Robust Unsupervised Audio-visual Speech Enhancement Using a Mixture of Variational Autoencoders

Mostafa Sadeghi   Xavier Alameda-Pineda*

Perception team, Inria Grenoble Rhône-Alpes, France

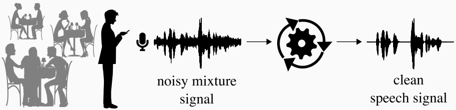
IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)

May 4-8, 2020

*Xavier Alameda-Pineda acknowledges ANR and the IDEX for funding the ML3RI project.

Introduction

Unsupervised speech enhancement



In the short-time Fourier transform (STFT) domain, for all $(f, n) \in \mathbb{B} = \{0, ..., F-1\} \times \{0, ..., N-1\}$, we observe:

$$x_{fn} = s_{fn} + b_{fn},$$

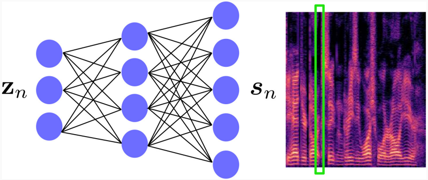
- s_{fn} is the clean speech signal.
- b_{fn} is the noise signal.
- f is the frequency index and n the time-frame index.

Separate the speech and noise signals from the observed mixture signal without training on noise.

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

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

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



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

Learning the parameters

- **Training dataset** of STFT speech time frames: $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N_{tr}-1}$
- **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 $\ln p(\mathbf{s};\boldsymbol{\theta})$:

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

where $q\left(\mathbf{z}_n|\mathbf{s}_n;\boldsymbol{\psi}\right) \approx p(\mathbf{z}_n|\mathbf{s}_n;\boldsymbol{\theta})$ is defined by an "encoding network" with parameters $\boldsymbol{\psi}$. $D_{\mathsf{KL}}(. \parallel .)$ is the Kullback–Leibler divergence.

Speech enhancement

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

Noise model: $\forall n: \quad \boldsymbol{b}_n \sim \mathcal{N}_c\left(\mathbf{0}, \text{diag}(\mathbf{W}_b \mathbf{H}_b[:, n])\right)$

Clean speech model: Trained VAE

▷ Observed variables: $\mathbf{x} = \{\mathbf{x}_n\}_{n=0}^{N-1}$. Latent variables: $\mathbf{z} = \left\{\mathbf{z}_n\right\}_{n=0}^{N-1}$

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

Monte-Carlo Expectation maximization (MCEM):

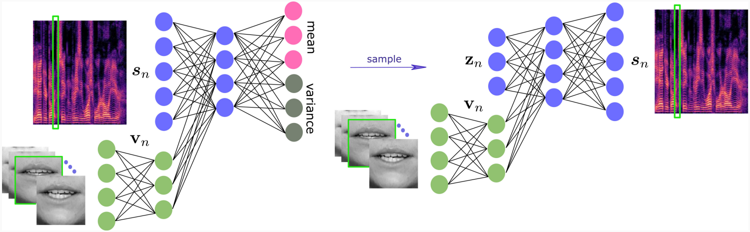
- **E-Step**: $Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)}[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \boldsymbol{\theta}_u)]$.
- **M-Step**: $\boldsymbol{\theta}_u^\star \leftarrow \arg\max_{\boldsymbol{\theta}_u}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

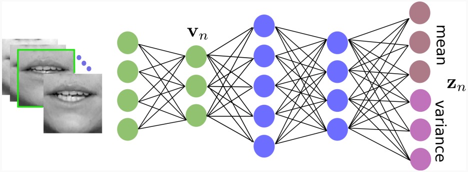
Speech estimation:

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

Audio-visual modeling of clean speech [Sadeghi et al., 2019]

- Generative model: $p(\boldsymbol{s}_n|\mathbf{z}_n, \mathbf{v}_n) = \mathcal{N}_c\Big(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}_s^{av}(\mathbf{z}_n, \mathbf{v}_n))\Big)$
- Encoder: $q(\mathbf{z}_n|\mathbf{s}_n, \mathbf{v}_n; \boldsymbol{\psi}) = \mathcal{N}\Big(\boldsymbol{\mu}_z^{av}(\mathbf{s}_n, \boldsymbol{v}_n), \text{diag}(\boldsymbol{\sigma}_z^{av}(\mathbf{s}_n, \mathbf{v}_n))\Big)$
- Prior of $\mathbf{z}_n$: $p(\mathbf{z}_n|\mathbf{v}_n) = \mathcal{N}\Big(\boldsymbol{\mu}_z(\boldsymbol{v}_n), \text{diag}(\boldsymbol{\sigma}_z(\mathbf{v}_n))\Big)$





Robust Audio-visual Speech Enhancement

AV-VAE usually yields better results than A-VAE, especially at low SNRs, provided clean (frontal, non-occluded) visual data [Sadeghi et al., 2019].

Noisy visual data:

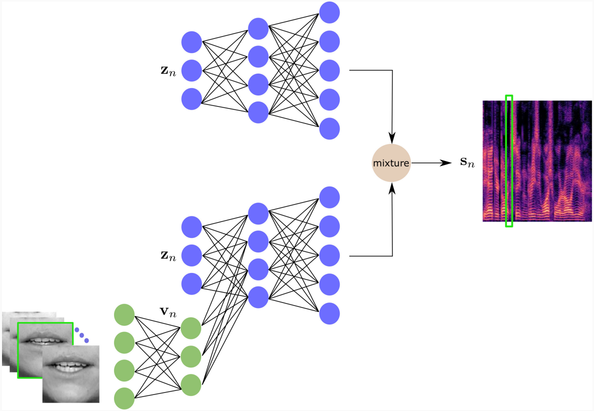
Some video frames might contain occluded and/or non-frontal lips region.

How to effectively benefit from AV-VAE for in-the-wild video recordings?

Our work: VAE mixture model (VAE-MM)

A mixture of A-VAE plus AV-VAE generative model:

- If the lips region is clean, use AV-VAE, otherwise use A-VAE.



The A-VAE and AV-VAE have been already trained on clean data.

Generative model

Mixture generative model: Combine A-VAE with AV-VAE

$$
\begin{cases}
p(\mathbf{s}_n|\mathbf{z}_n, \mathbf{v}_n, \alpha_n) & = \left[\mathcal{N}_c\Big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_s^a(\mathbf{z}_n))\Big)\right]^{\alpha_n} \times \left[\mathcal{N}_c\Big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_s^{av}(\mathbf{z}_n, \mathbf{v}_n)))\Big)\right]^{1-\alpha_n} \\
p(\mathbf{z}_n|\mathbf{v}_n, \alpha_n) & = \left[\mathcal{N}(\mathbf{0}, \mathbf{I})\right]^{\alpha_n} \times \left[\mathcal{N}\Big(\boldsymbol{\mu}_z^v(\boldsymbol{v}_n), \mathrm{diag}(\boldsymbol{\sigma}_z^v(\mathbf{v}_n))\Big)\right]^{1-\alpha_n}, \\
p(\alpha_n) & = \pi^{\alpha_n} \times (1-\pi)^{1-\alpha_n}.
\end{cases}
$$

$\alpha_n \in \{0, 1\}$ is a latent variable specifying the component of the mixture model that is used by the n-th frame.

Parameter estimation

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

Noise model: $\forall n: \quad \boldsymbol{b}_n \sim \mathcal{N}_c\Big(\mathbf{0}, \text{diag}(\mathbf{W}_b\mathbf{H}_b[:, n])\Big)$

Clean speech model: Trained A-VAE and AV-VAE generative networks

---

Inference:

- Observed variables: $\{\mathbf{x}_n, \mathbf{v}_n\}_{n=0}^{N-1}$
- Latent variables: $\big\{\mathbf{s}_n, \mathbf{z}_n, \alpha_n\big\}_{n=0}^{N-1}$
- Parameters to be estimated: $\boldsymbol{\theta}_u = \{\mathbf{W}_b, \mathbf{H}_b, \pi\}$

Parameter estimation

Variational Expectation-maximization (VEM)

Variational E-Step:

The intractable posterior $p(\mathbf{s}_n, \mathbf{z}_n, \alpha_n | \mathbf{x}_n, \mathbf{v}_n; \boldsymbol{\theta}_u)$ is approximated by a variational distribution factorizing as follows:

$$r(\mathbf{s}_n, \mathbf{z}_n, \alpha_n) = r(\mathbf{s}_n)\, r(\mathbf{z}_n)\, r(\alpha_n),$$

which are updated as follows [Bishop, 2006]:

VE $\mathbf{s}_n$-step: $$r(\mathbf{s}_n) \propto \exp\left(\mathbb{E}_{r(\mathbf{z}_n)\cdot r(\alpha_n)}\Big[\log p(\mathbf{x}_n, \mathbf{s}_n, \mathbf{z}_n, \alpha_n, \mathbf{v}_n; \boldsymbol{\theta}_u)\Big]\right)$$

VE $\mathbf{z}_n$-step: $$r(\mathbf{z}_n) \propto \exp\left(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\alpha_n)}\Big[\log p(\mathbf{x}_n, \mathbf{s}_n, \mathbf{z}_n, \alpha_n, \mathbf{v}_n; \boldsymbol{\theta}_u)\Big]\right)$$

VE α_n-step: $$r(\alpha_n) \propto \exp\left(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)}\Big[\log p(\mathbf{x}_n, \mathbf{s}_n, \mathbf{z}_n, \alpha_n, \mathbf{v}_n; \boldsymbol{\theta}_u)\Big]\right)$$

VE s_n-step

$$r(\mathbf{s}_n) = \mathcal{N}_c(\boldsymbol{m}_n, \mathrm{diag}(\boldsymbol{\nu}_n)), \quad \begin{cases} m_{fn} & = \frac{\gamma_{fn}}{\gamma_{fn} + (\mathbf{W}_b\mathbf{H}_b)_{fn}} \cdot x_{fn} \\ \nu_{fn} & = \frac{\gamma_{fn} \cdot (\mathbf{W}_b\mathbf{H}_b)_{fn}}{\gamma_{fn} + (\mathbf{W}_b\mathbf{H}_b)_{fn}} \end{cases}$$

which can be interpreted is an averaged Wiener filtering.

$\gamma_{fn}^{-1} = \sum_{\alpha_n \in \{0,1\}} r(\alpha_n) \cdot \eta_{fn}^{\alpha_n}$ (weighted precision over audio and audio-visual cases),

$\eta_{fn}^{\alpha_n} = \mathbb{E}_{r(\mathbf{z}_n)}\left[\frac{1}{\sigma_{s,f}^{\alpha_n}(\mathbf{z}_n, \mathbf{v}_n)}\right] \approx \frac{1}{D}\sum_{d=1}^{D} \frac{1}{\sigma_{s,f}^{\alpha_n}(\mathbf{z}_n^{(d)}, \mathbf{v}_n)}$ (average precision),

and $\{\mathbf{z}_n^{(d)}\}_{d=1}^{D}$ is a sequence sampled from $r(\mathbf{z}_n)$. Moreover:

$$\sigma_{s,f}^{\alpha_n}(\mathbf{z}_n, \mathbf{v}_n) = \begin{cases} \sigma_{s,f}^{a}(\mathbf{z}_n) & \alpha_n = 1 \\ \sigma_{s,f}^{av}(\mathbf{z}_n, \mathbf{v}_n) & \alpha_n = 0 \end{cases}.$$

VE z_n-step

For $r(\mathbf{z}_n)$ we obtain the following result:

$$r(\mathbf{z}_n) \propto \exp\Big(\sum_{\alpha_n \in \{0,1\}} r(\alpha_n) \cdot \Big[\log p(\mathbf{z}_n | \mathbf{v}_n, \alpha_n) + \sum_f -\log\Big(\sigma_{s,f}^{\alpha_n}(\mathbf{z}_n, \mathbf{v}_n)\Big) - \frac{|m_{fn}|^2 + \nu_{fn}}{\sigma_{s,f}^{\alpha_n}(\mathbf{z}_n, \mathbf{v}_n)} \Big] \Big).$$

The above distribution cannot be computed in closed-from. Nevertheless, we can draw samples from it using the Metropolis-Hastings (MH) algorithm (see our paper for more details).

VE α_n-step

To update the variational distribution of α_n, we can write:

$$r(\alpha_n) \propto \exp\Big(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)}\Big[\log p(\mathbf{s}_n|\mathbf{z}_n, \mathbf{v}_n, \alpha_n) + \log p(\mathbf{z}_n|\mathbf{v}_n, \alpha_n) + \log p(\alpha_n)\Big]\Big)$$

which is a Bernoulli distribution with

$$\pi_n = g\Big(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)}\Big[\log \frac{p(\mathbf{s}_n, \mathbf{z}_n|\mathbf{v}_n, \alpha_n = 1)}{p(\mathbf{s}_n, \mathbf{z}_n|\mathbf{v}_n, \alpha_n = 0)}\Big] + \log \frac{\pi}{1-\pi}\Big)$$

as the parameter, which is an averaged audio/audio-visual ratio. Here, $g(.)$ denotes the sigmoid function defined as $g(x) = 1/(1 + \exp(-x))$.

Parameters update and speech enhancement

M-Step:

Update parameters by optimizing the complete data log-likelihood:

$$
\begin{aligned}
Q(\boldsymbol{\theta}_u, \boldsymbol{\theta}_u^{\text{old}}) &\stackrel{c}{=} \mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)\cdot r(\alpha_n)}\Big[\log p(\mathbf{x}_n, \mathbf{s}_n, \mathbf{z}_n, \alpha_n, \mathbf{v}_n; \boldsymbol{\theta}_u)\Big] \\
&\stackrel{c}{=} \mathbb{E}_{r(\mathbf{s}_n)}\Big[\log p(\mathbf{x}_n|\mathbf{s}_n; \boldsymbol{\theta}_u)\Big] + \mathbb{E}_{r(\alpha_n)}\Big[\log p(\alpha_n)\Big] \\
&\stackrel{c}{=} \sum_{f,n} -\log\left(\mathbf{W}_b\mathbf{H}_b\right)_{fn} - \left(\frac{|x_{fn} - m_{fn}|^2 + \nu_{fn}}{\left(\mathbf{W}_b\mathbf{H}_b\right)_{fn}}\right) + \pi_n \log \pi + (1-\pi_n)
\end{aligned}
$$

Speech enhancement:

After the convergence of the VEM, the speech STFT frames are estimated using an averaged Wiener filtering:

$$
\hat{s}_{fn} = \mathbb{E}_{r(s_{fn})}[s_{fn}] = \frac{\gamma^*_{fn}}{\gamma^*_{fn} + \left(\mathbf{W}^*_b\mathbf{H}^*_b\right)_{fn}} \cdot x_{fn} \quad \forall (f,n)
$$

Experiments

Dataset

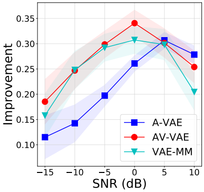
- **Noisy+clean speech**: NTCD-TIMIT database [Abdelaziz, 2017]
- **VAE models**: Pre-trained A-VAE and AV-VAE [Sadeghi et al., 2019]
- **Setup**:
  - Testing set of NTCD-TIMIT database;
  - $\sim$ 1 hours of speech;
  - 9 speakers;
  - Noise types: *LR*, *White*, *Cafe*, *Car*, *Babble*, and *Street*;
  - 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)

Results

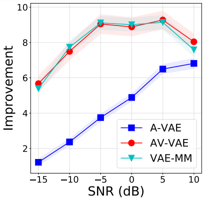
Objective measures (the higher, the better)

- Signal-to-distortion ratio (SDR).
- Perceptual evaluation of speech quality (PESQ) measure.

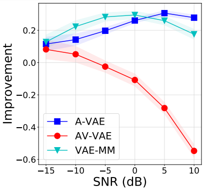
Improvement with respect to the input:



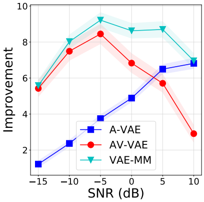
(a) PESQ(clean)



(b) SDR(clean)



(c) PESQ(noisy)



(d) SDR(noisy)

Conclusion and future work

The proposed robust technique can efficiently benefit from visual data for speech enhancement when some lips region frames are noisy (non-frontal, occluded).

- 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.

Thank you for your attention

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*," `https://arxiv.org/abs/1908.02590`, August 2019.
5. C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag Berlin, Heidelberg, 2006.
6. A. H. Abdelaziz, "*NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition*," in Proc. INTERSPEECH, 2017.