Prof. Dr.-Ing. Timo Gerkmann

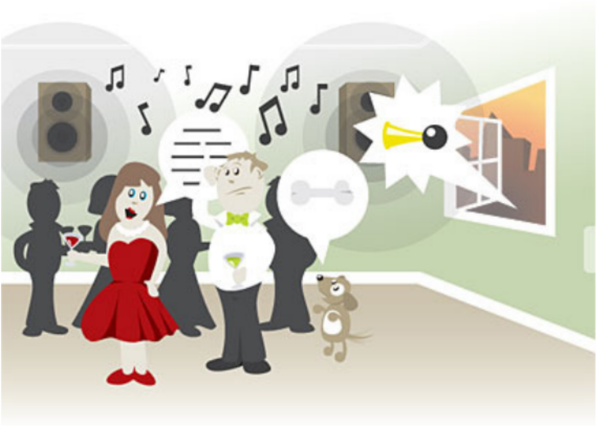
Statistical Signal Processing and Machine Learning for Speech Enhancement

Universität Hamburg
Department of Informatics
Signal Processing (SP)

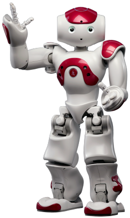
Sepember 30, 2021

- Speech communication disturbed by external noise sources

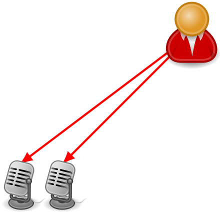
➡ Make information more easily accessible by humans and machines



Speech Communication Devices

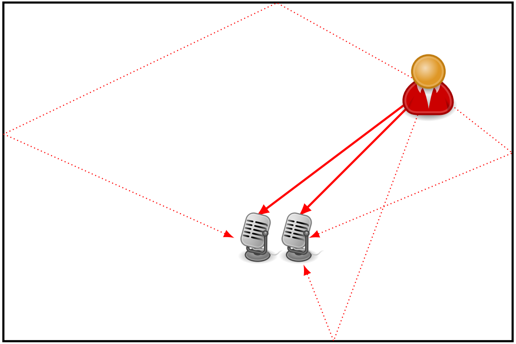


Signal Model

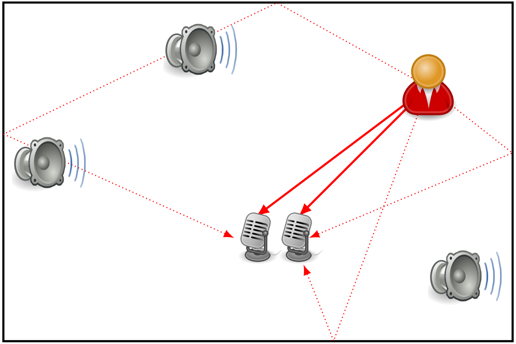


- Signal model: $y_m(t) = s_m(t)$

Signal Model



- Signal model: $y_m(t) = s(t) * h_m(t)$
- Conversation disturbed by
  - Reflections from the walls

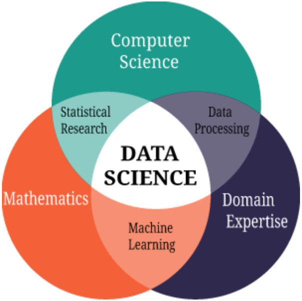


- Signal model: $y_m(t) = s(t) * h_m(t) + \sum_{i=1}^{I} n_{i,m}(t)$
- Conversation disturbed by
  - Reflections from the walls
  - Additive noise

➡ Signal model generalizes many data acquisition challenges

Research Methods

- Combine **statistical** methods, **machine learning** and **domain knowledge**
- Domain knowledge includes perceptive models, signal production models, and physical models.
- Practical constraints must be taken into account (complexity, storage, latency)



➡ Interdisciplinary exchange necessary

Overview

1. Single Channel Source Separation
2. Variational Autoencoders (VAEs) for Speech Enhancement
   - Conditional Variational Autoencoder for Speech Enhancement
   - Speech Enhancement with Stochastic Temporal Convolutional Networks
3. Nonlinear Multichannel Filtering

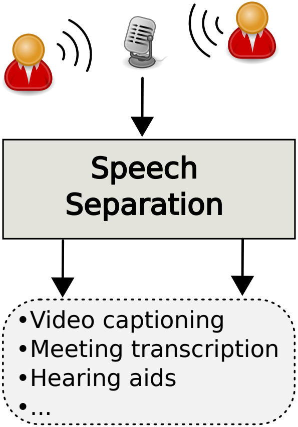
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

Single Channel Source Separation

David Ditter, Timo Gerkmann. "Influence of Speaker-Specific Parameters on Speech Separation Systems", ISCA Interspeech, Graz, Austria, Sep. 2019.

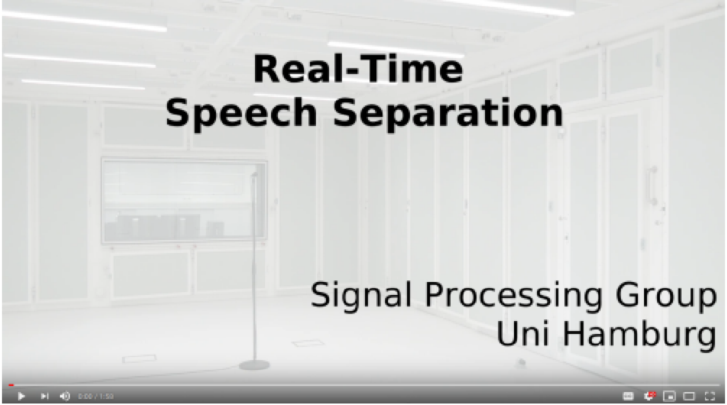
David Ditter, Timo Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation via TasNet", IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Barcelona, Spain, May 2020



Conditions:

- Undefined number of speakers
- Unknown speakers
- Single microphone

Real-Time Demo in Our Lab



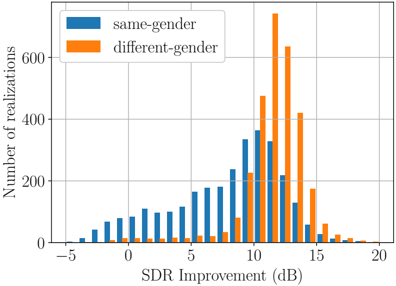
[1] [2]

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

[2] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," en, in *ISCA Interspeech*, Graz, Austria, Sep. 2019, pp. 4584–4588. [Online]. Available: `http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2459.html` (visited on 09/16/2019).

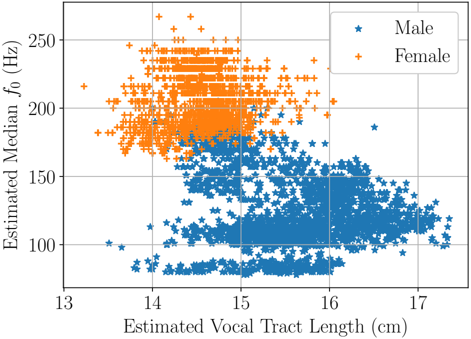
Problems with Speaker Constellations

- ✔ Shown system[3] shows good *average* performance
- ✖ But: Performance varies for certain speaker constellations



[3] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 686–690.

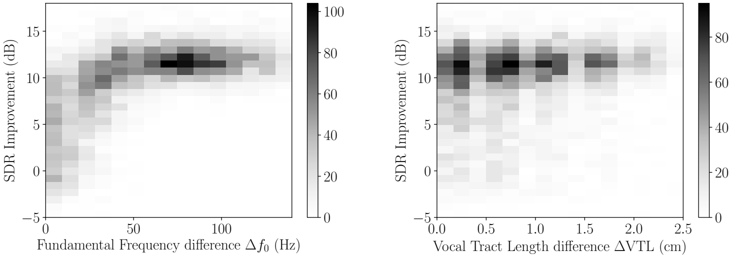
Which Speech Parameters are Gender-Specific?



➡ On average, females have shorter vocal tract lengths and higher fundamental frequencies

- Vocal tract length: Changes spectral envelope
- Fundamental frequency: Changes spectral fine structure

Is it the Pitch or the Vocal Tract Length?



- Measure performance for speaker pairs as a function of the difference in pitch and vocal tract length

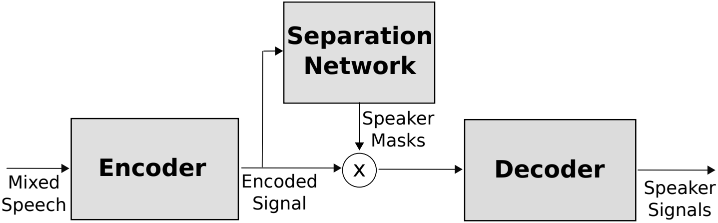
➡ Fundamental frequency difference Δf_0 is the dominant factor to predict separation quality[2]

➡ For speaker pairs with close f_0, source separation may be harmful

[2] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," en, in *ISCA Interspeech*, Graz, Austria, Sep. 2019, pp. 4584–4588. [Online]. Available: `http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2459.html` (visited on 09/16/2019).

Conv-TasNet Approach

Conv-TasNet approach[4] [1]

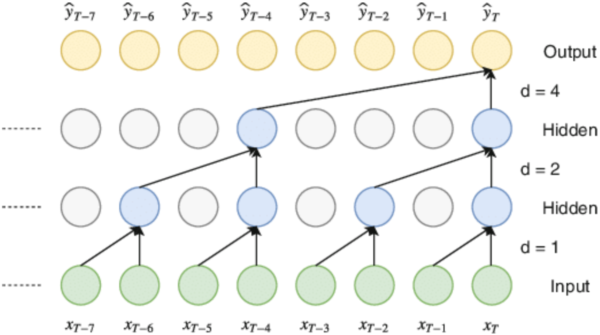


- Encoder and decoder are learned convolutional layers (i.e. filterbanks)
- Algorithmic latency defined by encoder window size
- Filterbank windows can be very small (e.g. ≤ 2 ms)
- Receptive field around 1 to 2 s

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

Separation Network

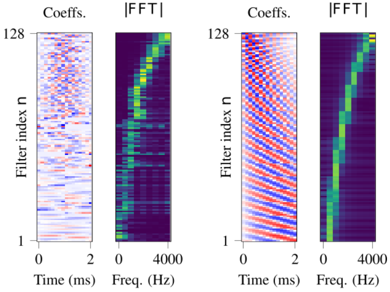


- Separation network is fully convolutional including non-linearites
- Use of dilated convolutions to enlarge receptive field
- Use of skip connections for easier training[4]

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

Learned Filterbank: Key to TasNet's Success?

Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)[1]:

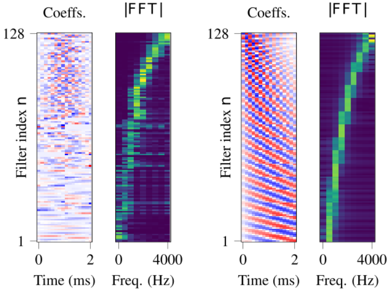


(a) Learned Filterbank (b) Proposed MP-GTF

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

Learned Filterbank: Key to TasNet's Success?

Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)[1]:



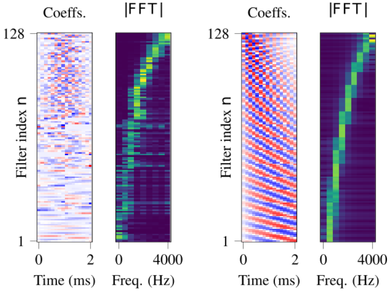
(a) Learned Filterbank (b) Proposed MP-GTF

Encoder / Filterbank	N	SI-SNRi (dB)
Learned	**512**	**15.4**
Learned	128	15.2
MP-GTF	512	15.9
MP-GTF	**128**	**16.1**

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

Learned Filterbank: Key to TasNet's Success?

Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)[1]:



(a) Learned Filterbank (b) Proposed MP-GTF

Encoder / Filterbank	N	SI-SNRi (dB)
Learned	**512**	**15.4**
Learned	128	15.2
MP-GTF	512	15.9
MP-GTF	**128**	**16.1**

- Motivation: Resembles human auditory system and structure of fully-learned encoder.
- ✔ Speeds up training time (less parameters)
- ✔ Slightly outperforms fully learned filterbanks

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

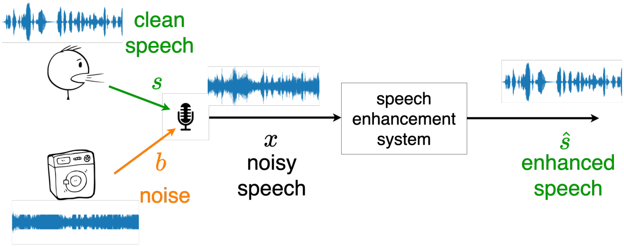
SP
Signal Processing

Variational Autoencoders (VAEs) for Speech Enhancement

Guillaume Carbajal (Ph.D.), Julius Richter (M.Sc), Huajian Fang (M.Sc.)

Contributions

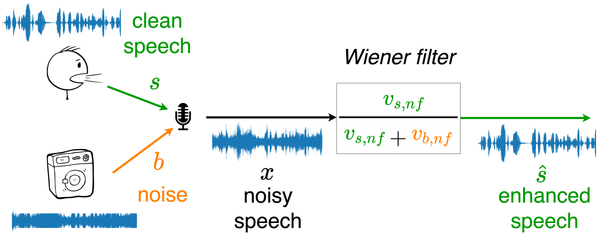
1. J. Richter, G. Carbajal, and T. Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks," in Interspeech, Oct. 2020, pp. 4516–4520.
2. H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, pp. 676–680.
3. G. Carbajal, J. Richter, and T. Gerkmann, "Guided Variational Autoencoder for Speech Enhancement with a Supervised Classifier," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, pp. 681–685.
4. G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement Learning for Variational Autoencoders Applied to Audio-Visual Speech Enhancement," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct. 2021.
5. H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Joint Reduction of Ego-Noise and Environmental Noise with a Partially-Adaptive Dictionary," in ITG Conference on Speech Communication.



Time-frequency domain: $x_{nf} = s_{nf} + b_{nf}$

Goal: Remove the noise b_{nf} without distorting the clean speech s_{nf}

Discriminative Vs. Generative Approaches



Discriminative

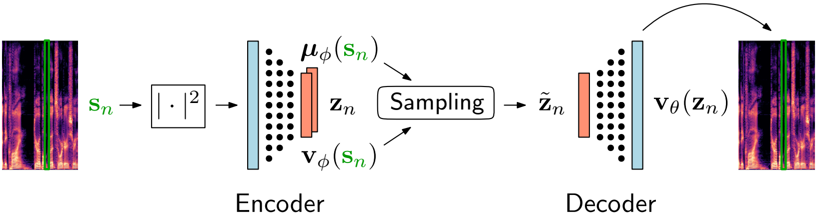
- Learn $p(s_{nf}|x_{nf})$
- Trained on pairs of (x_{nf}, s_{nf})
- ✖ Generalize to unseen situations not guaranteed

Generative

- Learn $p(s_{nf})$
- Trained on s_{nf} only
- ✔ Can generalize well to unseen situations

- $v_{s,nf} \rightarrow$ variational autoencoder (VAE)
- $v_{b,nf} \rightarrow$ nonnegative matrix factorization (NMF)

VAE as Speech Model – Training

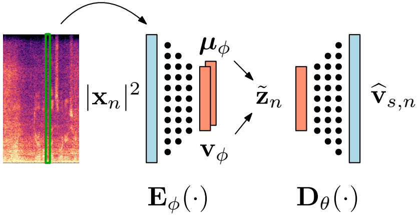


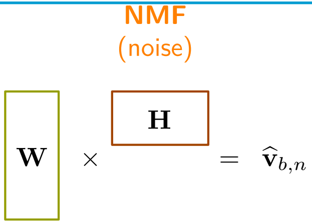
- Introduce latent variables $\mathbf{z} \in \mathbb{R}^L$ to help govern the distribution of the data $\mathbf{s} \in \mathbb{R}^F$, where often $L \ll F$
- Assume Gaussians for likelihood $p_\theta(\mathbf{s}|\mathbf{z})$ and posterior $q_\phi(\mathbf{z}|\mathbf{s})$ of $\mathbf{z}$
- Maximize the Evidence Lower Bound (ELBO)[5]

$$\text{ELBO}_{\theta,\phi}(\mathbf{s}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s})}[\log p_\theta(\mathbf{s}|\mathbf{z})]}_{\text{reconstruction accuracy}} - \underbrace{D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{s})||\,\mathcal{N}(0,\mathbf{I}))}_{\text{regularization}} \quad (1)$$

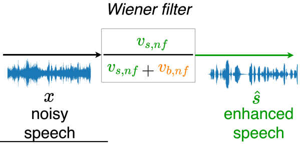
[5] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019. [Online]. Available: `https://arxiv.org/abs/1906.02691`.

VAE-NMF Framework – Test[6]





- Noisy speech as input to VAE
- Noise variance estimate: Nonnegative Matrix Factorization (NMF)
- Joint estimation of speech and noise PSD using Monte Carlo Expectation Maximization (MCEM)
- ✖ VAE encoder remains sensitive to noise



[6] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *MLSP*, Sep. 2018, pp. 1–6.

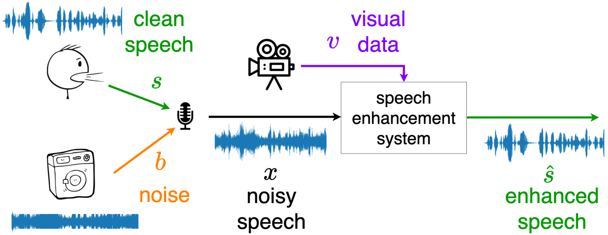
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP Signal Processing

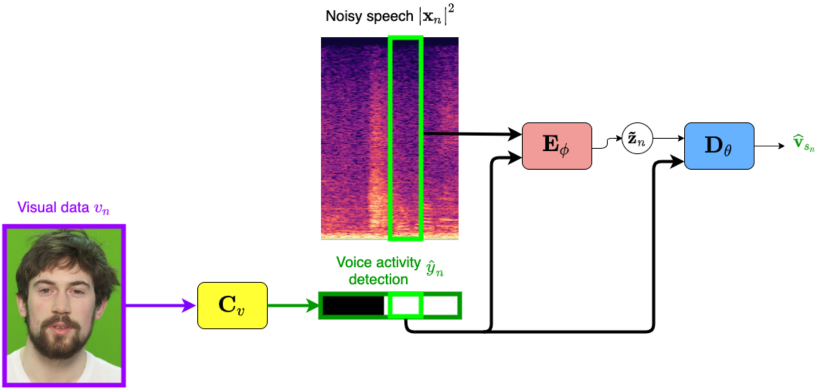
Conditional Variational Autoencoder for Speech Enhancement

G. Carbajal, J. Richter, and T. Gerkmann, "Guided Variational Autoencoder for Speech Enhancement with a Supervised Classifier," in *ICASSP*, Jun. 2021, pp. 681–685.

G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement Learning for Variational Autoencoders Applied to Audio-Visual Speech Enhancement," in *WASPAA*, Oct. 2021, accepted.



- **Advantage**: visual data v not affected by the noisy acoustic environment
- **Goals**:
  - Remove the noise b_{nf} without distorting the clean speech s_{nf}
  - Integrate visual data v as additional information

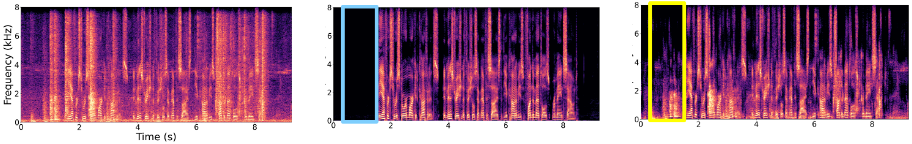


- Problem: In noise-only, VAE tries to reconstruct speech
- ✔ Voice activity y_n can be detected by a $\underbrace{\textbf{supervised}}_{=\text{learns } p(y|v)}$ visual-only classifier $\mathbf{C}_v$ [7]
- ✔ visual-only voice activity detection (VAD) robust to acoustic noise

[7] I. Ariav and I. Cohen, "An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks,", vol. 13, no. 2, pp. 265–274, May 2019.

CVAE – Limitations

Noisy speech x | Clean speech s | CVAE

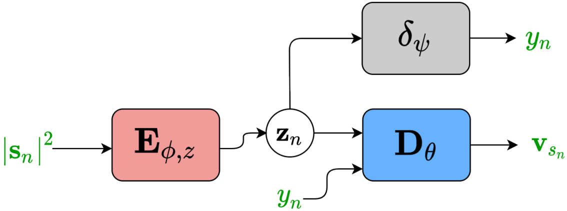


- ✖ CVAE still outputs signal when $\widehat{y}_n = 0$

Explanation:

- ✖ As DNN only sees clean speech in training ▶ does not learn role of y
  $\longrightarrow$ latent variable $\mathbf{z}_n$ already contains information of y_n
- ✖ ELBO does not guarantee $\underbrace{\textbf{disentanglement}}_{\substack{= \text{independence} \\ \text{between } \mathbf{z}_n \text{ and } y_n}}$ of $\mathbf{z}_n$ and y_n

Adversarial training[8] [9]

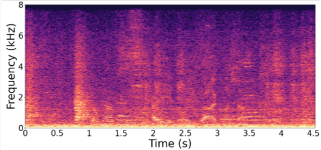


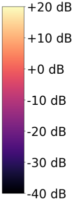
- ✔ Discriminator $\delta_\psi(\cdot)$ estimates y_n from latent variable $\mathbf{z}_n$
- ✔ Adversarial-encoder $\mathbf{E}_{\phi,z}(\cdot)$ makes discriminator $\delta_\psi(\cdot)$ unable to estimate y_n $\longrightarrow$ maximize entropy

[8] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5969–5978.

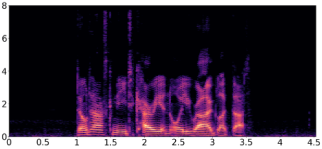
[9] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *Proc. WASPAA 2021*, Oct. 2021.

Noisy speech x

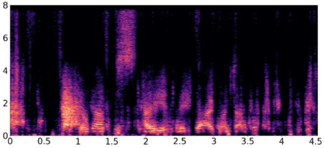




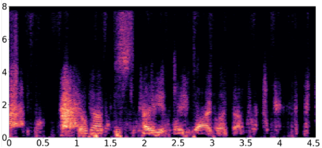
Clean speech s



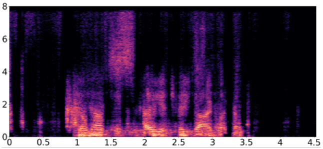
VAE



CVAE



DCVAE



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

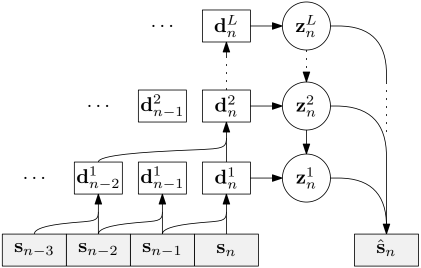
SP
Signal Processing

Speech Enhancement with Stochastic Temporal Convolutional Networks (STCNs)

J. Richter, G. Carbajal, and T. Gerkmann, “Speech Enhancement with Stochastic Temporal Convolutional Networks,” in Interspeech, Oct. 2020, pp. 4516–4520.

STCN model architecture

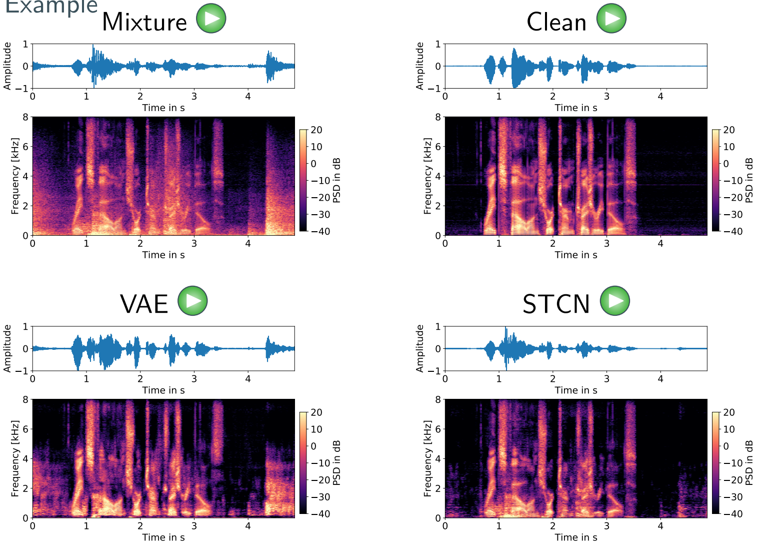
- ✖ Until now: no modeling of temporal dependencies
- ➡ Employ a stochastic temporal convolutional network (STCN)[10] [11]



[10] E. Aksan and O. Hilliges, "Stcn: Stochastic temporal convolutional networks," in *International Conference on Learning Representations*, 2018.

[11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," *Proc. Interspeech 2020*, pp. 4516–4520, 2020.

Audio Example



Conclusion: Modeling temporal dependencies → more robust VAE

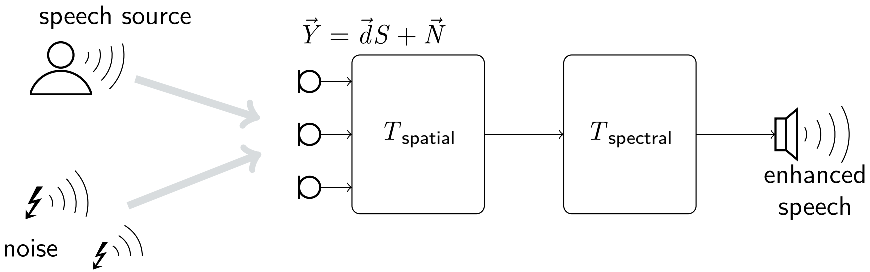
`https://uhh.de/inf-sp-stcn2020`

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

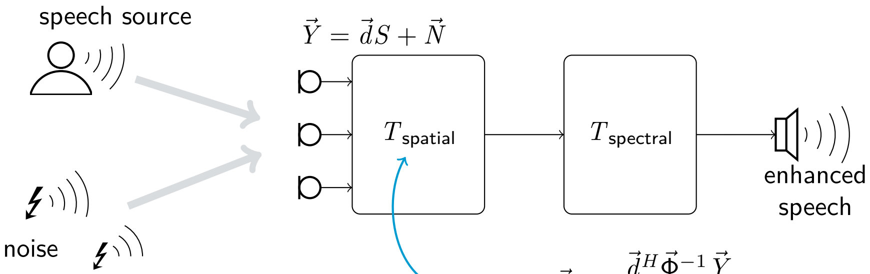
SP Signal Processing

Nonlinear Multichannel Filtering

Kristina Tesch, Timo Gerkmann, "Nonlinear Spatial Filtering in Multichannel Speech Enhancement", IEEE/ACM Trans. Audio, Speech, Language Proc., Vol. 29, pp. 1795-1805, 2021.



speech source
$\vec{Y} = \vec{d}S + \vec{N}$
T_{spatial}
T_{spectral}
enhanced speech
noise



$$T_{\text{MVDR}}(\vec{Y}) = \frac{\vec{d}^H \vec{\Phi}_n^{-1} \vec{Y}}{\vec{d}^H \vec{\Phi}_n^{-1} \vec{d}}$$

Minimum variance distortionless response (MVDR) beamformer

MVDR As Sufficient Statistic

The MVDR beamformer T_{MVDR} is a **sufficient statistic in the Bayesian sense** if

$$p_S(s|\vec{y}) = p_S(s|T_{\mathsf{MVDR}}(\vec{y}))$$

holds for every observation $\vec{y}$ and every prior distribution of S.

- ✔ Holds under a Gaussian noise assumption
- ➡ All information about S is retained in the output of the MVDR
- ➡ Separation of linear spatial filter and postfilter is optimal in the MMSE and MAP sense

Model the noise distribution by a multivariate complex Gaussian **mixture**, i.e., $\vec{N} \sim \sum_{m=1}^{M} c_m \mathcal{N}_{\mathbb{C}}(0, \vec{\Phi}_m)$.[12] [13]

[12] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

Joint Spatial-spectral Nonlinear MMSE Solution

Model the noise distribution by a multivariate complex Gaussian **mixture**, i.e., $\vec{N} \sim \sum_{m=1}^{M} c_m \mathcal{N}_{\mathbb{C}}(0, \vec{\Phi}_m)$.[12] [13]

$$T_{\mathsf{MMSE}}(\vec{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m \widetilde{Q}_m}{|\vec{\Phi}_m|} \exp\left\{-\vec{y}^H \vec{\Phi}_m^{-1} \vec{y}\right\} T_{\mathsf{MVDR}}^{(m)}(\vec{y}) \mathcal{M}_{\mathsf{n}} \left[T_{\mathsf{MVDR}}^{(m)}(\vec{y}) \right]}{\sum_{m=1}^{M} \frac{c_m Q_m}{|\vec{\Phi}_m|} \exp\left\{-\vec{y}^H \vec{\Phi}_m^{-1} \vec{y}\right\} \mathcal{M}_{\mathsf{d}} \left[T_{\mathsf{MVDR}}^{(m)}(\vec{y}) \right]}$$

- $T_{\mathsf{MVDR}}^{(m)}(\vec{y}) = \dfrac{\vec{d}^H \vec{\Phi}_m^{-1} \vec{y}}{\vec{d}^H \vec{\Phi}_m^{-1} \vec{d}}$
- $\mathcal{M}_{\mathsf{n}}$,$\mathcal{M}_{\mathsf{d}}$ related to confluent hypergeometric function
- $\widetilde{Q}_m$ and Q_m are functions of $\vec{d}$, $\vec{\Phi}_m$, ν and σ_s^2

[12] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

Joint Spatial-spectral Nonlinear MMSE Solution

Model the noise distribution by a multivariate complex Gaussian **mixture**, i.e., $\vec{N} \sim \sum_{m=1}^{M} c_m \mathcal{N}_{\mathbb{C}}(0, \vec{\Phi}_m)$.[12] [13]

$$T_{\mathsf{MMSE}}(\vec{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m \widetilde{Q}_m}{|\vec{\Phi}_m|} \exp\left\{-\vec{y}^H \vec{\Phi}_m^{-1} \vec{y}\right\} T_{\mathsf{MVDR}}^{(m)}(\vec{y}) \mathcal{M}_{\mathsf{n}}\left[T_{\mathsf{MVDR}}^{(m)}(\vec{y})\right]}{\sum_{m=1}^{M} \frac{c_m Q_m}{|\vec{\Phi}_m|} \exp\left\{-\vec{y}^H \vec{\Phi}_m^{-1} \vec{y}\right\} \mathcal{M}_{\mathsf{d}}\left[T_{\mathsf{MVDR}}^{(m)}(\vec{y})\right]}$$

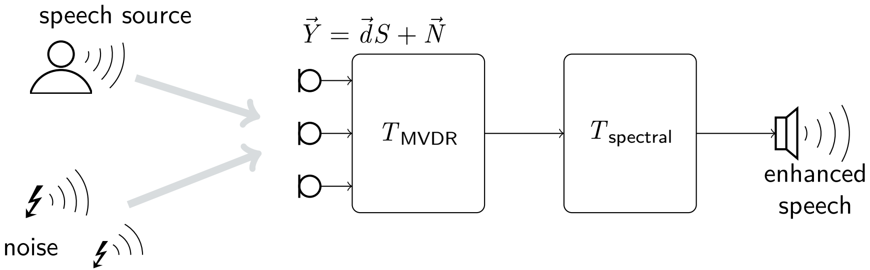
➡ **Cannot** be decomposed into a linear spatial filter and postfilter

- Dependency on the summation index m
- Quadratic term $\vec{y}^H \vec{\Phi}_m^{-1} \vec{y}$

---

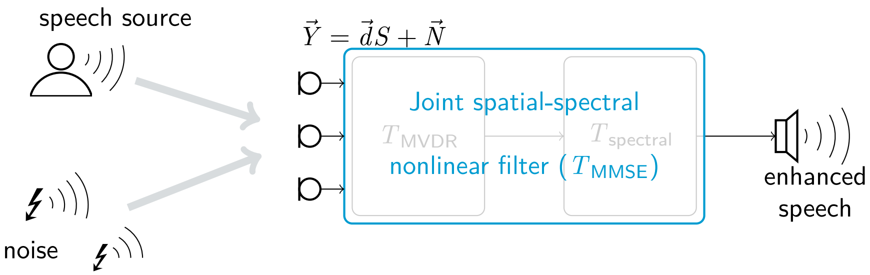
[12] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, “On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[13] K. Tesch and T. Gerkmann, “Nonlinear spatial filtering in multichannel speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.



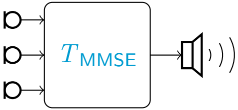
speech source
$\vec{Y} = \vec{d}S + \vec{N}$
T_{MVDR}
T_{spectral}
enhanced speech
noise

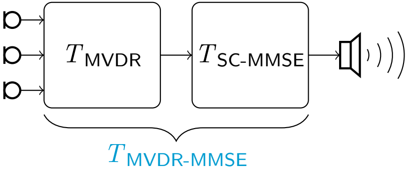
Research Questions



Should we replace the traditional approach with DNNs?

- How much can we gain from a joint spatial-spectral nonlinear filter?
- Where does the benefit of using a nonlinear spatial filter come from?





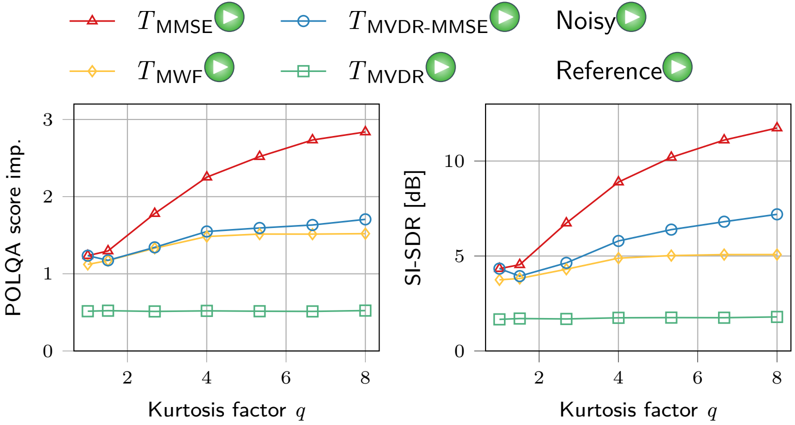
- MMSE optimal[12] joint spatial and spectral nonlinear processing

- MVDR beamformer combined with single channel MMSE estimator
- Derivation based on same assumptions[13]

[12] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

Complex Gaussian mixture distribution modelling diffuse noise



➡ Nonlinear filter improves upon the performance of the combined filter if noise is more heavy-tailed than a Gaussian

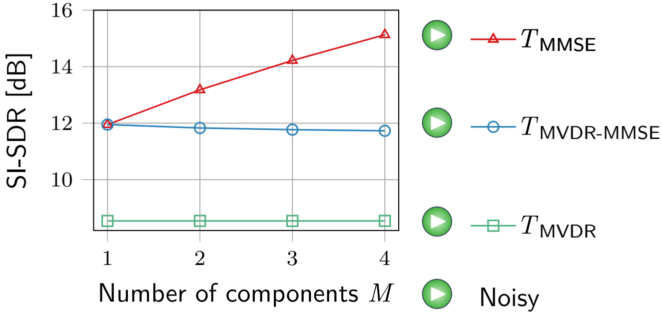
[14] K. Tesch, R. Rehr, and T. Gerkmann, "On Nonlinear Spatial Filtering in Multichannel Speech Enhancement," in *Interspeech 2019*, Graz, Austria, 2019, pp. 91–95.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

Real-world Noise Data (CHiME-3)[14]

Gaussian mixture distribution estimated with EM algorithm applied to segments

- Results for the cafeteria noise

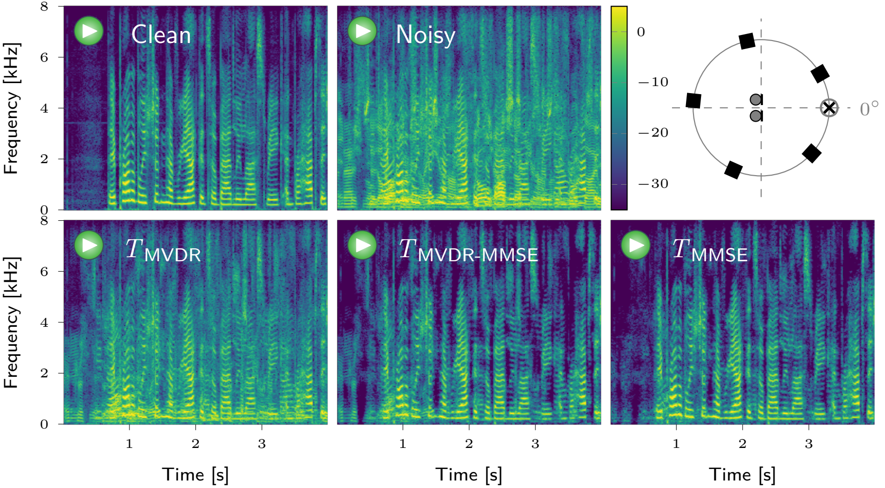


➡ Nonlinear spatial filter improves performance based on a non-Gaussian noise model

[14] K. Tesch, R. Rehr, and T. Gerkmann, "On Nonlinear Spatial Filtering in Multichannel Speech Enhancement," in *Interspeech 2019*, Graz, Austria, 2019, pp. 91–95.

Spatial Characteristics: Directional Interferences

Inhomogeneous noise field created by five interfering speakers[tesch2020inhomogeneous]

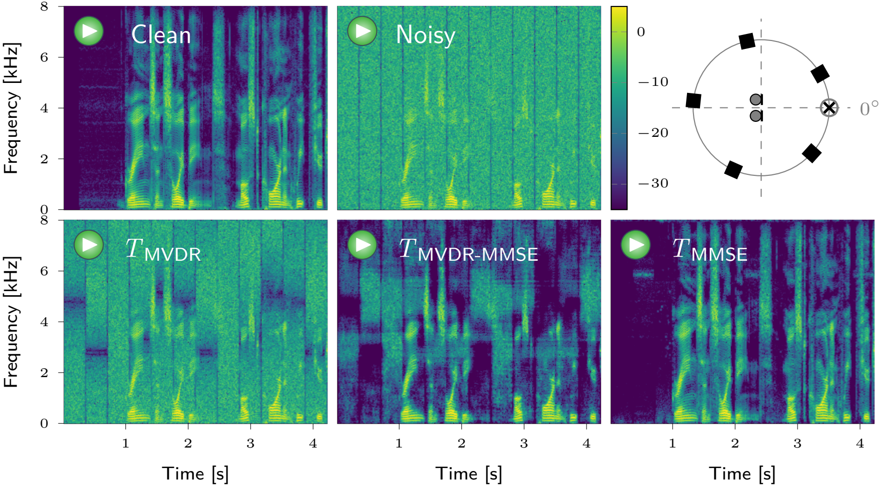


➡ Nonlinear spatial filter is beneficial

Δ POLQA: 0.84 ± 0.04

Δ SI-SDR: 4.63 ± 0.15

Inhomogeneous noise field created by five directional Gaussian noise sources



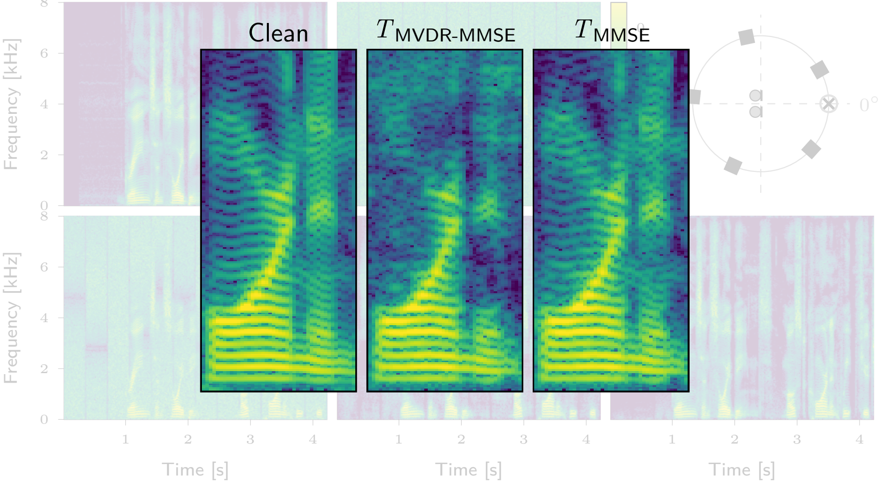
➡ Nonlinear spatial filter is beneficial

➡ Future: implementation using DNNs

Δ POLQA: 2.64 ± 0.08

Δ SI-SDR: 9.92 ± 0.30

Inhomogeneous noise field created by five directional Gaussian noise sources



➡ Nonlinear spatial filter is beneficial

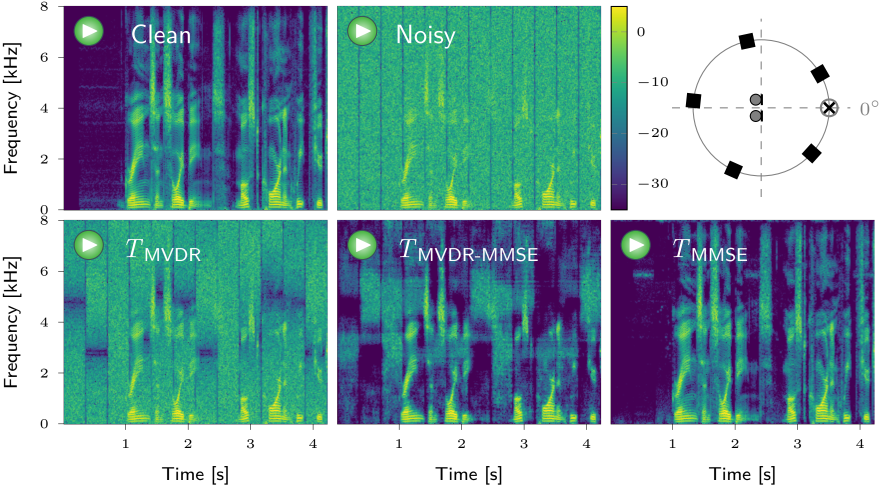
➡ Future: implementation using DNNs

Δ POLQA: 2.64 ± 0.08

Δ SI-SDR: 9.92 ± 0.30

Spatial Characteristics: Directional Interferences

Inhomogeneous noise field created by five directional Gaussian noise sources



➡ Nonlinear spatial filter is beneficial

➡ Future: implementation using DNNs

Δ POLQA: 2.64 ± 0.08

Δ SI-SDR: 9.92 ± 0.30

Conclusions

Conclusions

- Neural networks are a powerful tool for source separation[1]
- Variational Autoencoders
  - Elegant tool to combine statistical methods and machine learning
  - Noise robustness can be improved using
    - Conditioning on additional information (e.g. visual)[9]
    - Including temporal modelling[11]
- Neural networks: great potential also for multi-sensor signal processing[13]

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

[9] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *Proc. WASPAA 2021*, Oct. 2021.

[11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," *Proc. Interspeech 2020*, pp. 4516–4520, 2020.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

[1] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

[2] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," en, in *ISCA Interspeech*, Graz, Austria, Sep. 2019, pp. 4584–4588. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2459.html (visited on 09/16/2019).

[3] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 686–690.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[5] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019. [Online]. Available: https://arxiv.org/abs/1906.02691.

[6] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *MLSP*, Sep. 2018, pp. 1–6.

[7] I. Ariav and I. Cohen, "An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks,", vol. 13, no. 2, pp. 265–274, May 2019.

[8] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5969–5978.

References II

[9] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *Proc. WASPAA 2021*, Oct. 2021.

[10] E. Aksan and O. Hilliges, "Stcn: Stochastic temporal convolutional networks," in *International Conference on Learning Representations*, 2018.

[11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," *Proc. Interspeech 2020*, pp. 4516–4520, 2020.

[12] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[13] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

[14] K. Tesch, R. Rehr, and T. Gerkmann, "On Nonlinear Spatial Filtering in Multichannel Speech Enhancement," in *Interspeech 2019*, Graz, Austria, 2019, pp. 91–95.