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

Kuan-Lin Chen$^{12\dagger}$, Daniel D. E. Wong$^1$, Ke Tan$^1$, Buye Xu$^1$, Anurag Kumar$^1$, and Vamsi Krishna Ithapu$^1$

$^1$Meta Reality Labs Research  
$^2$Department of Electrical and Computer Engineering, University of California, San Diego

ICASSP 2023*  
April 25, 2023

$\dagger$Work done during an internship at Meta Reality Labs Research  
* A preprint is available at https://arxiv.org/abs/2211.08624
Outline

1. Single-channel speech enhancement

2. The conventional learning paradigm and popular loss functions

3. The proposed uncertainty-assisted learning framework
   - Multivariate Gaussian negative log-likelihood (NLL) functions
     - Homoscedastic uncertainty: An MSE loss
     - Heteroscedastic uncertainty: A diagonal case
     - Heteroscedastic uncertainty: A block diagonal case
   - On mitigating undersampling
     - Covariance regularization
     - Uncertainty weighting

4. Experiments

5. References
Speech enhancement (SE) aims at improving speech quality and intelligibility via recovering clean speech components from noisy recordings.

It is an essential part of many applications such as teleconferencing (Hsu et al., 2022), hearing aids (Pisha et al., 2019), and augmented hearing systems (Pisha et al., 2018).

Modern SE relies on deep learning (Lu et al., 2013; Xu et al., 2013, 2014; Wang and Chen, 2018; Tan and Wang, 2018; Pandey and Wang, 2019; Tan and Wang, 2019; Hu et al., 2020; Hao et al., 2021; Li et al., 2022).
The conventional learning paradigm and popular losses

Observation 1

Most SE models are trained without leveraging uncertainty. They assume the uncertainty is “homoscedastic.”

For example, the following loss functions

- Mean squared error (MSE)
- Mean absolute error (MAE)

are widely used in training SE models.

Question 1

What are the assumptions behind these loss functions?

Question 2

Can SE models achieve better performance if the assumptions are weakened?
Our contributions

- It was reported that minimizing a Gaussian NLL alone leads to inferior SE performance (Fang et al., 2022).

Contribution 1

*We propose a new uncertainty-assisted learning framework for SE and overcome the optimization difficulty that arises in the learning process.*

Contribution 2

*We show that, at no extra cost in terms of compute, memory, and parameters, directly minimizing a Gaussian 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.
Probabilistic models and assumptions

- Let the received signal in the STFT domain be $y_{t,f} + iy_{t,f} \in \mathbb{C}$ 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.

- We assume the clean signal is corrupted by additive noise, i.e.,

$$y = x + v$$

where $x$ and $v$ are the clean and noise random vectors, respectively.

- We assume a multivariate Gaussian model

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

where its conditional mean $\hat{\mu}_{\theta}(y)$ and covariance $\hat{\Sigma}_{\phi}(y)$ are directly learned from a dataset by a conditional density model $f_{\psi}$. 
The proposed uncertainty-assisted learning framework

Figure: We augment an SE model $f_\theta$ with a temporary submodel $f_\phi$ to estimate heteroscedastic uncertainty during training. The augmented model $f_\psi$ is defined by

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

(4)

- $f_\phi$ can be removed at inference time.

**Question 3**

*How to train the augmented model $f_\psi$?*
Given a dataset \( \{x_n, y_n\}_{n=1}^{N} \) containing pairs of target clean signal \( x_n \) and received noisy signal \( y_n \), we find the conditional mean \( \hat{\mu}_\theta(y) \) and covariance \( \hat{\Sigma}_\phi(y) \) maximizing the likelihood of the joint probability distribution

\[
p(x_1, x_2, \cdots, x_N \mid y_1, y_2, \cdots, y_N; \psi) = \prod_{n=1}^{N} p(x_n \mid y_n; \psi)
\]

where we assume the data points are independent and identically distributed.

- The maximization problem can be converted into minimizing the empirical risk using the following multivariate Gaussian NLL loss

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

- 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?*
Homoscedastic uncertainty: An MSE loss

If the covariance $\hat{\Sigma}_\phi(y)$ is assumed to be a scalar matrix

$$\hat{\Sigma}_\phi(y) = cI$$  \hspace{1cm} (7)

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 (6) 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

$$\ell_{\text{MSE}}(\theta) = \|x - \hat{\mu}_\theta(y)\|_2^2.$$  \hspace{1cm} (8)

- The submodel $f_\phi$ is not needed for an MSE loss so the optimization is performed only on $\theta$.
- Many SE works fall into this category, e.g., (Lu et al., 2013; Xu et al., 2013; Wang and Chen, 2018; Pandey and Wang, 2019; Tan and Wang, 2019).
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

$$\ell_{\text{Diagonal}}(\psi) = \sum_{t,f} \sum_{k \in \{r,i\}} \left( \frac{x_{k}^{t,f} - \hat{\mu}_{k;\theta}(y)}{\hat{\sigma}_{k;\phi}(y)} \right)^{2} + 2 \log \hat{\sigma}_{k;\phi}(y) \quad (9)$$

- The number of output units of the submodel $f_{\phi}$ is $2TF$.
- (9) allows the real and imaginary parts to have their own variance.
- This is a weaker assumption compared to the circularly symmetric complex Gaussian assumption used by Fang et al. (2022).

**Question 5**

Can we further weaken the assumption?
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

\[
\ell_{x,y}^{\text{Block}}(\psi) = \sum_{t,f} d_{t,f}^{t,f}(y) \left[ \hat{\Sigma}_{\phi}^{t,f}(y) \right]^{-1} d_{t,f}^{t,f}(y) + \log t_{\phi}^{t,f}(y) \tag{10}
\]

where

\[
t_{\theta}^{t,f}(y) = \left( \hat{\sigma}_{r;\phi}^{t,f}(y)\hat{\sigma}_{i;\phi}^{t,f}(y) \right)^2 - \left( \hat{\sigma}_{r;\phi}^{t,f}(y) \right)^2 , \tag{11}
\]

\[
d_{\theta}^{t,f}(y) = \begin{bmatrix} \hat{\mu}_{r;\theta}^{t,f}(y) \\ \hat{\mu}_{i;\theta}^{t,f}(y) \end{bmatrix} , \hat{\Sigma}_{\phi}^{t,f}(y) = \begin{bmatrix} \hat{\sigma}_{r;\phi}^{t,f}(y) \hat{\sigma}_{r;\phi}^{t,f}(y) \\ \hat{\sigma}_{r;\phi}^{t,f}(y) \hat{\sigma}_{i;\phi}^{t,f}(y) \end{bmatrix} . \tag{12}
\]

- The number of output units of the submodel \( f_{\phi} \) is \( 3TF \).

- The inference-time complexity of the SE model \( f_{\theta} \) remains the same as using an MSE loss or uncorrelated Gaussian NLL loss.
The undersampling problem

Taking the uncorrelated Gaussian NLL for example, the expected first-order derivative of $\ell_{x,y}^{\text{Diagonal}}$ with respect to $\hat{\mu}_{r;\theta}^{t,f}$ can be approximated by

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

(13)

- Given the unconstrained variance in the denominator, a larger variance makes the model $f_{\theta}$ harder to converge to a clean component compared to a loss component with a smaller variance.
- This undersampling issue was pointed out in a recent work by Seitzer et al. (2021), in which they proposed the $\beta$-NLL to mitigate undersampling.

Question 6

Can we generalize $\beta$-NLL to the multivariate case?
Let $\delta > 0$ be the lower bound of the eigenvalues of the Cholesy factor of the covariance matrix. The output of $f_{\phi}$ is modified by

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

(14)

for all $m \in \{1, 2, \cdots, 2TF\}$ where $\hat{L}^\delta_{\phi}(y)$ is now the regularized output of $f_{\phi}$.
To extend the $\beta$-NLL to a multivariate Gaussian NLL, we propose an \textit{uncertainty weighting} approach, which assigns a larger weight for a loss component according to the \textit{minimum eigenvalue} of the covariance matrix, leading to

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

(15)

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 $\beta = 0.5$. 

To extend the \textit{uncertainty weighting} approach, which assigns a larger weight for a loss component according to the \textit{minimum eigenvalue} of the covariance matrix, leading to
Experimental setup

- The DNS dataset (Reddy et al., 2021).
- We adopt the gated convolutional recurrent network (GCRN) (Tan and Wang, 2019) as the SE model $f_\theta$ for investigation.
- Given that the original GCRN has an encoder-decoder architecture with long short-term memory (LSTM) in between, we formulate the temporary submodel $f_\phi$ as an additional decoder that takes the output of the in-between LSTM as input.
- The augmented model $f_\psi$ formed by these two models is a GCRN with two distinct decoders.
Calibration of the probabilistic model

Question 7

Is it reasonable to assume the Gaussian probabilistic model?

(a) Real part.

(b) Imaginary part.

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

| SNR (dB) | $\delta$ | $\beta$ | WB-PESQ | STOI (%) | SI-SDR (dB) | NORESQA-MOS |
|----------|---------|---------|---------|----------|-------------|-------------|
|          | -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        | 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        |

| Gaussian NLL: Diagonal $\hat{\Sigma}_\phi$ | 0.0001 | 0 | 1.11 | 1.18 | 1.28 | 69.6 | 77.3 | 83.0 | 0.79 | 4.37 | 7.48 | 1.95 | 2.16 | 2.40 |
|                                           | 0.01   | 0 | 1.59 | 1.88 | 2.28 | 83.5 | 89.7 | 93.7 | 7.65 | 10.61 | 13.31 | 2.97 | 3.60 | 4.14 |
|                                           | 0.01   | 0.5| 1.74 | 2.08 | 2.48 | 86.2 | 91.3 | 94.6 | 9.83 | 12.55 | 15.01 | 3.14 | 3.77 | 4.25 |

| Gaussian NLL: Block diagonal $\hat{\Sigma}_\phi$ | 0.0001 | 0 | 1.07 | 1.08 | 1.11 | 59.4 | 66.5 | 72.0 | -6.46 | -4.20 | -2.82 | 1.56 | 1.47 | 1.44 |
|                                                | 0.001  | 0 | 1.53 | 1.80 | 2.19 | 82.6 | 89.1 | 93.3 | 7.08  | 10.08  | 13.01 | 2.71 | 3.33 | 3.97 |
|                                                | 0.01   | 0 | 1.61 | 1.92 | 2.33 | 83.9 | 90.1 | 94.0 | 7.82  | 10.73  | 13.51 | 2.98 | 3.60 | 4.15 |
|                                                | 0.001  | 0.5| 1.73 | 2.08 | 2.49 | 86.0 | 91.4 | 94.7 | 9.71  | 12.62  | 15.41 | 3.11 | 3.79 | 4.30 |
|                                                | 0.005  | 0.5| 1.75 | 2.11 | 2.52 | 86.4 | 91.6 | 94.8 | 10.09 | 13.05  | 15.88 | 3.07 | 3.75 | 4.22 |
|                                                | 0.01   | 0.5| 1.75 | 2.10 | 2.50 | 86.7 | 91.8 | 94.9 | 10.22 | 13.15  | 15.99 | 3.23 | 3.89 | 4.35 |
|                                                | 0.05   | 0.5| 1.72 | 2.08 | 2.49 | 86.3 | 91.6 | 94.8 | 10.12 | 13.09  | 15.86 | 2.96 | 3.63 | 4.15 |

**Table:** The methods of covariance regularization and uncertainty weighting effectively improve perceptual metric performance of multivariate Gaussian NLLs.

- The NLL using a block diagonal covariance with suitable $\delta$ and $\beta$ outperforms the MAE, MSE, and SI-SDR in terms of different metrics (Manocha and Kumar, 2022).
A hybrid loss performs better than the best single-task loss

Let a hybrid loss be defined as

$$\ell_{\text{Hybrid}} = \alpha \ell_{\beta}\text{-Block} + (1 - \alpha)\ell_{\text{SI-SDR}}$$

(16)

with $\alpha = 0.99$, $\delta = 0.01$, and $\beta = 0.5$.

| SNR (dB) | WB-PESQ | STOI (%) | SI-SDR (dB) |
|----------|---------|----------|-------------|
|          | -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** |

Table: Performance evaluation of the hybrid loss defined by (16).
Fang, H., Peer, T., Wermter, S., and Gerkmann, T. (2022). Integrating statistical uncertainty into neural network-based speech enhancement. In ICASSP, pages 386–390. IEEE.

Hao, X., Su, X., Horaud, R., and Li, X. (2021). Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In ICASSP, pages 6633–6637. IEEE.

Hsu, Y., Lee, Y., and Bai, M. R. (2022). Learning-based personal speech enhancement for teleconferencing by exploiting spatial-spectral features. In ICASSP, pages 8787–8791. IEEE.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., and Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In Interspeech, pages 2472–2476.

Li, A., You, S., Yu, G., Zheng, C., and Li, X. (2022). Taylor, can you hear me now? A Taylor-unfolding framework for monaural speech enhancement. In International Joint Conference on Artificial Intelligence.

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In Interspeech, pages 436–440.

Manocha, P. and Kumar, A. (2022). Speech quality assessment through MOS using non-matching references. In Interspeech, pages 654–658.

Pandey, A. and Wang, D. (2019). TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In ICASSP, pages 6875–6879. IEEE.

Pisha, L., Hamilton, S., Sengupta, D., Lee, C.-H., Vastare, K. C., Zubatiy, T., Luna, S., Yalcin, C., Grant, A., Gupta, R., Chockalingam, G., Rao, B. D., and Garudadri, H. (2018). A wearable platform for research in augmented hearing. In Asilomar Conference on Signals, Systems, and Computers, pages 223–227. IEEE.

Pisha, L., Warchall, J., Zubatiy, T., Hamilton, S., Lee, C.-H., Chockalingam, G., Mercier, P. P., Gupta, R., Rao, B. D., and Garudadri, H. (2019). A wearable, extensible, open-source platform for hearing healthcare research. IEEE Access, 7:162083–162101.

Reddy, C. K., Dubey, H., Koishida, K., Nair, A., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., and Srinivasan, S. (2021). INTERSPEECH 2021 deep noise suppression challenge. In Interspeech, pages 2796–2800.

Seitzer, M., Tavakoli, A., Antic, D., and Martius, G. (2021). On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In International Conference on Learning Representations.

Tan, K. and Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In Interspeech, pages 3229–3233.

Tan, K. and Wang, D. (2019). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:380–390.

Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10):1702–1726.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2013). An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters, 21(1):65–68.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1):7–19.