Leveraging Heteroscedastic Uncertainty in Learning Complex Spectral Mapping for Single-channel Speech Enhancement

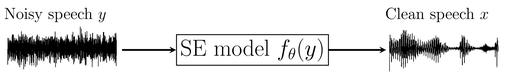
Kuan-Lin Chen[12†], **Daniel D. E. Wong**[1], **Ke Tan**[1], **Buye Xu**[1], **Anurag Kumar**[1], **Vamsi Krishna Ithapu**[1]
[1]Meta Reality Labs Research, [2]Department of ECE, University of California, San Diego [†]kuc029@ucsd.edu

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Abstract

- We propose **a new uncertainty-assisted learning framework** for speech enhancement (SE) and overcome the optimization difficulty that arises in the learning process.
- We show that, **at no extra cost in terms of compute, memory, and parameters**, directly minimizing a Gaussian negative log-likelihood (NLL) yields significantly better SE performance than minimizing a conventional loss such as the MAE or MSE, and slightly better SE performance than the SI-SDR loss.
- This is the **first successful study** that achieves improved perceptual metric performance by **directly** using heteroscedastic uncertainty for SE.

1 The conventional learning paradigm in SE



Most SE models are trained without leveraging uncertainty. Loss functions such as the mean squared error (MSE) and mean absolute error (MAE) are widely used in SE.

Question 1. *What are the assumptions behind these losses?*

Question 2. *Can SE models achieve better performance if the assumptions are weakened?*

2 Probabilistic models and assumptions

Definition 1. *Let the received signal in the STFT domain be $y_r^{t,f} + iy_i^{t,f} \in \mathbb{C}$ for all (t, f) with the time frame index $t \in \{1, 2, \cdots, T\}$ and frequency bin index $f \in \{1, 2, \cdots, F\}$. Let $y \in \mathbb{R}^{2TF}$ be the vector representing every real part and imaginary part of the STFT representation of the received signal.*

Assumption 1. *We assume a multivariate Gaussian model*

$$p\left(x|y;\psi\right) = \frac{\exp\left(-\frac{1}{2}\left[x-\hat{\mu}_\theta(y)\right]^\mathsf{T}\hat{\Sigma}_\phi^{-1}(y)\left[x-\hat{\mu}_\theta(y)\right]\right)}{\sqrt{(2\pi)^n \det\hat{\Sigma}_\phi(y)}} \quad (1)$$

where its conditional mean $\hat{\mu}_\theta(y)$ and covariance $\hat{\Sigma}_\phi(y)$ are learned from a dataset by a conditional density model f_ψ.

Model 1. *The conditional density model f_ψ is defined by*

$$\begin{bmatrix}\hat{\mu}_\theta(y)\\ vec\left[\hat{L}_\phi(y)\right]\end{bmatrix} = \begin{bmatrix}f_\theta(y)\\ f_\phi(\ddot{y})\end{bmatrix} = f_\psi(y) \quad (2)$$

where $\hat{\Sigma}_\phi(y) = \hat{L}_\phi(y)\hat{L}_\phi^\mathsf{T}(y)$ is the conditional covariance.

Remark 1. *f_ϕ is a temporary submodel that can be removed at inference time.*

Question 3. *How to train the augmented model f_ψ?*

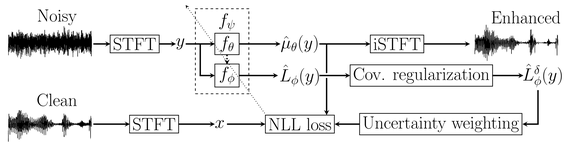


Figure 1: We augment an SE model f_θ with a temporary submodel f_ϕ to estimate heteroscedastic uncertainty during training.

Assumption 2. *Given a dataset $\{x_n, y_n\}_{n=1}^N$ containing pairs of target clean signal x_n and received noisy signal y_n, the likelihood of the joint probability distribution is assumed to be $p(x_1, x_2, \cdots, x_N|y_1, y_2, \cdots, y_N; \psi) = \prod_{n=1}^N p\left(x_n|y_n;\psi\right)$.*

3 Multivariate Gaussian NLL

The problem of maximum likelihood is equivalent to minimizing the empirical risk using the multivariate Gaussian NLL loss

$$\ell_{x,y}^{\text{Full}}(\psi) = \left[x-\hat{\mu}_\theta(y)\right]^\mathsf{T}\hat{\Sigma}_\phi^{-1}(y)\left[x-\hat{\mu}_\theta(y)\right] + \log\det\hat{\Sigma}_\phi(y). \quad (3)$$

The number of elements in $\hat{\Sigma}_\phi(y)$ is $4T^2F^2$, leading to exceedingly high training complexity.

Question 4. *How can we reduce the complexity and make the maximum likelihood tractable?*

3.1 Homoscedastic uncertainty: An MSE loss

If the covariance $\hat{\Sigma}_\phi(y)$ is assumed to be a scalar matrix $\hat{\Sigma}_\phi(y) = cI$ where c is a scalar constant and I is an identity matrix, then we actually assume the uncertainty is *homoscedastic*.

- The log-determinant term in (3) becomes a constant.
- The affinely transformed squared error reduces to an MSE.
- In this case, minimizing the Gaussian NLL is equivalent to the empirical risk minimization using an MSE loss

$$\ell_{x,y}^{\text{MSE}}(\theta) = \|x-\hat{\mu}_\theta(y)\|_2^2. \quad (4)$$

- The submodel f_ϕ is not needed for an MSE loss so the optimization is performed only on θ.

3.2 Heteroscedastic uncertainty: A diagonal case

If every random variable in the random vector drawn from $p(x|y)$ is assumed to be uncorrelated with the others, then the covariance reduces to a diagonal matrix.

- The Gaussian NLL ignores uncertainties across different T-F bins and between real and imaginary parts, leading to

$$\ell_{x,y}^{\text{Diagonal}}(\psi) = \sum_{t,f}\sum_{k\in\{r,i\}}\left[\frac{x_k^{t,f}-\hat{\mu}_{k;\theta}^{t,f}(y)}{\hat{\sigma}_{k;\phi}^{t,f}(y)}\right]^2 + 2\log\hat{\sigma}_{k;\phi}^{t,f}(y). \quad (5)$$

- The number of output units of the submodel f_ϕ is $2TF$.
- The real and imaginary parts have their own variance.
- This is a weaker assumption compared to the circularly symmetric complex Gaussian assumption used by [1].

Question 5. *Can we further weaken the assumption?*

3.3 Heteroscedastic uncertainty: A block diagonal case

We relax the uncorrelated assumption imposed between every real and imaginary part to take more uncertainty into account.

- The conditional covariance becomes a block diagonal matrix consisting of 2-by-2 blocks, giving the Gaussian NLL loss

$$\ell_{x,y}^{\text{Block}}(\psi) = \sum_{t,f}\underbrace{d_{\theta,x}^{t,f}(y)^\mathsf{T}\left[\hat{\Sigma}_\phi^{t,f}(y)\right]^{-1}d_{\theta,x}^{t,f}(y)}_{z_{x,y}^{t,f}(\psi)} + \log t_\phi^{t,f}(y) \quad (6)$$

where $t_\theta^{t,f}(y) = \left[\hat{\sigma}_{r;\phi}^{t,f}(y)\hat{\sigma}_{i;\phi}^{t,f}(y)\right]^2 - \left[\hat{\sigma}_{ri;\phi}^{t,f}(y)\right]^2$, $d_\theta^{t,f}(y) = \begin{bmatrix}x_r^{t,f}-\hat{\mu}_{r;\theta}^{t,f}(y)\\ x_i^{t,f}-\hat{\mu}_{i;\theta}^{t,f}(y)\end{bmatrix}$, and $\hat{\Sigma}_\phi^{t,f}(y) = \begin{bmatrix}\left[\hat{\sigma}_{r;\phi}^{t,f}(y)\right]^2 & \hat{\sigma}_{ri;\phi}^{t,f}(y)\\ \hat{\sigma}_{ri;\phi}^{t,f}(y) & \left[\hat{\sigma}_{i;\phi}^{t,f}(y)\right]^2\end{bmatrix}$.

- The number of output units of the submodel f_ϕ is $3TF$.
- The **inference-time complexity** of the SE model f_θ **remains the same** as using an **MSE** loss.

4 On mitigating undersampling

The expected first-order derivative of $\ell_{x,y}^{\text{Digonal}}$ w.r.t. $\hat{\mu}_{r;\theta}^{t,f}$ is

$$\mathbb{E}_{x,y}\left[\frac{\partial\ell_{x,y}^{\text{Diagonal}}}{\partial\hat{\mu}_{r;\theta}^{t,f}}\right] \approx \frac{2}{N}\sum_{n=1}^N\frac{\hat{\mu}_{r;\theta}^{t,f}(y_n)-x_{n,r}^{t,f}}{\left[\hat{\sigma}_{r;\phi}^{t,f}(y_n)\right]^2}. \quad (7)$$

- A larger variance makes the model f_θ harder to converge to a clean component.
- This undersampling issue was pointed out in a recent work [2], in which they proposed the β-NLL to mitigate undersampling.

Question 6. *Can we generalize β-NLL to the multivariate case?*

4.1 Covariance regularization

Let $\delta > 0$ be the lower bound of the eigenvalues of the Cholesy factor of the covariance matrix. The output of f_ϕ is modified by

$$\left[\hat{L}_\phi^\delta(y)\right]_{mm} = \max\left\{\left[\hat{L}_\phi(y)\right]_{mm}, \delta\right\} \quad (8)$$

for all m where $\hat{L}_\phi^\delta(y)$ is now the regularized output of f_ϕ.

4.2 Uncertainty weighting

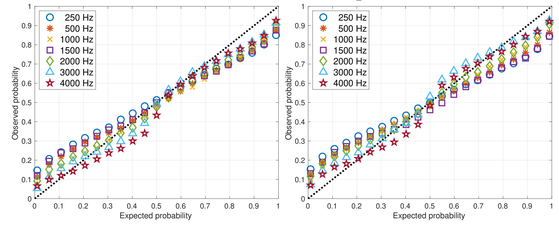
To extend the β-NLL to a multivariate Gaussian NLL, we propose an *uncertainty weighting* approach, which assigns a larger weight for a loss component according to the *minimum eigenvalue* of the covariance matrix, leading to

$$\ell_{x,y}^{\beta\text{-Block}}(\psi) = \sum_{t,f}\lambda_{\min}\left[\hat{\Sigma}_\phi^{t,f}(y)\right]^\beta z_{x,y}^{t,f}(\psi) \quad (9)$$

where $\lambda_{\min}[\cdot]$ gives the minimum eigenvalue which is treated as a constant. When $\beta = 0$, $\ell_{x,y}^{\beta\text{-Block}}(\psi)$ reduces to the original $\ell_{x,y}^{\text{Block}}(\psi)$. We pick $\beta = 0.5$.

5 Experiments

The DNS dataset [3] is used. We adopt the GCRN [4] as f_θ for investigation. f_ϕ is an additional decoder that takes the output of the in-between LSTM of the GCRN as input.



(a) Real part. (b) Imaginary part.

Figure 2: The quantile-quantile (Q-Q) plots suggest that the predictive Gaussian distributions reasonably capture the populations of the clean speech.

	δ	β	WB-PESQ			STOI (%)			SI-SDR (dB)		
SNR (dB)			-5	0	5	-5	0	5	-5	0	5
Unprocessed	n/a		1.11	1.15	1.24	69.5	77.8	85.2	-5.00	0.01	5.01
NLL ℓ^{Block}	0.0001	0	1.07	1.08	1.11	59.4	66.5	72.0	-6.46	-4.20	-2.82
	0.001	0	1.53	1.80	2.19	82.6	89.1	93.3	7.08	10.08	13.01
	0.01	0	1.61	1.92	2.33	83.9	90.1	94.0	7.82	10.73	13.51
	0.001	0.5	1.73	2.08	2.49	86.0	91.4	94.7	9.71	12.62	15.41
	0.005	0.5	**1.75**	**2.11**	**2.52**	86.4	91.6	94.8	10.09	13.05	15.88
	0.01	0.5	**1.75**	2.10	2.50	**86.7**	**91.8**	**94.9**	10.22	13.15	15.99
	0.05	0.5	1.72	2.08	2.49	86.3	91.6	94.8	10.12	13.09	15.86

Figure 3: The methods of covariance regularization and uncertainty weighting effectively improve the perceptual metric performance of the NLL loss.

	WB-PESQ			STOI (%)			SI-SDR (dB)			NORESQA-MOS		
SNR (dB)	-5	0	5	-5	0	5	-5	0	5	-5	0	5
Unprocessed	1.11	1.15	1.24	69.5	77.8	85.2	-5.00	0.01	5.01	2.32	2.36	2.45
MAE	1.50	1.76	2.09	84.4	90.4	93.9	9.83	12.63	15.02	2.77	3.27	3.65
MSE	1.63	1.94	2.29	85.1	90.6	94.0	10.24	13.21	15.97	2.86	3.52	4.02
SI-SDR	1.71	2.04	2.42	86.5	91.5	94.6	**10.96**	**13.92**	**16.80**	3.05	3.65	4.20
NLL ℓ^{Diagonal}	1.74	2.08	2.48	86.2	91.3	94.6	9.83	12.55	15.01	3.14	3.77	4.25
NLL ℓ^{Block}	**1.75**	**2.10**	**2.50**	**86.7**	**91.8**	**94.9**	10.22	13.15	15.99	**3.23**	**3.89**	**4.35**

Figure 4: The NLL using a block diagonal covariance with suitable δ and β outperforms the MAE, MSE, and SI-SDR in terms of different metrics [5].

	WB-PESQ			STOI (%)			SI-SDR (dB)		
SNR (dB)	-5	0	5	-5	0	5	-5	0	5
Unprocessed	1.11	1.15	1.24	69.5	77.8	85.2	-5.00	0.01	5.01
Best single-task	1.75	2.10	2.50	86.7	91.8	**94.9**	10.22	13.15	15.99
Hybrid	**1.77**	**2.14**	**2.53**	**86.9**	**91.9**	**94.9**	**10.62**	**13.58**	**16.30**

Figure 5: Evaluation of $\ell^{\text{Hybrid}} = \alpha\ell^{\beta\text{-Block}} + (1-\alpha)\ell^{\text{SI-SDR}}$ where $\alpha = 0.99$.

References

[1] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *ICASSP*. IEEE, 2022, pp. 386–390.

[2] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *ICLR*, 2021.

[3] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 deep noise suppression challenge," in *Interspeech*, 2021, pp. 2796–2800.

[4] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM TASLP*, 2019.

[5] P. Manocha and A. Kumar, "Speech quality assessment through MOS using non-matching references," in *Interspeech*, 2022, pp. 654–658.