Unsupervised Audio-visual Speech Enhancement

Mostafa Sadeghi

Multispeech team, Inria, France

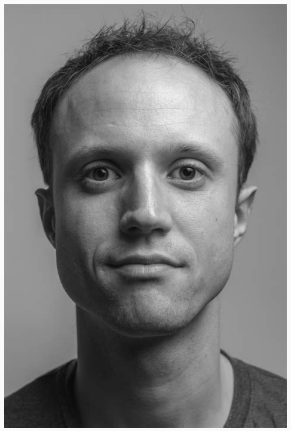
Capture Research Colloquium,

August 18-19, 2022, Redmond WA, USA.

Joint work with:

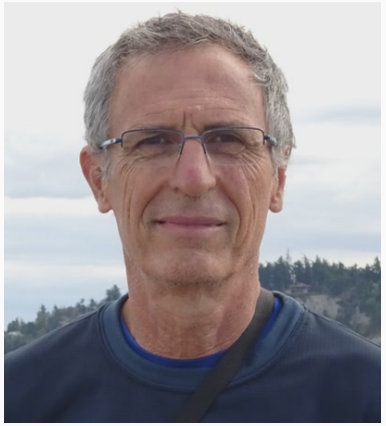
Xavier Alameda-Pineda

RobotLearn team, Inria



Simon Leglaive

IETR, CentraleSupélec



Radu Horaud

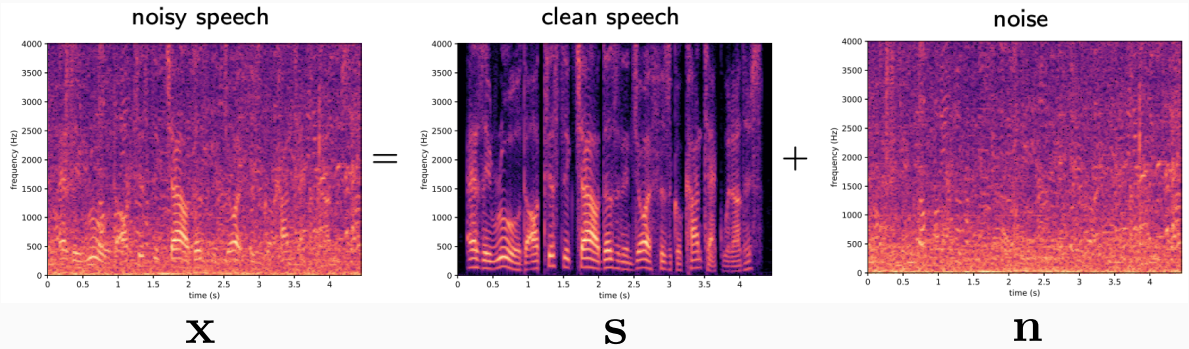
RobotLearn team, Inria

Laurent Girin

GIPSA-lab, CNRS

Speech Enhancement

Short-time Fourier transform (STFT) representation:



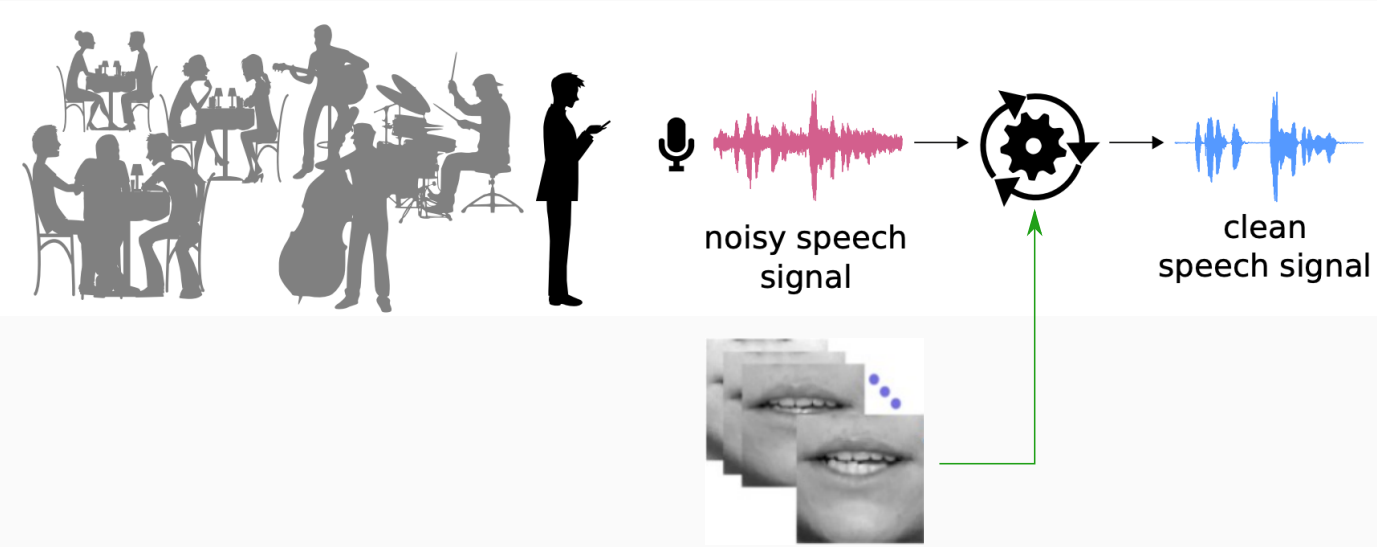
- $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$
- $\mathbf{x}_t = [x_{ft}]_{f=1}^{F} \in \mathbb{C}^F$ (similarly for $\mathbf{s}$ and $\mathbf{n}$).

Given noisy speech observation $\mathbf{x} = \mathbf{s} + \mathbf{n}$*, estimate the clean speech signal,* $\mathbf{s}$*.*

Audio-visual Speech Enhancement (AVSE)

Visual modality (lip movements):

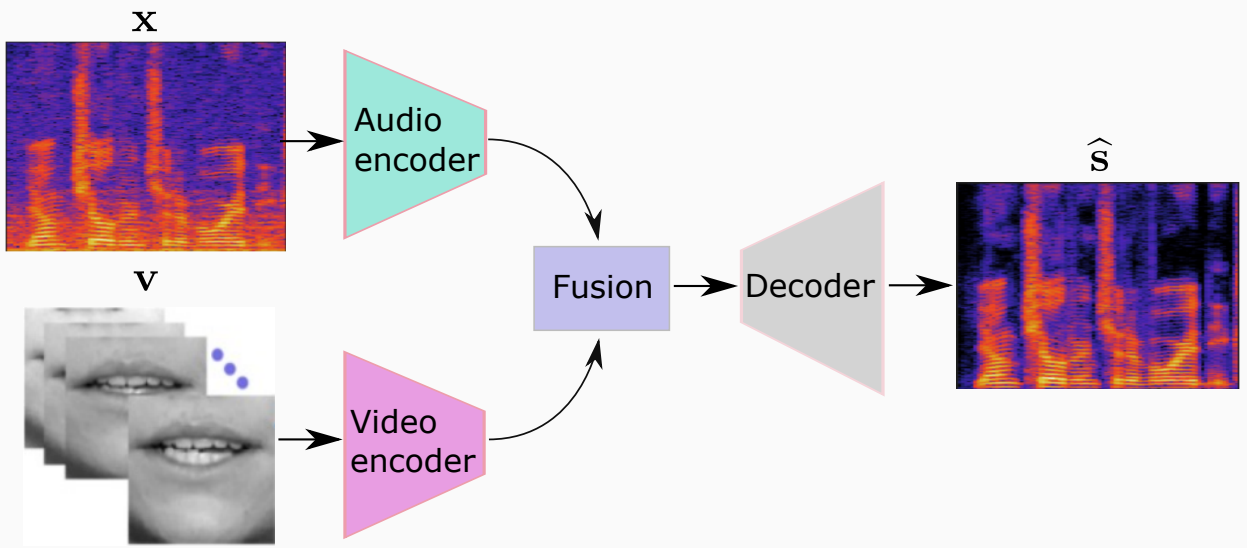
- Correlates well with speech signal (lip reading),
- Very helpful at highly noisy environments (unaffected by acoustic noise).



Given noisy speech observation $\mathbf{x} = \mathbf{s} + \mathbf{n}$ *& visual data* $\mathbf{v}$*, estimate the clean speech signal,* $\mathbf{s}$*.*

Supervised (discriminative) AVSE

Model $p_\Theta(\mathbf{s}|\mathbf{x},\mathbf{v})$, and learn Θ:



State-of-the-art performance, but ...

- Needs a huge audiovisual parallel (noise signal, clean speech) corpus
- Very deep and complex networks
- Lack of a systematic robust framework for noisy visual data

Unsupervised (generative) AVSE

Speech enhancement without training on noise.

Model $p_\Theta(\mathbf{s}|\mathbf{x},\mathbf{v}) \propto \underbrace{p_\psi(\mathbf{x}|\mathbf{s},\mathbf{v})}_{\text{Inference}} \cdot \underbrace{p_\theta(\mathbf{s}|\mathbf{v})}_{\text{Training}}$, and learn $\Theta = \theta \cup \psi$:

- Training - Learn speech's prior distribution $p_\theta(\mathbf{s}|\mathbf{v})$
- Inference - Model $p_\psi(\mathbf{x}|\mathbf{s},\mathbf{v})$, and infer $\mathbf{s}$ using $p_\theta(\mathbf{s}|\mathbf{v})$

▷ Advantages over supervised approaches:

- No need to huge parallel corpora $\rightarrow$ compact & lightweight models
- Potentially better generalization performance
- Flexibility to design robust frameworks

However, this approach is very recent, and significantly less explored.

Speech generative modeling

How to learn speech's prior distribution?

- Latent variable generative models: Variational autoencoder (VAE),[1] Normalizing Flow (NF),[2] etc.
- Score-based generative models:[3] Learn the score $\nabla_{\mathbf{s}} \log p_\theta(\mathbf{s}|\mathbf{v})$

We focus on VAE:

$$p_\theta(\mathbf{s}|\mathbf{v}) = \int p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v}) p_\theta(\mathbf{z}|\mathbf{v}) d\mathbf{z}$$

- $\mathbf{z} = \{\mathbf{z}_t\}$: (real-valued, low-dimensional) latent variables

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," ICLR, 2014.

[2] D. Rezende, D. Mohamed, "Variational inference with normalizing flows," ICML, 2015.

[3] Y. Song, S. Ermon, "Generative modeling by estimating gradients of the data distribution," NeurIPS, 2019

VAE-based speech modeling

A Gaussian generative model:[4][5]

$$\begin{cases} p_\theta(\mathbf{s}_t|\mathbf{z}_t, \mathbf{v}_t) = \mathcal{N}_c\Big(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}_\theta^{av}(\mathbf{z}_t, \mathbf{v}_t))\Big), \\ p_\theta(\mathbf{z}_t|\mathbf{v}_t) = \mathcal{N}\Big(\boldsymbol{\mu}_\theta^{p}(\mathbf{v}_t), \text{diag}(\boldsymbol{\sigma}_\theta^{p}(\mathbf{v}_t))\Big) \end{cases}$$

▷ $\boldsymbol{\sigma}_\theta^{av}(.,.), \boldsymbol{\mu}_\theta^{p}(.), \boldsymbol{\sigma}_\theta^{p}(.)$ are neural networks parameterized by θ

Given a training set $\{(\mathbf{s}_t, \mathbf{v}_t)\}_{t=1}^{T_{tr}} \longrightarrow$ *learn the generative model parameters, i.e.* θ*, using the maximum likelihood principle.*

[4] M. Sadeghi *et al.*, "Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1788 –1800, June 2020.

[5] S. Leglaive *et al.* "A variance modeling framework based on variational autoencoders for speech enhancement," MLSP, 2018.

VAE-based parameter learning

Need to maximize $\log p_\theta(\mathbf{s}|\mathbf{v})$, which is intractable. However:

$$\log p_\theta(\mathbf{s}|\mathbf{v}) = \log \int p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v}) p_\theta(\mathbf{z}|\mathbf{v}) d\mathbf{z}$$
$$\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s},\mathbf{v})}\left[\log \frac{p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v}) p_\theta(\mathbf{z}|\mathbf{v})}{q_\phi(\mathbf{z}|\mathbf{s},\mathbf{v})}\right] \triangleq \mathcal{L}(\theta,\phi)$$

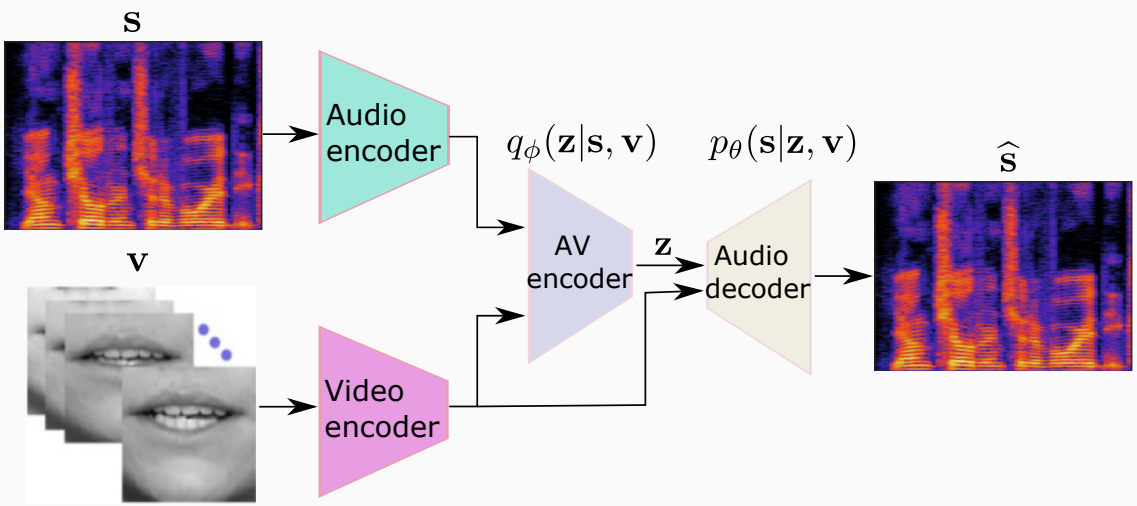
Using variational inference, maximize the lower bound of $\log p_\theta(\mathbf{s}|\mathbf{v})$:

$$\mathcal{L}(\theta,\phi) = \alpha \cdot \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s},\mathbf{v})}\Big[\log p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v})\Big] + (1-\alpha)\cdot \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{v})}\Big[\log p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{s},\mathbf{v}) \parallel p_\theta(\mathbf{z}|\mathbf{v})\Big)$$

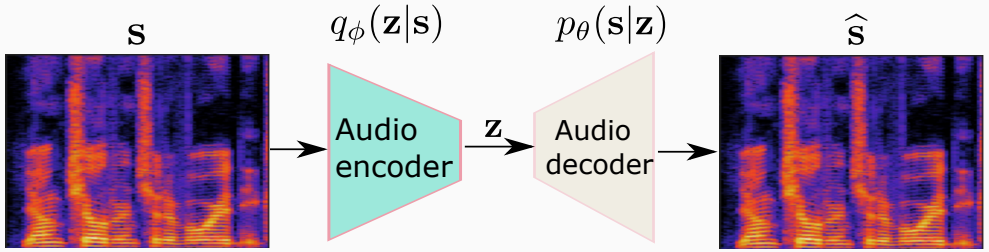
- $q_\phi(\mathbf{z}|\mathbf{s},\mathbf{v}) \approx p_\theta(\mathbf{z}|\mathbf{s},\mathbf{v})$: "encoder" with parameters ϕ.
- $D_{\mathrm{KL}}(. \parallel .)$ is the Kullback–Leibler divergence.
- $0 \leq \alpha \leq 1$ gives some reconstruction power to the prior network.

VAE architectures

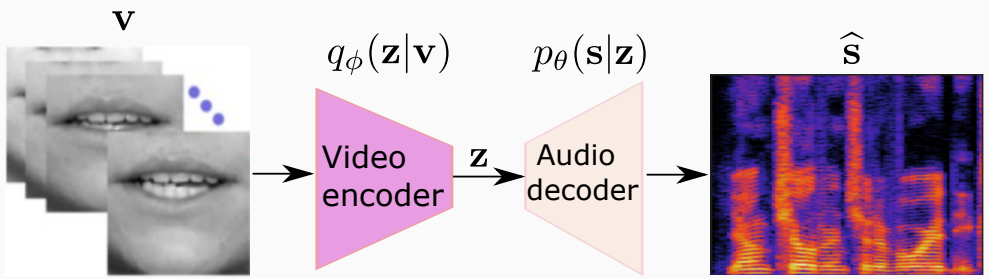
AV-VAE



A-VAE



V-VAE



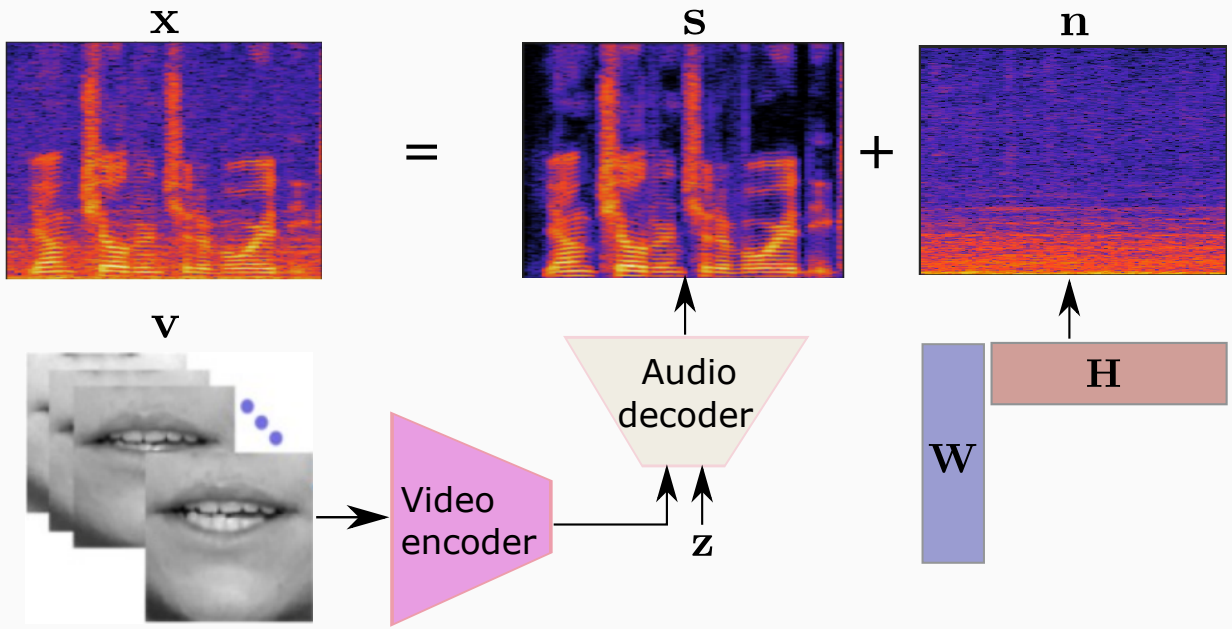
Speech Enhancement

Observation model: $\forall t: \quad \boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{n}_t$

Noise model: Non-negative Matrix Factorization (NMF)

$$\forall t: \quad \boldsymbol{n}_t \sim \mathcal{N}_c\left(\mathbf{0}, \mathrm{diag}(\mathbf{WH}[:,t])\right), \quad \mathbf{W}, \mathbf{H} \geq 0$$

Clean speech model: Trained generative (decoder) network.



Speech Enhancement

Inference:

- ▷ Parameters to be estimated: $\psi = \{\mathbf{W}, \mathbf{H}\}$
- ▷ Observed variables: $\{(\mathbf{x}_t, \mathbf{v}_t)\}_{t=1}^{T}$
- ▷ Latent variables: $\mathbf{z} = \left\{\mathbf{z}_t\right\}_{t=1}^{T}$

Parameter estimation:

$$\psi^* = \underset{\psi}{\operatorname{argmax}} \ \log p_\psi(\boldsymbol{x}|\mathbf{v}) = \underset{\psi}{\operatorname{argmax}} \int \log p_\psi(\boldsymbol{x}, \mathbf{z}|\mathbf{v}) d\mathbf{z}$$

Parameter Estimation

Monte Carlo Expectation Maximization (MCEM)

From an initialization $\psi^{(0)}$ of the parameters, iterate:

- **E-Step**: $Q(\psi|\psi^{(k)}) = \mathbb{E}_{p_{\psi^{(k)}}(\mathbf{z}|\mathbf{x},\mathbf{v})}[\log p_\psi(\mathbf{x},\mathbf{z},\mathbf{v})]$.

  Intractable expectation $\rightarrow$ Markov chain Monte Carlo method.

$$Q(\psi|\psi^{(k)}) \approx \frac{1}{R}\sum_{r=1}^{R} \log p_\psi(\mathbf{x},\mathbf{z}^{(r)},\mathbf{v})$$

  $\{\mathbf{z}^{(r)}\}_{r=1}^{R} \sim p(\mathbf{z}|\mathbf{x},\mathbf{v};\boldsymbol{\theta}_u^\star)$ using the Metropolis-Hastings method.

- **M-Step**: $\psi^{(k+1)} \leftarrow \operatorname{argmax}_\psi Q(\psi|\psi^{(k)})$.

  Standard multiplicative update rules.

Speech Estimation

Once the parameters are estimated, the speech STFT frames are estimated via a Wiener-like filtering ($\forall f, t$):

$$\begin{aligned}\hat{s}_{ft} &= \mathbb{E}_{p_{\psi^*}(s_{ft}|x_{ft},\mathbf{v}_t)}[s_{ft}] \\ &= \mathbb{E}_{p_{\psi^*}(\mathbf{z}_t|\mathbf{x}_t,\mathbf{v}_t)}\left[\frac{\sigma^{av}_{\theta,f}(\mathbf{z}_t,\mathbf{v}_t)}{\sigma^{av}_{\theta,f}(\mathbf{z}_t,\mathbf{v}_t)+(\mathbf{W}^*\mathbf{H}^*)_{f,t}}\right]\cdot x_{ft}.\end{aligned}$$

where ψ^* denotes the set of estimated parameters by the MCEM method.

- ▷ The intractable expectation is approximated by a Monte Carlo average.

Experiments

- ▷ NTCD-TIMIT dataset[6]
  - Audio-visual recordings in controlled conditions
  - Clean audio as well as noisy versions
  - Frontal video frames with 30 FPS- 67×67 lips images
  - ▷ Training set ($\sim$ 5 hours): 39 speakers $\times$ 98 sentences $\times$ 5 seconds
  - ▷ Test set ($\sim$ 1 hour): 9 speakers $\times$ 98 sentences $\times$ 5 seconds
  - ▷ Noise levels: -15 dB, -10 dB, -5 dB, 0 dB, 5 dB and 15 dB
  - ▷ Noise types: *Living Room (LR)*, *White*, *Cafe*, *Car*, *Babble*, and *Street*
- ▷ Baseline: Supervised[7]

[6] A.-H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," INTERSPEECH, 2017.

[7] A. Gabbay *et al.*, "Visual speech enhancement," INTERSPEECH, 2018.

Experiments

Networks architectures:

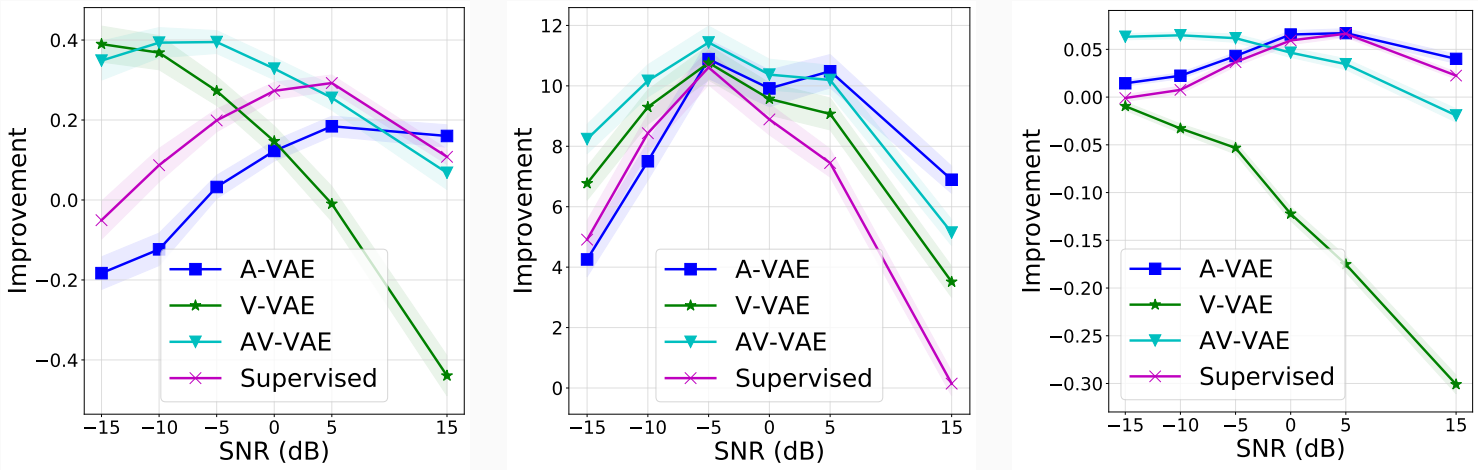
1. A-VAE:
   - Decoder: Single hidden layer, 128 nodes, hyperbolic tangent activations. Input dimension: 32 (latent space).
   - Encoder: Single hidden layer, 128 nodes, hyperbolic tangent activations. Input dimension: 513 (spectrogram time frame).
2. V-VAE:
   - Decoder: Same as A-VAE.
   - Encoder: ResNet-18 pre-trained model.
3. AV-VAE: Shares the same architecture as that of AV-VAE with visual embeddings being concatenated with the encoder's and decoder's input.

Results

Objective measures (the higher, the better)

- Signal-to-distortion ratio (SDR).
- Perceptual evaluation of speech quality (PESQ).
- Short-time objective intelligibility (STOI).

Improvement with respect to the input:



(a) PESQ **(b)** SDR **(c)** STOI

Audio Examples: `https://team.inria.fr/perception/research/av-vae-se/`

Conclusions

Unsupervised approaches have great potential and advantages for robust, generalizable, and interpretable AVSE. However, they come with some challenges, e.g., complex (iterative) inference.

Bridging the gap between the unsupervised and supervised approaches to benefit from the best of both worlds is an interesting future direction.

Thank you for your attention!