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

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

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

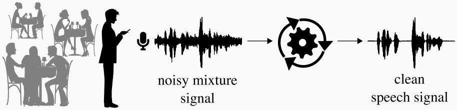
IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)

June 6-11, 2021

* 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).

Introduction

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.

▷ No training on noise, hence unsupervised.

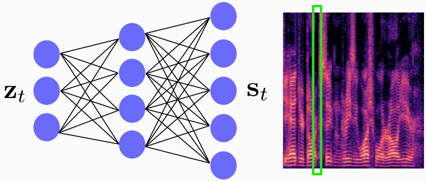
Generative speech model [Bando et al., 2018; Leglaive et al., 2018]

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 $\boldsymbol{s}_t$:

$$\boldsymbol{s}_t|\mathbf{z}_t \sim \mathcal{N}_c\Big(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}_s^a(\mathbf{z}_t))\Big), \qquad \text{with } \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$



- $\mathbf{z}_t \in \mathbb{R}^L$ 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 $\boldsymbol{\theta}$

Estimate the generative model parameters, i.e. $\boldsymbol{\theta}$.

Training: learning the parameters

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

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

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

where $q(\mathbf{z}_t|\mathbf{s}_t;\boldsymbol{\psi}) \approx p(\mathbf{z}_t|\mathbf{s}_t;\boldsymbol{\theta})$ is defined by an "encoding network" with parameters $\boldsymbol{\psi}$. $D_{\mathsf{KL}}(. \parallel .)$ 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: $\forall t: \quad \boldsymbol{b}_t \sim \mathcal{N}_c\left(\mathbf{0}, \text{diag}(\mathbf{WH}[:,t])\right)$

Clean speech model: Trained VAE

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

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

Monte-Carlo Expectation maximization (MCEM):

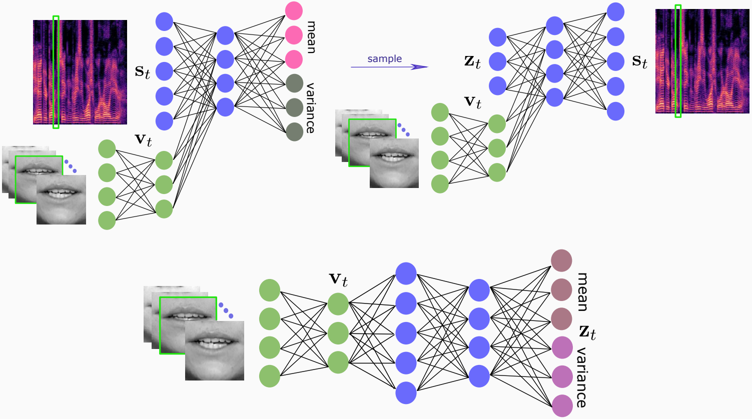
- **E-Step**: $Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)}[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \boldsymbol{\theta}_u)]$.
- **M-Step**: $\boldsymbol{\theta}_u^\star \leftarrow \arg\max_{\boldsymbol{\theta}_u} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star)$.

Speech estimation:

$$\hat{s}_{ft} = \mathbb{E}_{p(s_{ft}|x_{ft};\boldsymbol{\theta}^*)}[s_{ft}] = \mathbb{E}_{p(\mathbf{z}_t|\mathbf{x}_t;\boldsymbol{\theta}^*)}\left[\mathbb{E}_{p(s_{ft}|\mathbf{z}_t,\mathbf{x}_t;\boldsymbol{\theta}^*)}[s_{ft}]\right]$$

Audio-visual modeling of clean speech [Sadeghi et al., 2020]

- 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].

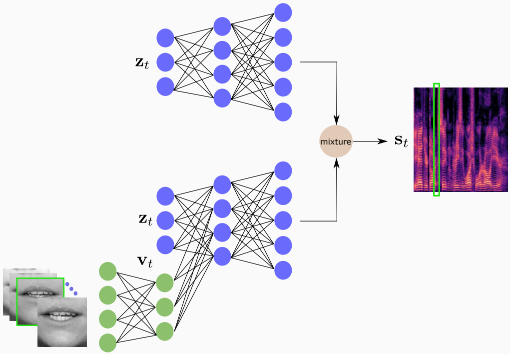


Robustness to noisy visual data

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.

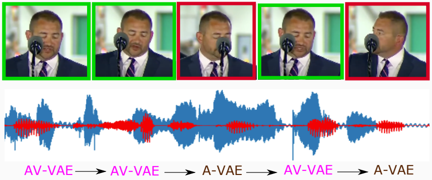
MIX-VAE [Sadeghi & Alameda-Pineda, 2020]:

- 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.



Switching Variational Auto-Encoders

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



Switching Variational Auto-Encoders (SwVAE):

- 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.
- A variational factorization of the posterior distribution of the latent variables is proposed.

Proposed model

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

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

- $\mathcal{MC}(\lambda, \tau)$ 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 $m_t \in \{1, \dots, M\}$ and realized as DNNs.

Testing: speech enhancement

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

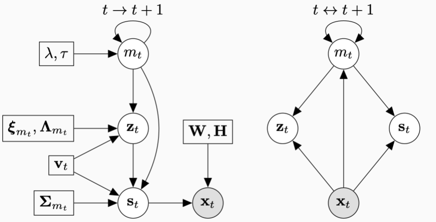
Noise model: $\forall t: \quad \boldsymbol{b}_t \sim \mathcal{N}_c\Big(\mathbf{0}, \text{diag}(\mathbf{WH}[:, t])\Big)$

Clean speech model: Trained VAE generative networks

Inference:

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

Parameter estimation

Variational Expectation-maximization (VEM)

Variational E-Step:

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

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

▷ r^s (and r^z) further factorize over time: $r^s(\boldsymbol{s}|\boldsymbol{m}) = \prod_t r^s(\boldsymbol{s}_t|m_t)$

▷ We set $r^z(\boldsymbol{z}_t|m_t) = \mathcal{N}(\boldsymbol{c}_{tm}, \boldsymbol{\Omega}_{tm})$, where $\boldsymbol{c}_{tm}$ 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 $\log p(\boldsymbol{x}, \boldsymbol{v})$:

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

VE s_t-step

$$r^s(\boldsymbol{s}_t|m_t) \propto p(\boldsymbol{x}_t|\boldsymbol{s}_t) \cdot \exp\left(\mathbb{E}_{r^z}\left[\log p(\boldsymbol{s}_t|\boldsymbol{z}_t, m_t; \boldsymbol{v}_t)\right]\right)$$

$$r^s(\boldsymbol{s}_t|m_t) = \mathcal{N}_c(\boldsymbol{\eta}_t^{m_t}, \text{diag}[\boldsymbol{\nu}_t^{m_t}]), \quad \begin{cases} \eta_{ft}^{m_t} = \frac{\gamma_{ft}^{m_t}}{\gamma_{ft}^{m_t} + (\mathbf{WH})_{ft}} \cdot x_{ft} \\ \nu_{ft}^{m_t} = \frac{\gamma_{ft}^{m_t} \cdot (\mathbf{WH})_{ft}}{\gamma_{ft}^{m_t} + (\mathbf{WH})_{ft}} \end{cases}$$

which can be interpreted is an averaged Wiener filtering. Also:

$$\gamma_{ft}^{m_t} = \left[\frac{1}{D}\sum_{d=1}^{D} \Sigma_{m_t,ff}^{-1}(\boldsymbol{z}_{m_t}^{(d)}, \boldsymbol{v}_t)\right]^{-1}$$

- $\Sigma_{m_t,ff}$ denotes the (f,f)-th entry of $\boldsymbol{\Sigma}_{m_t}$,
- $\{\boldsymbol{z}_{m_t}^{(d)}\}_{d=1}^{D}$ is a sequence sampled from $r^z(\boldsymbol{z}_t|m_t)$.

▷ The enhanced speech signal is the marginalisation over m_t:

$$\hat{\boldsymbol{s}}_t = \mathbb{E}_{r^m(m_t)}\left[\mathbb{E}_{r^s(\boldsymbol{s}_t|m_t)}[\boldsymbol{s}_t]\right] = \sum_{m_t} r^m(m_t)\boldsymbol{\eta}_t^{m_t}, \quad \forall t.$$

VE z_t-step

The set of parameters of $r^z(\boldsymbol{z}_t|m_t)$ is estimated by solving:

$$\max_{\boldsymbol{c}_{tm},\boldsymbol{\Omega}_{tm}} \mathbb{E}_{r^m(m_t)}\Big[\mathbb{E}_{r^z(\boldsymbol{z}_t|m_t)}\Big[\mathbb{E}_{r^s(\boldsymbol{s}_t|m_t)}\Big[\log p(\boldsymbol{s}_t|\boldsymbol{z}_t, m_t; \boldsymbol{v}_t)\Big]\Big]$$
$$- D_{\mathsf{KL}}(r^z(\boldsymbol{z}_t|m_t)\|p(\boldsymbol{z}_t|m_t;\boldsymbol{v}_t))\Big].$$

▷ Expectations over r^m and r^s, and the KL term can be evaluated in closed-form.

▷ Expectation over r^z is approximated with a single sample drawn from r^z.

▷ To back-propagate through the posterior parameters, the reparametrization trick is utilized

▷ A few iterations (of Adam optimizer) is enough for the convergence.

VE m_t-step

For $r^m(\boldsymbol{m})$, we obtain:

$$r^m(\boldsymbol{m}) \propto p(\boldsymbol{m}) \cdot \prod_{t=1}^{T} \exp(-g_t(m_t)) \tag{1}$$

with:

$$\begin{aligned} g_t(m_t) = &\mathbb{E}_{r^z}\Big[\mathrm{KL}(r^s(\boldsymbol{s}_t|m_t) \| p(\boldsymbol{s}_t|\boldsymbol{z}_t, m_t; \boldsymbol{v}_t))\Big] - \\ &\mathbb{E}_{r^s}\Big[\log p(\boldsymbol{x}_t|\boldsymbol{s}_t)\Big] + D_{\mathsf{KL}}(r^z(\boldsymbol{z}_t|m_t) \| p(\boldsymbol{z}_t|m_t; \boldsymbol{v}_t)) \end{aligned}$$

▷ Expectation over r^z is approximated by a Monte-Carlo estimate.

▷ To compute the marginal variational posterior $r^m(m_t)$, note that (1) has the same structure as standard HMM if we consider $\exp(-g_t(m_t))$ as the emission probability of the HMM.

→ We therefore use the forward-backward algorithm to compute $r^m(m_t)$.

M step

$\mathbf{W}$ and $\mathbf{H}$ are updated by optimizing the log-likelihood lower bound. Doing so, we obtain:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top \left(\mathbb{V} \odot (\mathbf{WH})^{\odot -2}\right)}{\mathbf{W}^\top (\mathbf{WH})^{\odot -1}},$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left(\mathbb{V} \odot (\mathbf{WH})^{\odot -2}\right) \mathbf{H}^\top}{(\mathbf{WH})^{\odot -1} \mathbf{H}^\top},$$

where $\mathbb{V} = \left[\sum_{m_t} r^m(m_t)(|x_{ft} - \eta_{ft}^{m_t}|^2 + \nu_{ft}^{m_t})\right]_{(f,t)}$, and $\odot$ signifies entry-wise operation.

▷ The parameters of the HMM, i.e. λ and τ, are updated by the standard formulae using the joint posterior probabilities computed by the forward-backward algorithm in the E-m step.

Experiments

Setup

- **Noisy+clean speech**: NTCD-TIMIT database [Abdelaziz, 2017]
  - Testing set of NTCD-TIMIT database;
  - $\sim$ 1 hour of speech;
  - 9 speakers;
  - Noise types: *LR*, *White*, *Cafe*, *Car*, *Babble*, and *Street*;
  - Noise levels: $\{-15, -10, -5, 0, 5, 10\}$ dB;
  - 270 noisy mixtures per noise level;
  - Different speakers and sentences than in the training set;
  - Clean lips region as well as noisy versions ($\sim$ 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]

Results

Objective measures (the higher, the better):

- Perceptual evaluation of speech quality (PESQ) measure in [-0.5,4.5],
- Signal-to-distortion ratio (SDR) in dB,
- Short-time objective intelligibility (STOI) in [0,1].

Results:

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**

Conclusion and future work

The proposed switching generative model provides a dynamic mechanism to make the performance robust with respect to noisy audio and visual data.

- The VEM framework is slow. Trying to re-use the trained encoders at inference time can reduce the complexity.
- Temporal modeling of the latent variables to benefit from time dependency between audio as well as visual frames.

References

1. D. P. Kingma and M. Welling, "*Auto-encoding variational Bayes*," ICLR, 2014.
2. Y. Bando et al., "*Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization*," in Proc. ICASSP, 2018, pp. 716–720
3. S. Leglaive et al., "*A variance modeling framework based on variational autoencoders for speech enhancement*," in Proc. MLSP, 2018.
4. 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.
5. 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.
6. C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag Berlin, Heidelberg, 2006.
7. A. H. Abdelaziz, "*NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition*," in Proc. INTERSPEECH, 2017.

Thank you for your attention