Audio-visual Speech Enhancement based on Deep Generative Models

Mostafa Sadeghi

MULTISPEECH team
Inria Nancy - Grand Est

January 2023

informatics mathematics
Inria

UNIVERSITÉ DE LORRAINE

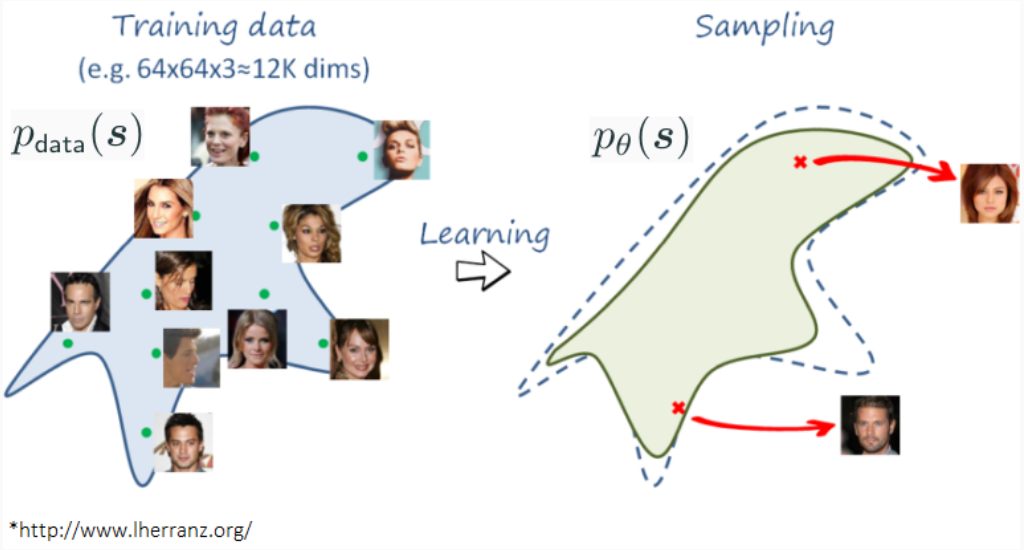
Loria

1. Deep Latent Variable Generative Models

2. Audio-visual Speech Enhancement

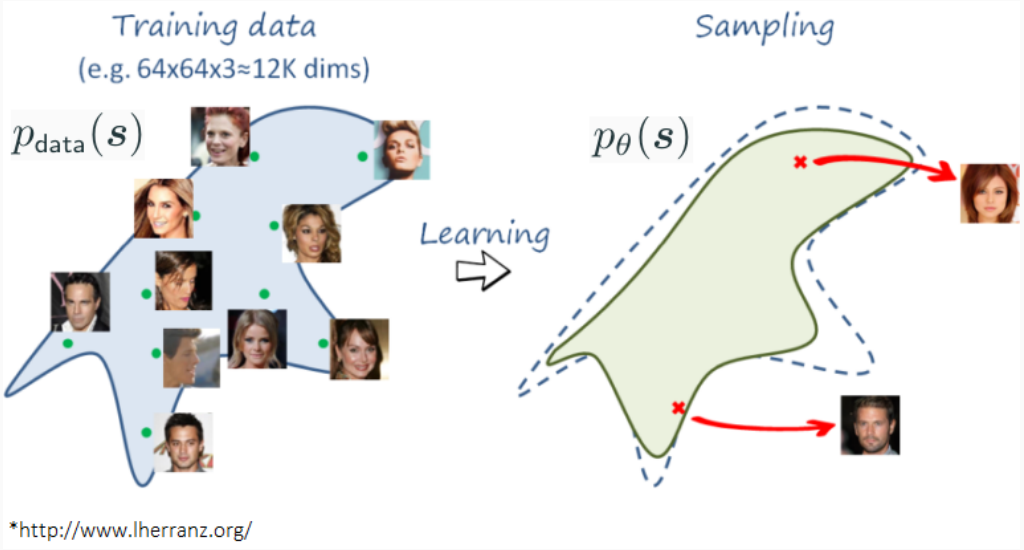
Deep Latent Variable Generative Models

Generative Models



Objective: Learning/simulating a complicated probability distribution of data, p_{data}, given some training samples: $\boldsymbol{s}_i \sim p_{\text{data}}(\boldsymbol{s}), \quad i = 1, \ldots, N.$

Generative Models



Objective: Learning/simulating a complicated probability distribution of data, p_{data}, given some training samples: $\boldsymbol{s}_i \sim p_{\text{data}}(\boldsymbol{s}), \quad i = 1, \ldots, N.$

Learn a parametric distribution $p_\theta(\boldsymbol{s})$ as close as possible to $p_{\text{data}}(\boldsymbol{s})$:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; D_{\text{KL}}\Big(p_{\text{data}}(\boldsymbol{s}) \parallel p_\theta(\boldsymbol{s})\Big)$$

$$= \underset{\theta}{\operatorname{argmax}} \; \mathbb{E}_{p_{\text{data}}}\Big[\log p_\theta(\boldsymbol{s})\Big] \approx \boxed{\underset{\theta}{\operatorname{argmax}} \; \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(\boldsymbol{s}_i)}$$

Latent Variable Generative Models

- $\boldsymbol{s} \in \mathbb{R}^n$: observed variable
- $\boldsymbol{z} \in \mathbb{R}^\ell$: latent variable, a concise representation of $\boldsymbol{s}$ ($\ell \ll n$)

$$\begin{cases} \boldsymbol{z} \sim p_\theta(\boldsymbol{z}) \\ \boldsymbol{s}|\boldsymbol{z} \sim p_\theta(\boldsymbol{s}|\boldsymbol{z}) \end{cases} \rightarrow \; p_\theta(\boldsymbol{s}) = \int p_\theta(\boldsymbol{s}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int p_\theta(\boldsymbol{s}|\boldsymbol{z}) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

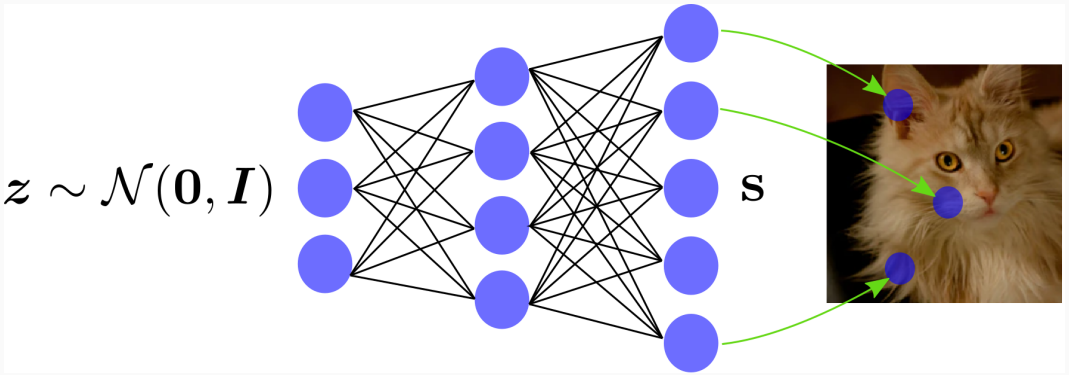
Latent Variable Generative Models

- $\boldsymbol{s} \in \mathbb{R}^n$: observed variable
- $\boldsymbol{z} \in \mathbb{R}^\ell$: latent variable, a concise representation of $\boldsymbol{s}$ $(\ell \ll n)$

$$\begin{cases} \boldsymbol{z} \sim p_\theta(\boldsymbol{z}) \\ \boldsymbol{s}|\boldsymbol{z} \sim p_\theta(\boldsymbol{s}|\boldsymbol{z}) \end{cases} \rightarrow \; p_\theta(\boldsymbol{s}) = \int p_\theta(\boldsymbol{s}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int p_\theta(\boldsymbol{s}|\boldsymbol{z}) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

Generating new samples:

Draw $\boldsymbol{z}_k \sim p_\theta(\boldsymbol{z})$, then draw a new sample $\boldsymbol{s}_k \sim p_\theta(\boldsymbol{s}|\boldsymbol{z}_k)$



Parameter Estimation: Variational Autoencoder

Variational Autoencoder (VAE)

Recall the generative model: $\begin{cases} p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\ p_\theta(\boldsymbol{s}|\boldsymbol{z}) = \mathcal{N}\Big(\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\Big) \end{cases}$

VAE approximates $p_\theta(\boldsymbol{z}|\boldsymbol{s})$ with a *parametric* Gaussian distribution:

$$q_\psi(\boldsymbol{z}|\boldsymbol{s}) = \mathcal{N}\Big(\boldsymbol{\mu}_\psi(\boldsymbol{s}), \boldsymbol{\Sigma}_\psi(\boldsymbol{s})\Big)$$

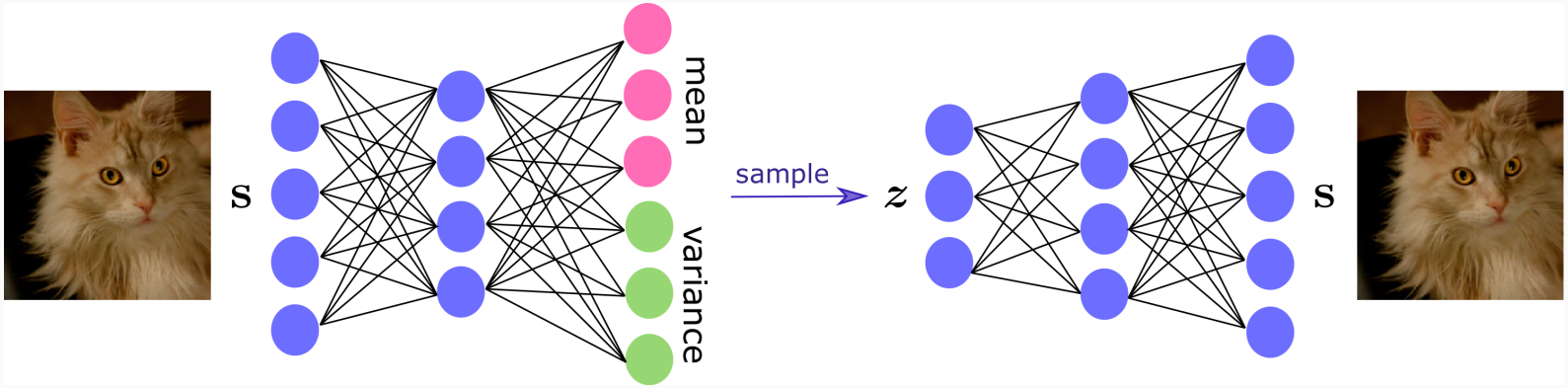
Parameter Estimation: Variational Autoencoder

Variational Autoencoder (VAE)

Recall the generative model: $\begin{cases} p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ p_\theta(\boldsymbol{s}|\boldsymbol{z}) = \mathcal{N}\Big(\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\Big) \end{cases}$

VAE approximates $p_\theta(\boldsymbol{z}|\boldsymbol{s})$ with a *parametric* Gaussian distribution:

$$q_\psi(\boldsymbol{z}|\boldsymbol{s}) = \mathcal{N}\Big(\boldsymbol{\mu}_\psi(\boldsymbol{s}), \boldsymbol{\Sigma}_\psi(\boldsymbol{s})\Big)$$

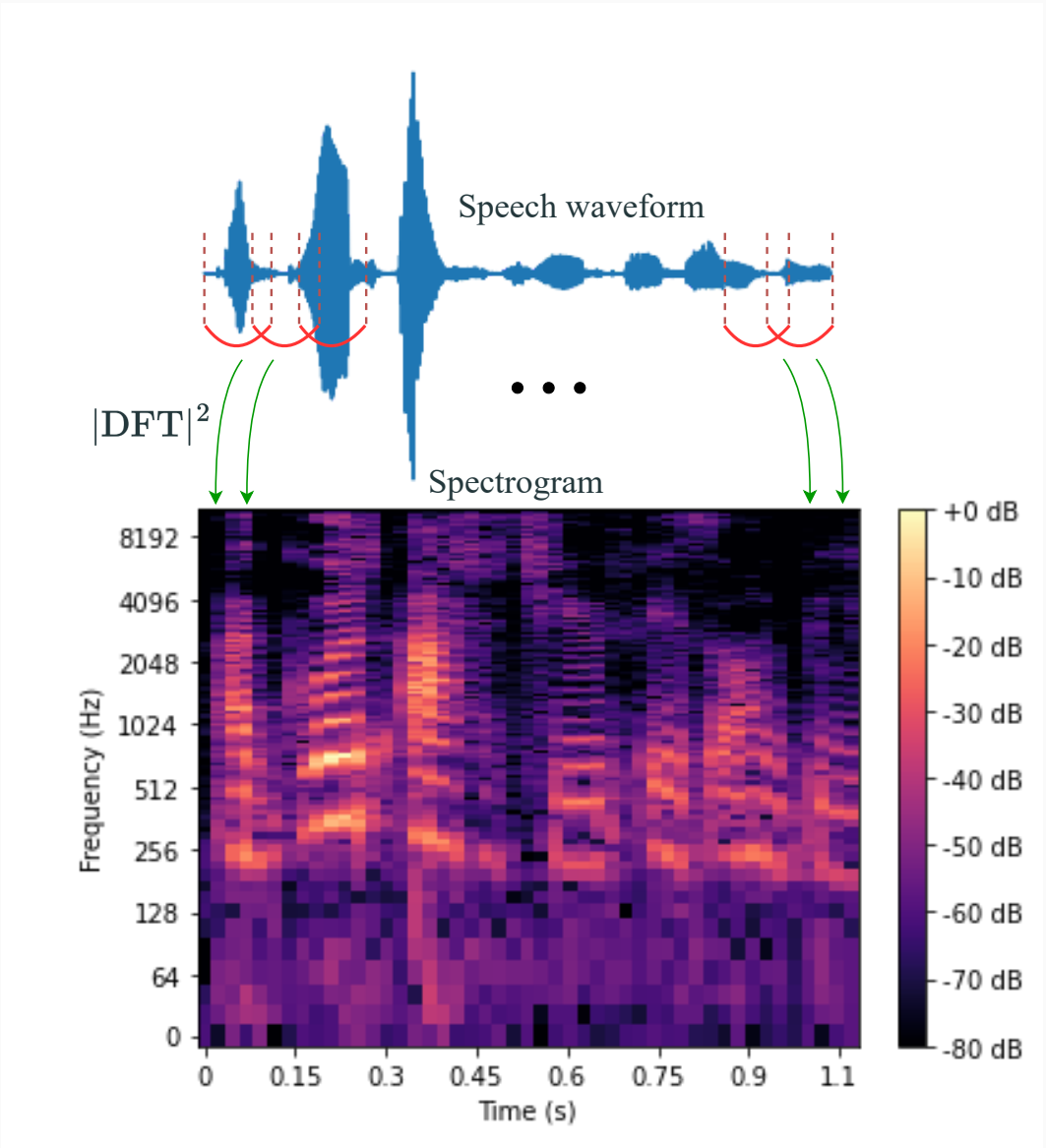


$$\theta^*, \psi^* = \arg\max_{\theta,\psi} \underbrace{\mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{s})}\Big[\log p_\theta(\boldsymbol{s}|\boldsymbol{z})\Big]}_{\text{Reconstruction term}} - \underbrace{D_{\mathsf{KL}}\Big(q_\psi(\boldsymbol{z}|\boldsymbol{s}) \parallel p(\boldsymbol{z})\Big)}_{\text{Regularization term}}$$

Audio-visual Speech Enhancement

Speech spectrogram

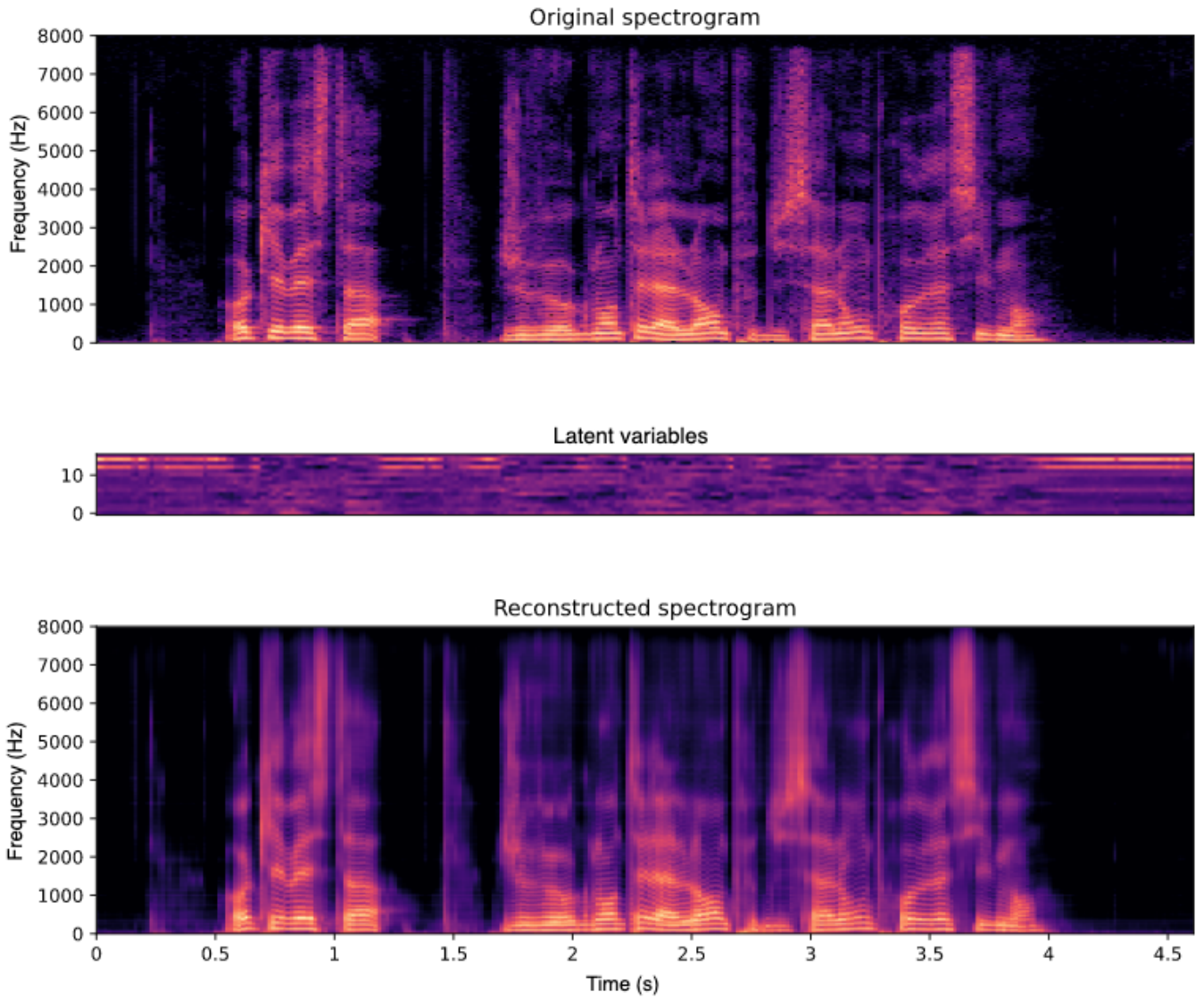
- A visual representation of the spectrum of frequencies based on Short-time Fourier transform (STFT).
- Apply discrete Fourier transform (DFT) to overlapping segments of speech waveform.
- Arrange the magnitude squared DFT vectors column-wise.



Speech auto-encoding using VAE

Original signal

Reconstructed signal



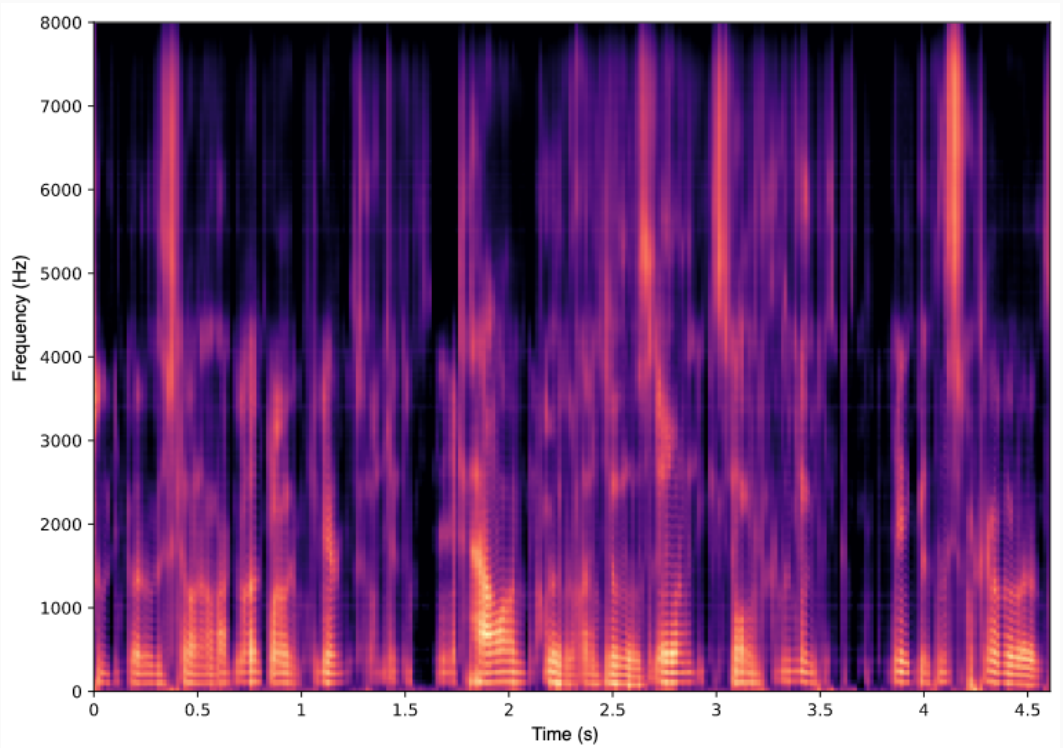
https://dynamicalvae.github.io/

Speech generation using VAE

Sample a latent code $\mathbf{z}$ from the prior, and give to the decoder to get a speech-like signal.

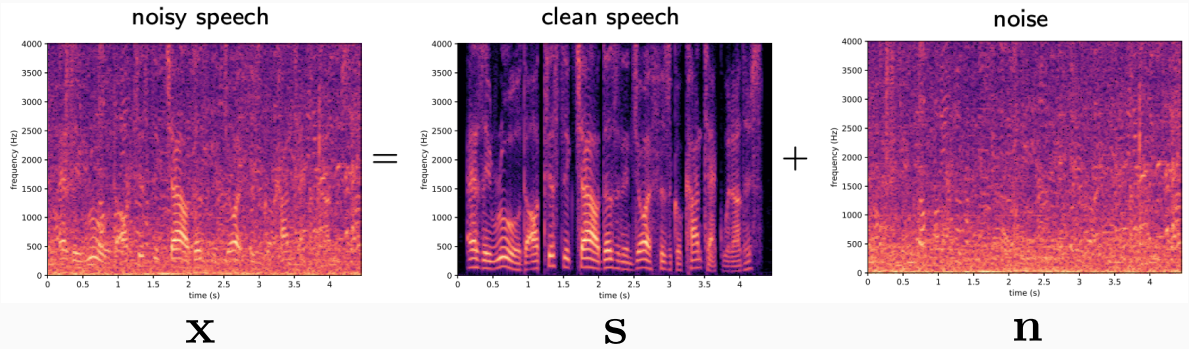
Generated signal

- Structured as a phoneme sequence, voiced/unvoiced phonemes
- Coarticulation, silences



`https://dynamicalvae.github.io/`

Speech Enhancement



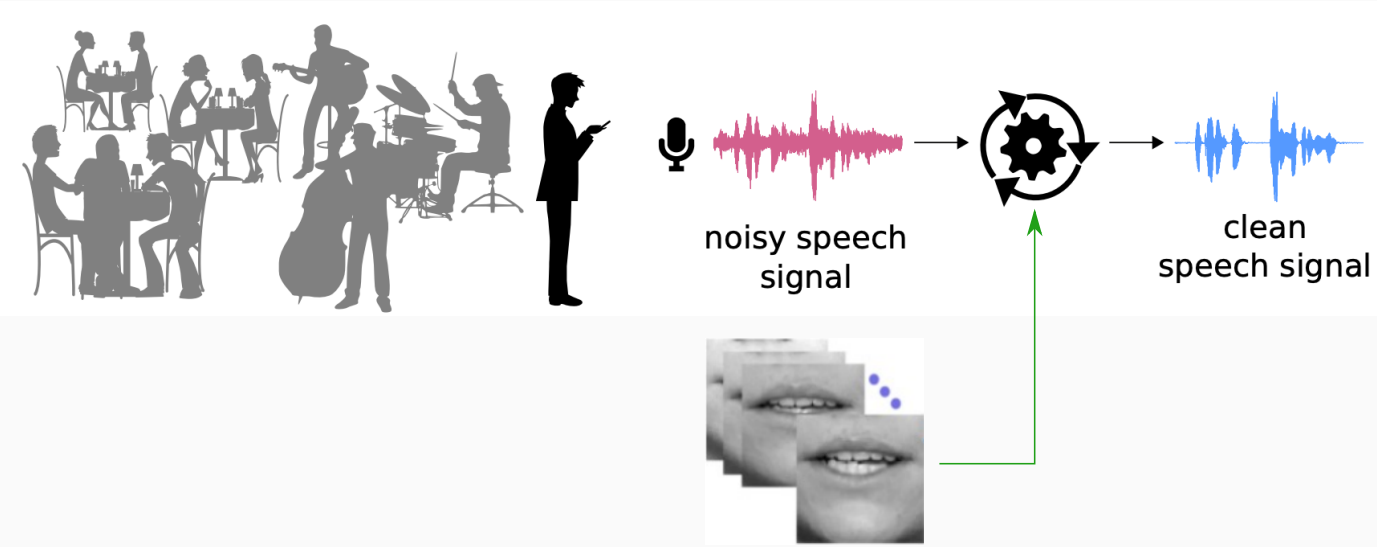
Improve the quality and intelligibility of the observed noisy speech signal $\mathbf{x}$.

- Close/distant conversations, listening comfort, hearing assistive devices.
- Automatic speech recognition for virtual assistants, social robots.

Audio-visual Speech Enhancement (AVSE)

Visual modality (lip movements):

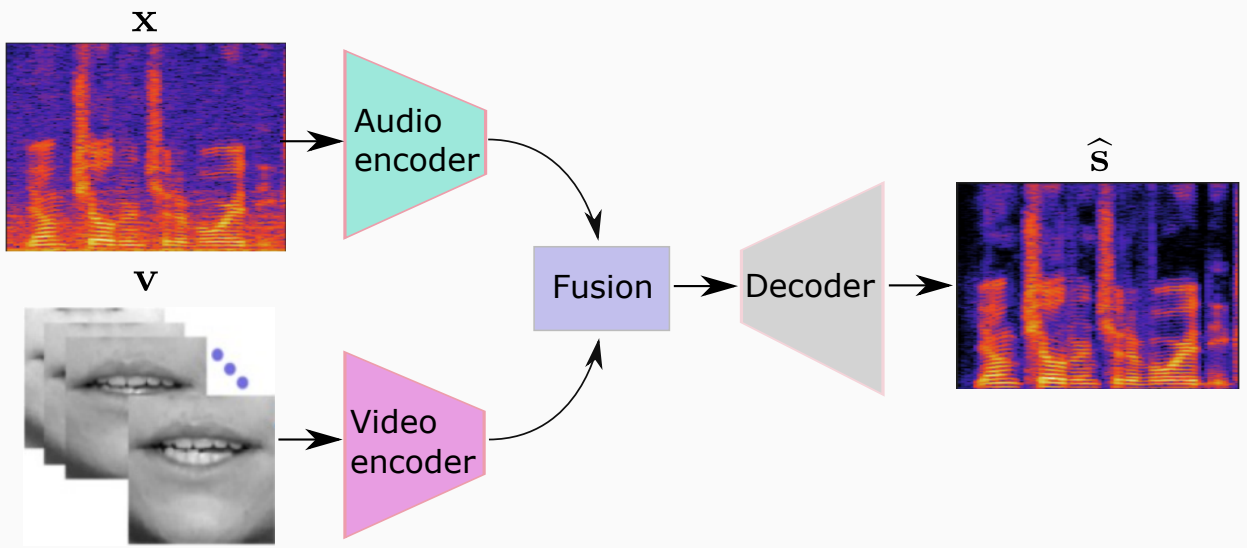
- Correlates well with speech signal (lip reading),
- Very helpful at highly noisy environments (unaffected by acoustic noise).



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

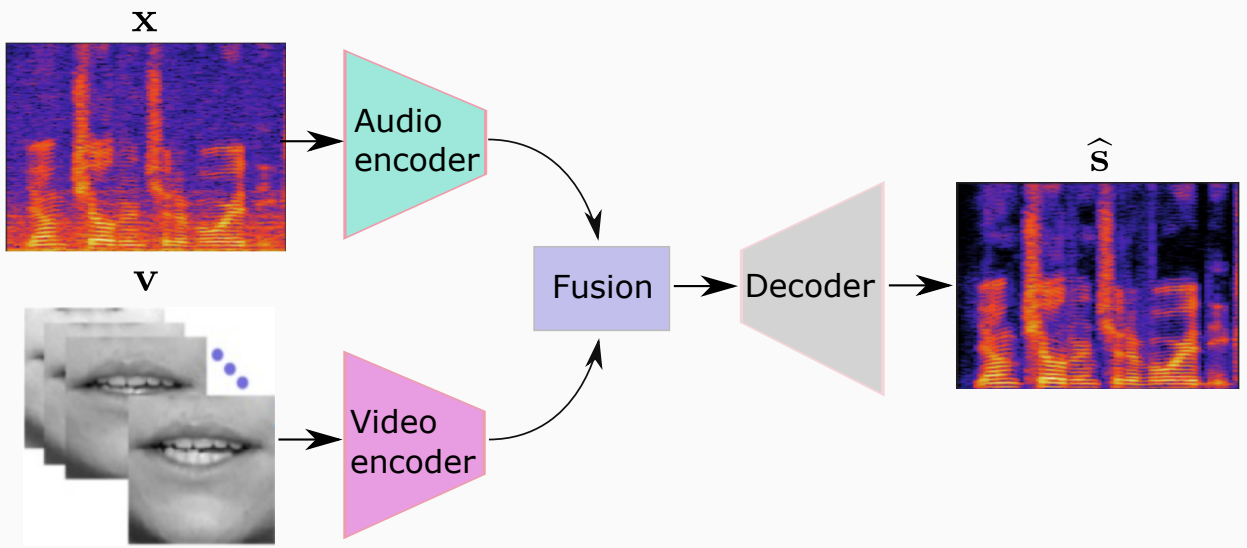
Supervised (discriminative) AVSE

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



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

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

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 $\Theta = \theta \cup \psi$:

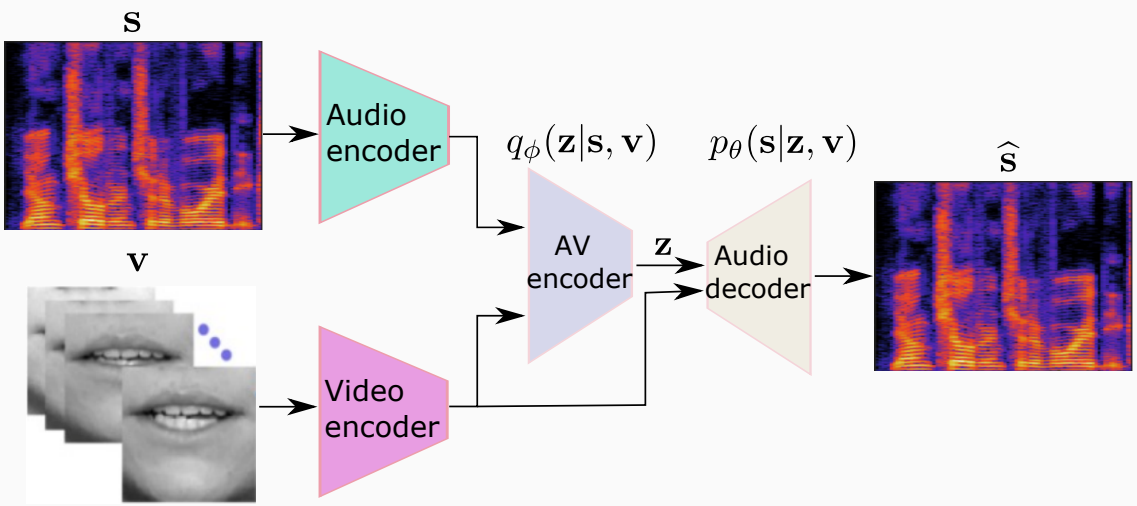
- Training - Learn speech prior distribution $p_\theta(\mathbf{s}|\mathbf{v})$
- Inference - Model $p_\psi(\mathbf{x}|\mathbf{s},\mathbf{v})$, and infer $\mathbf{s}$ using $p_\theta(\mathbf{s}|\mathbf{v})$

▷ Advantages over supervised approaches:

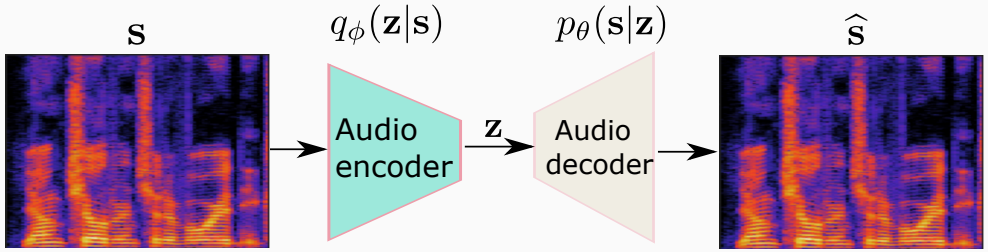
- No need to huge parallel corpora $\rightarrow$ compact & lightweight models
- Potentially better generalization performance

VAE architectures

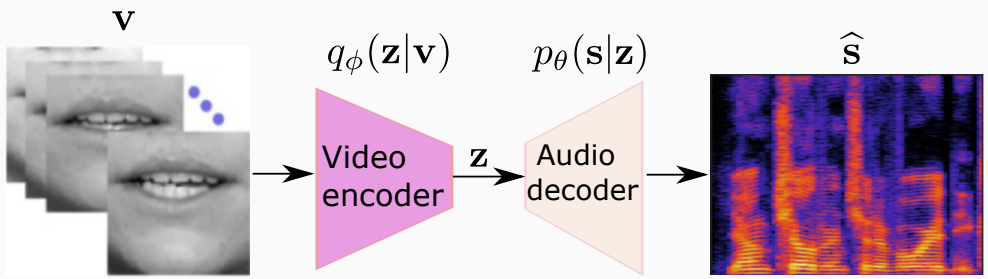
AV-VAE



A-VAE



V-VAE

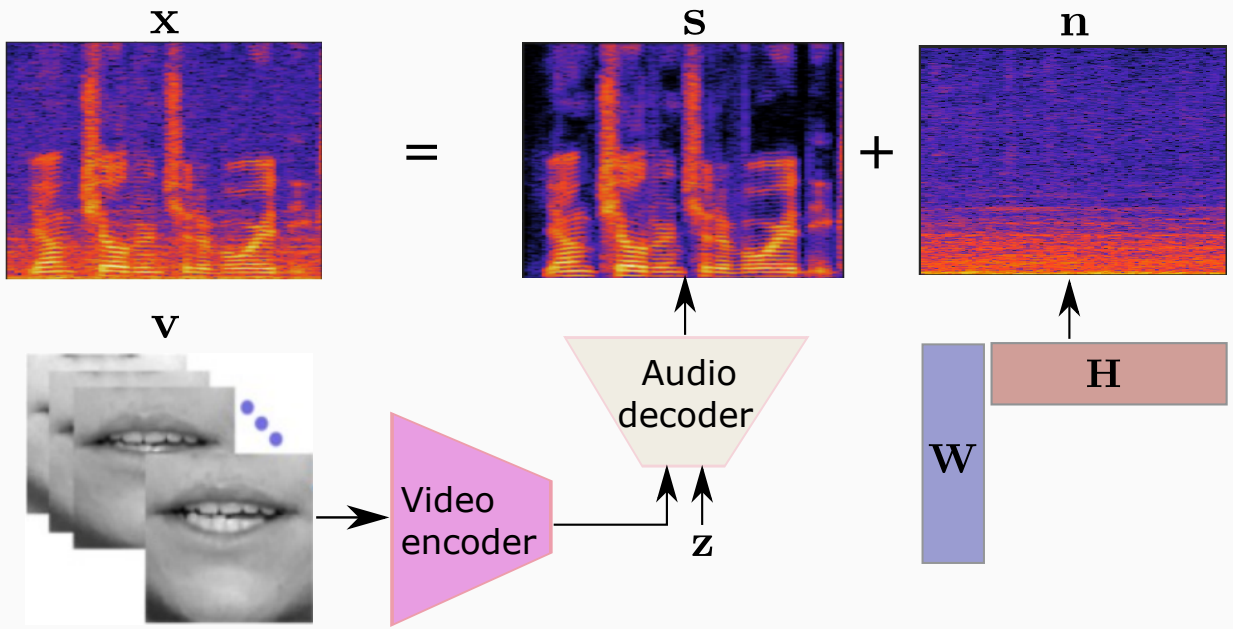


Speech Enhancement

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

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

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: $\mathbf{z} = \left\{\mathbf{z}_t\right\}_{t=1}^{T}$
- ▷ Likelihood:

$$p_\psi(\boldsymbol{x}_t|\mathbf{z}_t, \mathbf{v}_t) = \mathcal{N}_c\Big(\mathbf{0}, \mathsf{diag}\big(\boldsymbol{\sigma}_\theta^{av}(\mathbf{z}_t, \mathbf{v}_t)\big) + \mathsf{diag}(\mathbf{WH}[:, t])\Big)$$

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

Expectation Maximization (EM)

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.

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

  $\{\mathbf{z}^{(r)}\}_{r=1}^{R} \sim p(\mathbf{z}|\mathbf{x},\mathbf{v};\boldsymbol{\theta}_u^\star)$

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

Speech Estimation

Once the parameters are estimated, the speech STFT frames are estimated as follows ($\forall f, t$):

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

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

Examples

Noisy

A-VAE

V-VAE

AV-VAE

References

[1] Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. Foundations and Trends in Machine Learning, 12(4), pp.307-392.

[2] Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T. and Alameda-Pineda, X., 2021. Dynamical Variational Autoencoders: A Comprehensive Review. Foundations and Trends in Machine Learning, 15(1-2), pp.1-175.

[3] Sadeghi, M., Leglaive, S., Alameda-Pineda, X., Girin, L. and Horaud, R., 2020. Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, pp.1788-1800.

[4] Kang, Z., Sadeghi, M., Horaud, R., Donley, J., Kumar, A. and Alameda-Pineda, X., 2022. Expression-preserving face frontalization improves visually assisted speech processing. International Journal of Computer Vision (IJCV).

Thank you for your attention