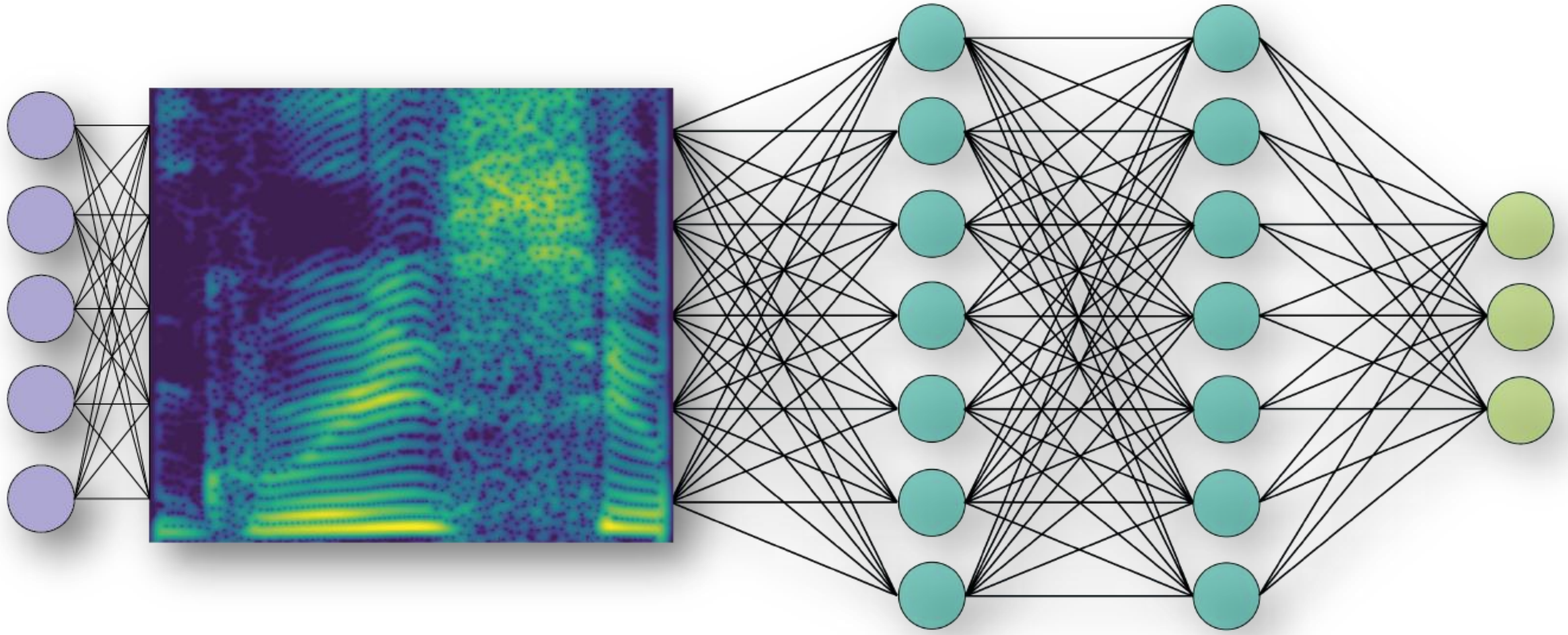


- **Probabilistic:** Monte-Carlo bias estimation



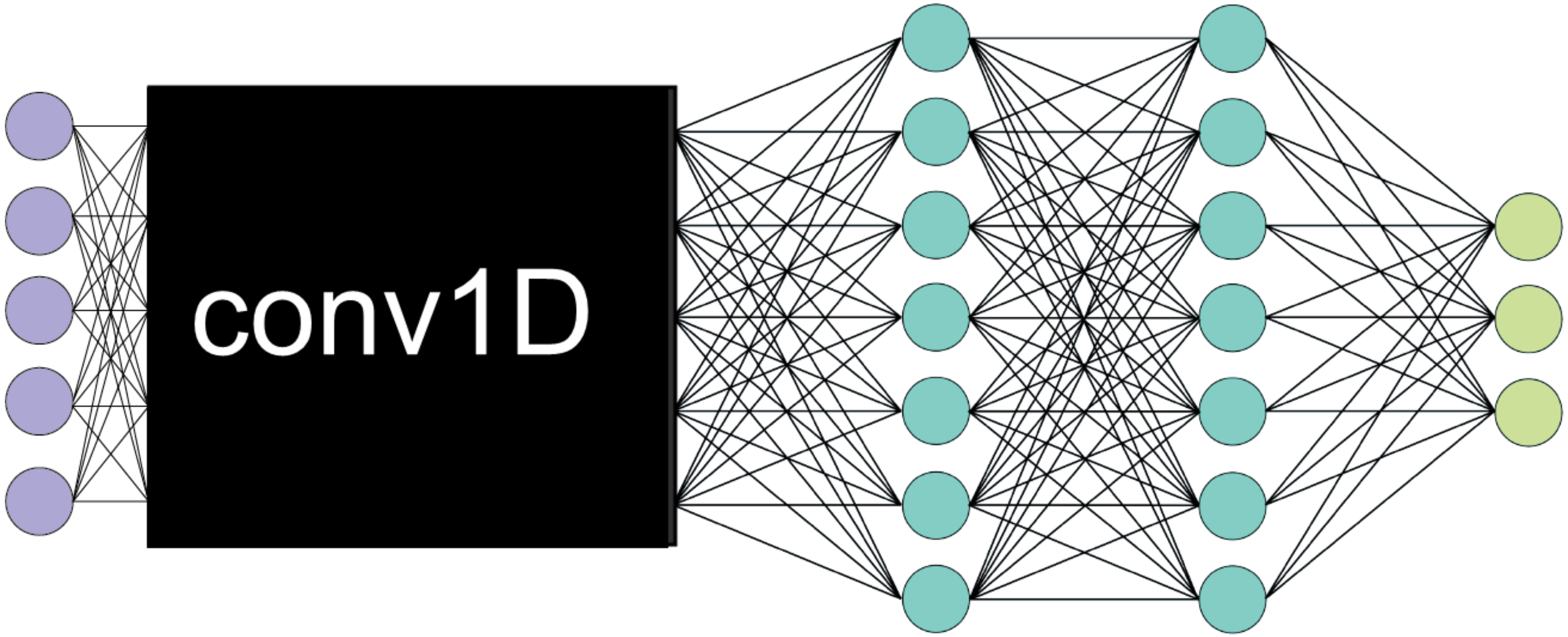
Encoding Audio in NNs

Classical Approach:



Encoding Audio in NNs

End-to-End:



~~Interpretable?~~ Stable?

Differentiable Regularization of the Condition Number for numerically stable DNNs

R. Nenov, D. Haider, P. Balazs

- Condition numbers (CN) measure stability of linear operators (output energy can be estimated by input energy) -> include in loss function.

- **Problems:**

  - Dependence of CN on operator is discontinuous

  - Trade-off between expressivity and stability?

Denoising results on MNIST data (Panels: High, Medium, Low SNR)
Top: Noisy data, Middle: Denoising without CN regularization,
Bottom: with CN regularization

Balazs P., Haider, D., Lostanlen, V., Perfler, F., "Trainable signal encoders that are robust against noise", Inter-Noise 2024
Nenov, R., Haider, D. & Balazs, P. (2024). “(Almost) Smooth Sailing: Towards Numerical Stability of Neural Networks Through Differentiable Regularization of the Condition Number”. ICML 2024 Workshop on Differentiable Almost Everything

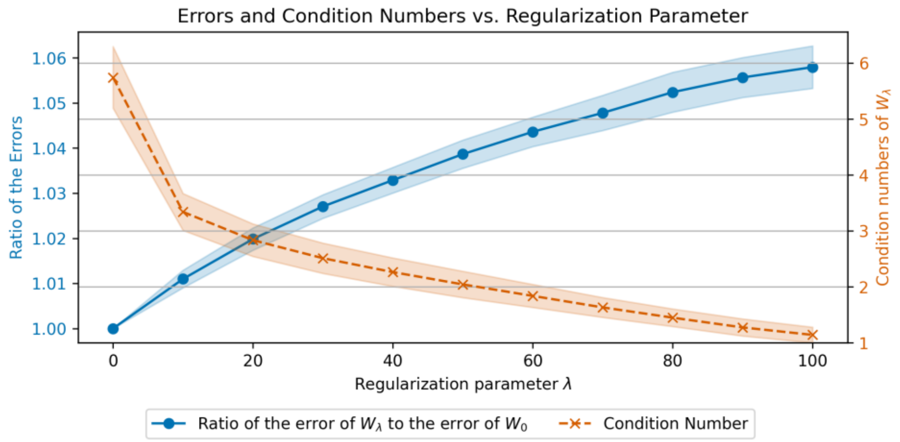
A differentiable regularizer

Since $\kappa(S) = \sigma_{\max}(S)/\sigma_{\min>0}(S)$ is not continuous in S, use instead:

$$r(S) := \frac{1}{2}\|S\|_2^2 - \frac{1}{2\nu}\|S\|_F^2 \,,$$

where $\nu = \min\{n, m\}$.

- Minima coincide
- Almost everywhere differentiable
- Gradient steps are guaranteed to reduce CN



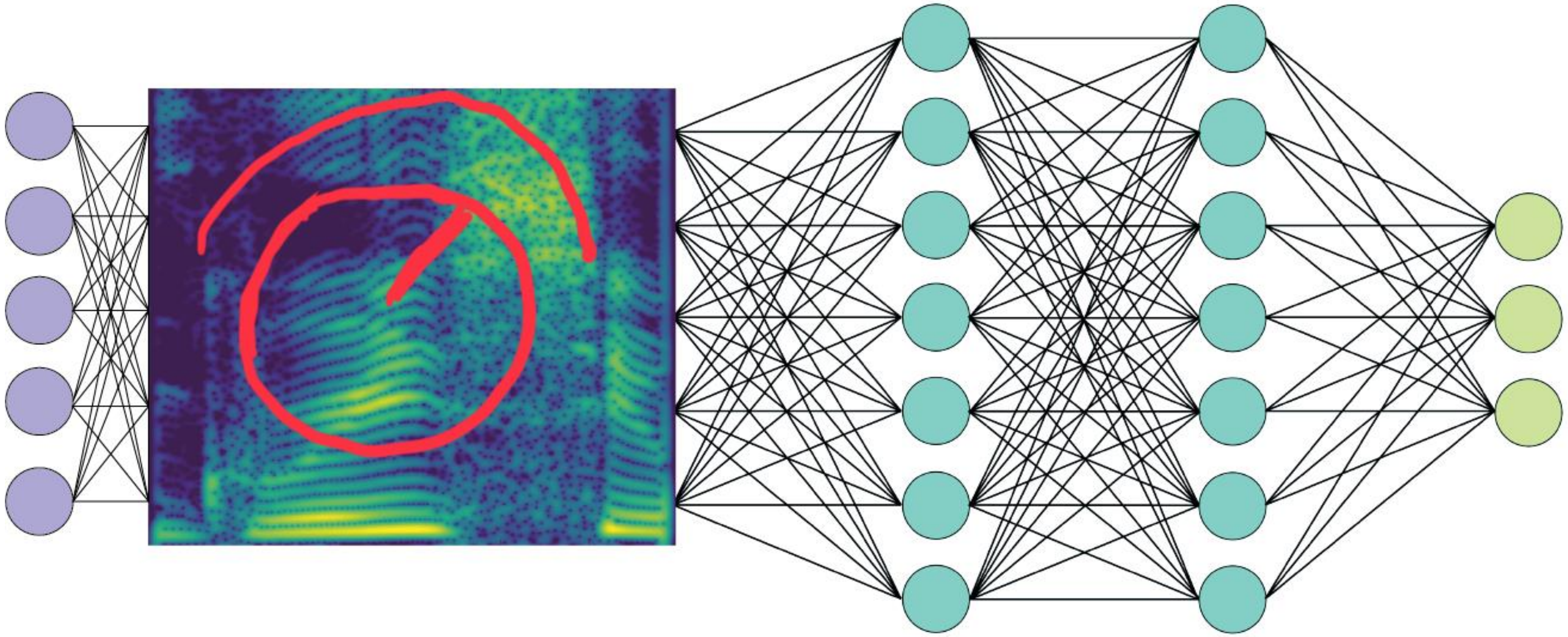
Regularization<–>Error Trade-off

λ	$\kappa(W_1)$	SNR ∞	SNR 1	SNR 0.5
0	43.12	**98.42 %**	**93.80 %**	71.91 %
10^{-3}	9.43	**98.38 %**	91.72 %	62.51 %
10^{-2}	4.62	98.11 %	91.95 %	63.61 %
10^{-1}	**1.45**	96.77 %	**93.68 %**	**74.27 %**
1	**1.53**	96.50 %	92.61 %	**84.25 %**

Table: Condition numbers of the first network layer and the classification accuracy on the test set for three SNRs.

Encoding Audio in NNs

Hybrid:

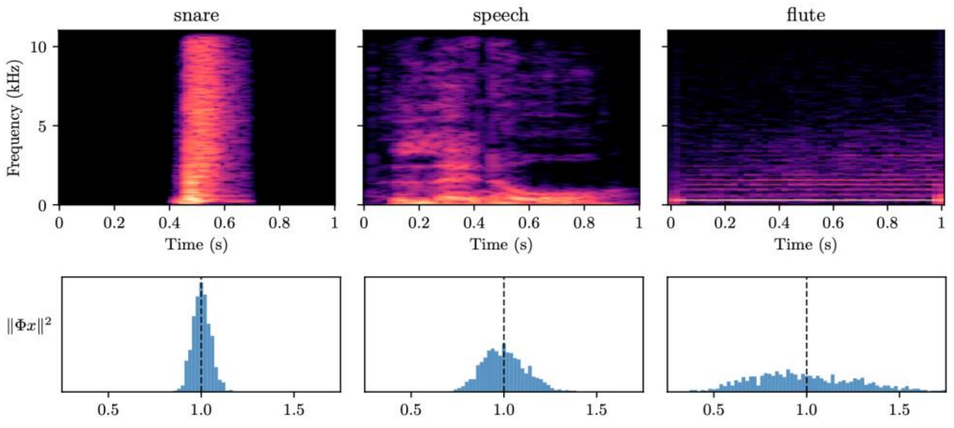


Interpretable? Stable?

Tightness for trainable audio encoders

D. Haider, F. Perfler, V. Lostanlen, M. Ehler, P. Balazs

- Analysis of Conv-Layers in audio encoders as oversampled FIR filter banks: Tightness = Small condition number
- Stability analysis of Gaussian *random filterbanks* (random initialization of network weights)
- Construction of tight *hybrid filterbanks* via perceptually-motivated inductive bias

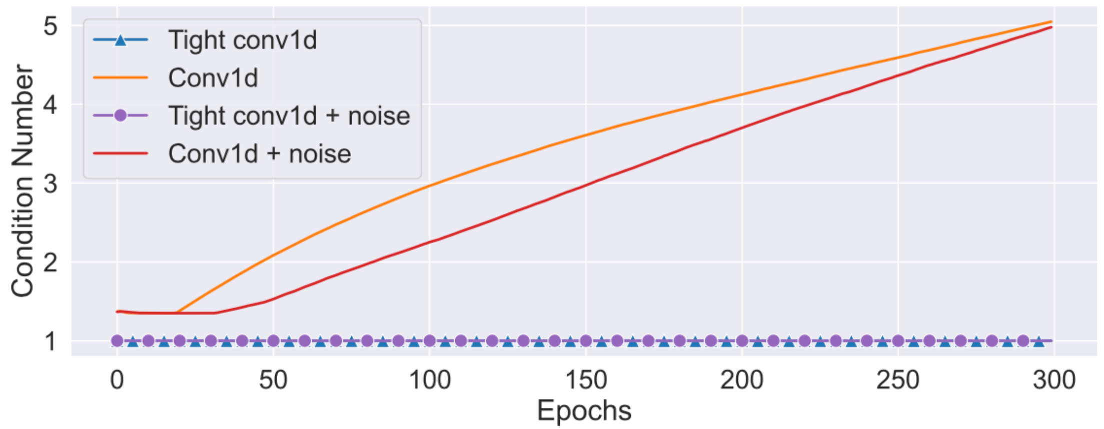


Energy deviation for audio signals with different auto-correlation characteristics

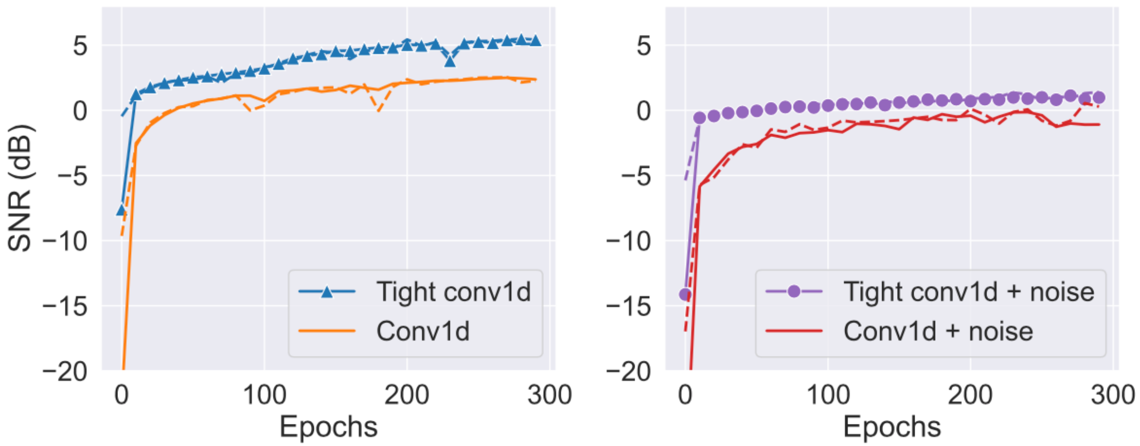
Haider, D., Lostanlen, V., Ehler, M., Balazs, P. , "Instabilities in Convnets for Raw Audio", IEEE Signal Processing Letters (2024)
Haider∗, D., Perfler∗, F., Lostanlen, V., Ehler, M., Balazs, P., "Hold Me Tight: Stable Encoder–Decoder Design for Speech Enhancement", Interspeech 2024

Results

- Effect of stabilization mechanisms



- Improved SNR in a denoising task



Machine Learning @ ARI

Conclusion

Machine Learning @ ARI - Summary

- Varied interests in ML as a tool or a field of study
- Interdisciplinary cooperation in projects and via ML Team
- Different groups/projects require different tools and expertise
- Project-based structure naturally leads to fluctuation
- We believe we're doing pretty well, though.

Thank you for listening! Join the tour of our lab if interested (directly after the talk).