ANR JCJC **REAVISE** (2023-2026)

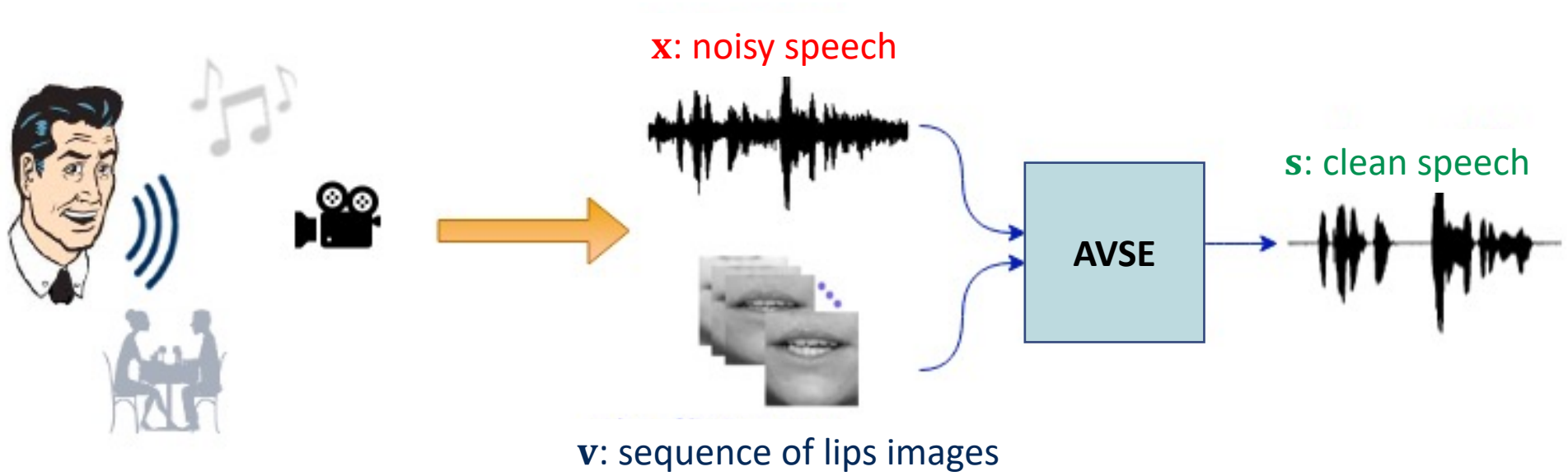
Robust and Efficient Deep Learning based Audio-visual Speech Enhancement

Mostafa Sadeghi

Multispeech Team

Inria Nancy – Grand Est

Audoi-visual Speech Enhancement (AVSE)



» AVSE: Incorporate *visual modality* (lips movements) to improve speech quality in noisy environments.

- Human-machine interaction (virtual assistant, social robots)
- Human-human interaction (listening comfort, hearing aids)

Two main AVSE categories:

Supervised (discriminative)	Unsupervised (generative)
Model $p(\mathbf{s}\|\mathbf{x},\mathbf{v})$ with a deep neural network.	Model $p(\mathbf{s}\|\mathbf{v})$, then combine it with $p(\mathbf{x}\|\mathbf{s},\mathbf{v})$
Model trained on noise ($\mathbf{x} = \mathbf{s} + \mathbf{n}, \mathbf{v}, \mathbf{s}$)	Model trained on only clean data $(\mathbf{s}, \mathbf{v})$
❌ Generalization issue	✅ Potentially better generalization
❌ Complex and big models	✅ Lightweight models
✅ Fast inference	❌ Costly, iterative inference

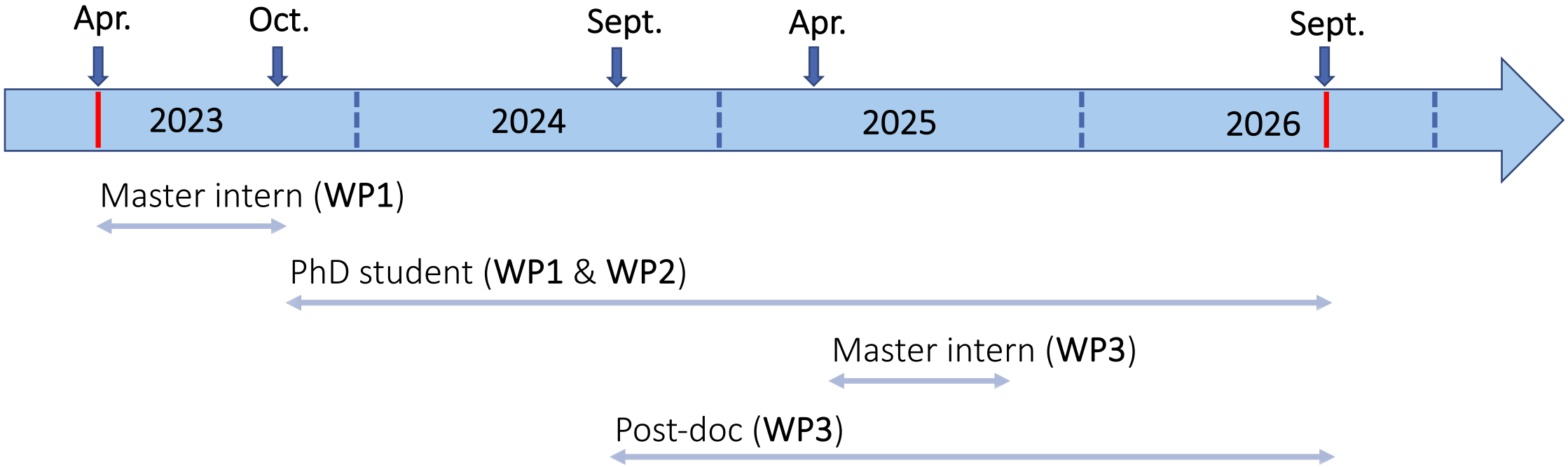
REAVISE aims to bridge the gap between the supervised and unsupervised AVSE approaches, benefiting from the best of both worlds

REAVISE objectives

1. Robust, efficient neural architectures for audio-visual fusion (**WP1**)
2. Data-efficient, generalizable models and frameworks for AVSE (**WP2**)
3. Fast and efficient inference algorithms with convergence guarantees (**WP3**)

REAVISE will achieve these objectives by leveraging recent methodological and theoretical breakthroughs in *deep neural networks*, *computer vision*, *statistical signal processing*, and *optimization*.

ANR JCJC **REAVISE** – Project organization



Collaborators

- **Romain Serizel** (University of Lorraine) – *Audio signal processing*
- **Xavi Alameda-Pineda** (Inria Grenoble) – *Computer vision & multimodal ML*
- **Timo Gerkmann** (University of Hamburg) – *Speech signal processing*
- **Franck Iutzeler** (University of Grenoble) – *Numerical optimization*

Thank you