Audio-visual Speech Enhancement with a Deep Kalman Filter Generative Model

Ali Golmakani, Mostafa Sadeghi, and Romain Serizel

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

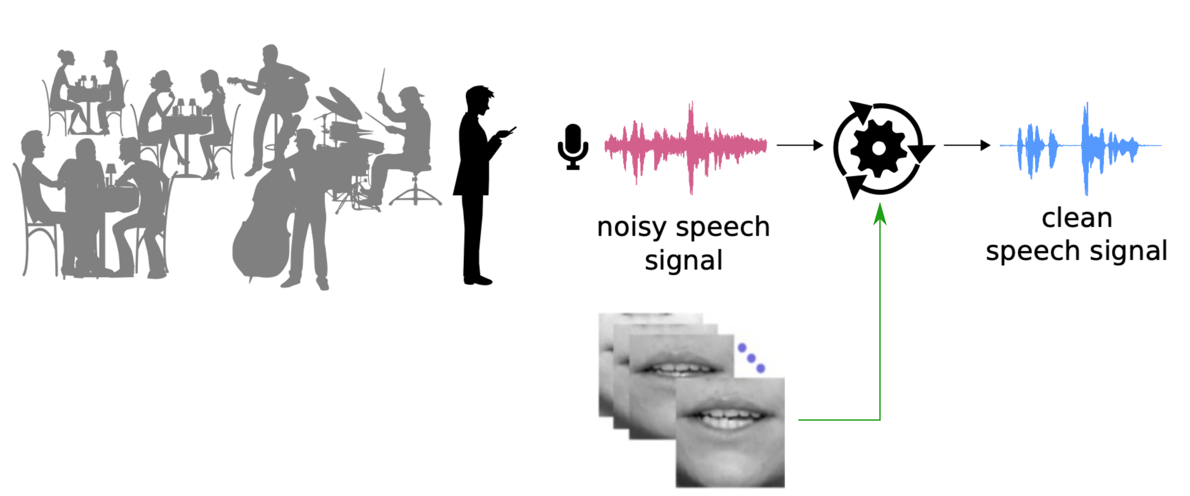
ICASSP'23

Poster number: 4421

Overview

- Generative models based on, e.g. VAEs [1], are promising for audio-visual speech enhancement (AVSE).
- Existing VAE-based AVSE models [2] overlook speech data's sequential nature and underutilize visual data.
- This work introduces AV-DKF, a generative model that effectively fuses audio-visual data using a first-order Markov chain model for latent variables.
- An efficient inference methodology is developed for estimating speech signals at test time.
- Experimental results show the superiority of AV-DKF over audio-only and non-sequential VAE-based audio-visual model.

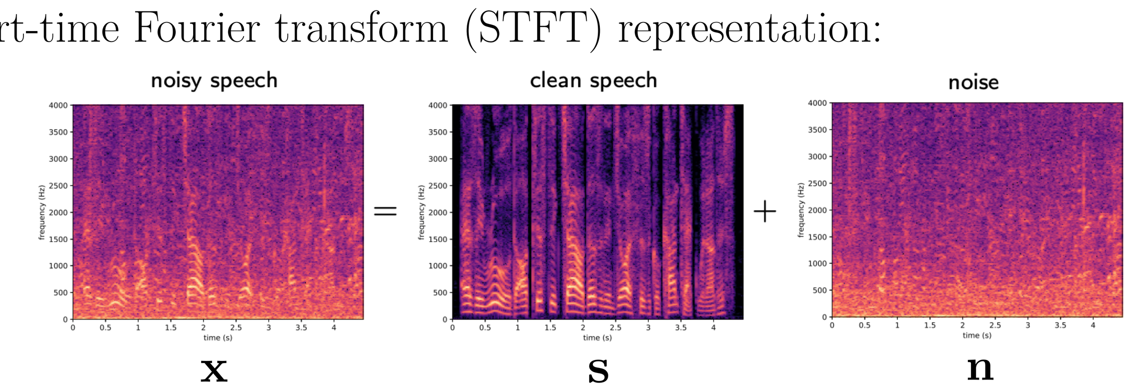
Audio-visual speech enhancement



Visual modality (lip movements):

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

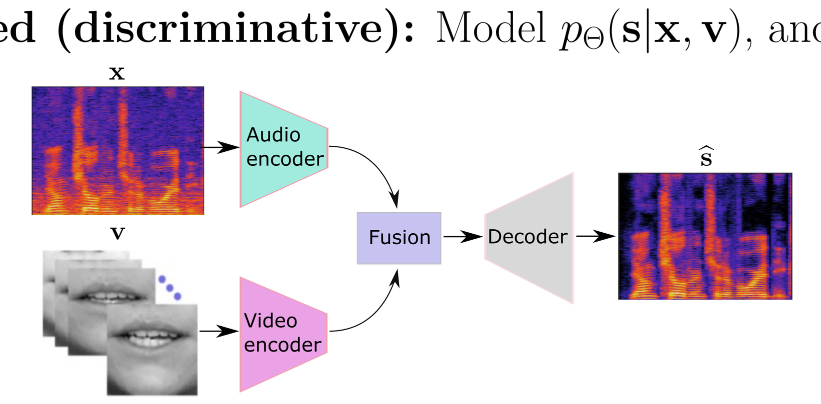
Short-time Fourier transform (STFT) representation:



- $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ (similarly for $\mathbf{s}$, $\mathbf{v}$ (visual features), and $\mathbf{n}$).

AVSE approaches

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



▷ **Unsupervised (generative):** *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 $p_\theta(\mathbf{s}|\mathbf{v})$

Speech generative modeling & enhancement

We focus on VAE [1]: $p_\theta(\mathbf{s}|\mathbf{v}) = \int p_\theta(\mathbf{s}|\mathbf{z},\mathbf{v})p_\theta(\mathbf{z}|\mathbf{v})d\mathbf{z}$

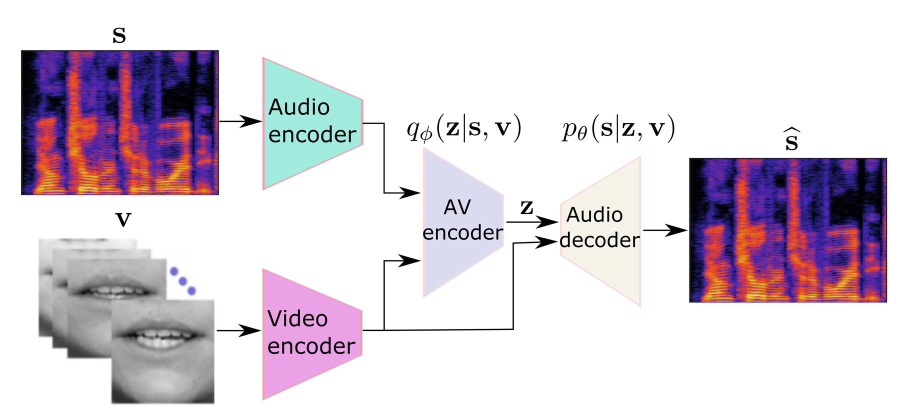
- $\mathbf{z} = \{\mathbf{z}_t \in \mathbb{R}^L\}$: (real-valued, low-dimensional, $L \ll F$) latent variables

A Gaussian generative model for *individual time-frames*:

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

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

Given a training set $\{(\mathbf{s}_t,\mathbf{v}_t)\}_{t=1}^T \longrightarrow$ learn θ using the maximum likelihood principle.

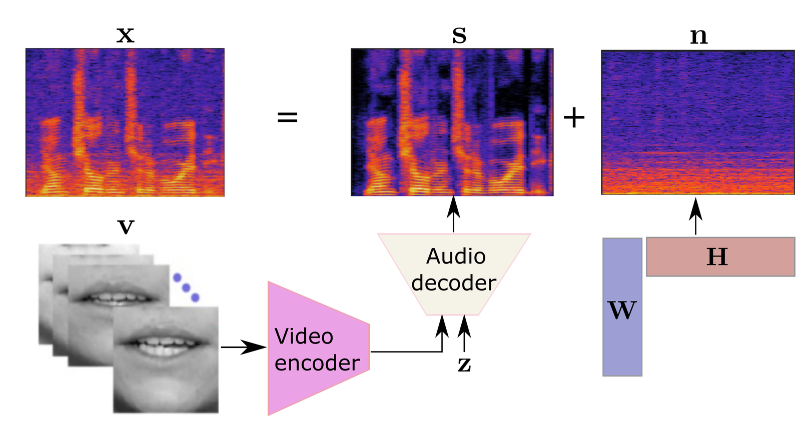


Observation model: $\forall t: \quad \boxed{\boldsymbol{x}_t = \sqrt{g_t}\boldsymbol{s}_t + \boldsymbol{n}_t} \quad g_t > 0$

Noise model: Non-negative Matrix Factorization (NMF)

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

Clean speech model: Trained generative (decoder) network.



▷ **Standard AV-VAE [2]:**

$$p_\theta(\mathbf{s}_{1:T},\mathbf{z}_{1:T},\mathbf{v}_{1:T}) = \prod_{t=1}^T p_\theta(\mathbf{s}_t,\mathbf{z}_t,\mathbf{v}_{1:T}) = \prod_{t=1}^T p_\theta(\mathbf{s}_t|\mathbf{z}_t,\mathbf{v}_{1:T})p_\theta(\mathbf{z}_t|\mathbf{v}_{1:T})$$

⚠ No temporal modeling → not realistic for STFT time frames.

Proposed methodology: AV-DKF

Time-dependent factorization: $p_\theta(\mathbf{s}_{1:T},\mathbf{z}_{1:T},\mathbf{v}_{1:T}) = \prod_{t=1}^T p_\theta(\mathbf{s}_t|\mathbf{z}_t,\mathbf{v}_{1:T}) \times \underbrace{p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{v}_{1:T})}_{\text{first-order Markov model}}$

$$\begin{cases} p_\theta(\mathbf{s}_t|\mathbf{z}_t,\mathbf{v}_{1:T}) = \mathcal{N}_c\Big(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\theta_s}(\mathbf{z}_t,\mathbf{v}_{1:T}))\Big), \\ p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{v}_{1:T}) = \mathcal{N}\Big(\boldsymbol{\mu}_{\theta_z}(\mathbf{z}_{t-1},\mathbf{v}_{1:T}), \operatorname{diag}(\boldsymbol{\sigma}^2_{\theta_z}(\mathbf{z}_{t-1},\mathbf{v}_{1:T}))\Big), \end{cases}$$

AV-DKF inference:

$$q_\psi(\mathbf{z}_{1:T}|\mathbf{u}_{1:T}) = \prod_{t=1}^T q_\psi(\mathbf{z}_t|\mathbf{r}_t) = \prod_{t=1}^T \mathcal{N}\Big(\boldsymbol{\mu}_\psi(\mathbf{r}_t), \operatorname{diag}(\boldsymbol{\sigma}^2_\psi(\mathbf{r}_t))\Big), \quad \mathbf{u}_{1:T} = \{\mathbf{s}_t,\mathbf{v}_t\}_{t=1}^T, \quad \mathbf{r}_t = \{\mathbf{z}_{t-1},\mathbf{u}_{t:T}\}$$

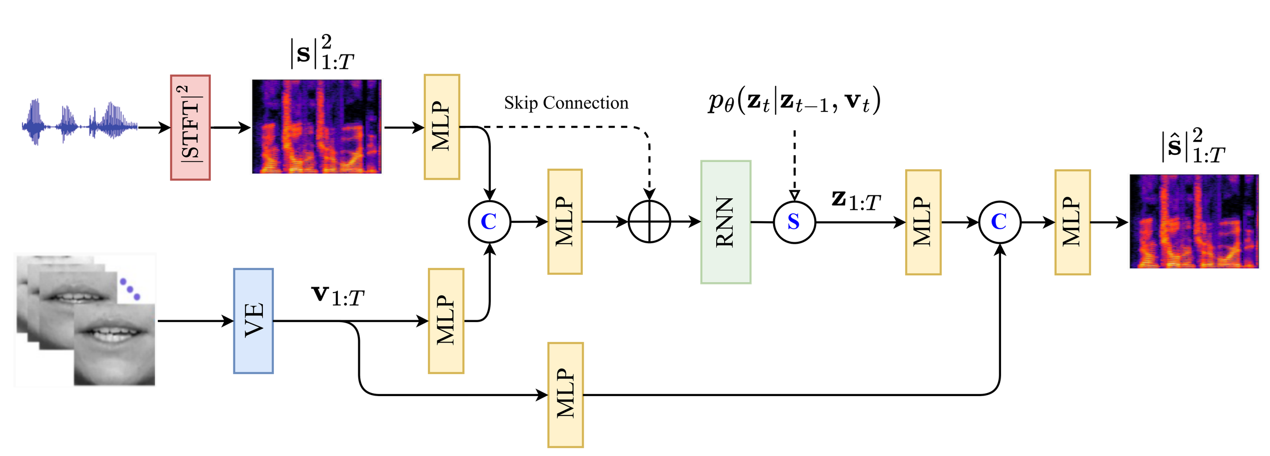


Figure 1:Schematic diagram of the proposed AV-DKF generative model (without explicit architecture of the prior network). **MLP**: multilayer perception, **RNN**: recurrent neural network, **VE**: video encoder, ⊕: addition, **C**: concatenation, **S**: sampling in the latent space.

Speech enhancement

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

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

  $$\mathbf{z}^*,\mathbf{g}^* = \arg\max_{\mathbf{z},\mathbf{g}} \sum_{t=1}^T \log p_\phi(\mathbf{x}_t|\mathbf{z}_t,g_t,\mathbf{v}_t) + \log p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{v}_t) + \log p(g_t),$$

  As opposed to the previous works, here $\mathbf{g}$ is treated as a latent variable with a Gamma prior distribution.

- **M-Step**: $\psi^{(k+1)} \leftarrow \arg\max_\psi Q(\psi|\psi^{(k)})$

Speech estimation (posterior mean):

$$\hat{\mathbf{s}} = \mathbb{E}_{p_{\psi^*}(\mathbf{s}|\mathbf{x},\mathbf{v})}\{\mathbf{s}\} \approx \Big\{\frac{g_t^*\boldsymbol{\sigma}^2_\theta(\mathbf{z}_t^*,\mathbf{v}_{1:T})}{g_t^*\boldsymbol{\sigma}^2_\theta(\mathbf{z}_t^*,\mathbf{v}_{1:T}) + [\mathbf{W}^*\mathbf{H}^*]_t} \odot \mathbf{x}_t\Big\}_{t=1}^T.$$

Experiments

- **Corpus**: NTCD-TIMIT [3]: 56 English speakers (39 training, 8 validation, 9 test), 98 sentences (~ 5 s) per speaker.
- **Noise types**: *Living Room, White, Cafe, Car, Babble, Street.*
- **STFT parameters**: 64 ms sine window, 75% overlap → STFT frames of length $n = 513$.
- **Baselines**: A-VAE [4], AV-VAE [2], A-DKF [5].

Performance measures: Signal-to-distortion ratio (**SDR**), Perceptual evaluation of speech quality (**PESQ**) [-0.5,4.5], Short-time objective intelligibility (**STOI**) [0,1].

Table 1:For each method, **top row**: g_t updated by multiplicative rules [4,5], **bottom row**: g_t proposed update.

Metric	SI-SDR (dB)					PESQ					STOI				
SNR (dB)	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
Input	-12.80	-7.72	-2.91	2.04	7.25	1.51	1.76	2.05	2.37	2.85	0.20	0.30	0.43	0.56	0.69
A-VAE	**-7.37**	**-1.92**	**3.78**	**8.65**	**13.07**	1.63	1.91	2.20	2.50	2.85	0.21	0.32	0.45	0.59	0.72
	-8.46	-2.60	3.02	8.11	13.01	**1.67**	**1.95**	**2.25**	**2.58**	**2.90**	**0.22**	0.32	**0.47**	**0.60**	**0.73**
AV-VAE	-6.86	**-0.83**	**4.70**	**9.38**	**13.90**	1.74	2.00	2.31	2.61	2.90	0.20	0.31	0.45	0.59	0.72
	-6.65	-0.86	4.47	9.26	13.77	**1.75**	**2.03**	**2.34**	**2.65**	**2.93**	**0.22**	**0.33**	**0.47**	**0.61**	**0.73**
A-DKF	**-6.50**	-1.41	1.99	4.36	5.55	1.48	1.67	1.87	2.02	2.13	0.22	0.33	0.45	0.55	0.64
	-7.02	**-0.92**	**4.76**	**10.39**	**14.96**	**1.78**	**2.08**	**2.41**	**2.75**	**3.03**	0.22	**0.35**	**0.50**	**0.65**	**0.77**
AV-DKF	-5.04	-0.21	2.93	4.92	5.48	1.39	1.61	1.82	1.97	2.07	0.22	0.33	0.44	0.55	0.63
	-3.78	**1.78**	**7.19**	**11.66**	**15.81**	**1.94**	**2.24**	**2.54**	**2.80**	**3.05**	**0.25**	**0.38**	**0.52**	**0.66**	**0.77**

Conclusions:

▷ Audio-visual models outperform their audio-only counterparts.
▷ The proposed **g**-update works much better than the multiplicative one.
▷ The proposed AV-DKF model provides higher metrics than AV-VAE.

References

1. D. P. Kingma and M. Welling, "*Auto-encoding variational Bayes*," ICLR, 2014.
2. 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.
3. A.-H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," INTERSPEECH, 2017.
4. S. Leglaive *et al.* "A variance modeling framework based on variational autoencoders for speech enhancement," MLSP, 2018.
5. X. Bie *et al.*, "Unsupervised speech enhancement using dy- namical variational autoencoders," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2993–3007, 2022.