Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

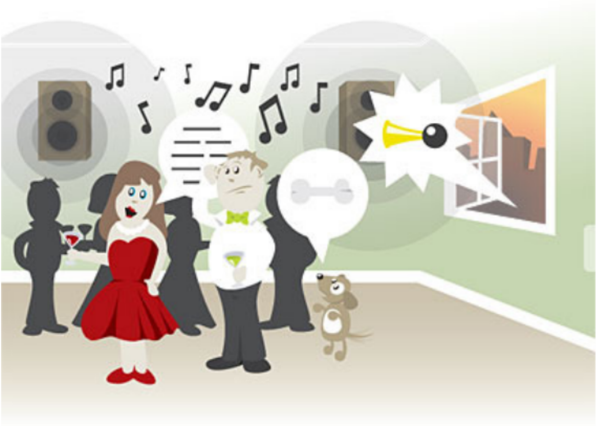
Prof. Dr.-Ing. Timo Gerkmann

Machine Learning for Speech Signal Processing on Hearing Devices

Universität Hamburg
Department of Informatics
Signal Processing (SP)

August 13, 2022

- How can Machine Learning help to make information more easily accessible by humans and machines



Overview

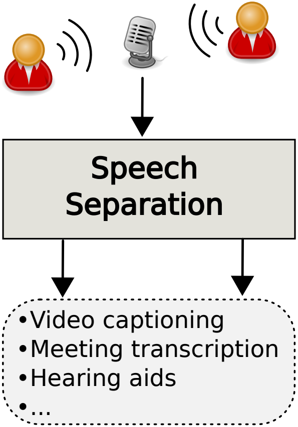
1. Single Channel Source Separation
2. Phase Estimation Enables High Quality at Low Latency
3. Non-linear Multi-channel Filtering
4. Diffusion-based Generative Models for Speech Enhancement

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

Single Channel Source Separation

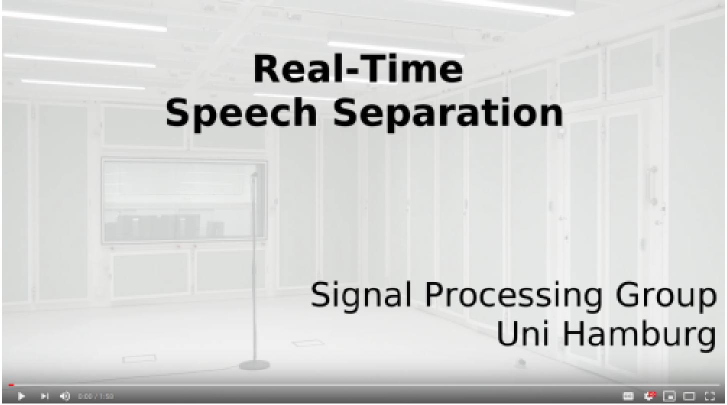
Cocktail-Party Problem



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).

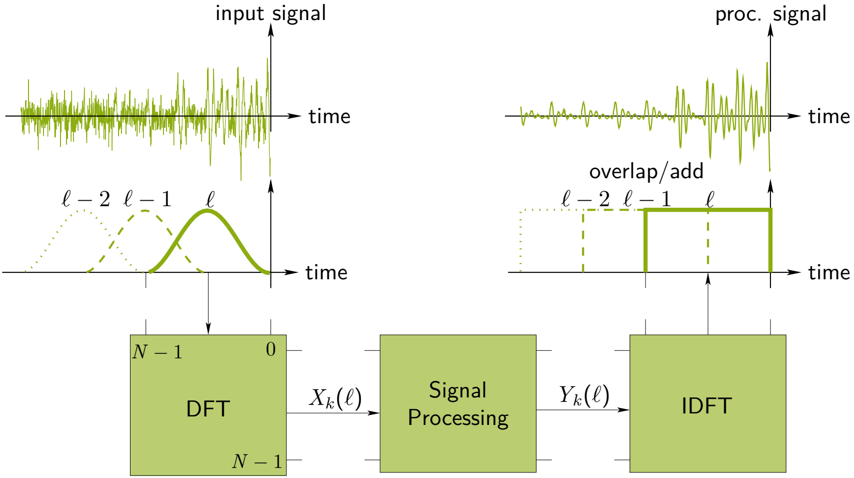
SP Signal Processing

Conclusions

- Machine Learning enables separating sources recorded with only one microphone
- As traditional approaches, these algorithms can be made real-time capable
- The algorithmic latency depends on the chosen frame sizes

Phase Estimation Enables High Quality at Low Latency

Tal Peer, M.Sc.


input signal
time
proc. signal
time
overlap/add
ℓ − 2
ℓ − 1
ℓ
time
ℓ − 2
ℓ − 1
ℓ
time
N − 1
0
DFT
N − 1
X_k(ℓ)
Signal Processing
Y_k(ℓ)
IDFT

Frame Length in STFT Speech Processing

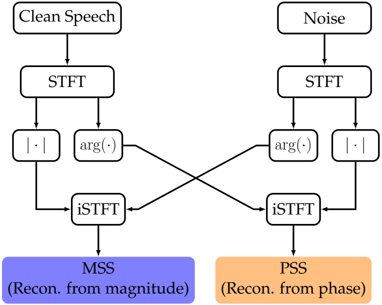
- STFT-based speech processing traditionally uses frames of around 32ms
  - ✔ Short enough to capture non-stationarity of speech
  - ✔ Long enough to admit a reasonable spectral resolution

- **Is this optimal?**
  - ✖ Frame length imposes a lower bound on algorithmic latency
  - ✖ The justification for 32ms is mainly based magnitude and ignores phase

➡ As traditional enhancement methods are magnitude centric, 32ms appears a well motivated choice

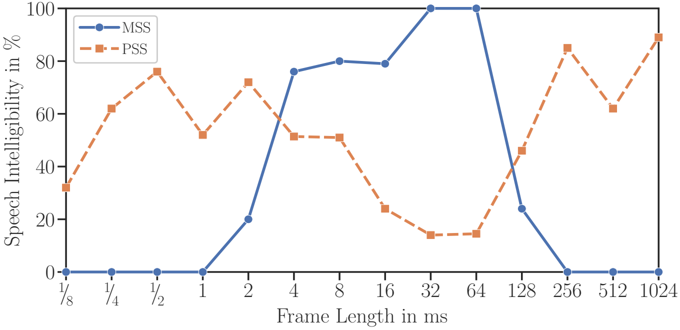
- But what if we were able to also estimate phase?
- Kazama et al.[3]: listening experiment on intelligibility under variation of frame length

➡ The information contained in magnitude and phase varies with frame length

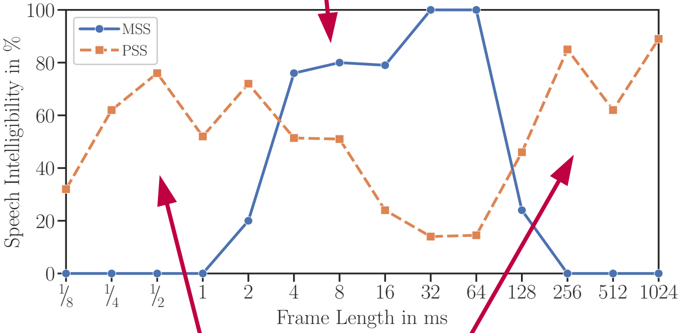


[3] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1432–1439, 2010.

Phase, Magnitude and Frame Length (2)



Medium frames: magnitude suffices for good reconstruction



Short and long frames: magnitude loses relevance, good reconstruction is possible from phase alone

Frame Length in Deep Speech Enhancement

➡ Phase information gets important for short frames

✖ Model-based phase estimation methods exist only for long frames[4,5]

Research Questions

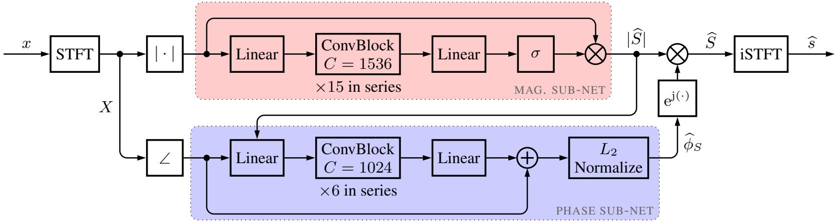
- Can we use modern machine learning approaches to estimate phase when using short frames?
- Which frame length for phase-aware STFT-based networks?

[4] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[5] T. Peer, K.-J. Ziegert, and T. Gerkmann, "Plosive Enhancement Using Phase Linearization and Smoothing," in *Speech Communication; 14th ITG Conference*, Kiel (online), Sep. 2021, pp. 1–5.

We use a DNN with explicit magnitude and phase estimation[6,7] to

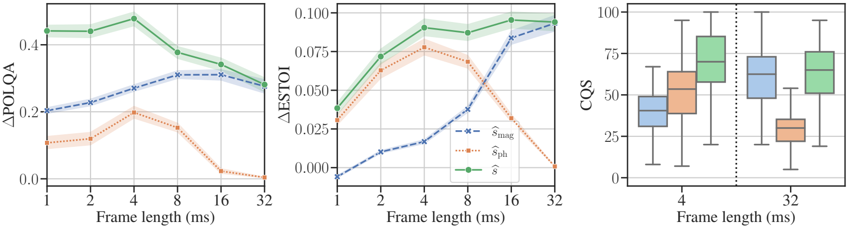
- Quantify the contribution of phase and magnitude estimation for different frame lengths
- Quantify overall performance of joint network for different frame length



[6] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Interspeech 2018*, ISCA, Sep. 2, 2018, pp. 3244–3248.

[7] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *arXiv preprint arXiv:2203.16222*, 2022.

Results



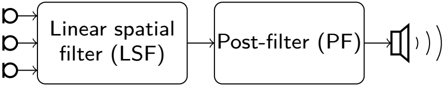
- ✔ Trend observed on oracle data carries over to DNN-based magnitude and phase estimation
- ✔ Machine Learning can be used to estimate phase also with short frames
- ✔ Phase-aware processing is particularly beneficial with short frames
- ✔ Short frames of 4ms enable **short latency and high quality**

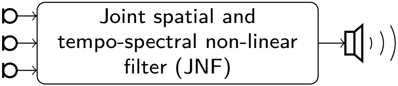
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

Non-linear Multi-channel Filtering

Kristina Tesch, M.Sc.





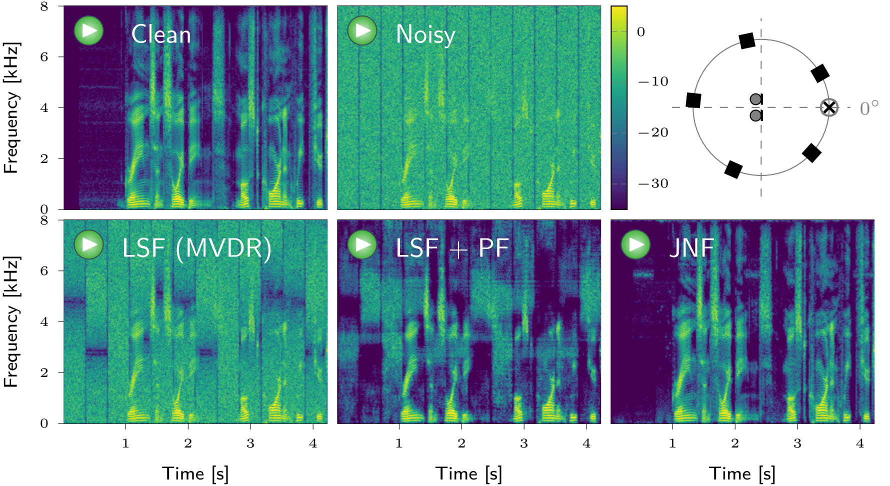
- Analytic solutions
- Computationally lightweight
- MMSE optimal for Gaussian noise[8]

- Drops linearity assumption
- Integrates spatial and tempo-spectral processing
- ➡ More powerful processing model
- ➡ Parameter estimation challenging

[8] 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.

Proof of Concept with Oracle Data

Inhomogeneous noise field created by five directional Gaussian noise sources



➡ Joint non-linear spatial-spectral filter is a more powerful processing model

Δ POLQA: 2.64 ± 0.08

Δ SI-SDR: 9.92 ± 0.30

From Theory to Practice: DNN-based JNF[9]

- We saw: Joint nonlinear spatial spectral filtering (JNF) is more powerful than traditional beamformer + postfilter
- Above examples provided a proof of concept, but estimating the required higher-order statistics is very difficult in practice

[9] K. Tesch and T. Gerkmann, *Insights into deep non-linear filters for improved multi-channel speech enhancement*, submitted to IEEE Trans. Audio, Speech, and Language Proc., 2022. [Online]. Available: `https://arxiv.org/abs/2206.13310`.

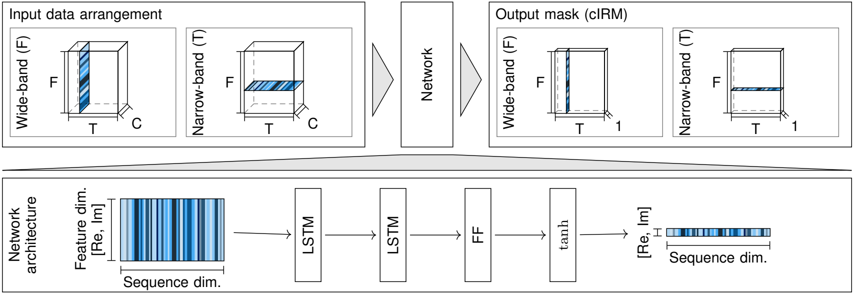
From Theory to Practice: DNN-based JNF[9]

- We saw: Joint nonlinear spatial spectral filtering (JNF) is more powerful than traditional beamformer + postfilter
- Above examples provided a proof of concept, but estimating the required higher-order statistics is very difficult in practice

Research Questions

- Do our theoretical findings carry over when learning a JNF using DNNs?
  - Such DNN-based JNFs are fundamentally different to *DNN-guided* beamformers!
- How important are the interdependencies between different sources of information?
- What are the implications that arise for the design of network architectures?

[9] K. Tesch and T. Gerkmann, *Insights into deep non-linear filters for improved multi-channel speech enhancement*, submitted to IEEE Trans. Audio, Speech, and Language Proc., 2022. [Online]. Available: https://arxiv.org/abs/2206.13310.



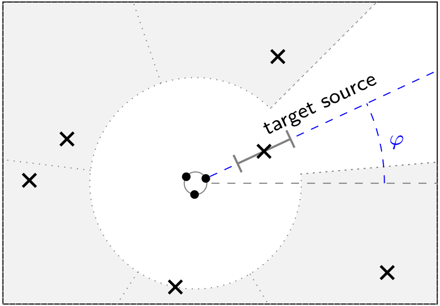
- ➡ Network structure that allows to easily **control the integration of different sources of information**
  - Combine spatial with spectral (F-JNF) or temporal (T-JNF) information or both (FT-JNF)

Approach - Dataset

Speaker extraction focusing on spatial filtering capabilities

Task:

- Speaker extraction scenario
- 2-5 microphones in a circular array
- 1 target speaker
- 5 interfering speakers



Dataset generation:

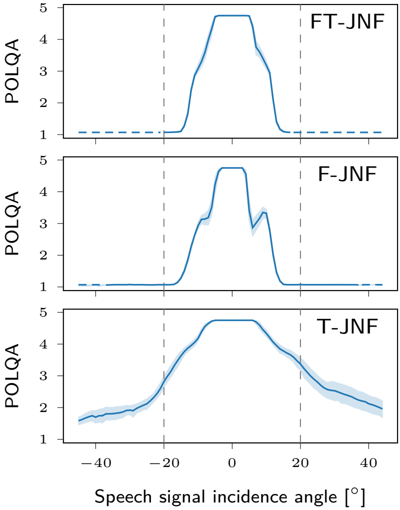
- Clean speech from WSJ0 (75 male and 74 female speakers)
- 6000 training examples (25 hours of training data)
- Simulation using the source-image model
- SNR between -9 and 2 dB
- Room dimensions between $(2.5 \times 3 \times 2.2)$ and $(5 \times 9 \times 3.5)$ meters
- T60: $0.2 - 0.5$ seconds

Interdependency Between Information Sources

	Δ POLQA	ESTOI
F-JNF	1.15	0.70
T-JNF[10]	0.74	0.63
FT-JNF (ours)	**1.43**	**0.76**

➡ Addional spectral information is more valuable than temporal information

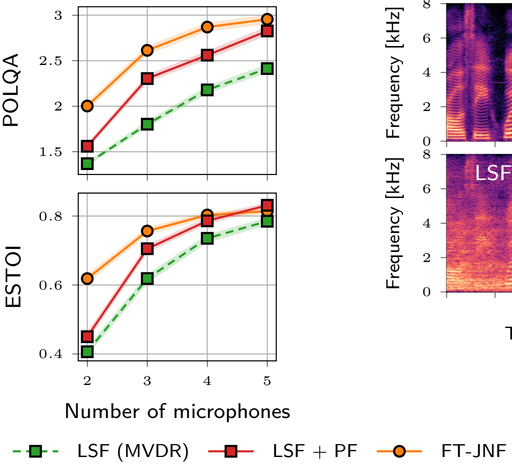
➡ Spectral information increases the spatial selectivity

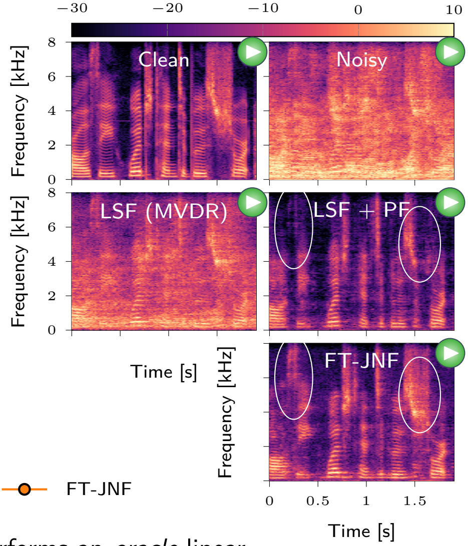


[10] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory,", 2019, pp. 298–302.

Non-Linear Versus Linear Spatial Filter

Blind estimation using DNNs



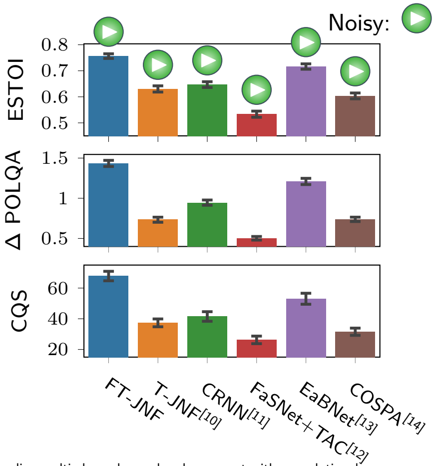


➡ A joint non-linear filter outperforms an *oracle* linear spatial filter plus post-filter

Comparison with State-of-the-art Methods

- Complex masked-based:
  FT-JNF, T-JNF, CRNN
- Beamformer-inspired:
  FaSNet+TAC, EaBNet, COSPA
- FT-JNF and T-JNF have the same lowest number of parameters

➡ Proposed FT-JNF outperforms all other methods



[11] S. Chakrabarty and E. A. P. Habets, “Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks,”, vol. 13, no. 4, pp. 787–799, 2019.

[12] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,”, Barcelona, Spain, 2020, pp. 6394–6398.

[13] A. Li, W. Liu, C. Zheng, and X. Li, “Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement,”, Singapore, 2022, pp. 6487–6491.

[14] M. M. Halimeh and W. Kellermann, “Complex-valued spatial autoencoders for multichannel speech enhancement,”, Singapore, 2022,

Conclusions

Deep non-linear filters **overcome the linear processing model** and **exploit dependencies** between spatial and tempo-spectral information

- ➡ Spectral information increases the spatial selectivity of the filter
- ➡ The proposed scheme that exploits spatial, spectral and temporal information outperforms state-of-the-art network architectures

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP Signal Processing

Diffusion-based Generative Models for Speech Enhancement

Julius Richter, M.Sc. and Simon Welker, M.Sc.

Speech Enhancement with Deep Learning

- Speech enhancement algorithms are nowadays dominated by the use of deep neural networks (DNNs)

✔ Exploit temporal-spectral structure to distinguish speech from noise

Discriminative model vs. generative model

- Statistical models can be classified as generative or discriminative
- Discriminative models dominate the task of speech enhancement
- Recently, there is a trend towards generative approaches

Why Generative Modeling?

Discriminative models

- Learn to directly map noisy speech to the corresponding clean speech
- Trained with a variety of clean/noisy speech pairs
- ✖ No guarantee of robustness in unseen situations
- ✖ Unpleasant speech distortions may outweigh the benefits of noise reduction

Why Generative Modeling?

Discriminative models

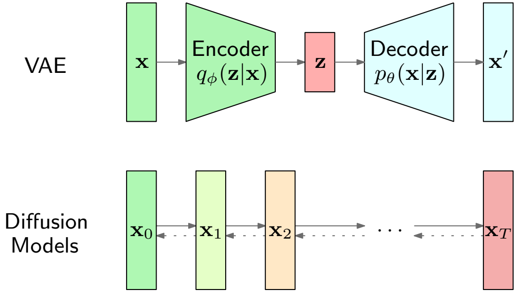
- Learn to directly map noisy speech to the corresponding clean speech
- Trained with a variety of clean/noisy speech pairs
- ✖ No guarantee of robustness in unseen situations
- ✖ Unpleasant speech distortions may outweigh the benefits of noise reduction

Generative models

- Learn a prior distribution over clean speech data
- Infer clean speech from noisy signals that are assumed to lie outside the learned distribution
- ✔ Generalize well to unseen acoustic situations
- ✔ Aim to produce natural sounding speech

Popular deep generative models

- Variational autoencoders (VAEs)
- Generative adversarial networks (GANs)
- Auto-regressive models
- Diffusion-based generative models



Our contributions:

- Incorporate temporal dependencies into the VAE[15]
- Improve the robustness with a noise-aware encoder[16]
- Guide the VAE with a supervised classifier trained on voice acticity or ideal binary mask prediction[17]
- Disentanglement learning of the latent variables applied to audio-visual voice activity detection[18]

Limitations:

- ✖ VAE needs additional noise estimator to form a Wiener filter
- ✖ Limited by the bottleneck of the latent representation

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

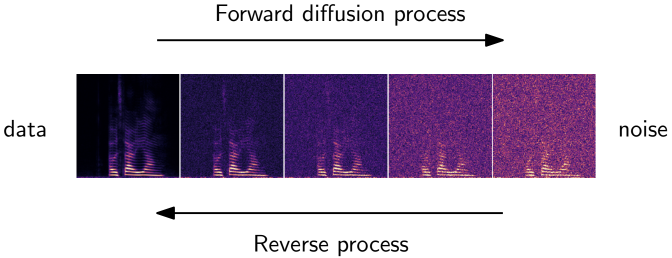
[16] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 676–680, 2021.

[17] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 681–685, 2021.

[18] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, pp. 126–130, 2021.

Generative diffusion models[19,20] consist of two processes:

- Forward diffusion process that gradually adds noise to the input
- Reverse process that learns to generate data by denoising



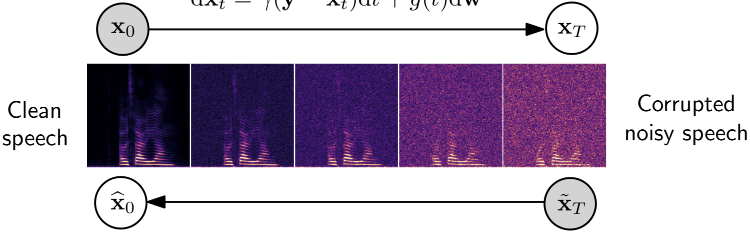
[19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, 2015, pp. 2256–2265.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

Stochastic Diffusion Process

- Model the corruption of clean speech as a diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ [21]
- Define the diffusion process as a solution to a stochastic differential equation (SDE)[22]

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$



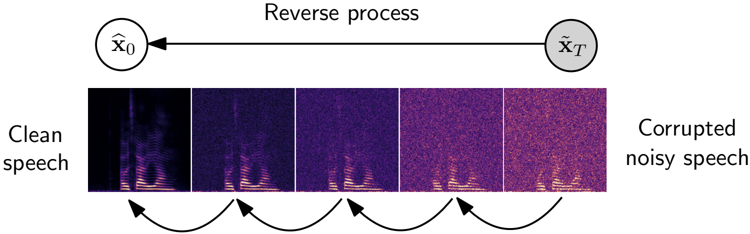
$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y} - \mathbf{x}_t) + g(t)^2 \underbrace{\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)}_{\approx \mathbf{s}_\theta(\mathbf{x}, \mathbf{y}, t)}]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

- The learned *score model* $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)$ predicts the added Gaussian noise

[21] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *Proc. Interspeech 2022*, 2022.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Representations (ICLR)*, 2021.

- Initialize reverse process with $\tilde{\mathbf{x}}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma(T)^2\mathbf{I})$
- Solve reverse SDE with general-purpose SDE solvers (e.g. Euler-Maruyama)
- Necessary reverse steps $N \approx 30 \Rightarrow$ 30 model calls



$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y} - \mathbf{x}_t) + g(t)^2\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

Datasets:

- WSJ0-CHiME3
  - Clean speech utterances from Wall Street Journal (WSJ0)[23]
  - Noise signals from CHiME3[24]
  - SNR uniformly sampled between 0 and 20 dB
- Voicebank-Demand[25]
  - Standardized dataset often used as a benchmark

Matched and mismatched conditions:

condition	train/valid	test
matched	WSJ0-CHiME3	WSJ0-CHiME3
mismatched	Voicebank-Demand	WSJ0-CHiME3

[23] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete*, May 2007.

[24] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015.

[25] C. V. Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," *9th ISCA Speech Synthesis Workshop*, pp. 159–165, 2016.

Matched condition: Training and test on same datasets

Method	Type	POLQA	ESTOI	SI-SDR [dB]
Mixture	-	2.63 ± 0.67	0.78 ± 0.14	10.0 ± 5.7
RVAE	G	2.97 ± 0.63	0.85 ± 0.11	15.8 ± 5.0
CDiffuSE	G	2.77 ± 0.52	0.80 ± 0.09	7.3 ± 1.9
SGMSE+ (ours)	G	$\mathbf{3.71 \pm 0.50}$	$\mathbf{0.92 \pm 0.05}$	17.2 ± 4.6
Conv-TasNet	D	3.65 ± 0.54	$\mathbf{0.93 \pm 0.05}$	$\mathbf{19.9 \pm 4.3}$
MetricGAN+	D	3.52 ± 0.61	0.88 ± 0.08	10.5 ± 4.5

Matched condition: Training and test on same datasets

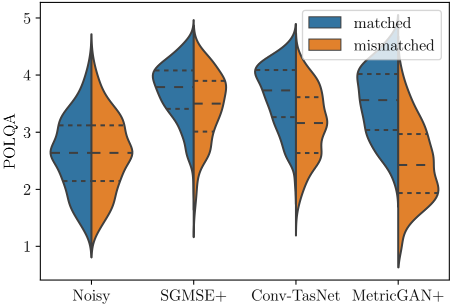
Method	Type	POLQA	ESTOI	SI-SDR [dB]
Mixture	-	2.63 ± 0.67	0.78 ± 0.14	10.0 ± 5.7
RVAE	G	2.97 ± 0.63	0.85 ± 0.11	15.8 ± 5.0
CDiffuSE	G	2.77 ± 0.52	0.80 ± 0.09	7.3 ± 1.9
SGMSE+ (ours)	G	$\mathbf{3.71 \pm 0.50}$	$\mathbf{0.92 \pm 0.05}$	17.2 ± 4.6
Conv-TasNet	D	3.65 ± 0.54	$\mathbf{0.93 \pm 0.05}$	$\mathbf{19.9 \pm 4.3}$
MetricGAN+	D	3.52 ± 0.61	0.88 ± 0.08	10.5 ± 4.5

Mismatched condition: Training and test on different datasets

RVAE	G	2.84 ± 0.61	0.82 ± 0.11	13.9 ± 4.8
CDiffuSE	G	2.20 ± 0.50	0.71 ± 0.10	3.8 ± 2.5
SGMSE+ (ours)	G	$\mathbf{3.43 \pm 0.61}$	$\mathbf{0.90 \pm 0.07}$	$\mathbf{16.2 \pm 4.1}$
Conv-TasNet	D	3.13 ± 0.60	0.88 ± 0.08	15.2 ± 3.9
MetricGAN+	D	2.47 ± 0.67	0.76 ± 0.12	6.8 ± 3.1

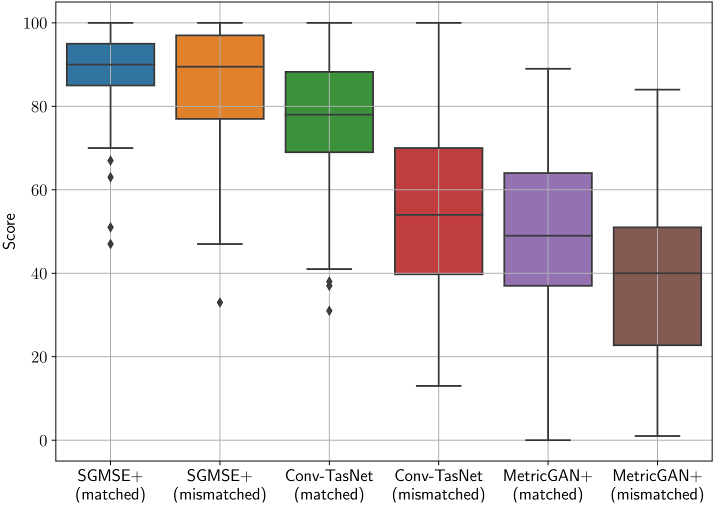
Matched vs. mismatched condition

- Performance for the matched condition is on par or even slightly better than discriminative baselines
- ✔ Proposed approach is more robust in unseen situations



Listening experiment

- 10 participants rate 12 random examples from the test set



Audio Examples

	Matched	Mismatched
Clean		
Noisy		
RVAE		
SGMSE+		
Conv-TasNet		
Clean		
Noisy		
RVAE		
SGMSE+		
Conv-TasNet		

Dereverberation - Audio Examples

Clean	▶
Reverberant	▶
SGMSE+	▶
Clean	▶
Reverberant	▶
SGMSE+	▶

- Interestingly, the same architecture can also be used very well to dereverberate signals!

Conclusions

Our proposed approach:

- Performs on par with state-of-the art discriminative methods
- Can be applied both to denoising and dereverberation
- Generalizes better under unmatched training conditions

1. S. Welker, J. Richter, and T. Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in Interspeech, Sept. 2022.

2. J. Richter, S. Welker, J.-M. Lemercier, B. Lay and T. Gerkmann, "Speech Enhancement and Dereverberation with Diffusion-based Generative Models," submitted to TASLP, 2022.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP Signal Processing

Conclusions

Conclusions

SP Signal Processing

- Deep Neural Networks (DNNs) are very powerful tools for single channel enhancement and source separation
- DNNs allow to estimate spectral phases to allow for high quality speech at low algorithmic latencies
- For multichannel speech enhancement, DNNs can be used to learn joint nonlinear spatial-spectral filters that may outperform the traditional beamformer + postfilter framework
- Diffusion-based generative models are an exciting upcoming field that may increase the robustness in unseen environments