Semi-Supervised Multichannel Speech Enhancement with Variational Autoencoders and Non-Negative Matrix Factorization

Simon LEGLAIVE[1] Laurent GIRIN[1,2] Radu HORAUD[1]

1: Inria Grenoble Rhône-Alpes 2: Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab

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

May 15, 2019 – Bighton, UK

Introduction

Multichannel speech enhancement

Multichannel speech enhancement

Semi-supervised approach:

- ◇ Training from clean speech signals only.
- ◇ Free of generalization issues regarding the noisy recording environment.

We want the method to be speaker-independent.

Speech enhancement as a source separation problem

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

$$\mathbf{x}_{fn} = \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{1}$$

- $\mathbf{s}_{fn} \in \mathbb{C}^I$ is the clean speech signal.
- $\mathbf{b}_{fn} \in \mathbb{C}^I$ is the noise signal.
- f is the frequency index and n the time-frame index.
- I is the number of microphones.

Objective

Separate the speech and noise signals from the observed mixture signal.

Multichannel local Gaussian model (Vincent et al. 2010; Duong et al. 2010)

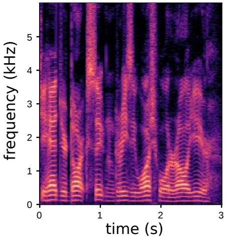
Local Gaussian model: Independently for all $(f, n) \in \mathbb{B}$,

$$\mathbf{s}_{fn} \sim \mathcal{N}_c(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{s},fn}) \qquad \text{and} \qquad \mathbf{b}_{fn} \sim \mathcal{N}_c(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b},fn}). \tag{2}$$

Covariance matrix model:

$$\boldsymbol{\Sigma}_{\mathbf{j},fn} = v_{j,fn} \times \mathbf{R}_{\mathbf{j},f}, \qquad j \in \{s, b\}. \tag{3}$$

$v_{j,fn}$ is the short-term power spectral density



$\mathbf{R}_{\mathbf{j},f}$ is the spatial covariance matrix.

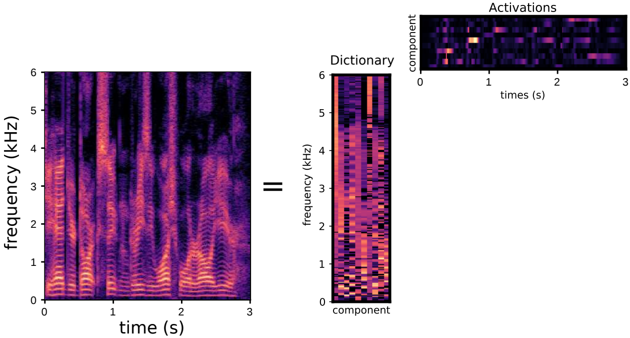
It encodes spatial cues and room properties.

Spectral modeling with non-negative matrix factorization (NMF)

NMF-based spectro-temporal model (Arberet et al. 2010):

$$v_{j,fn} = (\mathbf{W}_j \mathbf{H}_j)_{f,n}, \qquad j \in \{s, b\}, \tag{4}$$

- $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ is a dictionary matrix of spectral templates.
- $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$ is the activation matrix.
- K_j is the rank of the factorization (usually $K_j(F + N) \ll FN$).



Semi-supervised setting (Smaragdis et al. 2007)

- **Training**: Learn $\mathbf{W}_s$ from a dataset of clean speech signals.

$$\min_{\mathbf{W}_s \in \mathbb{R}_+^{F \times K_s}} \sum_{(f,n) \in \mathbb{B}} d_{\mathrm{IS}}\left(|s_{fn}|^2,\; v_{s,fn} = (\mathbf{W}_s \mathbf{H}_s)_{f,n}\right), \tag{5}$$

where $d_{\mathrm{IS}}(\cdot,\cdot)$ is the Itakura-Saito (IS) divergence (Févotte et al. 2009).

- **Test**: Estimate the remaining speech and noise model parameters from the noisy mixture signal.

Semi-supervised setting (Smaragdis et al. 2007)

- **Training**: Learn $\mathbf{W}_s$ from a dataset of clean speech signals.

$$\min_{\mathbf{W}_s \in \mathbb{R}_+^{F \times K_s}} \sum_{(f,n)\in\mathbb{B}} d_{\text{IS}}\left(|s_{fn}|^2,\ v_{s,fn} = (\mathbf{W}_s\mathbf{H}_s)_{f,n}\right), \tag{5}$$

where $d_{\text{IS}}(\cdot,\cdot)$ is the Itakura-Saito (IS) divergence (Févotte et al. 2009).

- **Test**: Estimate the remaining speech and noise model parameters from the noisy mixture signal.

In this work, we explore the use of neural networks as an alternative to this supervised NMF-based variance model.

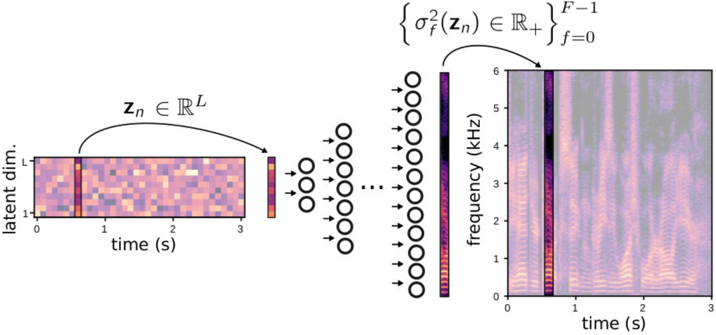
Deep generative speech model

Single-channel deep generative speech model (Bando et al. 2018)

Independently for all $(f, n) \in \mathbb{B}$,

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(0, \sigma_f^2(\mathbf{z}_n)\right), \qquad \text{with } \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \tag{6}$$

and $\sigma_f^2 : \mathbb{R}^L \mapsto \mathbb{R}_+$ corresponds to a neural network of parameters $\boldsymbol{\theta}_s$.

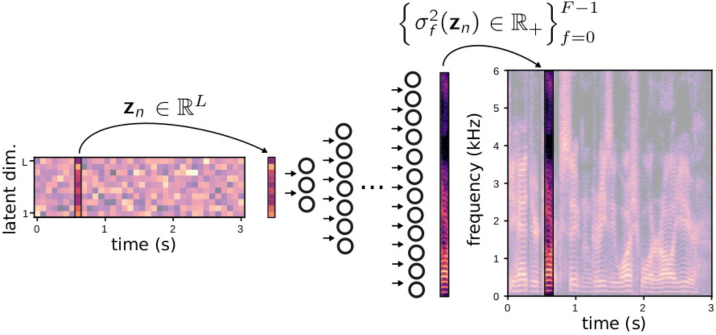


Single-channel deep generative speech model (Bando et al. 2018)

Independently for all $(f, n) \in \mathbb{B}$,

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(0, \sigma_f^2(\mathbf{z}_n)\right), \qquad \text{with } \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \tag{6}$$

and $\sigma_f^2 : \mathbb{R}^L \mapsto \mathbb{R}_+$ corresponds to a neural network of parameters $\boldsymbol{\theta}_s$.



How to learn the parameters $\boldsymbol{\theta}_s$ of this generative neural network?

Learning the model parameters with variational autoencoders

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

Learning the model parameters with variational autoencoders

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

Taking ideas from variational inference, maximize a lower bound of $\ln p(\mathbf{s}; \boldsymbol{\theta}_s)$, which can be recast as:

$$\min_{\boldsymbol{\theta}_s} \sum_{(f,n)\in\mathbb{B}} \mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)} \left[d_{IS}\left(|s_{fn}|^2 ; \sigma_f^2(\mathbf{z}_n) \right) \right], \tag{7}$$

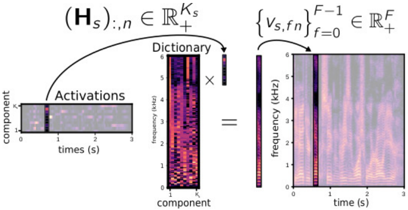
where $q(\mathbf{z}_n|\mathbf{s}_n;\phi)$ is an approximation of $p(\mathbf{z}_n|\mathbf{s}_n;\boldsymbol{\theta}_s)$ and is defined by an "encoding network" of parameters ϕ (see paper for more details).

NMF- vs VAE-based spectro-temporal speech modeling

NMF-based model

$$v_{s,fn} = (\mathbf{W}_s\mathbf{H}_s)_{f,n} = (\mathbf{W}_s)_{f,:}^\top (\mathbf{H}_s)_{:,n}$$

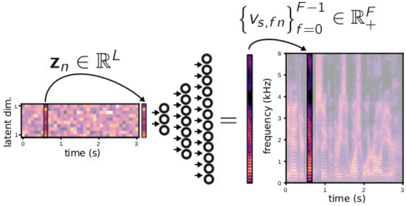
- ▷ linear function of $(\mathbf{H}_s)_{:,n} \in \mathbb{R}_+^{K_s}$.
- ▷ $\#$ trainable parameters $= F \times K_s$.
- ▷ IS divergence minimization.
- ▷ Interpretability.



VAE-based model

$$v_{s,fn} = \sigma_f^2(\mathbf{z}_n)$$

- ▷ non-linear function of $\mathbf{z}_n \in \mathbb{R}^L$.
- ▷ $\#$ trainable parameters is free.
- ▷ IS divergence minimization.
- ▷ Lack of (direct) interpretability.



Multichannel speech enhancement

Models for semi-supervised multichannel speech enhancement

Supervised multichannel speech model

$$\mathbf{s}_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(\mathbf{0}, \sigma_f^2(\mathbf{z}_n)\mathbf{R}_{\mathbf{s},f}\right), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \tag{8}$$

where $\sigma_f^2(\cdot)$ was trained during the training stage.

Models for semi-supervised multichannel speech enhancement

Supervised multichannel speech model

$$\mathbf{s}_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(\mathbf{0}, \sigma_f^2(\mathbf{z}_n)\mathbf{R}_{\mathbf{s},f}\right), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \tag{8}$$

where $\sigma_f^2(\cdot)$ was trained during the training stage.

Unsupervised multichannel noise model

$$\mathbf{b}_{fn} \sim \mathcal{N}_c\left(\mathbf{0}, (\mathbf{W}_b\mathbf{H}_b)_{f,n}\mathbf{R}_{\mathbf{b},f}\right), \tag{9}$$

where $\mathbf{W}_b \in \mathbb{R}_+^{F\times K_b}$ and $\mathbf{H}_b \in \mathbb{R}_+^{K_b\times N}$.

Mixture model

$$\mathbf{x}_{fn} = \sqrt{g_n}\mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{10}$$

where $g_n \in \mathbb{R}_+$ is a gain parameter (Leglaive et al. 2018).

Unsupervised model parameters estimation

Likelihood

$$\mathbf{x}_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(\mathbf{0}, g_n \sigma_f^2(\mathbf{z}_n)\mathbf{R}_{\mathbf{s},f} + (\mathbf{W}_b \mathbf{H}_b)_{f,n} \mathbf{R}_{\mathbf{b},f}\right). \tag{11}$$

- Unsupervised model parameters to be estimated:

$$\boldsymbol{\theta}_u = \left\{\mathbf{W}_b, \mathbf{H}_b, \mathbf{R}_{\mathbf{s},f}, \mathbf{R}_{\mathbf{b},f}, \mathbf{g} = [g_0, ..., g_{N-1}]^\top\right\}.$$

- Intractable marginal likelihood:

$$p(\mathbf{x}_{fn}; \boldsymbol{\theta}_u) = \int p(\mathbf{x}_{fn} | \mathbf{z}_n; \boldsymbol{\theta}_u) p(\mathbf{z}_n) d\mathbf{z}_n. \tag{12}$$

Unsupervised model parameters estimation

Likelihood

$$\mathbf{x}_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(\mathbf{0}, g_n \sigma_f^2(\mathbf{z}_n)\mathbf{R}_{\mathbf{s},f} + (\mathbf{W}_b\mathbf{H}_b)_{f,n}\mathbf{R}_{\mathbf{b},f}\right). \qquad (11)$$

▷ Unsupervised model parameters to be estimated:

$$\boldsymbol{\theta}_u = \left\{\mathbf{W}_b, \mathbf{H}_b, \mathbf{R}_{\mathbf{s},f}, \mathbf{R}_{\mathbf{b},f}, \mathbf{g} = [g_0, ..., g_{N-1}]^\top\right\}.$$

▷ Intractable marginal likelihood:

$$p(\mathbf{x}_{fn}; \boldsymbol{\theta}_u) = \int p(\mathbf{x}_{fn}|\mathbf{z}_n; \boldsymbol{\theta}_u)p(\mathbf{z}_n)d\mathbf{z}_n. \qquad (12)$$

▷ Expectation-maximization (EM) algorithm.

Observed data:

$$\mathbf{x} = \left\{\mathbf{x}_{fn} \in \mathbb{C}^I\right\}_{(f,n)\in\mathbb{B}}$$

Latent data:

$$\mathbf{z} = \left\{\mathbf{z}_n \in \mathbb{R}^L\right\}_{n=0}^{N-1}$$

Monte Carlo EM algorithm and speech estimation

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) \quad = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)} \left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_u)\right]$$

Monte Carlo EM algorithm and speech estimation

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$\begin{aligned} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)} \left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_u)\right] \\ &\approx \frac{1}{R} \sum_{r=1}^{R} \ln p\left(\mathbf{x}, \mathbf{z}^{(r)}; \boldsymbol{\theta}_u\right), \end{aligned} \tag{13}$$

where the samples $\left\{\mathbf{z}^{(r)}\right\}_{r=1,\dots,R}$ are i.i.d. and asymptotically drawn from $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_u^\star)$ using a Markov chain Monte Carlo method.

Monte Carlo EM algorithm and speech estimation

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$\begin{aligned} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)} \left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_u)\right] \\ &\approx \frac{1}{R} \sum_{r=1}^{R} \ln p\left(\mathbf{x}, \mathbf{z}^{(r)}; \boldsymbol{\theta}_u\right), \end{aligned} \tag{13}$$

  where the samples $\left\{\mathbf{z}^{(r)}\right\}_{r=1,\dots,R}$ are i.i.d. and asymptotically drawn from $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_u^\star)$ using a Markov chain Monte Carlo method.

- **M-Step.**

$$\boldsymbol{\theta}_u^\star \leftarrow \underset{\boldsymbol{\theta}_u}{\arg\max} \;\; Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star). \tag{14}$$

  Minimize-majorize approach similar to (Sawada et al. 2013).

Monte Carlo EM algorithm and speech estimation

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$\begin{aligned} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u^\star)}\left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_u)\right] \\ &\approx \frac{1}{R}\sum_{r=1}^{R} \ln p\left(\mathbf{x}, \mathbf{z}^{(r)}; \boldsymbol{\theta}_u\right), \end{aligned} \quad (13)$$

where the samples $\left\{\mathbf{z}^{(r)}\right\}_{r=1,\ldots,R}$ are i.i.d. and asymptotically drawn from $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_u^\star)$ using a Markov chain Monte Carlo method.

- **M-Step.**

$$\boldsymbol{\theta}_u^\star \leftarrow \arg\max_{\boldsymbol{\theta}_u} \; Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star). \quad (14)$$

Minimize-majorize approach similar to (Sawada et al. 2013).

- Posterior mean speech estimate with multichannel Wiener-like filtering.

Experiments

Dataset

- **Clean speech signals**: TIMIT database.
- **Noise signals**: DEMAND database (domestic environment, nature, office, indoor public spaces, street and transportation).
- **Training**:
  - training set of TIMIT database;
  - $\sim$ 4 hours of speech;
  - 462 speakers.
- **Test**:
  - 168 stereo noisy mixtures at 0 dB signal-to-noise ratio;
  - Different speakers and sentences than in the training set.

Semi-supervised baseline method (Sawada et al. 2013)

Supervised multichannel speech model

$$\mathbf{s}_{fn} \sim \mathcal{N}_c\left(\mathbf{0}, (\mathbf{W_s H_s})_{f,n} \mathbf{R}_{\mathbf{s},f}\right), \tag{15}$$

where $\mathbf{W_s} \in \mathbb{R}_+^{F \times K_s}$ is learned during the training stage.

Unsupervised multichannel noise model

$$\mathbf{b}_{fn} \sim \mathcal{N}_c\left(\mathbf{0}, (\mathbf{W}_b \mathbf{H}_b)_{f,n} \mathbf{R}_{\mathbf{b},f}\right). \tag{16}$$

Test time: Maximum-likelihood estimation of the unsupervised model parameters and multichannel Wiener filtering.

Results

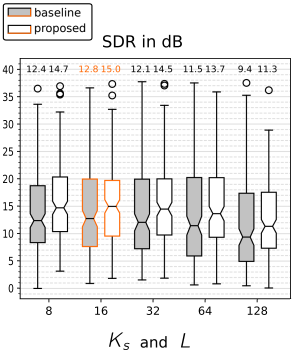
Objective measures (the higher, the better)

- ▷ Signal-to-distortion ratio (SDR).
- ▷ Perceptual evaluation of speech quality (PESQ) measure.
- ▷ Short-time objective intelligibility (STOI) measure.

Results

Objective measures (the higher, the better)

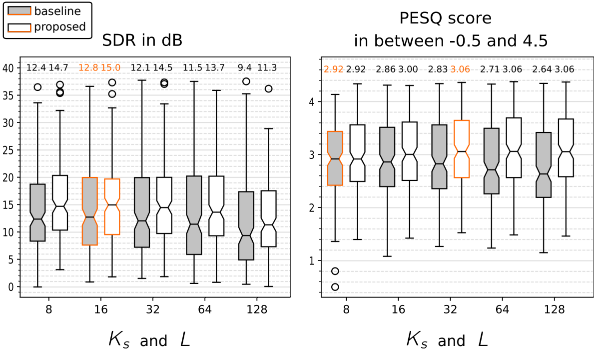
- ▷ Signal-to-distortion ratio (SDR).
- ▷ Perceptual evaluation of speech quality (PESQ) measure.
- ▷ Short-time objective intelligibility (STOI) measure.



Results

Objective measures (the higher, the better)

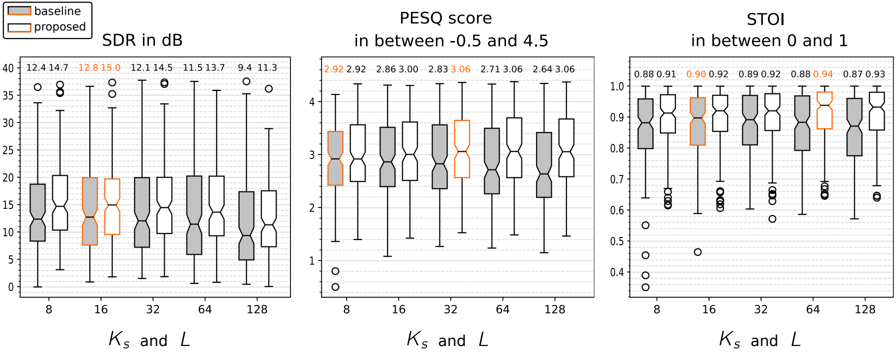
- ▷ Signal-to-distortion ratio (SDR).
- ▷ Perceptual evaluation of speech quality (PESQ) measure.
- ▷ Short-time objective intelligibility (STOI) measure.



Results

Objective measures (the higher, the better)

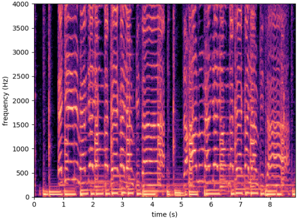
- ▷ Signal-to-distortion ratio (SDR).
- ▷ Perceptual evaluation of speech quality (PESQ) measure.
- ▷ Short-time objective intelligibility (STOI) measure.



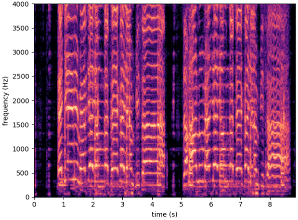
Singing voice separation in a stereo mixture

- ▷ VAE model trained on speaking and not singing voice.
- ▷ Unsupervised noise model $\rightarrow$ flexibility.

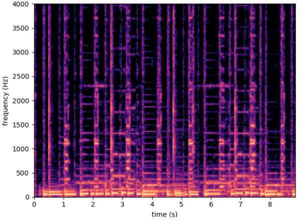
Mixture



Estimated voice



Estimated accompaniment



Song: "Ana" by Vieux Farka Toure

Conclusion

Conclusion

For a semi-supervised multichannel speech enhancement application, VAE-based generative speech models are an interesting alternative to supervised NMF models.

Limitations and future work:

- MCEM algorithm is slow ($\sim 7\times$ slower than the baseline method).
- Variational EM algorithm.
- Temporal modeling of the latent variables.

Thank you for your attention

Audio examples and code:

`https://sleglaive.github.io`

References

Arberet, S., A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst (2010). "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation". In: *Proc. IEEE Int. Conf. Information Sciences, Signal Processing and their Applications (ISSPA)*.

Bando, Y., M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara (2018). "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*.

Duong, N. Q. K., E. Vincent, and R. Gribonval (2010). "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 18.7.

Févotte, C., N. Bertin, and J.-L. Durrieu (2009). "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis". In: *Neural computation* 21.3.

Kingma, D. P. and M. Welling (2014). "Auto-encoding variational Bayes". In: *Proc. Int. Conf. Learning Representations (ICLR)*.

Leglaive, S., L. Girin, and R. Horaud (2018). "A variance modeling framework based on variational autoencoders for speech enhancement". Proc. IEEE Int. Workshop Machine Learning Signal Process. (MLSP).

Sawada, H., H. Kameoka, S. Araki, and N. Ueda (2013). "Multichannel extensions of non-negative matrix factorization with complex-valued data". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 21.5.

Smaragdis, P., B. Raj, and M. Shashanka (2007). "Supervised and semi-supervised separation of sounds from single-channel mixtures". In: *Proc. Int. Conf. Indep. Component Analysis and Signal Separation*.

Vincent, E., M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies (2010). "Probabilistic modeling paradigms for audio source separation". In: *Machine Audition: Principles, Algorithms and Systems*. Ed. by W. Wang. IGI Global.