Switching Variational Auto-Encoders for Noise-Agnostic Audio-visual Speech Enhancement

Mostafa Sadeghi[1] and Xavier Alameda-Pineda[2]

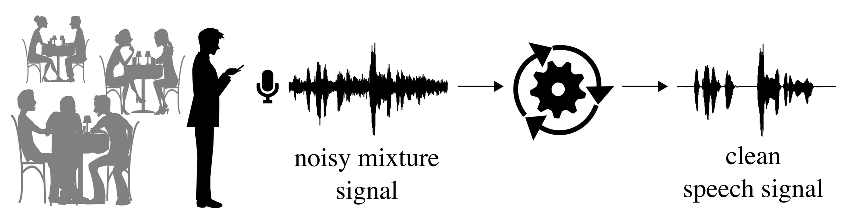
[1]Multispeech team, Inria Nancy - Grand Est, France, [2]Perception team, Inria Grenoble Rhône-Alpes, France

Poster number: 3517

Overview

- Unsupervised audio-visual speech enhancement is addressed.
- A switching generative model (VAE) is proposed for clean speech.
- The model provides noise-agnostic speech enhancement.

Unsupervised speech enhancement



In the short-time Fourier transform (STFT) domain, for all $(f,t) \in \mathbb{B} = \{0,...,F-1\} \times \{0,...,T-1\}$, we observe: $\boxed{x_{ft} = s_{ft} + b_{ft}}$

- $s_{ft} \rightarrow$ clean speech signal, and $b_{ft} \rightarrow$ noise signal
- $(f,t) \rightarrow$ frequency and time-frame indices.

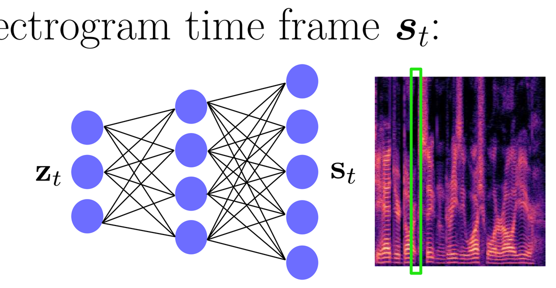
Separate the speech and noise signals without training on noise.

Training: Learn $p(\boldsymbol{s}_t) = \int p(\boldsymbol{s}_t|\mathbf{z}_t)p(\mathbf{z}_t)d\mathbf{z}_t$

Testing: Using $p(\boldsymbol{s}_t)$ and $p(\boldsymbol{x}_t|\boldsymbol{s}_t)$ estimate $\boldsymbol{s}_t$, $\forall t$.

Generative model for each clean spectrogram time frame $\boldsymbol{s}_t$:

$$\begin{cases} \boldsymbol{s}_t|\mathbf{z}_t \sim \mathcal{N}_c\Big(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}_s^a(\mathbf{z}_t))\Big), \\ \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{cases}$$



- $\mathbf{z}_t \in \mathbb{R}^L$ is a latent random variable ($L \ll F$).
- $\boldsymbol{\sigma}_s^a(.) : \mathbb{R}^L \mapsto \mathbb{R}_+^F$ is a neural network parameterized by $\boldsymbol{\theta}$.

Training: learning the parameters

- **Training dataset:** $\mathbf{s} = \{\mathbf{s}_t \in \mathbb{C}^F\}_{t=0}^{T_{tr}-1}$
- **Difficulty**: Intractable likelihood $p_{\boldsymbol{\theta}}(\mathbf{s}) = \int p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$
- **Solution**: Variational autoencoder (VAE) [Kingma and Welling 2014]

Using variational inference, maximize a lower bound of $\ln p_{\boldsymbol{\theta}}(\mathbf{s})$:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{1}{T_{tr}} \sum_{t=0}^{T_{tr}-1} \mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}_t|\mathbf{s}_t)}\Big[\ln p_{\boldsymbol{\theta}}(\mathbf{s}_t|\mathbf{z}_t)\Big] - D_{\mathrm{KL}}\Big(q_{\boldsymbol{\psi}}(\mathbf{z}_t|\mathbf{s}_t) \parallel p(\mathbf{z}_t)\Big)$$

where $q_{\boldsymbol{\psi}}(\mathbf{z}_t|\mathbf{s}_t) \approx p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{s}_t)$ is defined by an "encoding network" with parameters $\boldsymbol{\psi}$. $D_{\mathrm{KL}}(. \parallel .)$ is the Kullback–Leibler divergence.

Testing: speech enhancement

Noisy speech model: $\forall t: \quad \boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{b}_t$

Noise model: $\forall t: \quad \boldsymbol{b}_t \sim \mathcal{N}_c\big(\mathbf{0}, \operatorname{diag}(\mathbf{WH}[:,t])\big)$

Clean speech model: Trained VAE

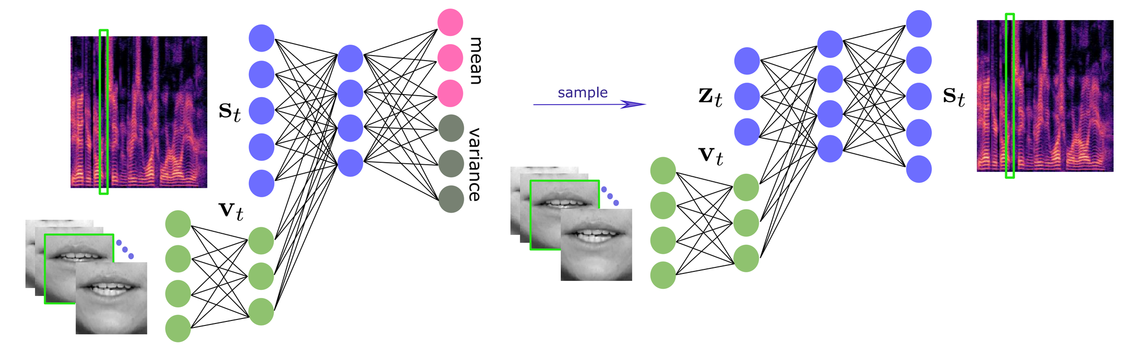
▷ Observed variables: $\mathbf{x} = \{\mathbf{x}_t\}_{t=0}^{T-1}$. Latent variables: $\mathbf{z} = \{\mathbf{z}_t\}_{t=0}^{T-1}$.

▷ Parameters to be estimated: $\boldsymbol{\theta}_u = \{\mathbf{W}, \mathbf{H}\}$.

Monte-Carlo Expectation maximization (MCEM) is used for inference.

Audio-visual modeling of clean speech

- Visual modality (lip movements) provides complementary information about speech.
- Audio-visual VAE (AV-VAE) model outperforms audio-only VAE (A-VAE) [Sadeghi et al., 2020].

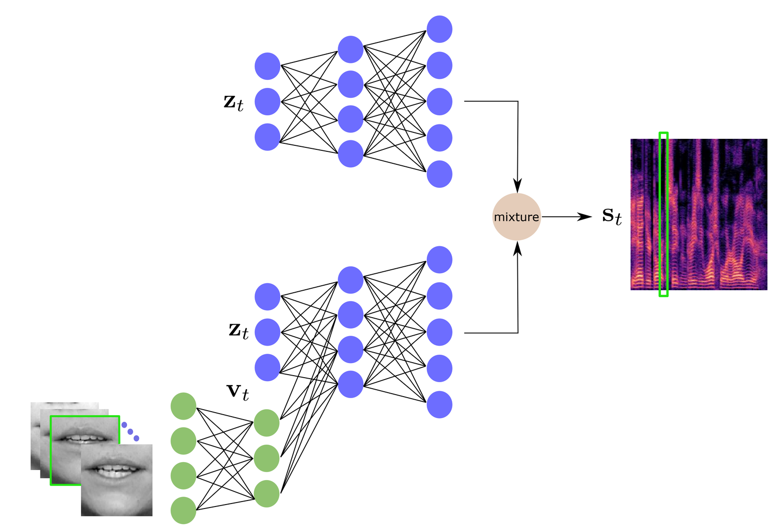


Robustness to noisy visual data

AV-VAE yields poor performance when the visual modality is not clean, e.g., mouth area is occluded or speaker's face is not frontal.

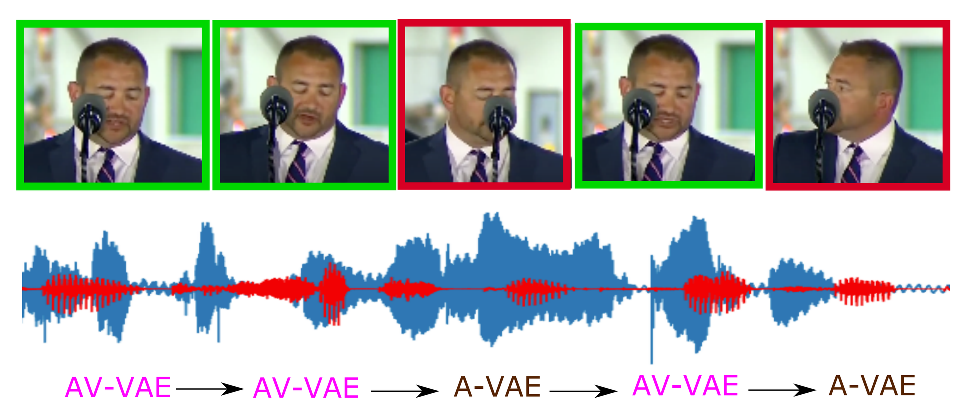
MIX-VAE [Sadeghi & Alameda-Pineda, 2020]:

- A mixture of pre-trained A-VAE and AV-VAE generative models.
- If the lip region is clean, use AV-VAE, otherwise use A-VAE.



Our work: Switching VAEs

Objective: To devise a robust generative modeling framework for speech enhancement using several VAEs with a dynamic selection mechanism.



Switching Variational Auto-Encoder (SwVAE):

A set of M already trained VAEs with a switching variable $m_t \in \{1,\ldots,M\}$ modeled with a Markov chain:

$$\begin{cases} p(m_1,\ldots,m_T) \sim \mathcal{MC}(\lambda,\tau), \\ p(\mathbf{z}_t|m_t;\mathbf{v}_t) \sim \mathcal{N}\Big(\boldsymbol{\xi}_{m_t}(\mathbf{v}_t), \boldsymbol{\Lambda}_{m_t}(\mathbf{v}_t)\Big), \\ p(\mathbf{s}_t|\mathbf{z}_t,m_t;\mathbf{v}_t) \sim \mathcal{N}_c\Big(\mathbf{0}, \boldsymbol{\Sigma}_{m_t}(\mathbf{z}_t,\mathbf{v}_t)\Big), \end{cases} \tag{1}$$

- A Markovian dependency is assumed to switch between different VAE-based generative models.
- The model can be understood as a Hidden Markov Model (HMM) with emission probabilities given by the decoder of VAEs.
- $\mathcal{MC}(\lambda,\tau)$ is short for a Markov chain with initial distribution λ and transition distribution τ.
- $\boldsymbol{\xi}_{m_t}(.)$, $\boldsymbol{\Lambda}_{m_t}(.)$, and $\boldsymbol{\Sigma}_{m_t}(.,.)$ are non-linear transformations of their inputs indexed by $m_t \in \{1,\ldots,M\}$.

Testing: speech enhancement

The observation (noise) model is like before. The clean speech model is the SwVAE consisting of several trained VAEs (here A-VAE and AV-VAE).

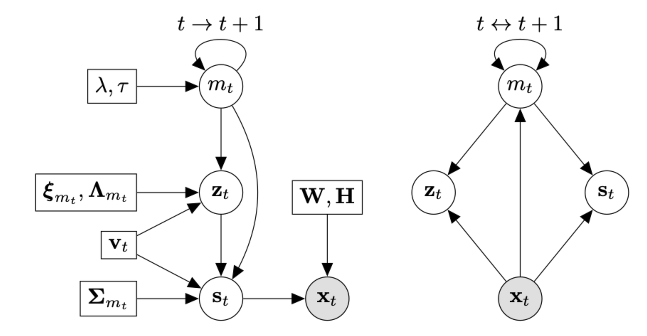
Inference:

▷ Observed variables: $\{\mathbf{x}_t, \mathbf{v}_t\}_{t=0}^{T-1}$

▷ Latent variables: $\{\mathbf{s}_t, \mathbf{z}_t, m_t\}_{t=0}^{T-1}$

▷ Parameters to be estimated: $\{\lambda, \tau, \mathbf{W}, \mathbf{H}\}$

▷ Once the parameters are learned, estimate the clean speech $\{\mathbf{s}_t\}_{t=0}^{T-1}$.



(a) Graphical model (b) Variational approximation

Variational Expectation-maximization (VEM)

Defining $\mathbf{x} = \{\mathbf{x}_t\}_{t=0}^{T-1}$ (analogously $\mathbf{s}, \mathbf{z}, \mathbf{m}, \mathbf{v}$), the intractable posterior of the latent variables is approximated by a variational distribution:

$$p(\mathbf{s},\mathbf{z},\mathbf{m}|\mathbf{x},\mathbf{v}) \approx r^s(\mathbf{s}|\mathbf{m})r^z(\mathbf{z}|\mathbf{m})r^m(\mathbf{m}),$$

▷ We set $r^z(\mathbf{z}_t|m_t) = \mathcal{N}(\mathbf{c}_{tm}, \boldsymbol{\Omega}_{tm})$, where $\mathbf{c}_{tm}$ and $\boldsymbol{\Omega}_{tm}$ (diagonal) are to be learned along with r^s and r^m.

▷ We optimize a lower-bound of the data log-likelihood $\log p(\mathbf{x},\mathbf{v})$:

$$\mathbb{E}_{r^s r^z r^m}\left[\log \frac{p(\mathbf{x},\mathbf{v},\mathbf{s},\mathbf{z},\mathbf{m})}{r^s(\mathbf{s}|\mathbf{m})r^z(\mathbf{z}|\mathbf{m})r^m(\mathbf{m})}\right] \leq \log p(\mathbf{x},\mathbf{v}) \tag{2}$$

Variational E-Step: Optimize (2) over r^s, r^m and parameters of r^z.

M Step: Optimize (2) over $\mathbf{W}, \mathbf{H}$, leading to multiplicative rules.

Clean speech estimation: The enhanced speech signal is the marginalisation over m_t:

$$\hat{\mathbf{s}}_t = \mathbb{E}_{r^m(m_t)}\Big[\mathbb{E}_{r^s(\mathbf{s}_t|m_t)}[\mathbf{s}_t]\Big], \quad \forall t.$$

Experiments

- **Corpus**: NTCD-TIMIT [Abdelaziz, 2017]
  - ∼ 1 hours of speech, 9 speakers;
  - Noise types: *LR*, *White*, *Cafe*, *Car*, *Babble*, and *Street*;
  - Noise levels: $\{-15,-10,-5,0,5,10\}$ dB;
  - Clean lips region as well as noisy versions (∼ one-third of total video frames/sample)
- **VAE models**: Pre-trained A-VAE and AV-VAE [Sadeghi et al., 2020]
- **Baseline**: MIX-VAE [Sadeghi & Alameda-Pineda, 2020]

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

Measure	PESQ					SDR (dB)					STOI				
SNR (dB)	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
Input	1.44	1.67	2.04	2.30	2.72	-12.30	-7.30	-3.45	1.88	6.73	0.22	0.32	0.45	0.56	0.68
MIX-VAE - clean	**1.70**	1.92	2.29	2.48	2.66	**-3.51**	1.67	5.38	9.22	12.07	0.24	0.35	0.47	0.55	0.65
SwVAE - clean	1.67	**1.97**	**2.39**	**2.62**	**2.83**	-3.59	**2.00**	**6.24**	**10.73**	**14.12**	**0.25**	**0.36**	**0.51**	**0.61**	**0.72**
MIX-VAE - noisy	**1.66**	1.91	2.22	2.41	2.51	**-3.78**	1.50	5.18	8.72	10.88	0.23	0.34	0.45	0.53	0.63
SwVAE - noisy	1.65	**1.94**	**2.36**	**2.60**	**2.81**	-3.97	**1.84**	**6.14**	**10.51**	**14.06**	**0.24**	**0.35**	**0.50**	**0.59**	**0.67**

▷ Both methods exhibit robustness to noisy visual data.

▷ SwVAE performs better when changing noise level.

References

1. D. P. Kingma and M. Welling, "*Auto-encoding variational Bayes*," ICLR, 2014.
2. S. Leglaive et al., "*A variance modeling framework based on variational autoencoders for speech enhancement*," in Proc. MLSP, 2018.
3. M. Sadeghi et al., "*Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoder*," IEEE Transactions on Audio, Speech and Language Processing, vol. 28, pp. 1788-1800, May 2020.
4. M. Sadeghi and X. Alameda-Pineda, "Robust Unsupervised Audio-visual Speech Enhancement Using a Mixture of Variational Autoencoders," in Proc. ICASSP, Barcelona, Spain, May 2020.
5. A. H. Abdelaziz, "*NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition*," in Proc. INTERSPEECH, 2017.

Xavier Alameda-Pineda acknowledges ANR JCJC ML3RI project (ANR-19-CE33-0008-01). This work has been partially supported by MIAI @ University Grenoble Alpes, (ANR-19-P3IA-0003).