# AES EUROPE 2026

## From DSP to AI Audio Engineering

Cumhur Erkut | Aalborg University Copenhagen

**28-30 May 2026** • Technical University of Denmark • Copenhagen, Denmark

# OVERVIEW

## TALK ROADMAP

### 1 · The Classical Foundation
•DSP and physical modeling as the bedrock of musical acoustics

### 2 · The Machine Learning Turn
•How data-driven methods are transforming the pipeline

### 3 · Two Demonstrators
•Measurement-driven studies + Physical modeling in XR

### 4 · The AI Audio Engineering Framework
•Harnessing the current practice in Agentic AI for Audio

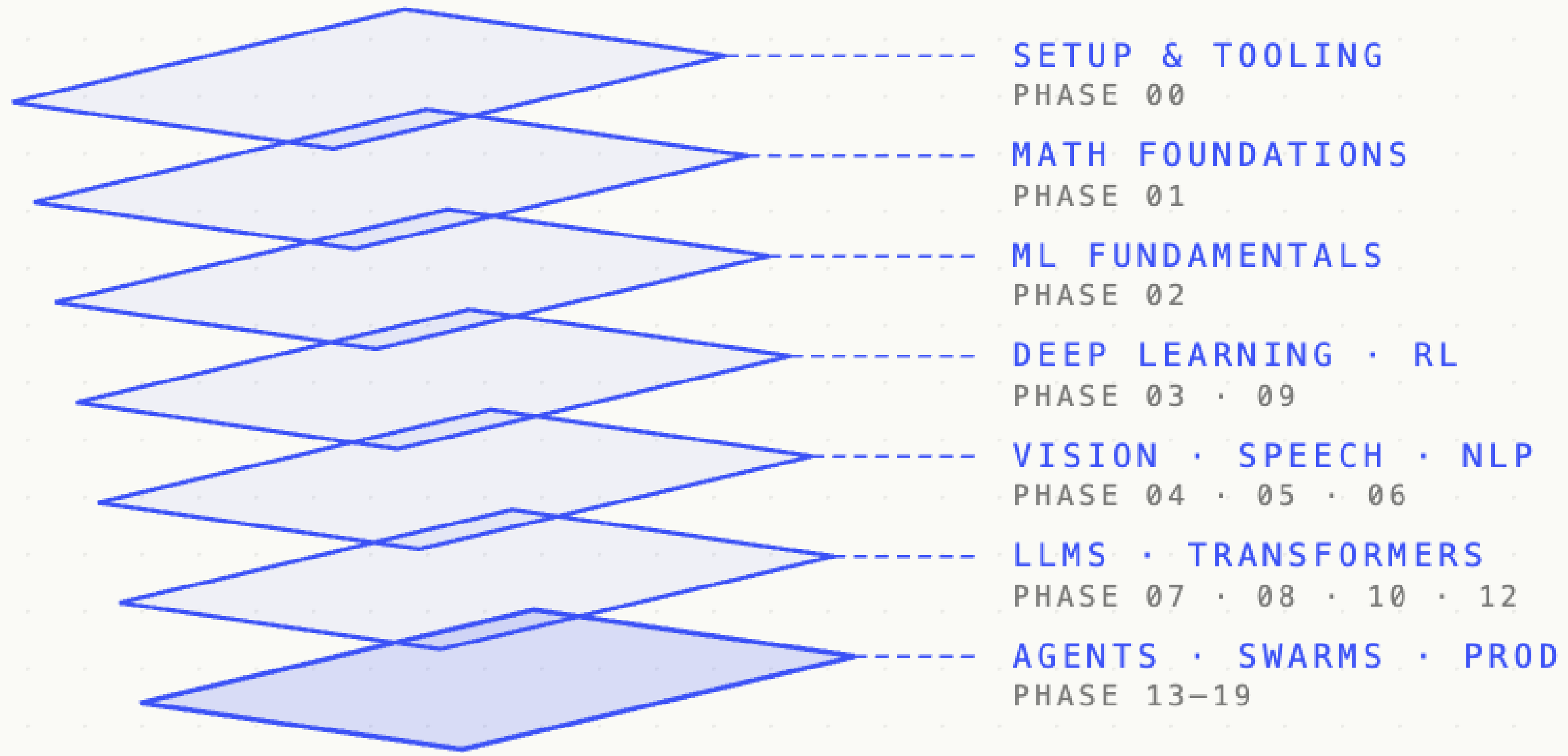
AES EUROPE 2026

# AI ENGINEERING FROM SCRATCH.

A reference manual for people who want to design and build AI systems from first principles.

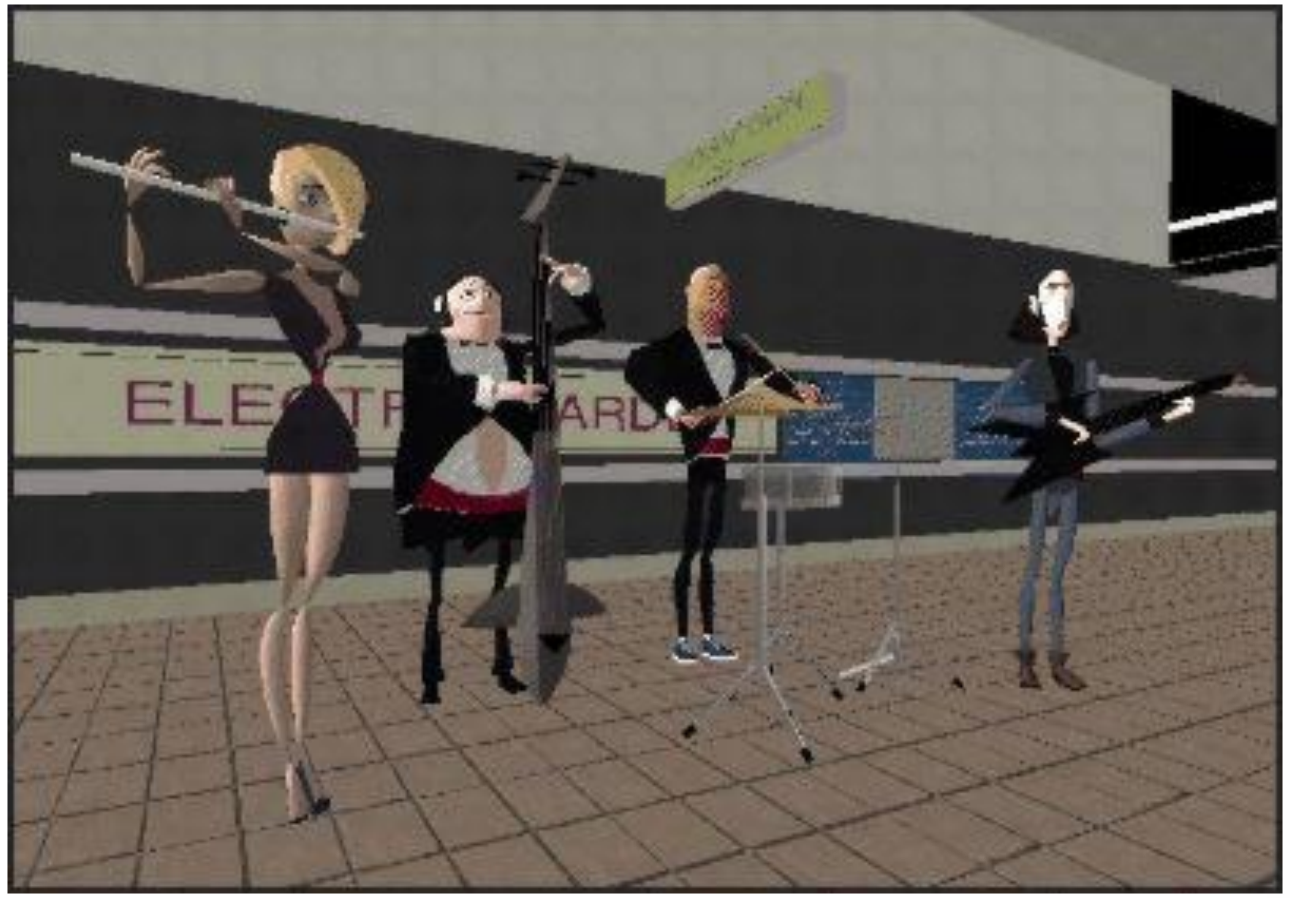
20 PHASES · 280+ LESSONS · PYTHON · TYPESCRIPT · RUST · JULIA

FIG_000.A — CURRICULUM STACK



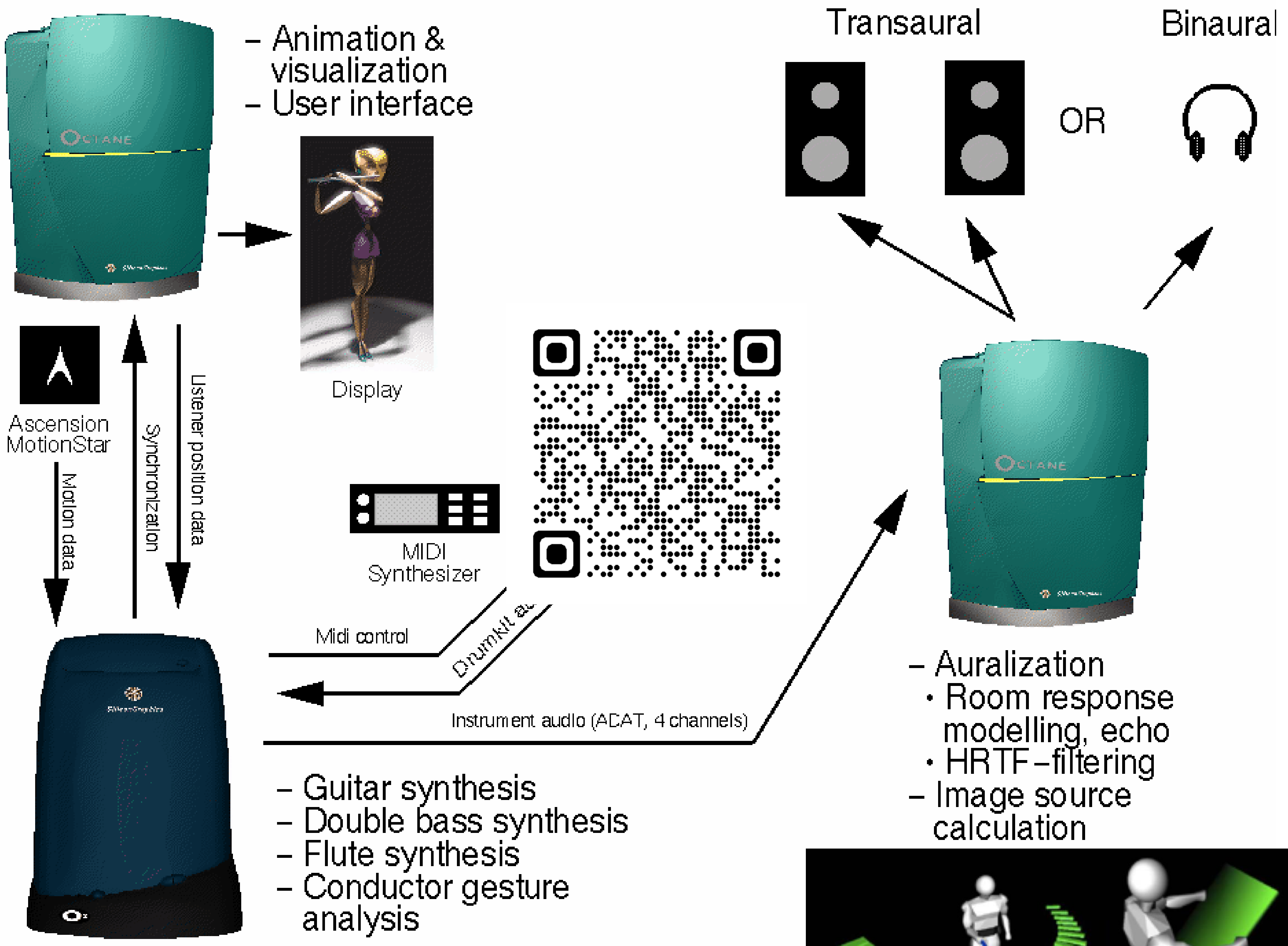

https://github.com/rohitg00/ai-engineering-from-scratch

# A short history: ‘97 MSc (TR), ‘98 to FI, ‘02 PhD, ‘07-’12 ARF (FI), ‘13 to DK-SMC, '16 DK-EDU, '20 COVID, '22 AI/ML, '24 AI Eng

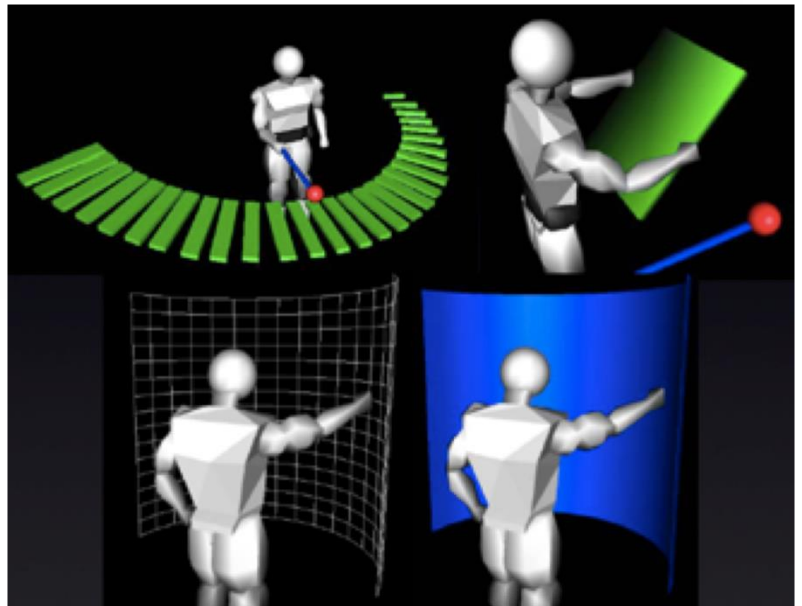



*Snapshot of [SIGGRAPH 97 demo]*

**DIVA is a collaborative research group of the following topics:**
•Real-time automatic character animation
•Interaction through **motion analysis,** especially conductor following.
•Sound generation with physical instrument models
•Acoustics modelling and auralization
•EVE - The Experimental Virtual Environment




Old is the new new.



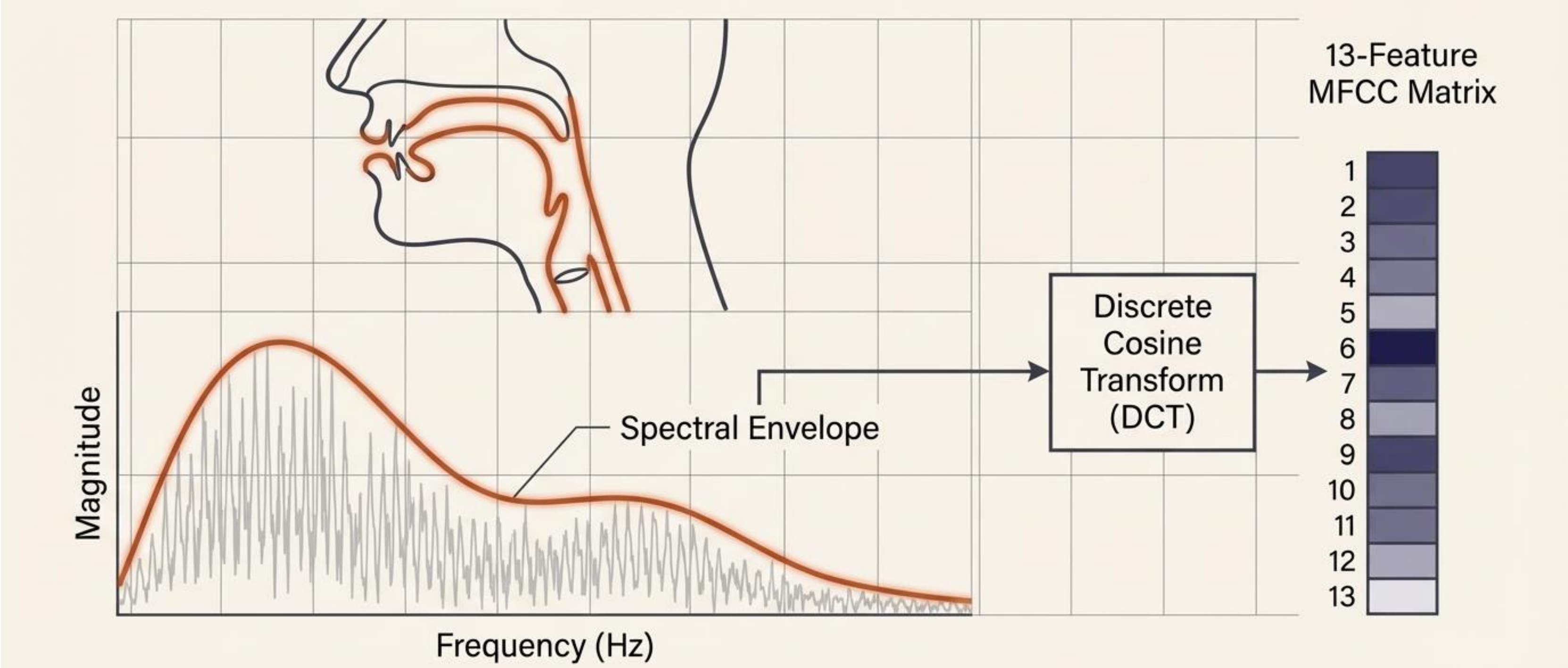
# THE CLASSICAL DSP FOUNDATION

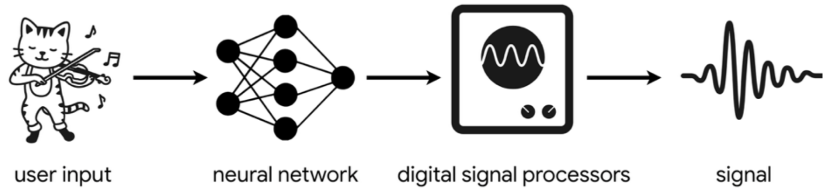
## CLASSICAL DSP METHODS

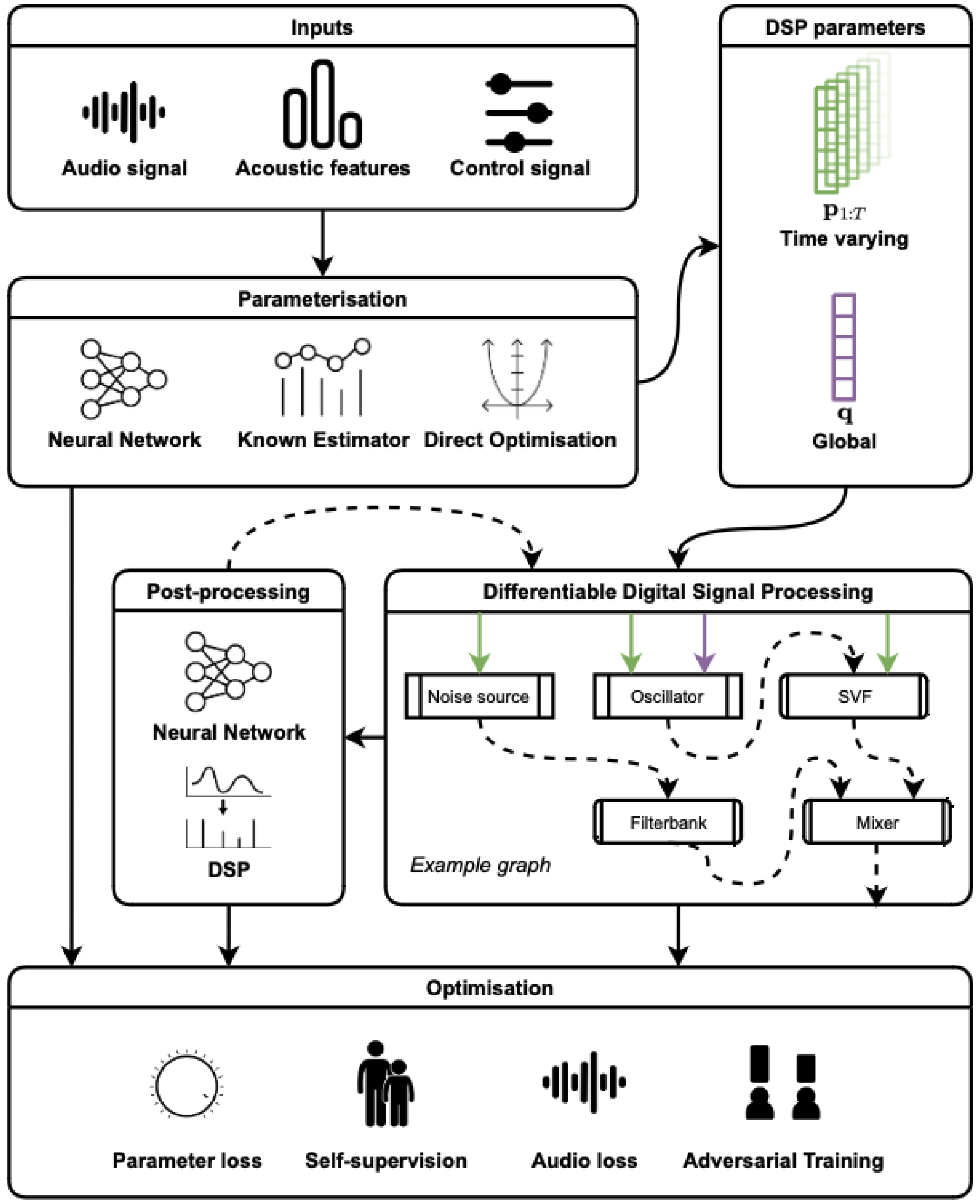
- Filtering & spectral analysis
- Modal analysis => IIR Filter Banks
- Convolution and reverb
- Delay networks & waveguides
- Transfer function modeling

## WHAT DSP GAVE US

- Tight analytical control over system behavior
- Reproducible studies of timbre, articulation, resonance
- Integral to real-time synthesis engines
- Foundation for interactive sound design
- Algorithms with clear physical interpretation





Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). DDSP: Differentiable Digital Signal Processing. In , Proc. Intl. Conf. Learning Representations (ICLR)

Hayes, et al (2024). A review of differentiable digital signal processing for music and speech synthesis. Frontiers in Signal Processing, 3 http://dx.doi.org/10.3389/frsip.2023.1284100

# PHYSICAL MODELING: A MATURE DISCIPLINE

## A MATURE SYNTHESIS & SIMULATION TECHNOLOGY

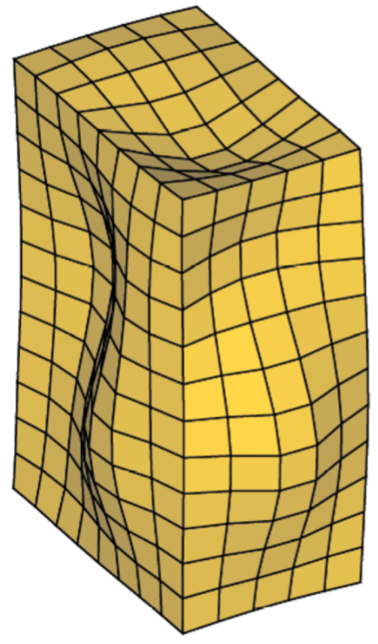
- Vibro-acoustic systems simulated in real time
- Digital prototyping for traditional and unconventional instruments
- Perceptually meaningful control interfaces support expressivity
- Computational tractability with physical interpretability

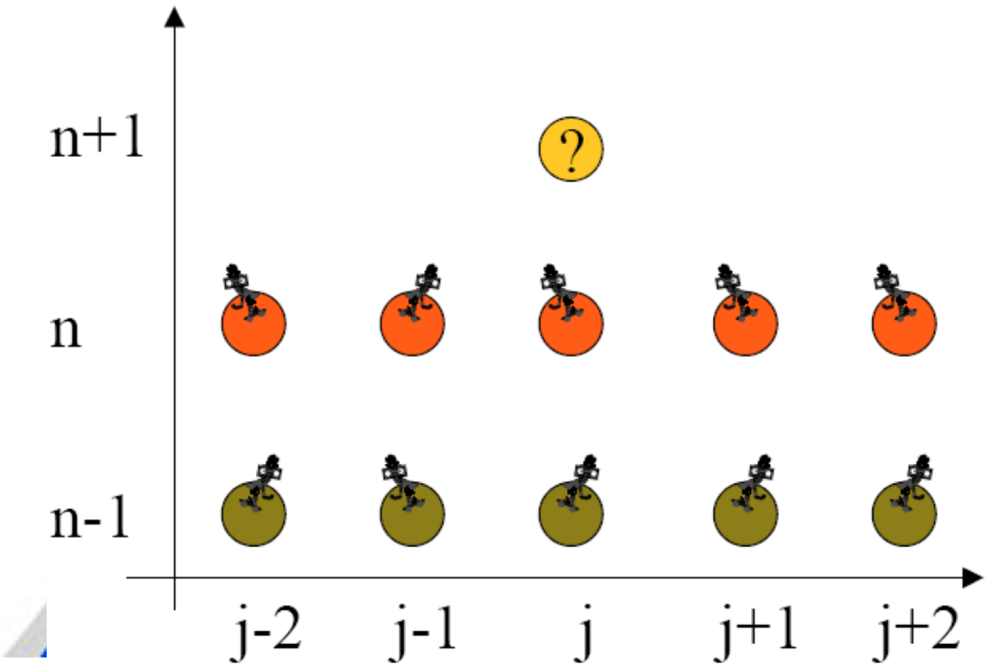
## KEY PARADIGMS

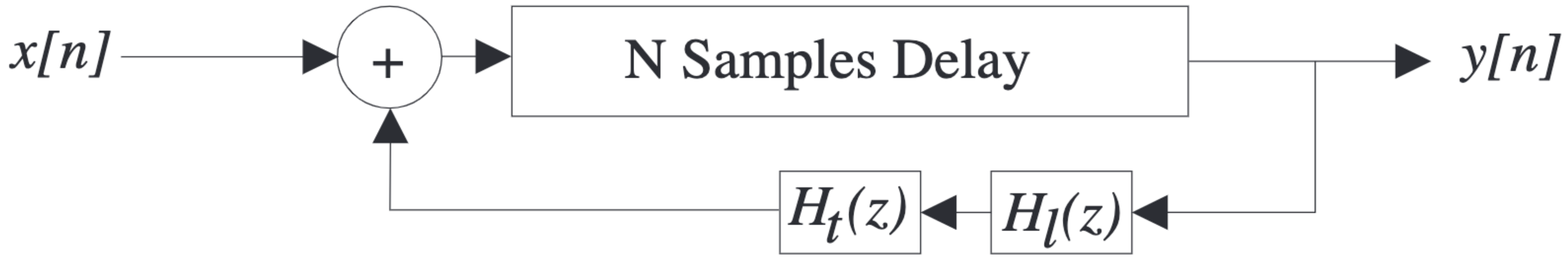
- Finite-difference time-domain (FDTD) simulation
- Digital waveguide synthesis
- Modal synthesis and spectral modeling
- Coupling vibro-acoustic models with haptic/gestural control

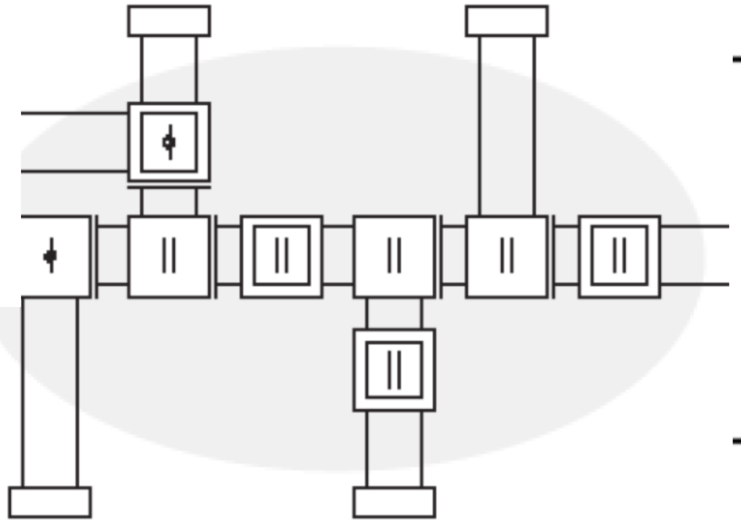
$$K\mathbf{P} + i\omega C\mathbf{P} - \omega^2 M\mathbf{P} = i\rho\omega\mathbf{F}$$

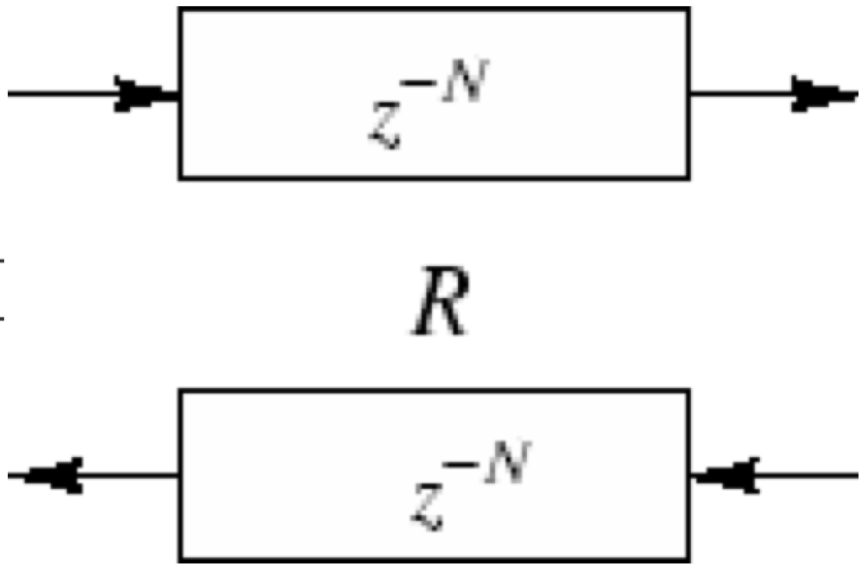
M, C, and K are mass, damping, and stiffness matrices








AES EUROPE 2026

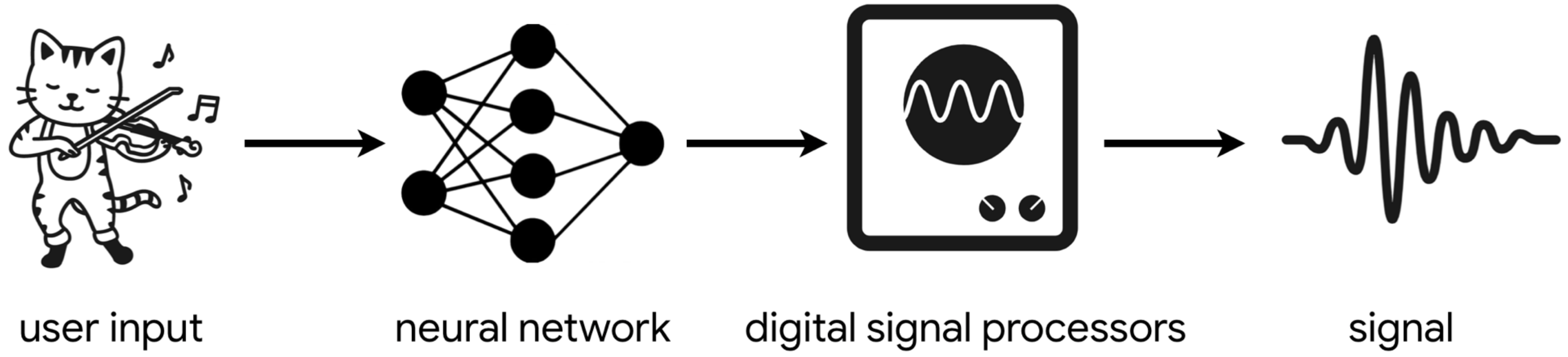
# THE MACHINE LEARNING TRANSFORMATION

## WHAT WE CAN NOW LEARN

- Impulse and frequency responses
- Transfer functions from data
- High-level parameter relationships
- Residual behavior beyond analytical models

## NEURAL APPROACHES

- Differentiable digital waveguides
- Neural reverberation
- Differentiable modal synthesis
- Learned excitation models




## THE KEY INSIGHT

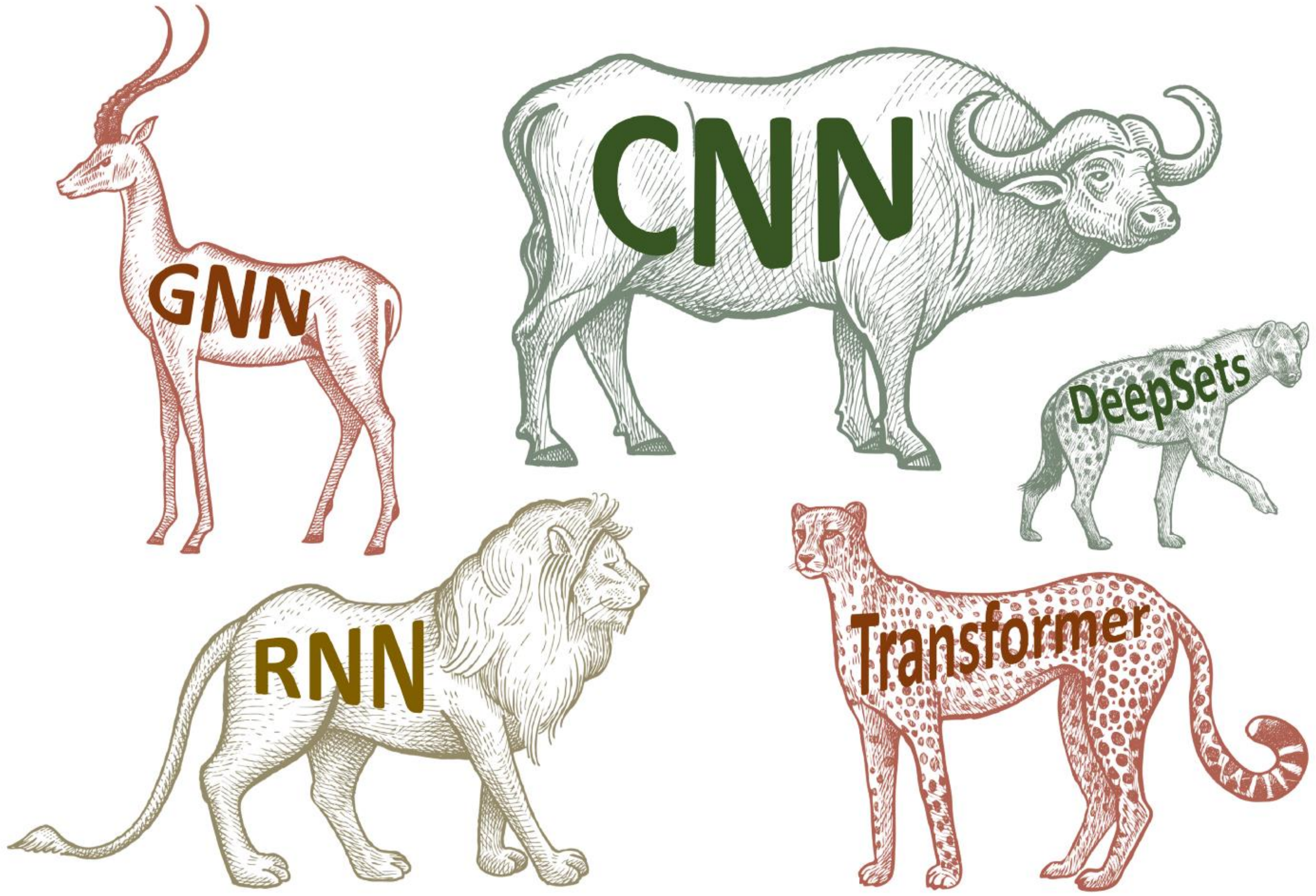
Data-driven methods complement physical

**DSP handles:**

- Interpretable, latency-critical stages

**Neural components handle:**

- Ill-posed inference tasks
- Residual and nonlinear behaviors



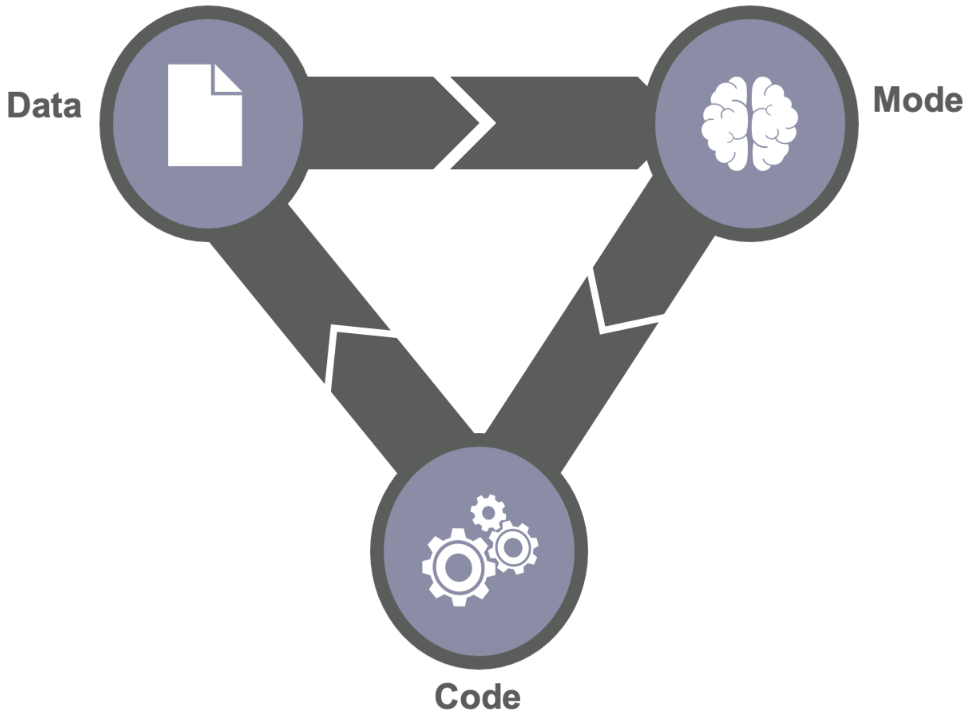
# AI AUDIO ENGINEERING: AN EMERGING DISCIPLINE

## A PARADIGM SHIFT

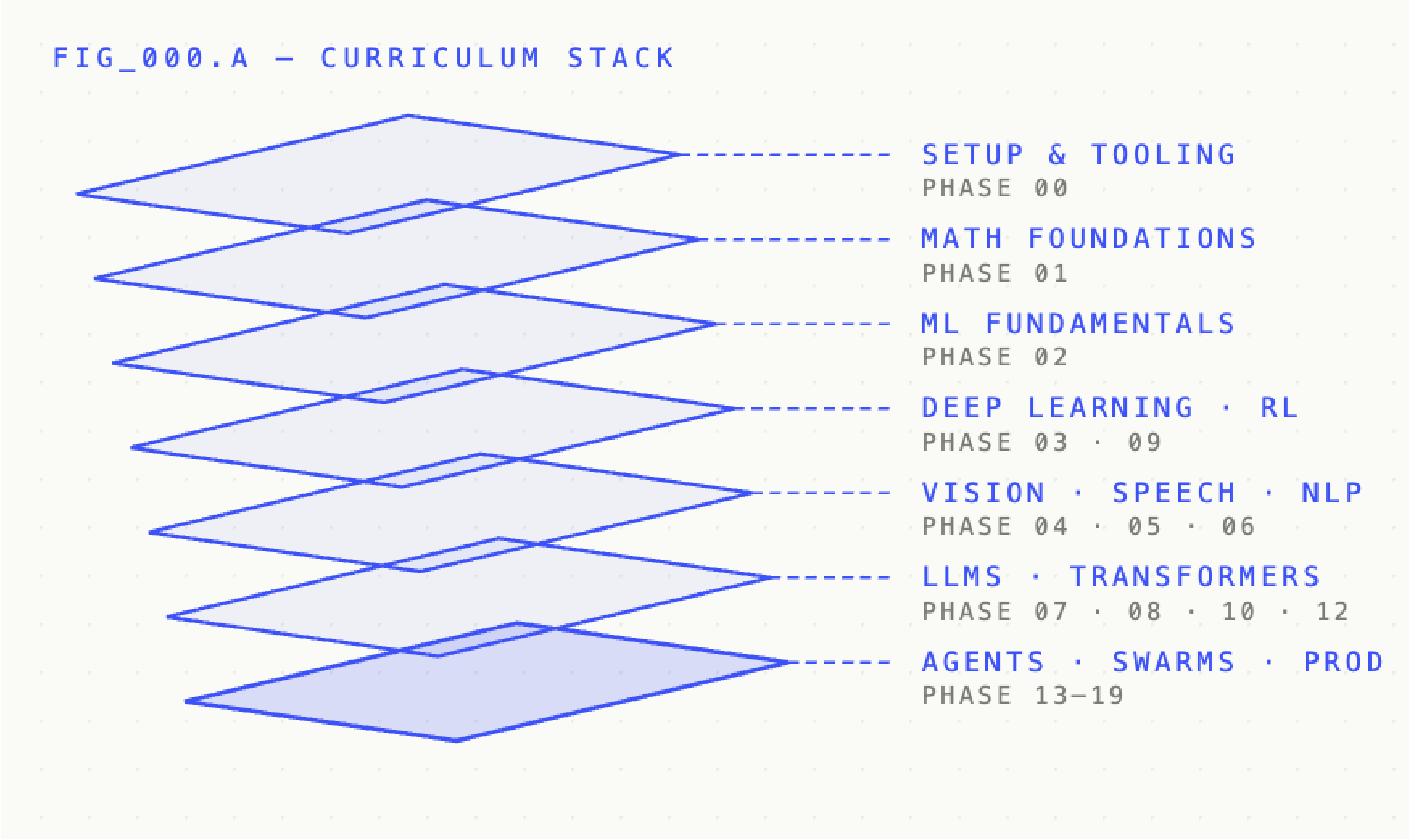
Not just new algorithms — new workflows, evaluation practices, and deployment contexts

## THE SHIFT IN PRACTICE

- Classical DSP: algorithms terminate at release, AI Audio Engineering assumes post-deployment evolution
- Perceptual feedback integration and dataset governance
- Realism and expressivity judged by performer experience








AES EUROPE 2026

https://github.com/rohitg00/ai-engineering-from-scratch

DEMONSTRATOR 1

# MEASUREMENT-DRIVEN STUDIES

From high-resolution instrument data through model constraints to Simulation Intelligence

# SIMULATION INTELLIGENCE — THE NINE MOTIFS

Building blocks of the SI “operating-system” stack — a roadmap merging scientific computing, simulation & AI (Lavin et al., 2021)

## MODULES

Application-layer methods that compose into scientific workflows

1. Multi-physics & multi-scale modeling
2. Surrogate modeling & emulation
3. Simulation-based inference
4. Causal modeling & inference
5. Agent-based modeling

## ENGINES

Inference & differentiation that power the modules above

6. Probabilistic programming
7. Differentiable programming

## FOUNDATION

Self-improving search & code synthesis

8. Open-ended optimization
9. Machine programming

Motifs are interconnected & interdependent — like layers of an operating system; coordinated, they accelerate scientific discovery in silico.

# DATA AS MODELING CONSTRAINT for SIMULATION INTELLIGENCE

## MEASUREMENT EXTRACTS

- Partial frequencies and decay rates
- Linear and nonlinear modal features
- Structural coupling and radiation patterns

## → CLASSICAL MODEL

Physical models parameterized by measurement-derived priors

- Reduced parameter search space
- Constrained model architecture

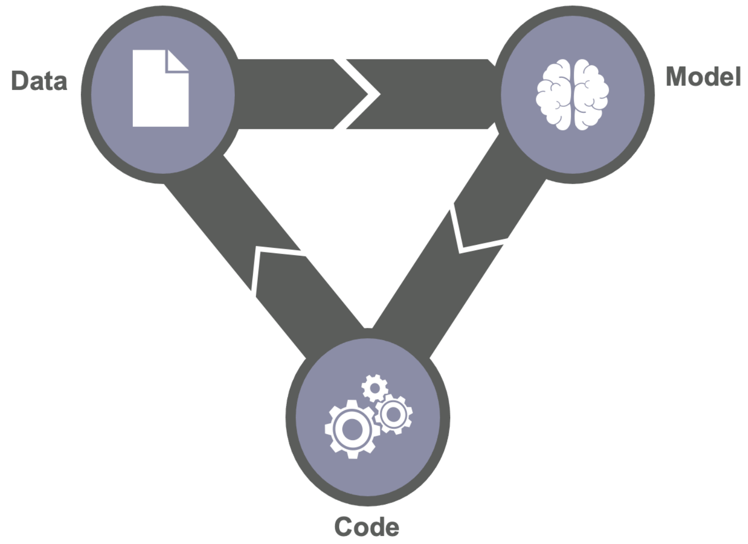
## THE HYBRID APPROACH

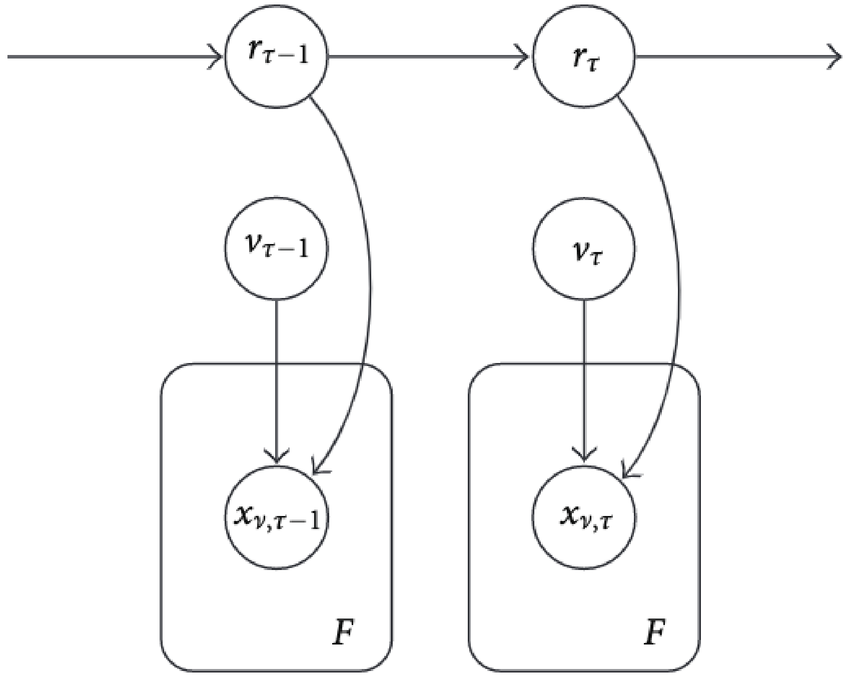
- Analytical model captures the physics
- Neural surrogate learns residuals
- Residuals = behavior not captured analytically

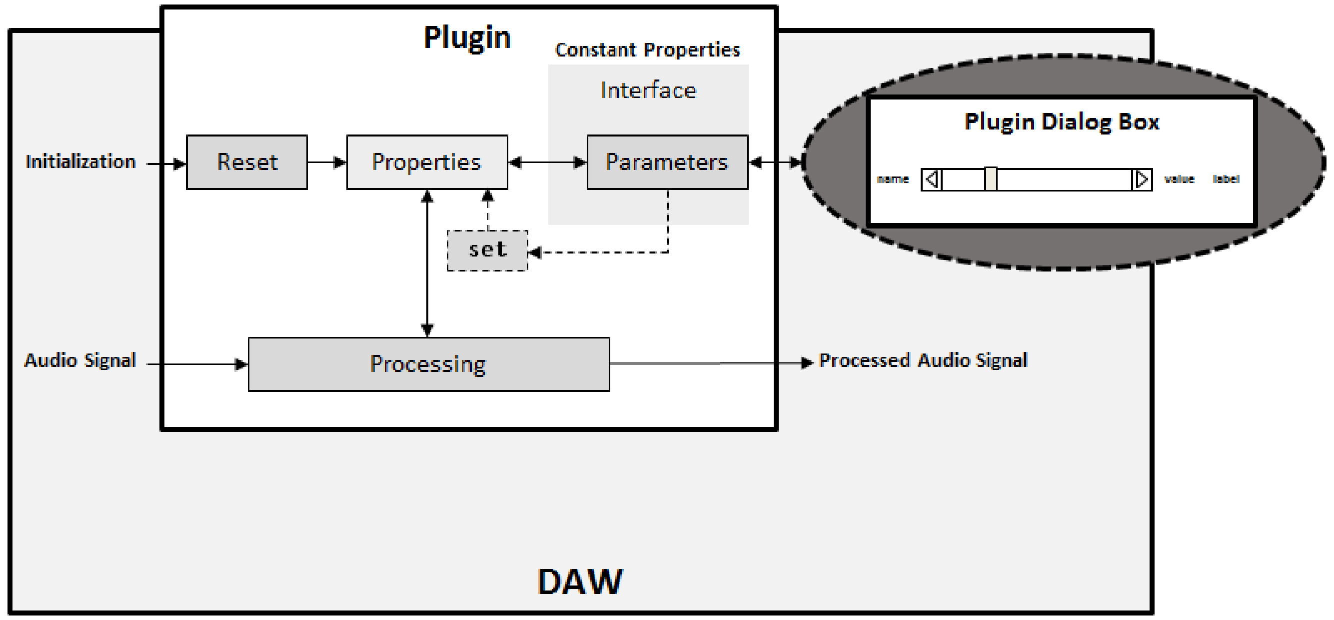
## KEY PRINCIPLE

**Data as a modeling constraint for simulation intelligence, not as an end-to-end solution**

- Measurement narrows the hypothesis space
- Models remain interpretable and physically grounded

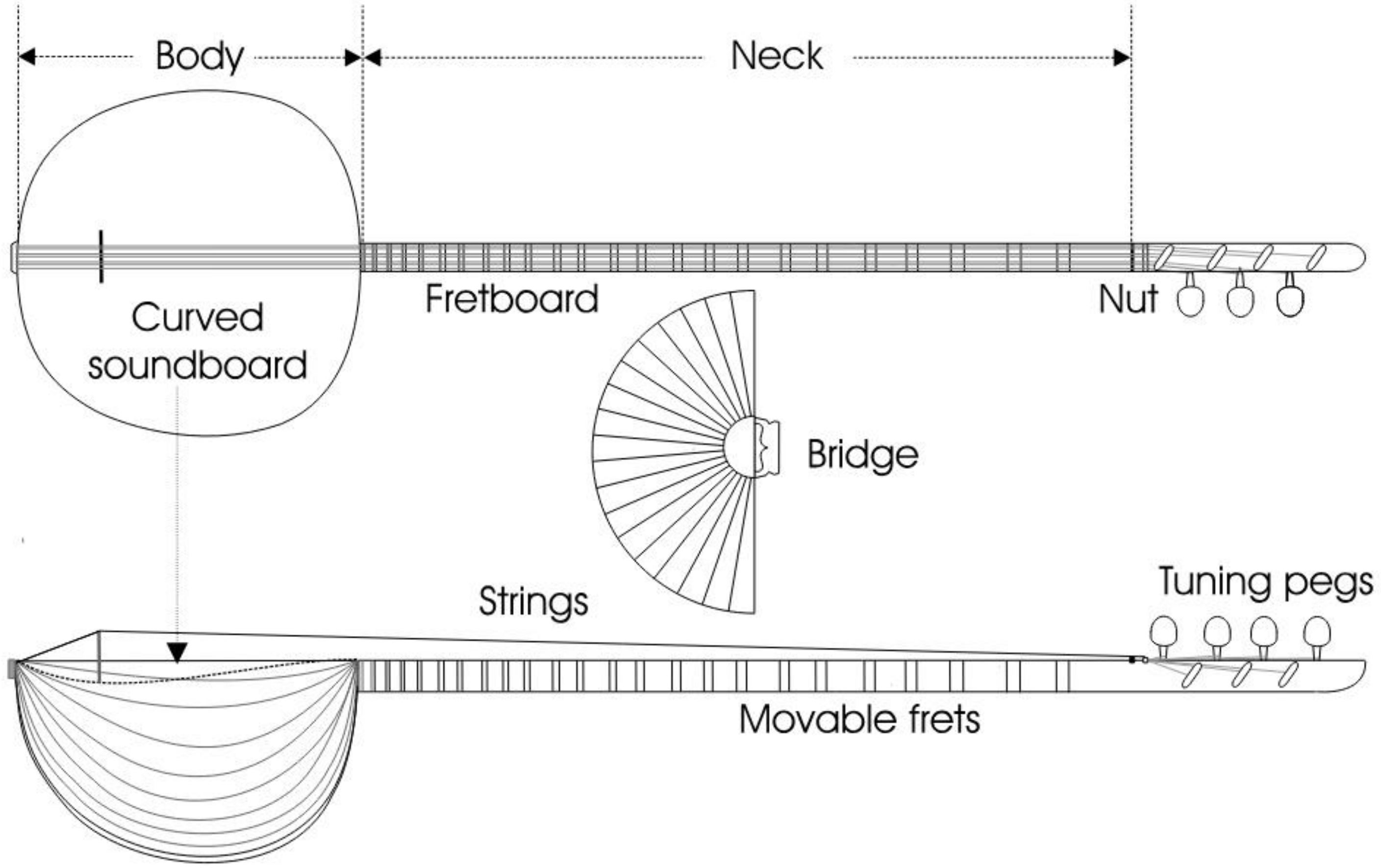

AES EUROPE 2026

Şimşekli, Umut, Antti Jylhä, Cumhur Erkut, and A T Cemgil. "Real-Time Recognition of Percussive Sounds by a Model-Based Method." *EURASIP Journal on Advances in Signal Processing* **2011** (1): 291860.

# CASE STUDY: HISTORICAL INSTRUMENT ACOUSTICS

## CASE STUDY: TURKISH TANBUR / FINNISH KANTELE

A historical plucked-string instrument with rich timbral complexity



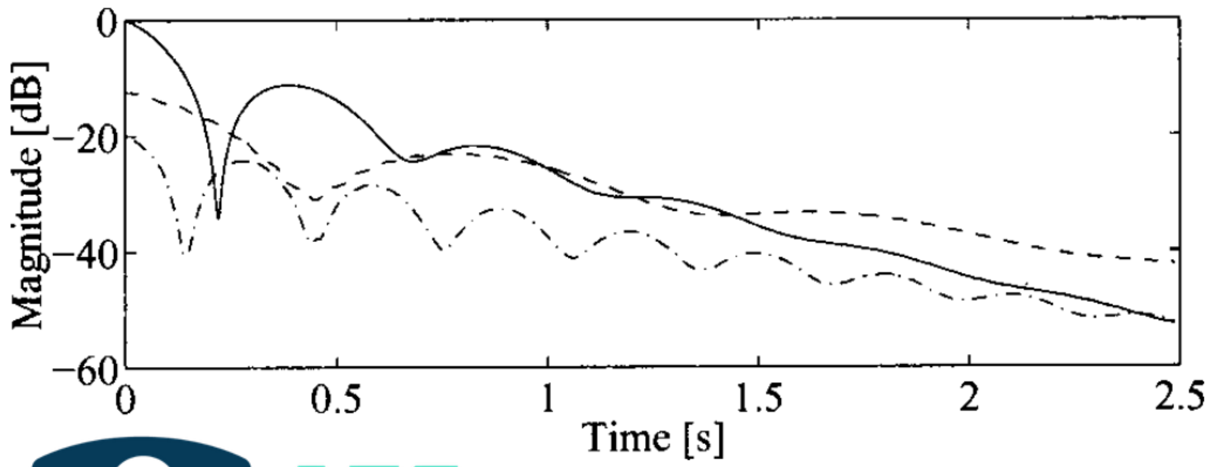

## METHODOLOGY

- High-resolution acoustical and vibration analysis
- Linear and nonlinear vibration features extracted from measurement data
- Calibrating nonlinear digital waveguide against measurement
- Neural surrogates + transfer learning for related instruments

## OUTCOME

Measurement-derived priors inform both classical modeling and neural surrogates

Robustness to measurement noise evaluated; remaining challenges identified



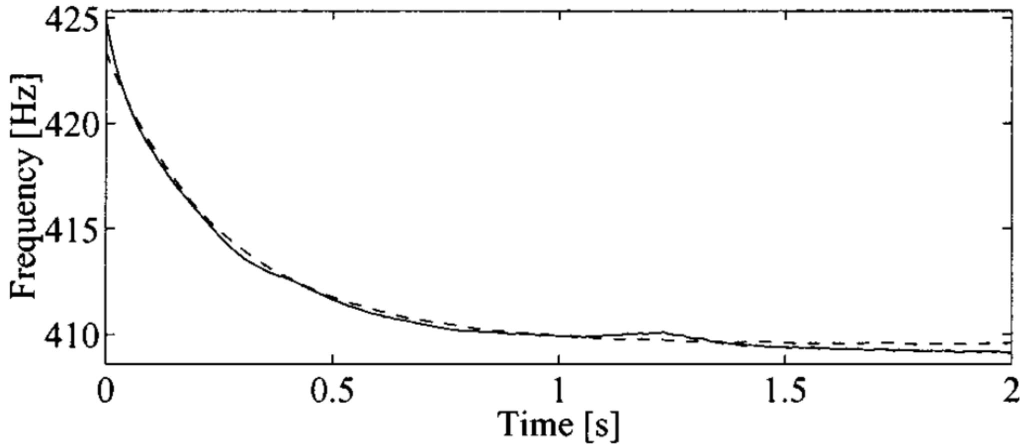
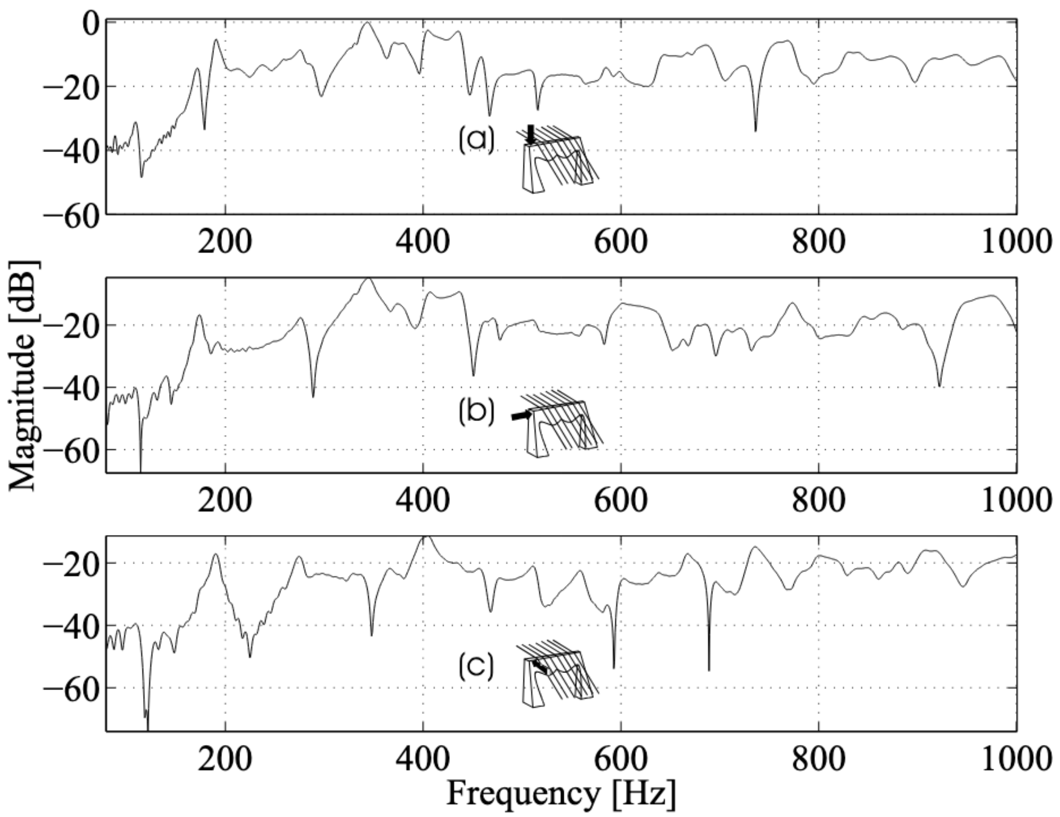




FIG. 15. Variation of the fundamental frequency in a kantele tone (obtained from the modified kantele) resulting from a fortissimo pluck (solid line) and a synthetic tone (dashed line).

Erkut *et al.*: Analysis and synthesis of the kantele 1689



DEMONSTRATOR 2

# PHYSICAL MODELING IN XR

Virtual instruments for experimentation, pedagogy, and performance

# VIRTUAL INSTRUMENTS IN EXTENDED REALITY

## THE XR INSTRUMENT SYSTEM

- Virtual strings, membranes, plates
- Driven by gesture or haptic interfaces
- Physically-informed or machine-augmented synthesis
- Binaural and room acoustic spatialization

## HAPTIC CONTROL

- Perceptual feedback loop with the player
- Spatial audio creates convincing sonic interactions
- Ecological validity approximating real playing

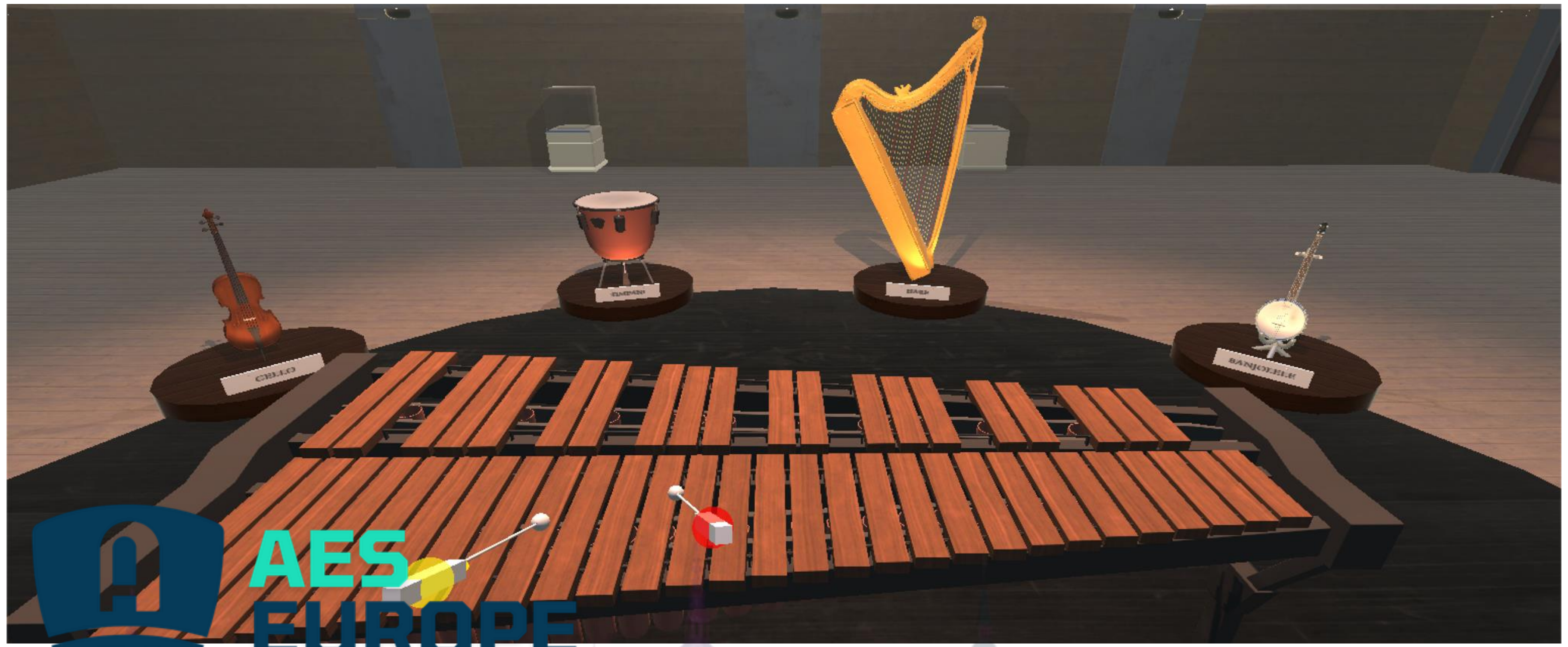
## XR ENABLES

- Immersive instrument experiences
- Novel pedagogical scenarios
- Performance and composition tools
- Rapid prototyping of instrument designs

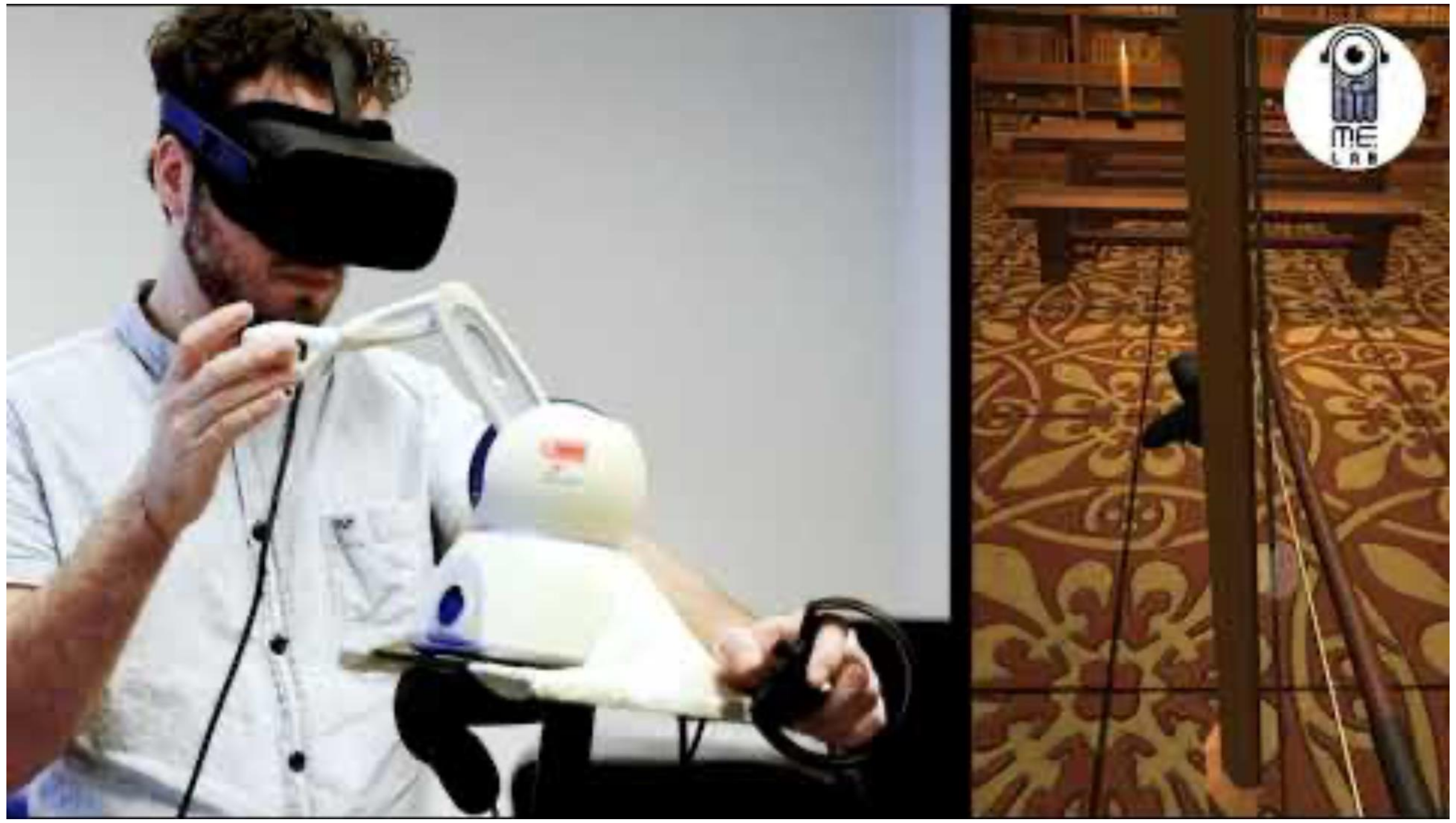
## REAL-TIME PIPELINE

Gesture detection → physical model → synthesis → spatialization

- Latency-critical: DSP handles time-sensitive stages
- Neural components: auxiliary inference tasks



https://github.com/SilvinWillemsen/SIVEtoolkit



# XR AS PERCEPTUAL EVALUATION PLATFORM

## XR AS A PERCEPTUAL DATA COLLECTION PLATFORM

Beyond ecological validity — XR instruments actively generate evaluation data

## CAPABILITIES

- In-context subjective testing (vs. isolated lab listening)
- Preference modeling and perceptual rating at scale
- Player-in-the-loop adaptation of synthesis parameters
- Performance capture: gesture, timing, expressive nuance

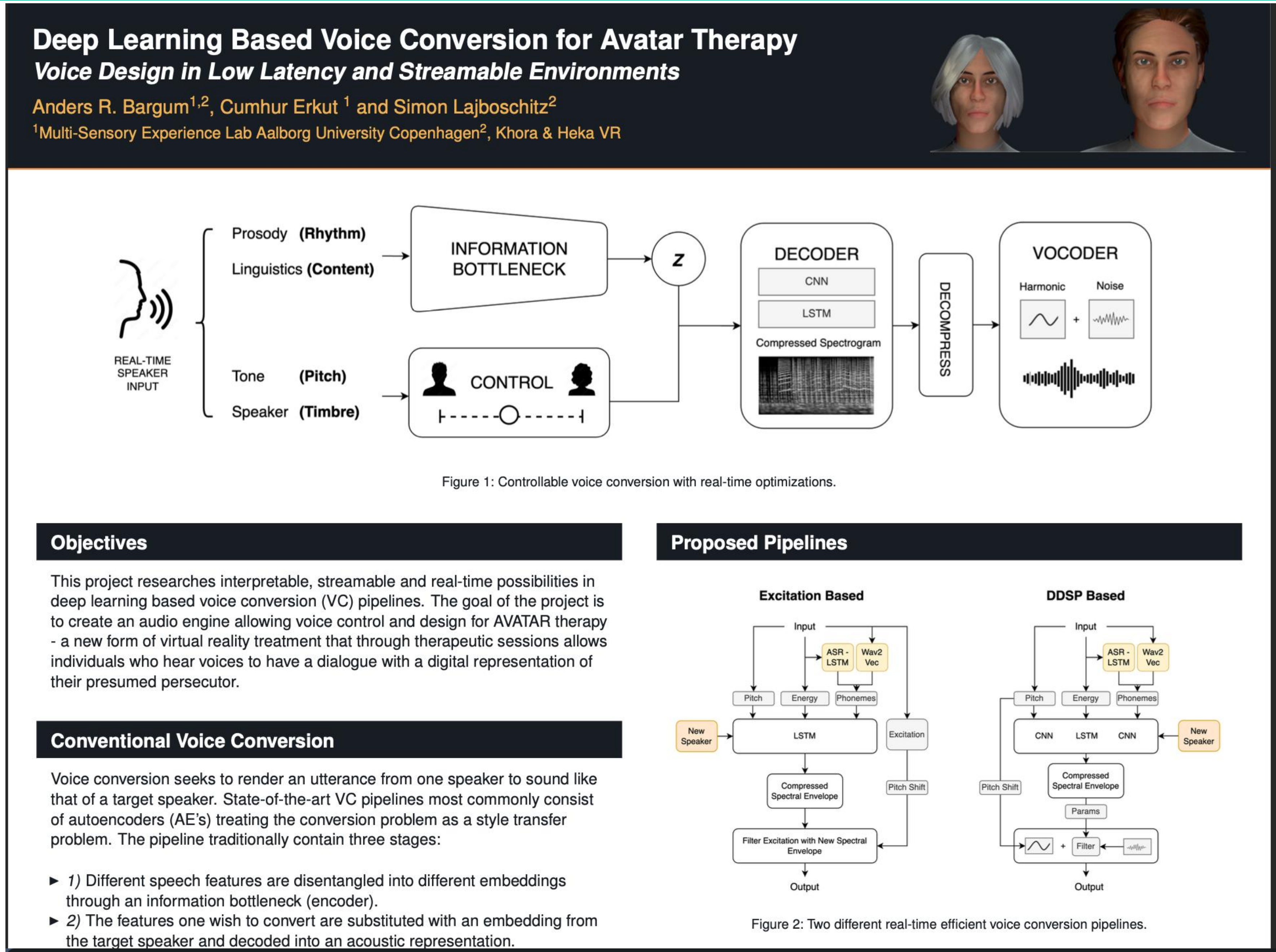
## WHY THIS MATTERS

These tasks are historically expensive or impractical to run at scale

XR lowers the barrier while increasing validity and sample diversity

Perceptual data feeds directly back into model refinement


AES EUROPE 2026



**Deep Learning Based Voice Conversion for Avatar Therapy**

***Voice Design in Low Latency and Streamable Environments***

Anders R. Bargum[1,2], Cumhur Erkut [1] and Simon Lajboschitz[2]

[1]Multi-Sensory Experience Lab Aalborg University Copenhagen[2], Khora & Heka VR



Figure 1: Controllable voice conversion with real-time optimizations.

**Objectives**

This project researches interpretable, streamable and real-time possibilities in deep learning based voice conversion (VC) pipelines. The goal of the project is to create an audio engine allowing voice control and design for AVATAR therapy - a new form of virtual reality treatment that through therapeutic sessions allows individuals who hear voices to have a dialogue with a digital representation of their presumed persecutor.

**Conventional Voice Conversion**

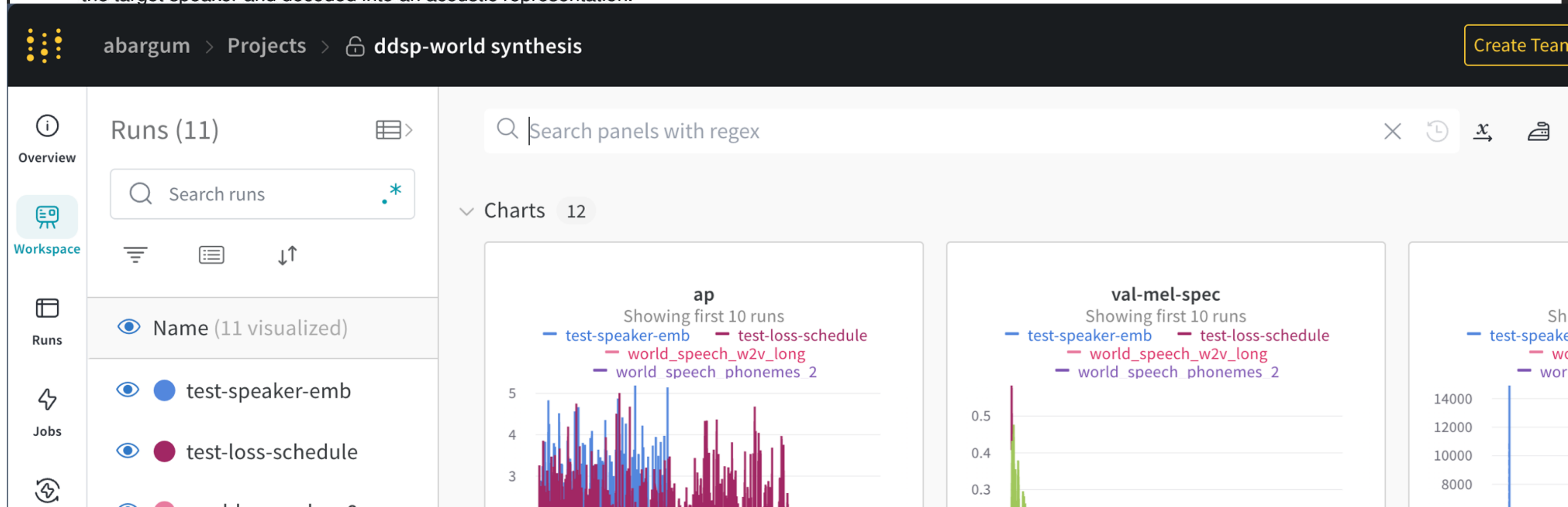
Voice conversion seeks to render an utterance from one speaker to sound like that of a target speaker. State-of-the-art VC pipelines most commonly consist of autoencoders (AE's) treating the conversion problem as a style transfer problem. The pipeline traditionally contain three stages:

- *1)* Different speech features are disentangled into different embeddings through an information bottleneck (encoder).
- *2)* The features one wish to convert are substituted with an embedding from the target speaker and decoded into an acoustic representation.

**Proposed Pipelines**



Figure 2: Two different real-time efficient voice conversion pipelines.



THE FRAMEWORK

# THE AI AUDIO ENGINEERING WORKFLOW

Measurement, modeling, learning, and perceptual evaluation as a continuous loop

# AI AUDIO ENGINEERING: A FORMAL DEFINITION

*AI Audio Engineering considers audio models to be*
**deployable, maintainable,**
**and continuously improvable artifacts**
*governed by data acquisition, inference, simulation intelligence,*
*perceptual evaluation, and lifecycle operations*

In contrast to classical DSP — where algorithms terminate at release — AI Audio Engineering assumes continuous evolution, feedback integration, and governance by performer experience and perceptual criteria.

# THE CONTINUOUS WORKFLOW LOOP

**MEASURE**
High-resolution audio + vibration data

→

**MODEL**
Classical + neural physical models

→

**EVALUATE**
XR perceptual data collection

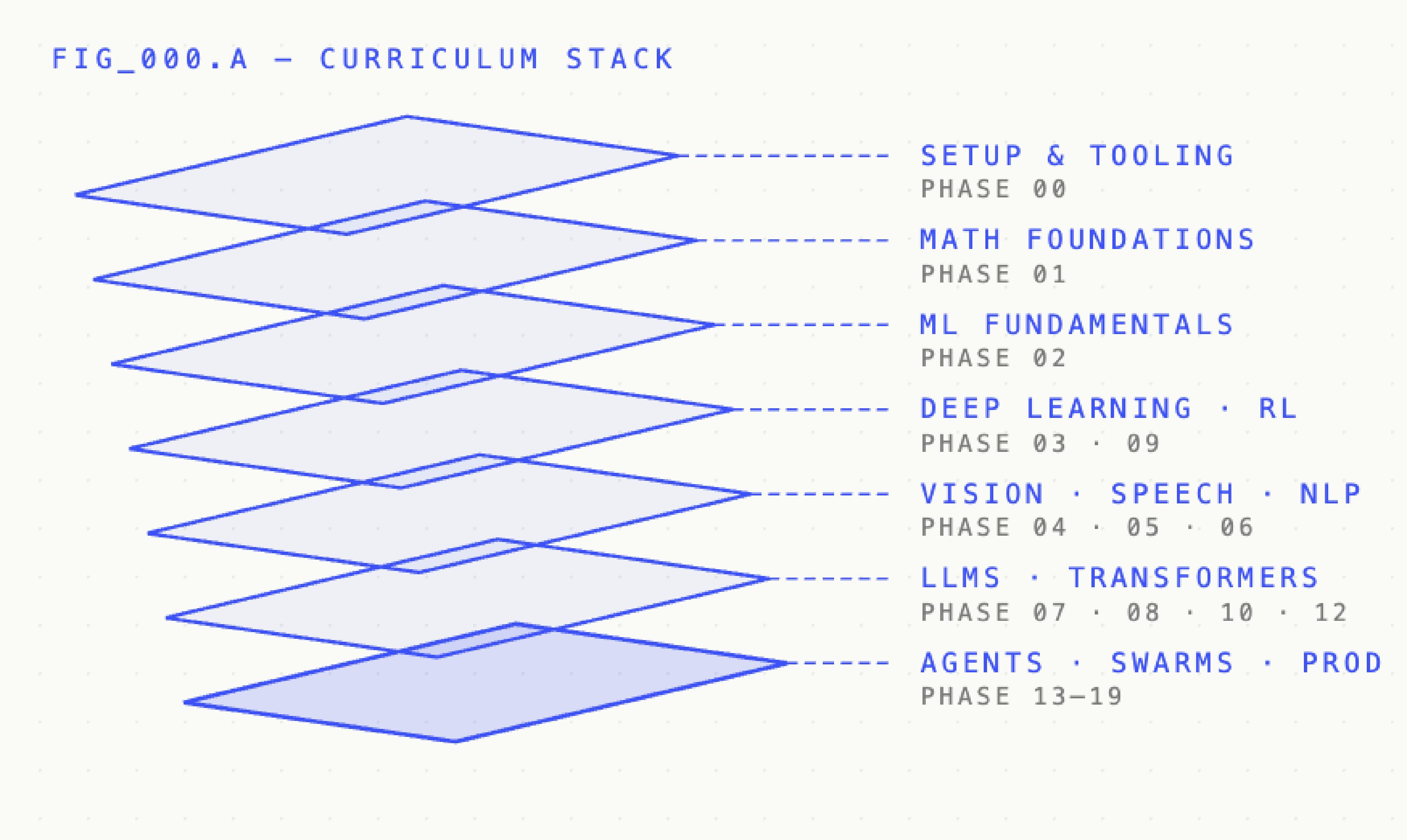
→

**LEARN**
Refine, adapt, improve

Measurement constrains model space • Models generate real-time audio • XR collects perceptual data • Learning refines models


Many tools available for all researchers / students in Denmark
See https://hpc.aau.dk

AES EUROPE 2026



# CONCLUSION

## AN OPEN INVITATION TO COLLABORATE

**This framework invites collaboration across:**

•Acoustics • Digital Signal Processing • Spatial Audio • AI & Machine Learning

## OPEN QUESTIONS FOR THE COMMUNITY

- How do we evaluate “realism” in AI Audio Engineering systems?
- What dataset governance standards apply to trained audio models?
- How should perceptual feedback be integrated into training pipelines?
- Can simulation intelligence extend to non-musical vibro-acoustic domains?

Thank you • Cumhur Erkut • Aalborg University Copenhagen • AES Europe 2026

https://moduler.aau.dk/course/2026-2027/TECHSPAI26

### FACTS ABOUT THE MODULE

| | |
|---|---|
| Danish title | Signalbehandling for AI-ingeniører |
| Module code | TECHSPAI26 |
| Module type | Course |
| Category | Separate module |
| Semester | November 2026. Copenhagen, Building A, 2.0.023<br><br>04.11 08:15 - 11:45<br>11.11 08:15 - 11:45<br>18.11 08:15 - 11:45<br>25.11 08:15 - 16:15 |
| ECTS | 5 |
| Language of instruction | English |
| Location of the lecture | Campus Copenhagen |
| Responsible for the module | **Markus Löchtefeld**, **Cumhur Erkut**, **Razvan Paisa**, **Nicola Walker** |

AES EUROPE 2026

https://danishsoundcluster.dk/ai-engineering-in-audio-signal-processing/

## AI Engineering In Audio Signal Processing

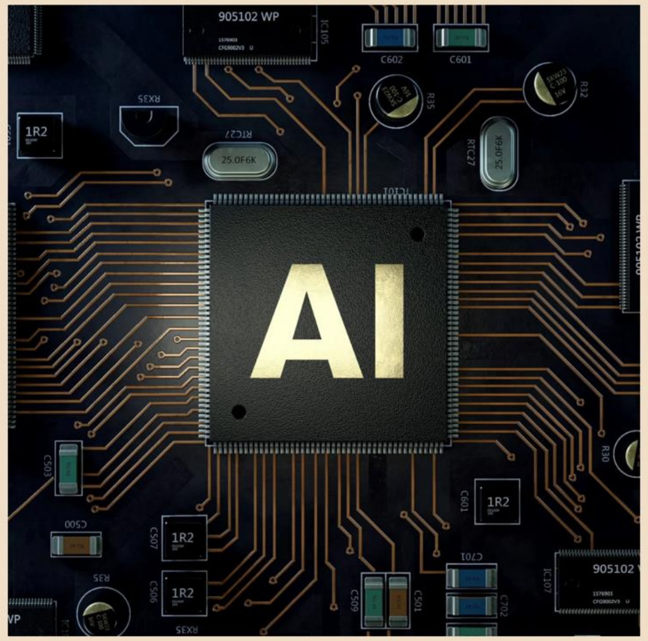
**SOUND DISCOVERY: A forum for sound exploration**

AI in audio moves rapidly from research prototypes to deployed real-world systems. 2026 is coined as the year of Voice AI Interaction, and Large Language Models (LLM) can now be inferred from small, embedded devices like mobile phones, hearing aids, and headphones.

However, the growing size of LLM knowledge caches are still too heavy for these platforms to support, restricting the full potential of Generative and Agentic AI to big GPU clusters.

What if we would transfer the best AI Engineering practices of LLMs to the domain of audio signal processing, without the LLMs themselves?

**Join us for a free-for-members tasting session that introduces participants to the core ingredients of modern AI audio engineering, starting with Data, Models, and Algorithmic Code, and expanding toward the full lifecycle management used in production environments.**



# ANTICIPATED OBJECTIONS

## OBJECTION 1
### “Where’s the empirical evidence?”

- This talk proposes an architectural framework — quantitative benchmarking is the natural next phase
- Tanbur/kantele demonstrator has published measurement-derived priors with validated nonlinear waveguide calibration (Şimşekli et al., EURASIP 2011)
- XR demonstrator (SIVE toolkit, open-source) has been used in peer-reviewed perceptual studies
- We invite collaborative benchmarking across the AES community

## OBJECTION 2
### “How do you handle real-time latency?”

- Strict architectural separation: DSP runs the synthesis loop at sub-5 ms latency: Pruning (Sudholdt et al, 2023)
- Neural components operate asynchronously on parameter inference — not in the audio callback
- SIVE toolkit demonstrates this split on commodity VR hardware (Meta Quest) with stable performance
- Neural surrogates handle offline tasks (transfer learning, residual modeling) — not sample-by-sample synthesis

## OBJECTION 3
### “Isn’t this just MLOps for audio?”

- Standard MLOps assumes stationary data and objective metrics — audio has neither
- Integrates simulation intelligence (9 motifs, Lavin et al.) with perceptual-in-the-loop evaluation — absent from generic MLOps
- Domain-specific: measurement-constrained model search, physics-informed priors, XR ecological evaluation
- Perceptual quality is subjective, context-dependent, and performer-specific — no parallel in standard ML pipelines