Integrating Uncertainty Into Neural Network-Based Speech Enhancement

Huajian Fang , *Student Member, IEEE*, Dennis Becker , Stefan Wermter , *Member, IEEE*, and Timo Gerkmann , *Senior Member, IEEE*

Abstract—Supervised masking approaches in the time-frequency domain aim to employ deep neural networks to estimate a multiplicative mask to extract clean speech. This leads to a single estimate for each input without any guarantees or measures of reliability. In this paper, we study the benefits of modeling uncertainty in clean speech estimation. Prediction uncertainty is typically categorized into *aleatoric uncertainty* and *epistemic uncertainty*. The former refers to inherent randomness in data, while the latter describes uncertainty in the model parameters. In this work, we propose a framework to jointly model aleatoric and epistemic uncertainties in neural network-based speech enhancement. The proposed approach captures aleatoric uncertainty by estimating the statistical moments of the speech posterior distribution and explicitly incorporates the uncertainty estimate to further improve clean speech estimation. For epistemic uncertainty, we investigate two Bayesian deep learning approaches: Monte Carlo dropout and Deep ensembles to quantify the uncertainty of the neural network parameters. Our analyses show that the proposed framework promotes capturing practical and reliable uncertainty, while combining different sources of uncertainties yields more reliable predictive uncertainty estimates. Furthermore, we demonstrate the benefits of modeling uncertainty on speech enhancement performance by evaluating the framework on different datasets, exhibiting notable improvement over comparable models that fail to account for uncertainty.

Index Terms—Speech enhancement, Bayesian estimator, uncertainty estimation, deep neural networks.

I. INTRODUCTION

SPEECH recorded in noisy environments is often corrupted by background noise, which renders it difficult to understand by either humans or machines via automatic speech recognition systems. These problems call for robust speech enhancement algorithms, which extract desired clean speech from noisy mixtures to improve speech quality and intelligibility of recordings. In this paper, we consider single-channel speech enhancement.

Speech enhancement algorithms typically utilize the short-time Fourier transform (STFT) to transfer the recorded signal into the time-frequency domain, where multiplicative filters can be applied to obtain an estimate of clean speech [1], [2]. Various Bayesian estimators, e.g., maximum a posteriori (MAP) and minimum mean squared error (MMSE) estimators, have been developed based on different statistical distributions about speech and noise, aiming to restore either the spectral coefficients of the STFT or the spectral magnitudes [3], [4], [5], [6]. Given the assumption that speech is degraded by uncorrelated additive noise and that both follow complex Gaussian distributions with zero mean, the well-known Wiener filter can be derived. Traditionally, the speech and noise variances estimated by statistical model-based methods [1], [7] can be used to construct the MMSE-optimal Wiener filter.

Recently, neural networks have been widely used in speech enhancement methods due to their flexibility and effectiveness in nonlinear modeling. Depending on their application, varying degrees of success are reported [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Specifically, deep neural networks have been utilized to replace some of the building blocks of conventional speech enhancement methods. For instance, a neural network-based speech presence probability estimator has been proposed in [8] and combined with a single-channel multi-frame approach [9]. In [10], [11], neural networks are employed to estimate speech and noise power spectrum densities that are required in various Bayesian estimators. Additionally, recent work has leveraged the probabilistic modeling of generative networks for speech enhancement. For example, the variational autoencoder (VAE) has been used to estimate the clean speech distribution, which is then combined with a separate noise model to construct a noise reduction Wiener filter [12], [14]. The robustness of this filter can be further improved by injecting noise information [16], temporal dependencies [20], [21], [22], and information from other modalities, such as vision [17], [23]. Besides, speech enhancement approaches based on perceptual metric-guided adversarial training [24], [25] and diffusion-based generative models [26], [27] have also been presented. In contrast, supervised masking approaches [18] aim to learn a mapping from the noisy input to a masking filter. It

allows neural networks to directly estimate a time-frequency filter by training on a large amount of noisy-clean speech pairs using an appropriate cost function [19]. In this work, we focus on supervised masking approaches.

While the time-frequency noise-removing filter aims to remove noise with minimum speech distortions, the algorithm's robustness and reliability are not guaranteed, especially when speech is corrupted by previously unobserved noise. To alleviate this shortcoming, research has been conducted to investigate how to generalize to unseen situations by, e.g., developing more sophisticated network architectures, improved features, or including more training data that covers a wide variety of acoustic scenarios [28], [29], [30]. The first is often accompanied by a tremendous increase in model parameters, while the latter is rather time-demanding. Still, improving the generalization ability of neural networks in unseen scenarios is an unsolved problem considering the black-box nature of neural networks. It is thus necessary and beneficial to obtain the associated uncertainty as an indicator of reliability besides the point estimate, especially when the model is processing out-of-distribution samples that are insufficiently represented by training data.

In machine learning, predictive uncertainty is typically decomposed into two categories [31], [32], [33]: *aleatoric* uncertainty and *epistemic* uncertainty. The term aleatoric uncertainty is used to describe the uncertainty of an estimate due to the intrinsic randomness of noisy observations. For speech enhancement, it originates from the stochastic nature of both speech and noise and is reflected in the variance of the clean speech posterior predictive distribution. Epistemic uncertainty is of different nature: If the parameters of a neural network are trained, e.g., using different training data, different initialization, or a different number of epochs, different parameters result. Therefore, also the parameters of a neural network used to estimate clean speech are uncertain. This uncertainty of the parameters is called epistemic uncertainty (also known as *model uncertainty*). For a general introduction to uncertainty modeling, readers are suggested to refer to a review article by Hüllermeier et al. [31]. Various uncertainty measures have been employed in the deep regression setting, such as confidence intervals, differential entropy, and variance. Depeweg et al. [34] propose to measure uncertainty based on the entropy of the predictive distribution, which represents the information level of random variables. Pearce et al. [35] use confidence intervals (which state how certain the estimate is within a certain range) in a distribution-free setting. In this paper, we address uncertainty modeling in a probabilistic way following [33], [36], [37] and measure the uncertainty in terms of the *variance*.

Aleatoric uncertainty: Due to the stochastic nature of speech and noise, a mapping from noisy speech to clean speech is uncertain as reflected by the posterior predictive distribution of clean speech. We can model this posterior using a specific conditional distribution, such as a Gaussian or a Laplacian [33], [36], [38], and employ a neural network to directly estimate the statistical moments of this distribution. While the predicted mean is the MMSE estimate of the target [2], the associated variance can be used to quantify the data inherent uncertainty, i.e., aleatoric uncertainty [33].

Few studies in neural network-based speech enhancement have incorporated the uncertainty of aleatoric nature. Chai et al. propose to use a generalized Gaussian distribution to model the prediction error on a logarithmic scale [39]. In [40], a neural network is used to estimate the parameters of a Gaussian mixture model, which then serves as the basis of an extra statistical model-based speech enhancement approach. This results in only a slight improvement over the baseline optimized with the MMSE criterion. Siniscalchi [41] leverages neural networks to learn a histogram distribution to approximate the conditional target speech distribution, which is assumed to be a truncated Gaussian distribution with a fixed variance in each frequency band. However, the fixed variance does not help to capture data-dependent uncertainty.

Epistemic uncertainty: Estimating the statistical moments of the speech posterior predictive distribution allows capturing aleatoric uncertainty, but fails to account for epistemic uncertainty, which corresponds to the uncertainty in neural network parameters [31], [32], [33]. Epistemic uncertainty can be captured using Bayesian inference approaches, which instead of modeling the parameters of a neural network as *deterministic* values, place a distribution over the network parameters and estimates the posterior distribution of the *stochastic* network parameters [33]. By sampling from the posterior network parameter distribution, multiple sets of neural network parameter realizations can be obtained, thus producing multiple output predictions for each input sample. Uncertainty in predictions due to epistemic uncertainty can be empirically quantified by the variance in these output predictions [31], [33]. While the true posterior network distribution is intractable [42], it can be approximated using 1) Markov Chain Monte Carlo (MCMC) methods [43], [44], which are sampling-based approaches that construct a Markov Chain with the posterior network parameter distribution as its stationary distribution, 2) variational inference [42], [45], [46], which approximates the true posterior network parameter distribution with a tractable variational distribution, and 3) ensemble approaches [36], [47], [48], which were proposed from the frequentist perspective but are considered as an approximate Bayesian approach [37], [49]. For instance, Gal et al. [42] perform variational inference and interpret the dropout regularization technique [50] as imposing Bernoulli distributions on the neural network's weights. This method, referred to as Monte Carlo dropout (*MC dropout*), provides a set of target estimates from multiple forward passes by activating dropout at inference. This set of predictions can empirically approximate the outcome distribution for each input sample and allows inference of the variance (i.e., epistemic uncertainty). In contrast, *Deep ensembles* proposed in [36] can quantify epistemic uncertainty by training multiple neural networks with random weight initialization [37], [38].

Recent studies attempt to consider the uncertainty of epistemic nature in, e.g., speech emotion recognition [51], [52] and speech recognition [53], [54], [55]. In [51], epistemic uncertainty is captured in a speech emotion recognition model for selective prediction, where samples with low confidence (high uncertainty) are rejected. Braun et al. [53] apply a Gaussian distribution to the weights of an end-to-end speech recognition

model to capture uncertainty of neural network parameters, which is then used for parameter pruning. In a recent publication [54], epistemic uncertainty is employed to improve the robustness of domain adaptation for speech recognition. However, quantifying epistemic uncertainty in neural network-based speech enhancement remains unexplored.

Contributions: Capturing overall predictive uncertainty, which reflects both aleatoric and epistemic uncertainties, is challenging, especially for deep neural networks, but crucial for an understanding of the model's prediction behaviour. In this work, we propose a method that allows capturing aleatoric uncertainty and combining it with epistemic uncertainty approximations to quantify overall predictive uncertainty. In the context of neural network-based speech enhancement, to the best of our knowledge, this is the first work to study different sources of uncertainty in a joint framework and provides for systematic analyses.

We follow the complex Gaussian speech-plus-noise assumption and propose to train a neural network to estimate the Wiener filter and its variance, which quantifies aleatoic uncertainty, based on the MAP inference of *complex spectral coefficients*. To regularize the variance estimation, we build an approximate MAP (AMAP) estimator of *spectral magnitudes* using the estimated Wiener filter (mean of the complex clean speech posterior predictive distribution) and uncertainty (variance of the complex clean speech posterior distribution) explicitly. The resulting AMAP estimator is in turn used in conjunction with the MAP inference of complex spectral coefficients to form a novel hybrid loss function. Rather than discarding uncertainty information at inference, the proposed scheme allows us to explicitly incorporate aleatoric uncertainty approximations into clean speech estimation in a principled way to further correct erroneous speech estimates.

Previous studies on modeling epistemic uncertainty have focused on other tasks than speech enhancement, e.g., [38], [51], [52], [53], [54], [55], [56]. Yet, questions such as how reliable and accurate the estimates of epistemic uncertainty are in speech enhancement, and how modeling epistemic uncertainty affects enhancement performance, have not been addressed. To this end, we investigate two Bayesian deep learning techniques: MC dropout [42] and Deep ensembles [36] to capture epistemic uncertainty in clean speech estimation due to their efficiency in approximating Bayesian inference. Although previous works have explored ensemble-based speech enhancement methods [57], [58], they did not investigate the effectiveness of ensemble-based methods for uncertainty estimation.

Moreover, we propose to estimate overall predictive uncertainty reflecting both aleatoric and epistemic uncertainties by combining the proposed hybrid loss function with the ensemble-based method. Finally, we present a comprehensive analysis of uncertainty from different sources and show their impacts on speech enhancement performance over different datasets, which we hope lays the foundation for further use of uncertainties.

This paper extends our previous conference publication [59], which studied aleatoric uncertainty. Here, we propose to additionally capture epistemic uncertainty and combine them to quantify overall predictive uncertainty in clean speech estimation. Furthermore, we provide a more detailed analysis with respect to uncertainty estimates from different sources in a joint framework. Section II describes the signal model. In Section III, we propose to estimate the uncertainty of aleatoric nature following the complex Gaussian-distributed speech posterior and present how this uncertainty can be incorporated into clean speech estimation. In Section IV, we show how to capture epistemic uncertainty and quantify overall predictive uncertainty that combines different sources of uncertainty. We introduce the experimental setting in Section V, analyze uncertainty estimates in Section VI, and present enhancement performance in Section VII. Section VIII summarizes the findings.

II. SIGNAL MODEL

In the single-channel speech enhancement problem, the noisy mixture consists of clean speech and additive noise. We apply the STFT to obtain the representation in the time-frequency domain as:

$$X_{ft} = S_{ft} + N_{ft}, \tag{1}$$

where X_{ft}, S_{ft}, and N_{ft} represent the complex spectral coefficients of mixture, speech, and noise, at the time frame $t \in \{1, 2, \ldots, T\}$ and the frequency bin $f \in \{1, 2, \ldots, F\}$. T and F denote the number of time frames and frequency bins respectively. The objective is to recover clean speech in the time-frequency domain by applying a multiplicative filter. To derive such a filter, various assumptions are made according to different signal characteristics. By assuming that the speech and noise coefficients are uncorrelated and follow a circularly symmetric complex Gaussian distribution,

$$S_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{s,ft}^2), \quad N_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{n,ft}^2), \tag{2}$$

where $\sigma_{s,ft}^2$ and $\sigma_{n,ft}^2$ represent the variances of speech and noise respectively, the likelihood $p(X_{ft}|S_{ft})$ follows a complex Gaussian distribution with mean S_{ft} and variance $\sigma_{n,ft}^2$, given by

$$p(X_{ft}|S_{ft}) = \frac{1}{\pi\sigma_{n,ft}^2} \exp\left(-\frac{|X_{ft} - S_{ft}|^2}{\sigma_{n,ft}^2}\right). \tag{3}$$

With the likelihood in (3) and the prior in (2), we can apply Bayes' theorem to obtain the posterior distribution of clean speech as a complex Gaussian of the form [2]:

$$p(S_{ft}|X_{ft}) = \frac{1}{\pi\lambda_{ft}} \exp\left(-\frac{|S_{ft} - W_{ft}^{\text{WF}}X_{ft}|^2}{\lambda_{ft}}\right), \tag{4}$$

$$W_{ft}^{\text{WF}} = \frac{\sigma_{s,ft}^2}{\sigma_{s,ft}^2 + \sigma_{n,ft}^2}, \quad \lambda_{ft} = \frac{\sigma_{s,ft}^2\sigma_{n,ft}^2}{\sigma_{s,ft}^2 + \sigma_{n,ft}^2}. \tag{5}$$

W_{ft}^{WF} is recognized as the *Wiener filter* and λ_{ft} is the variance of the posterior distribution. Under this assumption, the MMSE estimator, which corresponds to the expectation of the posterior distribution, leads to the Wiener filter applied as:

$$\widetilde{S}_{ft} = W_{ft}^{\text{WF}} \cdot X_{ft}. \tag{6}$$

Due to the symmetry of the complex Gaussian distribution, the MAP estimator of complex speech coefficients is identical to the MMSE estimator.

III. ALEATORIC UNCERTAINTY ESTIMATION

Although speech enhancement is typically formulated as a problem with a single output, the dependency between input and output can be modeled stochastically by means of a speech posterior predictive distribution $p(S_{ft}|X_{ft})$, i.e., a variance λ_{ft} is associated with the clean speech estimate and can be interpreted as a measure of uncertainty of the Wiener estimate [2]. This uncertainty accounts for random effects in data and is referred to as *aleatoric uncertainty* [33], [36]. When properly captured, aleatoric uncertainty can reflect the expected estimation error in the absence of ground truth.

A. Deep Aleatoric Uncertainty Estimation

In contrast to traditional signal processing techniques [1], [2], [60], where the Wiener filter is constructed by separately estimating the variances of speech and noise from the noisy mixture X_{ft}, neural network-based supervised masking methods allow direct estimation of multiplicative filters. Besides the Wiener filter W_{ft}^{WF}, one can further estimate the data-dependent aleatoric uncertainty λ_{ft} if the neural network is optimized using the speech posterior predictive distribution (4), i.e., by minimizing the negative logarithm of the posterior distribution of clean speech $p(S_{ft}|X_{ft})$ (the logarithm does not affect the optimization problem due to monotonicity) and averaging over time-frequency bins:

$$\widetilde{W}_{ft}^{\text{WF}}, \widetilde{\lambda}_{ft}$$
$$= \underset{W_{ft}^{\text{WF}}, \lambda_{ft}}{\arg\min} \underbrace{\frac{1}{FT} \sum_{f,t} \log(\lambda_{ft}) + \frac{|S_{ft} - W_{ft}^{\text{WF}} X_{ft}|^2}{\lambda_{ft}}}_{\mathcal{L}_{p(S|X)}}, \quad (7)$$

where $\widetilde{W}_{ft}^{\text{WF}}$, $\widetilde{\lambda}_{ft}$ denote estimates of the Wiener filter and associated aleatoric uncertainty [33], [36].

In contrast, if we assume a constant uncertainty for all time-frequency bins, i.e., $\lambda_{ft} = \lambda^*$, and refrain from explicitly optimizing for λ^*, $\mathcal{L}_{p(S|X)}$ degenerates into the well-known mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{FT} \sum_{f,t} |S_{ft} - W_{ft}^{\text{WF}} X_{ft}|^2, \quad (8)$$

which is widely used in neural network-based regression tasks including speech enhancement [19]. However, neural networks trained to perform point estimation do not necessarily output reliable estimates for clean speech when processing out-of-distribution samples that are underrepresented by the training data [28]. In this work, we discard the assumption of constant uncertainty; instead, we propose to treat uncertainty estimation as an additional task by training a neural network with the negative log speech posterior $\mathcal{L}_{p(S|X)}$. Consequently, this method not only allows us to obtain a noise-removing mask, but

also empowers the model to capture the uncertainty of aleatoric nature associated with predictions.

Modeling aleatoric uncertainty by minimizing the logarithm of the posterior predictive distribution results in an improvement over baselines that fail to account for uncertainty in computer vision tasks [33]. However, directly using $\mathcal{L}_{p(S|X)}$ as the loss function is prone to overfitting [59] and may result in reduced estimation performance of the Wiener filter. A recent publication [61] also reveals that directly minimizing the logarithm of the conditional probability hinders the training of mean estimation, which leads to premature convergence. To tackle this problem, we propose an additional regularization of the loss function by incorporating the estimated uncertainty into clean speech estimation as described next.

B. Joint Enhancement and Uncertainty Estimation

Estimating uncertainty λ_{ft} associated with the Wiener filter is challenging since ground truth of uncertainty is not readily available. Instead, uncertainty estimation is an unsupervised task with an unspecified search space, which can potentially lead to unstable training [62], [63]. In this work, we propose to incorporate a subsequent speech enhancement task that explicitly uses both the Wiener filter and its uncertainty λ_{ft} during the training procedure. The speech enhancement task provides additional coupling between the outputs (Wiener filter and uncertainty). In this manner, the neural network is guided to estimate the uncertainty values that are relevant to the speech enhancement task, as well as to enhance the estimation of the Wiener filter.

Considering complex coefficients with a symmetric posterior (4), the MAP and MMSE estimators both lead directly to the Wiener filter W_{ft}^{WF} and do not require an uncertainty estimate. However, this situation changes if we consider spectral magnitude estimation. The magnitude posterior $p(|S_{ft}| \,|X_{ft})$, derived by integrating the phase out of (4), follows a Rician distribution [4]

$$p(|S_{ft}| \,|X_{ft})$$
$$= \frac{2|S_{ft}|}{\lambda_{ft}} \exp\left(-\frac{|S_{ft}|^2 + (W_{ft}^{\text{WF}})^2 |X_{ft}|^2}{\lambda_{ft}}\right)$$
$$\times I_0\left(\frac{2|X_{ft}| |S_{ft}| W_{ft}^{\text{WF}}}{\lambda_{ft}}\right), \quad (9)$$

where $I_0(\cdot)$ is the modified zeroth-order Bessel function of the first kind.

In order to compute the MAP estimate for the spectral magnitude, the mode of the Rician distribution has to be estimated, which is difficult to do analytically. However, it can be approximated by substituting a Bessel function approximation following [64] into (9) and maximizing with respect to the spectral magnitude, yielding a simple closed-form expression [2], [4]:

$$|\widehat{S}_{ft}| \approx W_{ft}^{\text{AMAP}}|X_{ft}|$$
$$= \left(\frac{1}{2}W_{ft}^{\text{WF}} + \sqrt{\left(\frac{1}{2}W_{ft}^{\text{WF}}\right)^2 + \frac{\lambda_{ft}}{4|X_{ft}|^2}}\right) |X_{ft}|, \quad (10)$$
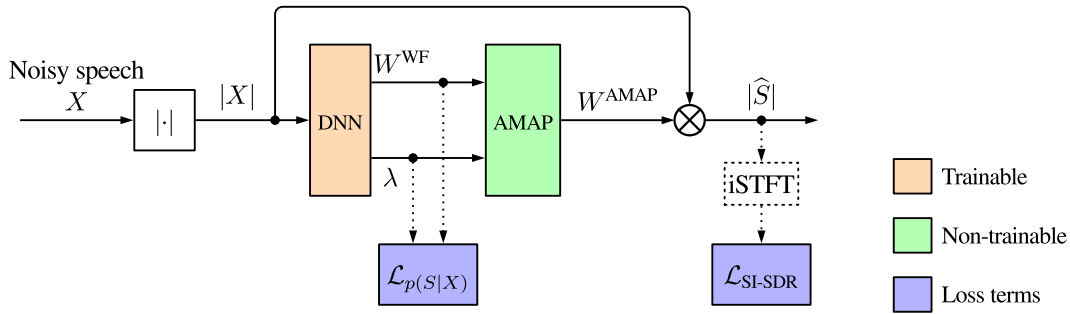
Fig. 1. Block diagram of the proposed neural network-based aleatoric uncertainty estimation.

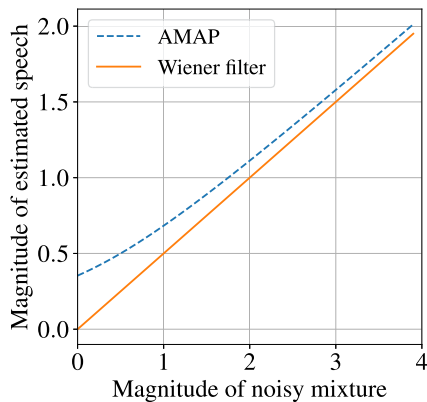Fig. 2. Input-output characteristics of the AMAP estimator W_{ft}^{AMAP} and Wiener filter W_{ft}^{WF} (setting $\sigma_{s,ft}^2 = \sigma_{n,ft}^2 = 1$ in this example).

where $|\widehat{S}_{ft}|$ is an estimate of the clean spectral magnitude $|S_{ft}|$ using the AMAP estimator of spectral magnitudes W_{ft}^{AMAP}. It can be noticed that the estimator W_{ft}^{AMAP} utilizes both the Wiener filter W_{ft}^{WF} and the associated uncertainty λ_{ft}. Fig. 2 illustrates the input-output estimation characteristics of the AMAP estimator and Wiener filter [2]. We can see that W_{ft}^{AMAP} is *nonlinear* with respect to the noisy input and tends to cause less target attenuation than the Wiener filter especially for low inputs. This indicates that incorporating the associated uncertainty λ_{ft} may increase the robustness of the estimator by potentially preserving more speech at the slight cost of noise removal.

After combining the estimated magnitude $|\widehat{S}_{ft}|$ with the noisy phase, we can apply the inverse STFT to obtain an estimate of the time-domain speech signal, denoted as $\widehat{s}$. Afterwards, the estimated time-domain signal is used to compute the negative scale-invariant signal-to-distortion ratio (SI-SDR) metric [65]:

$$\mathcal{L}_{\text{SI-SDR}} = -10 \log_{10}\left(\frac{||\alpha s||^2}{||\alpha s - \widehat{s}||^2}\right), \quad \alpha = \frac{\widehat{s}^T s}{||s||^2}, \quad (11)$$

which is leveraged as an additional term in the loss function that forces the speech estimate (computed with W_{ft}^{AMAP}) to be similar to the time-domain clean speech target s. While a spectrum loss like (8) is a straightforward solution to regularize the uncertainty estimation, the time-domain loss is expected to be more effective

since it is directly related to the raw waveform, implicitly taking phase information into account and thus promoting speech reconstruction for better perceptual performance [66].

Eventually, we propose to combine the SI-SDR loss $\mathcal{L}_{\text{SI-SDR}}$ with the negative log-posterior $\mathcal{L}_{p(S|X)}$ given in (7), and train the neural network using a hybrid loss function

$$\mathcal{L} = \beta \mathcal{L}_{p(S|X)} + (1 - \beta)\mathcal{L}_{\text{SI-SDR}}, \quad (12)$$

with the weighting factor $\beta \in [0, 1]$. By explicitly using the estimated uncertainty for the speech enhancement task, the hybrid loss guides both mean and variance estimation to improve speech enhancement performance. Fig. 1 depicts a block diagram of this approach.

IV. BAYESIAN UNCERTAINTY ESTIMATION

While neural networks performing point estimation have demonstrated effectiveness in speech enhancement, it is not guaranteed that neural networks can generalize well to unfamiliar acoustic situations. Therefore, to quantify the overall predictive confidence regarding the estimated clean speech, it is necessary to also assess the uncertainty of the neural network parameters (i.e., *epistemic uncertainty*). Note that a single neural network optimized using the proposed hybrid loss (12) allows capturing aleatoric uncertainty but is unaware of epistemic uncertainty. To solve this, we can utilize Bayesian deep learning approaches, assuming that the weights of a neural network follow some probability distribution rather than deterministic values. Furthermore, when combined with the loss (12), an ensemble of networks can provide both aleatoric uncertainty and epistemic uncertainty estimates.

A. Epistemic Uncertainty Estimation

Bayesian deep learning provides a set of principled methods to capture epistemic uncertainty [36], [42], [43], [44], [46], [48]. Early work on MCMC methods [43], [44] constructs a Markov chain with the posterior network parameter distribution as its stationary distribution and generates multiple network parameter realizations by sampling from this distribution. However, MCMC methods are computationally inefficient and do not scale well to neural networks with a large number of parameters [37], [38]. Recent work based on variational inference allows approximating the true posterior network parameter distribution with a tractable distribution [45], [46], while at the

same time ensemble-based methods are proposed as simple and scalable frequentist alternatives to model uncertainty [36], [47], [48]. Among the existing Bayesian deep learning methods, MC dropout and Deep ensembles have shown their scalability in large neural network-based problems, such as semantic segmentation [33] and depth estimation [37]. Here, we investigate their effectiveness for uncertainty estimation in speech enhancement.

We define a neural network as a function parameterized by θ and a training dataset that contains noisy-clean speech pairs $\mathcal{D} = \{(S_{11}, X_{11}), \ldots, (S_{FT}, X_{FT})\}$. Hereafter we omit the indices ft, since all time-frequency bins are treated independently in (4). Since the posterior network parameter distribution $p(\theta|\mathcal{D})$ is computationally intractable in a high dimensional space, variational inference approximates the true posterior network parameter distribution by a pre-specified variational distribution $q(\theta)$ and the speech posterior predictive distribution at inference time is obtained by marginalizing out $q(\theta)$ as:

$$
\begin{aligned}
p(S|X, \mathcal{D}) &= \int p(S|X, \theta) p(\theta|\mathcal{D}) d\theta \\
&\approx \frac{1}{M} \sum_{m=1}^{M} p(S|X, \theta_m), \quad \theta_m \sim q(\theta),
\end{aligned}
\tag{13}
$$

where θ_m represents m-th sampling from $q(\theta)$ [67]. MC dropout approximates the posterior network parameter distribution using the Bernoulli distribution and samples neural network weights by activating dropout at inference time. Gal et al. provide further details on the derivations in [42]. This allows obtaining M target speech estimates from multiple stochastic forward passes for each input. In contrast, Deep ensembles repeatedly train the same model M times with random initialization and random data shuffling [36], generating M neural networks with deterministic network parameter estimates $\{\theta_m\}_{m=1}^{m=M}$. Since θ_m can be viewed as independent samples from a certain approximate distribution $q(\theta)$, Deep ensembles can be considered equivalent to approximate Bayesian inference [37]. Therefore, the predictive distribution is obtained similarly to (13). Furthermore, neural networks usually contain a large number of parameters, which makes them multi-modal in the parameter space. Different initialization starting points in Deep ensembles allow the neural network to converge to different local optima, thus potentially capturing multiple modes of $p(\theta|\mathcal{D})$ [37], [48].

Epistemic uncertainty can be approximated by building an ensemble of neural networks using either MC dropout or Deep ensembles, where each network is trained to estimate the Wiener filter only with the loss function $\mathcal{L}_{\text{MSE}}$ (8). With the results of M forward passes, we can approximate the mean and variance of the distribution $p(S|X)$ by the empirical mean and variance of the prediction set [38], [42]:

$$
\widetilde{S} = \frac{1}{M} \sum_{m=1}^{M} \widetilde{S}_{\theta_m}, \quad \widetilde{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} |\widetilde{S}_{\theta_m} - \widetilde{S}|^2,
\tag{14}
$$

where $\widetilde{S}_{\theta_m}$ denotes clean speech estimated using the neural network with parameters θ_m. $\widetilde{S}$ represents the average clean speech estimate and $\widetilde{\Sigma}$ quantifies the epistemic uncertainty.

B. Overall Predictive Uncertainty

In the case of optimizing the network using (12), besides the Wiener estimate $\widetilde{S}_{\theta_m}$, each neural network with weights θ_m can produce the associated variance $\widetilde{\lambda}_{\theta_m}$. The overall predictive uncertainty, which reflects both aleatoric and epistemic uncertainties, can be computed using the law of total variance [33], [38]:

$$
\widetilde{S} = \frac{1}{M} \sum_{m=1}^{M} \widetilde{S}_{\theta_m}, \quad \widehat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} \left(|\widetilde{S}_{\theta_m} - \widetilde{S}|^2 + \widetilde{\lambda}_{\theta_m} \right),
\tag{15}
$$

where $\widetilde{S}$ denotes the average Wiener estimate, and $\widehat{\Sigma}$ quantifies the overall predictive uncertainty.

For each neural network with weights θ_m, we can further generate the AMAP clean speech estimate $\widehat{S}_{\theta_m}$ by explicitly incorporating the associated uncertainty $\widetilde{\lambda}_{\theta_m}$ as in (10). Therefore, given an ensemble of networks, besides the average Wiener estimate $\widetilde{S}$, the average AMAP estimate can be obtained by:

$$
\widehat{S} = \frac{1}{M} \sum_{m=1}^{M} \widehat{S}_{\theta_m}.
\tag{16}
$$

V. EXPERIMENTAL SETTING

A. Datasets

For training and validation, we use a subset of the Deep Noise Suppression (DNS) Challenge's training set [68], which contains synthetic audio samples of 100 hours with signal-to-noise ratios (SNRs) uniformly distributed between -5 dB and 20 dB. The dataset is randomly split into 80 and 20 hours for training and validation respectively. The model is evaluated on two different unseen datasets. The first is the reverb-free synthetic test set released by DNS Challenge. This evaluation dataset is disjoint from the training and validation datasets and is created by adding noise signals sampled from 12 categories [68] to speech signals from [69] at SNRs distributed between 0 dB and 25 dB [68]. The second unseen evaluation dataset is created using clean speech from the evaluation subset of WSJ0 (`si_et_05`) [70] and four types of noise from CHiME3 (`cafe`, `street`, `pedestrian`, and `bus`) [71]. The SNRs are randomly selected from $\{-10\,\text{dB}, -5\,\text{dB}, 0\,\text{dB}, 5\,\text{dB}, 10\,\text{dB}\}$.

B. Architecture and Hyperparameters

To ensure a fair comparison, all experiments are performed based on the same U-Net neural network architecture [72], [73]. The U-Net structure with skip connections between the encoder and the decoder is comprised of several blocks, each of which consists of: 2D convolution layer + instance normalization [74] + Leaky ReLU with slope 0.2. The encoder contains 6 blocks that increase the feature channel from 1 to 512 progressively $(1 - 16 - 32 - 64 - 128 - 256 - 512)$, while the decoder reduces it back to 16 $(512 - 256 - 128 - 64 - 32 - 16 - 16)$, followed by a 1×1 convolution layer that outputs a mask of the same shape as the input. For all blocks, the kernel size is set to $(5, 5)$ with stride $(1, 2)$ and padding $(2, 2)$, processing a 2-D input with a dimension of (T, F). For the model estimating

aleatoric uncertainty, the output layer is split into two heads that predict both the Wiener filter and associated uncertainty.[1] We applied the sigmoid activation function to the estimated Wiener filter, while using the *log-exp* technique to constrain the uncertainty output to be greater than 0, i.e., the network outputs the logarithm of the variance, which is then recovered by the exponential term in the loss function. The batch size is 64; the learning rate is 0.001; the weight decay parameter is set to 0.0005. All neural networks are trained with the Adam optimizer [75]. The training process is stopped if the validation loss fails to decrease for 10 consecutive epochs and the learning rate is halved when the validation loss does not decrease for 3 epochs.

The noisy-clean speech pairs have a sampling rate of 16 kHz, and the STFT is computed using a 32 ms Hann window with 50% overlap.

C. Methods

The algorithms considered in this work include:
1) *Baseline WF:* The U-Net architecture was trained on noisy-clean speech pairs using loss function (8). This serves as a baseline, assuming a constant variance for all time-frequency bins and estimating the Wiener filter for each input only.
2) *Baseline SI-SDR:* Following the same constant variance assumption as *Baseline WF*, the U-Net network was trained to output a multiplicative filter and optimized using the time-domain loss function (11). This serves as another baseline that fails to account for uncertainty.
3) *Aleatoric-WF & Aleatoric-AMAP:* The hybrid loss function (12) allows us to generate two possible clean estimates for each input, i.e., by using the estimated Wiener filter (6) or by applying the AMAP estimator (10) that incorporates both the Wiener filter and its associated uncertainty. They are denoted as *Aleatoric-WF* and *Aleatoric-AMAP* respectively. We observe experimentally that the performance of Aleatoric-AMAP only fluctuates slightly with different β values, while the performance of Aleatoric-WF decreases when the value of β is large. The weighting factor β was empirically chosen to be 0.001 to achieve a good trade-off between the performance of Aletoric-WF and Aleaotirc-AMAP.
4) *MC dropout:* Inserting dropout after each convolution layer regularizes too strongly and impacts the model performance [56], which was confirmed in our preliminary experiments. We thus studied several variants of the U-Net by inserting the dropout layer at different positions of the architecture, and selected the variant with three dropout layers (drop probability of 0.5 [37], [50], [56]) inserted after the three deepest blocks of the encoder. The same cost function as *Baseline WF* is used. This method captures epistemic uncertainty by activating the dropout layers at inference.

5) *Deep ensembles:* The same setup as *Baseline WF* was trained M times with random initialization. This allows the model to capture epistemic uncertainty.
6) *DE-Aleatoric-WF & DE-Aleatoric-AMAP:* The same setup as *Aleatoric-WF/AMAP* was trained M times with random intialiation. This allows capturing aleatoric and epistemic uncertainties simultaneously. We average over the estimates according to (15) and (16) to obtain two clean speech estimates: *DE-Aleatoric-WF* and *DE-Aleatoric-AMAP* respectively.

VI. ANALYSIS OF UNCERTAINTY ESTIMATION

In this section, we introduce the evaluation metrics for uncertainty and then analyze the captured aleatoric and epistemic uncertainties. Finally, we show that combining two types of uncertainty yields more reliable predictive uncertainty.

A. Uncertainty Evaluation Metrics

To evaluate the captured uncertainty, the sparsification plot and the sparsification error are used as evaluation metrics [37], [38], [76]. The sparsification plot illustrates the correlation between the uncertainty measure and the true error. The error of a time-frequency bin is defined as the absolute square between the estimated spectral coefficient and the ground-truth. For this plot, the errors in the time-frequency domain are first sorted according to their corresponding uncertainty measures. The residual error should gradually decrease when the time-frequency bins with large uncertainties are removed. This leads to a plot of the root mean squared error (RMSE) versus the fraction of removed time-frequency bins. Normalization is applied to ensure that the plot is initialized at 1. The best ordering of uncertainty measures is determined by ranking the true errors [38], [76]. This provides a lower bound of each sparsification plot, denoted as the *oracle* curve, i.e., when the uncertainty estimates and errors are perfectly correlated, the sparsification plot and the oracle curve coincide. The sparsficiation error is computed as the difference between the sparsification plot and the corresponding oracle curve, and the area under the sparsification error (AUSE) curve provides a single value that enables comparison of different uncertainty modeling techniques. A lower AUSE value (i.e., the closer the sparsification plot is to its oracle curve) indicates a more accurate estimate of uncertainty.

B. Analysis of Aleatoric Uncertainty Estimation

In this part, we analyze the captured data-dependent aleatoric uncertainty associated with the Wiener estimate. For this, an audio example from the DNS challenge test set is selected to illustrate the effectiveness of the proposed optimization metric in modeling uncertainty. Aleatoric-WF in Fig. 3(c) shows the spectrogram of the clean speech obtained by applying the estimated Wiener filter. By computing the absolute square between the clean reference and estimated spectral coefficients, we can obtain the estimation error as depicted in Fig. 3(d). It can be observed that large errors occur when the speech is heavily disturbed by noise, as in the region marked by the green box,

---

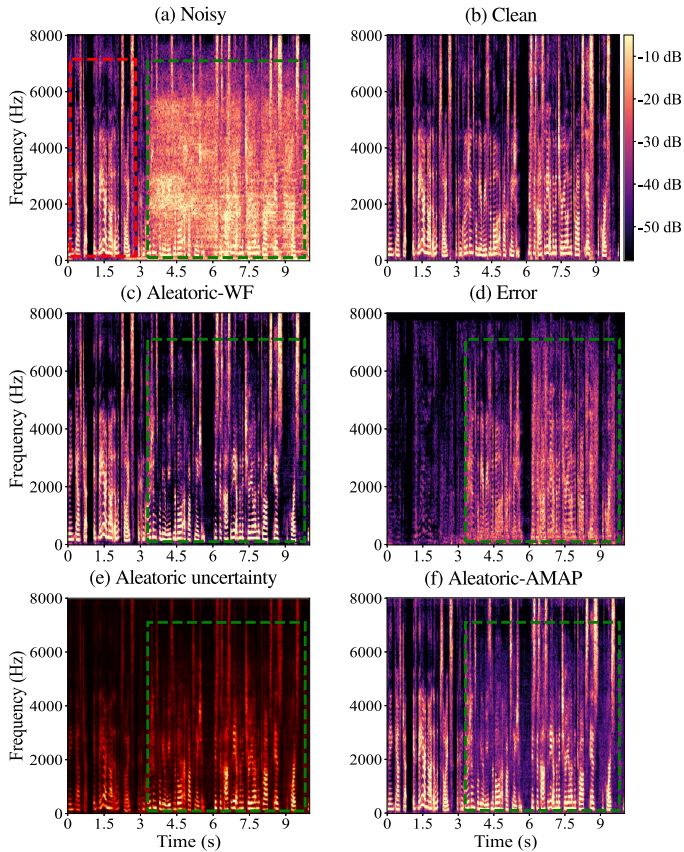[1]The model is available at: https://github.com/sp-uhh/uncertainty-SE.

Fig. 3.   Aleatoric uncertainty (shown in (e)) captured by the proposed loss function (12) for an excerpt from the DNS test dataset. The uncertainty is visualized as a heatmap. The black color indicates low uncertainty, whereas the brighter color indicates higher uncertainty. (a) Noisy (b) Clean (c) Aleatoric-WF (d) Error (e) Aleatoric uncertainty (f) Aleatoric-AMAP.

while for inputs with less distortion, such as the first three seconds, the model produces smaller errors. Meanwhile, the proposed loss function enables the estimation of uncertainty associated with the Wiener filter, as shown in Fig. 3(e), denoted as aleatoric uncertainty. It shows that aleatoric uncertainty prevails in speech presence regions. By relating Fig. 3(d) to (e), the model outputs relatively large uncertainty (e.g., the green box-marked part) when large errors are produced. This suggests that the neural network is able to produce reasonable uncertainty estimates when dealing with complex unseen inputs. Furthermore, we can incorporate the estimated uncertainty into clean speech inference, as in (10), which leads to a clean speech estimate shown in Fig. 3(f), denoted as Aleatoric-AMAP. It is observed that more speech is preserved than Aleatoric-WF in the highly-uncertain green box-marked region at some cost of noise reduction, i.e., Aleatoric-AMAP leads to less speech distortion with a slight tendency of retaining more noise. The reason for this is that with reliable uncertainty estimates, Aleatoric-AMAP can increase the estimator's value in (10) under high uncertainty (as the AMAP estimator's value is positively correlated with the uncertainty estimate when other terms are fixed), thus causing less target attenuation.

Besides the qualitative analysis, we can associate the captured uncertainty with the corresponding prediction errors on the time-frequency bin scale and use sparsification plots to analyze
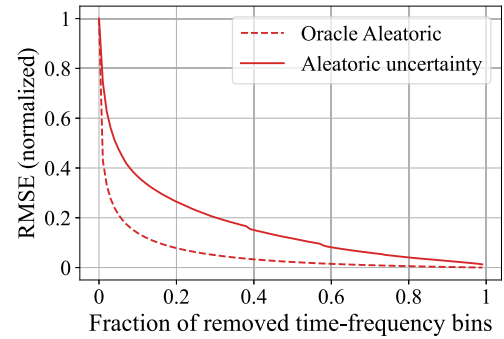

Fig. 4.   Sparsification plot of aleatoric uncertainty $\tilde{\lambda}$ evaluated on the DNS test dataset. The dashed line denotes the lower bound of the sparsification plot of aleatoric uncertainty. A smaller distance of the sparsification plot to the oracle curve indicates a more accurate uncertainty estimation.

the reliability of the uncertainty estimates. The sparsification plot shown in Fig. 4 is computed based on all audio samples in the DNS reverb-free test dataset. We observe a rapid decrease at the beginning in Fig. 4, implying that large errors come with large uncertainty estimates. By removing 20 percent of time-frequency bins with high uncertainty (i.e., 0.2 in the horizontal axis), the RMSE value drops by around two-thirds. Thus, the monotonically decreasing sparsification plot in Fig. 4 again suggests that the predicted aleatoric uncertainty measurement is closely related to the estimation error.

C. Analysis of Epistemic Uncertainty Estimation

Next, we ignore aleatoric uncertainty and analyze separately epistemic uncertainty in the model parameters. For this, the neural networks are trained to perform only point estimation, i.e., trained with the loss function (8). An ensemble of models is collected by applying Deep ensembles or MC dropout to approximate the predictive mean and variance.

In Fig. 5, we present the same audio example as in Fig. 3 to illustrate the uncertainty measures based on MC dropout and Deep ensembles. MC dropout and Deep ensembles provide the clean speech estimates as shown in the first row of Fig. 5. The estimation error for each method is obtained similarly by calculating the absolute square between the estimated and clean spectral coefficients, shown in the second row. As can be observed, both methods produce large errors as well as associated large uncertainties when the signal is heavily corrupted by noise, i.e., the green box-marked region. While the noise corruption is less severe, i.e., the region marked with a red box, the model generates low prediction errors and also a relatively low level of uncertainty. From the visual analysis, the uncertainty generated by Deep ensembles is more correlated with the error, while MC dropout appears to underestimate the uncertainty of incorrect predictions. To objectively assess the reliability of uncertainty measures, we also utilize the sparsification plots and the sparsification errors, as illustrated in Figs. 6 and 7 respectively.

In Fig. 6, we show the sparsification plots of Deep ensembles and MC dropout for a different number of forward passes $M \in \{2, 4, 8, 16, 32\}$. It can be observed that both MC dropout and Deep ensembles yield decreasing sparsification plots, suggesting that they produce accurate uncertainties that correlate well with
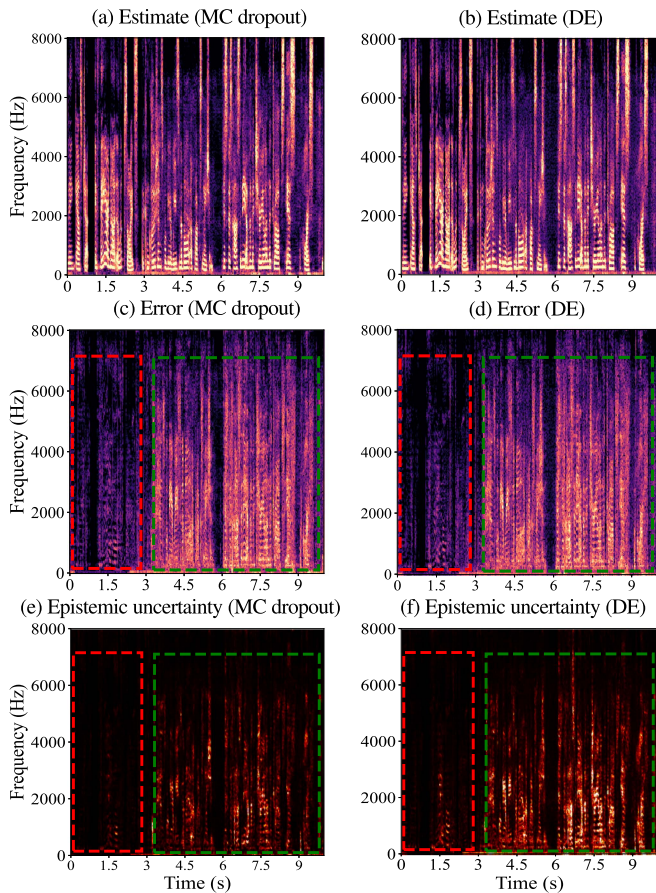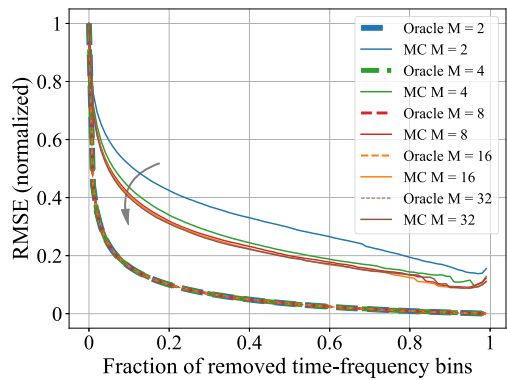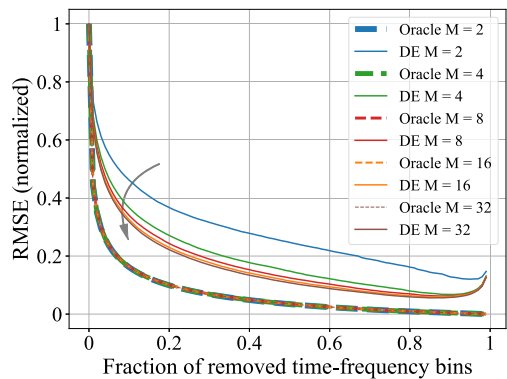
Fig. 5. The same excerpt as in Fig. 3 illustrates the captured epistemic uncertainty obtained by applying Bayesian deep learning methods ($M = 16$). *Estimate (MC dropout)* and *Estimate (DE)* represent clean speech estimated using MC dropout and Deep ensembles. (a) Estimate (MC dropout) (b) Estimate (DE) (c) Error (MC dropout) (d) Error (DE) (e) Epistemic uncertainty (MC dropout) (f) Epistemic uncertainty (DE).

Fig. 6. Sparsification plots of epistemic uncertainty $\widetilde{\Sigma}$ for the DNS test dataset. The dashed line denotes the lower bound of the corresponding sparsification plot, denoted as Oracle M. A smaller distance of the sparsification plot to the oracle curve indicates a more accurate uncertainty estimation. Note that all oracle curves are visually *overlapping*.
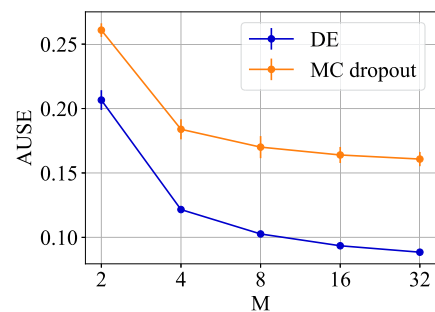

Fig. 7. AUSE for the DNS test dataset. AUSE is plotted relative to a different number of forward passes M. The markers denote the mean and the vertical bars indicate the standard deviation. Lower values indicate a smaller deviation from the oracle curve, and thus more reliable uncertainty estimation.

the estimation errors. It also shows that a large M leads to a sparsification plot closer to its corresponding oracle curve, i.e., improves the performance of the uncertainty estimation, and this improvement becomes saturated when M is sufficiently large, e.g., from $M = 16$ to $M = 32$.

To comprehensively compare MC dropout and Deep ensembles in terms of uncertainty modeling, AUSE is plotted as a function of different numbers of forward passes M. Multiple models for each M are used to provide mean and standard deviation to account for variations resulting from random factors in training. 16 MC dropout models are trained and used to compute the mean of AUSE and its standard deviation for each possible M. For Deep ensembles, 16 disjoint sets of M models are randomly selected from the 33 trained models to compute the mean and standard deviation of AUSE. The AUSE plot in Fig. 7 provides an alternative and more informative evaluation than a single sparsification plot. It indicates that Deep ensembles generally produce more accurate uncertainty than MC dropout, which may fail to produce reliable uncertainties for some erroneous predictions. This coincides with our visual observation in the green box-marked region in Fig. 5.

D. Prediction Uncertainty Combining Aleatoric and Epistemic Uncertainties

In this part, we investigate the overall prediction uncertainty obtained by combining aleatoric uncertainty and epistemic uncertainty as in (15). To obtain the overall prediction uncertainty, we use an ensemble of models trained with the optimization metric (12) such that both aleatoric and epistemic uncertainty are captured. It has been shown in Section VI-C that Deep ensembles
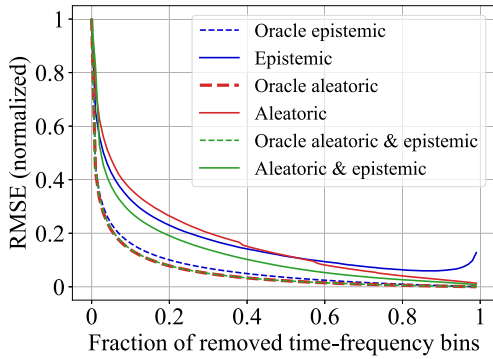
Fig. 8. Sparsification plots of aleatoric $\tilde{\lambda}$, epistemic $\widetilde{\Sigma}$, and overall predictive uncertainty $\widehat{\Sigma}$ (i.e., aleatoric & epistemic) on the DNS test dataset. Note that *Oracle aleatoric* and *Oracle aleatoric & epistemic* overlap.

TABLE I
AUSE VALUES OF *ALEATORIC*, *EPISTEMIC*, AND *ALEATORIC & EPISTEMIC* IN FIG. 8

	Aleatoric	Epistemic	Aleatoric & epistemic
AUSE	0.110	0.094	0.067

yield more accurate epistemic uncertainty than MC dropout and, therefore, are selected for the estimation of the overall predictive uncertainty. Although a larger number of models M could potentially improve the mean and variance estimation, we restrict M to 16 as further improvements become subtle while the computation time increases considerably.

In Fig. 8, we use sparsification plots to analyze the quality of prediction uncertainty estimates combining aleatoric and epistemic uncertainties. The corresponding AUSE values are provided in Table I. The plot illustrates that the overall predictive uncertainty estimates correlate stronger with the estimation error than either of the two uncertainties alone. This suggests that two sources of uncertainty may complement each other and combining both leads to more reliable uncertainty estimates. For example, Deep ensembles do not seem to capture sufficient uncertainty for less distorted input (e.g., first three seconds) as shown in Fig. 5, while aleatoric uncertainty shown in Fig. 3 could be able to compensate for this shortcoming.

VII. INFLUENCE OF MODELING UNCERTAINTY FOR SPEECH ENHANCEMENT PERFORMANCE

In this section, we show how modeling different sources of uncertainty affects the performance of speech enhancement. To evaluate the speech enhancement performance, we employ perceptual evaluation of speech quality (PESQ) [77] to measure speech quality, extended short-time objective intelligibility (ESTOI) [78] to measure speech intelligibility, and SI-SDR to account for both noise reduction and speech distortion.

To show the impact of modeling aleatoric uncertainty on speech enhancement performance, we compare the performance of the model trained with the proposed loss function (12) with that of Baseline WF and Baseline SI-SDR. The proposed method enables speech estimation via either the Wiener filter, which

TABLE II
EVALUATION RESULTS ON THE DNS TEST DATASET. ALL RESULTS ARE STATED AS MEAN ± 95%-CONFIDENCE INTERVAL. *UNC.* STANDS FOR *UNCERTAINTY*

	Unc.	PESQ	ESTOI	SI-SDR
Noisy (DNS)	-	1.58 ± 0.07	0.81 ± 0.02	9.07 ± 0.89
Baseline WF	✗	2.48 ± 0.10	0.90 ± 0.01	16.84 ± 0.74
Baseline SI-SDR	✗	2.63 ± 0.10	0.91 ± 0.01	17.49 ± 0.78
MC dropout	✓	2.53 ± 0.10	0.90 ± 0.01	16.88 ± 0.74
Deep ensembles	✓	2.66 ± 0.10	0.91 ± 0.01	17.16 ± 0.73
Aleatoric-WF	✓	2.62 ± 0.11	0.91 ± 0.01	17.54 ± 0.78
Aleatoric-MAP	✓	2.69 ± 0.10	0.91 ± 0.01	17.54 ± 0.78
DE-Aleatoric-WF	✓	2.77 ± 0.11	0.92 ± 0.01	17.88 ± 0.78
DE-Aleatoric-AMAP	✓	2.83 ± 0.10	0.92 ± 0.01	17.90 ± 0.78

implicitly takes uncertainty into account during the training process, or the approximated MAP filter, which explicitly includes uncertainty to estimate speech, denoted as Aleatoric-WF and Aleatoric-AMAP respectively. Table II shows the average evaluation results on the DNS synthetic non-reverb test set. Aleatoric-WF shows improvements in PESQ, ESTOI, and SI-SDR compared to the Baseline WF, indicating the benefit of weighting Wiener estimates with uncertainty during training. Further PESQ improvements over both Baseline WF and Baseline SI-SDR can be observed when explicitly incorporating uncertainty into clean speech estimation, that is, Aleatoric-AMAP. This demonstrates the advantage of modeling uncertainty associated with the Wiener estimate rather than directly estimating optimal points. When evaluated on another dataset with speech from WSJ and noise from CHiME3, the performance gap between Aleatoric-AMAP and the baselines in terms of PESQ is further increased, as shown in Fig. 9, indicating that the model that takes uncertainty into account has improved generalization capacities for speech enhancement. This can be attributed to the nonlinear estimation characteristics of the uncertainty-based AMAP estimator with respect to noisy inputs and the resulting better speech preservation properties. We observe larger improvements over the baselines at high SNRs, which might be explained by the fact that, at high SNRs, speech quality (and thus PESQ) is mainly affected by speech distortions, while at low SNRs the main factor is residual noise. Overall, these evaluation results demonstrate the notable benefits of modeling aleatoric uncertainty in the algorithm.

To show the impact of modeling epistemic uncertainty on speech enhancement performance, we compare the performance of Deep ensembles and MC dropout with Baseline WF. We again restrict M to 16 as in Section VI-D. MC dropout performs comparably to Baseline WF on the DNS test set, while a larger improvement can be observed when using Deep ensembles. This improvement is even more pronounced in PESQ. Similarly, the results on the second test set are shown in Fig. 9, where Deep ensembles and MC dropout improve over Baseline WF in terms of PESQ for all considered SNRs and provide higher ESTOI scores, especially at low SNRs. We observe that Deep ensembles not only provide more accurate uncertainty estimates than MC dropout but also lead to a better speech enhancement performance. A possible explanation is that while MC dropout only captures local uncertainty around a single mode, Deep ensembles trained with different initialization points are capable
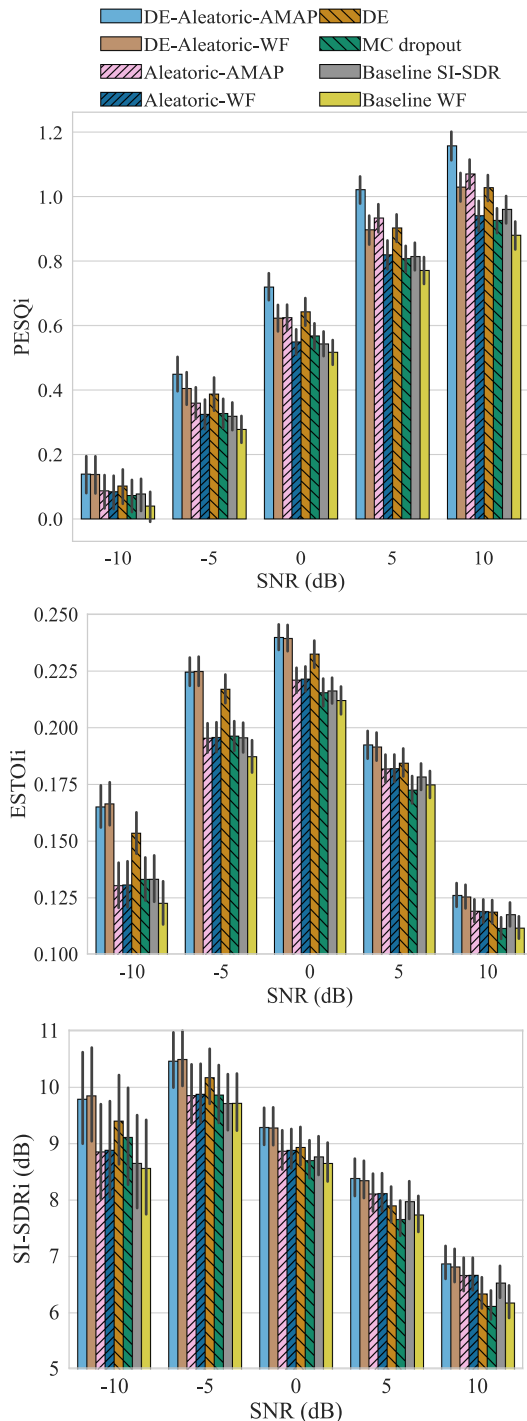
Fig. 9. Performance improvement on the dataset with speech from WSJ0 and noise from CHiME3. PESQi denotes PESQ improvement with respect to noisy mixtures. ESTOIi and SI-SDRi are defined similarly. Markers and vertical bars indicate the mean and 95% confidence interval.

of exploring multiple modes in the function space to account for training data, see, e.g., [48], [49]. This may allow the neural network to generalize better to complex acoustic scenarios.

To show the impact of modeling predictive uncertainty that combines both aleatoric and epistemic uncertainties on speech enhancement performance, we use the same set of models as described in Section VI-D. We take the average of estimates

as in (15) and (16) and obtain two speech estimates, called DE-Aleatoric-WF and DE-Aleatoric-AMAP respectively. They both provide better ESTOI and SI-SDR scores than the baselines, the epistemic uncertainty-only model, and the aleatoric uncertainty-only model, especially at low SNRs. Moreover, DE-Aleatoric-AMAP yields higher scores in PESQ likely due to the uncertainty-dependent regularization and exploration of multiple modes in the function space. This indicates that combining the model that accounts for aleatoric uncertainty with the ensemble-based method can take advantage of the benefits of both approaches and further improve the performance. Overall, the evaluation results across different datasets show that quantifying uncertainty in neural network-based speech enhancement leads to a considerable improvement in enhancement performance over the baseline models.

VIII. CONCLUSION

In this paper, besides estimating clean speech, we quantified predictive uncertainty in neural network-based speech enhancement. For this, aleatoric uncertainty, which describes inherent uncertainty in data, and epistemic uncertainty, which accounts for uncertainty of the model, were captured and analyzed in a joint framework. We investigated the reliability of uncertainty estimates from different sources, and how it affects the enhancement performance. Our proposed hybrid loss function based on MAP inference of complex spectral coefficients and an AMAP estimator of spectral magnitudes has demonstrated the effectiveness in modeling aleatoric uncertainty. In addition, the proposed scheme provided a principled way to create a noise-removing mask that explicitly incorporates uncertainty to further improve speech enhancement performance. The evaluation results on different datasets have shown increased generalization capacities when modeling aleatoric uncertainty.

To empirically approximate the predictive distribution and capture epistemic uncertainty, we employed two Bayesian deep learning methods, MC dropout and Deep ensembles. We showed that Deep ensembles not only provide more accurate estimates of epistemic uncertainty than MC dropout, but also lead to more prominent improvements in speech enhancement. A reason may be that Deep ensembles can potentially converge to different local minima in the loss landscape due to random initialization. Furthermore, we combined the proposed hybrid function with Deep ensembles to quantify overall prediction uncertainty, which reflects both data uncertainty and model uncertainty. An analysis using sparsification plots showed that combining different types of uncertainties further improves the reliability of predictive uncertainty estimation, indicating the complementary nature of the two sources of uncertainty. Finally, our experiments indicated that the performance of clean speech estimation can be considerably improved over the baselines while additionally obtaining predictive uncertainty estimates.

In summary, this work investigated capturing predictive uncertainty in neural network-based speech enhancement and showed the noticeable benefits of modeling uncertainty for clean speech estimation. Uncertainty can indicate the algorithm's confidence in the output in the absence of ground truth, which

is essential for assessing the reliability of speech estimates. With this work, we hope to enlighten discussions on modeling uncertainty in the speech enhancement task, while facilitating future research on how to take advantage of uncertainty.

REFERENCES

[1] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech*, E. Enhancement, T. Vincent Virtanen, and S. Gannot, Eds. Hoboken, NJ, USA: Wiley, 2018, pp. 65–85.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 10, pp. 1–9, 2003.

[5] I. Andrianakis and P. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, 2006, p. III.

[6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4037–4040.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[8] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 191–195.

[9] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[10] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[11] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deep-MMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[12] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 716–720.

[13] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. Eur. Signal Conf.*, 2021, pp. 436–440.

[14] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2018, pp. 1–6.

[15] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 681–685.

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

[17] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, 2021, pp. 126–130.

[18] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[19] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. IEEE Int. Conf. Telecommun. Signal Process.*, 2021, pp. 72–76.

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

[21] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 371–375.

[22] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2993–3007, 2022.

[23] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1788–1800, 2020.

[24] S.-W. Fu et al., "MetricGAN : An improved version of metricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[25] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7412–7416.

[26] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7402–7406.

[27] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.

[28] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1937–1949, 2021.

[29] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Proc. Interspeech*, 2016, pp. 3738–3742.

[30] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[31] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021.

[32] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?," *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, 2009.

[33] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2017, vol. 30, pp. 5580–5590.

[34] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1184–1193.

[35] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4075–4084.

[36] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6405–6416.

[37] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable Bayesian deep learning methods for robust computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 318–319.

[38] E. Ilg et al., "Uncertainty estimates and multi-hypotheses networks for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 652–667.

[39] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1919–1931, Dec. 2019.

[40] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi, and T. Nakatani, "Deep mixture density network for statistical model-based feature enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 251–255.

[41] S. M. Siniscalchi, "Vector-to-vector regression via distributional loss for speech enhancement," *IEEE Signal Process. Lett.*, vol. 28, pp. 254–258, 2021.

[42] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[43] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 681–688.

[44] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1683–1691.

[45] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[46] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," in *Proc. Int. Conf. Learn. Representations Workshop Track*, 2016.

[47] J. Nixon, B. Lakshminarayanan, and D. Tran, "Why are bootstrapped deep ensembles not better?," in *Proc. "I Can't Believe It's Not Better!" Neural Inf. Process. Syst. Workshop*, 2020.

[48] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," in *Proc. "Bayesian Deep Learn." Neural Inf. Proc. Syst. Workshop*, 2019.

[49] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 4697–4708.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[51] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 8384–8388.

[52] N. R. Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and G. Timo, "End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks," in *Proc. Interspeech*, 2022, pp. 151–155.

[53] S. Braun and S.-C. Liu, "Parameter uncertainty for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5636–5640.

[54] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6553–6557.

[55] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing uncertainties in speech recognition using dropout," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6730–6734.

[56] K. Alex, B. Vijay, and C. Roberto, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 57.1–57.12.

[57] J. Le Roux, S. Watanabe, and J. R. Hershey, "Ensemble learning for speech enhancement," in *Proc. IEEE Workshop Appl. Signal Proc. Audio, Acoust.*, 2013, pp. 1–4.

[58] H. Zhang, X. Zhang, and G. Gao, "Multi-target ensemble learning for monaural speech separation," in *Proc. Interspeech*, 2017, pp. 1958–1962.

[59] H. Fang, T. Peer, S. Wermter, and T. Gerkmann, "Integrating statistical uncertainty into neural network-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 386–390.

[60] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[61] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *Proc. Int. Conf. Learn. Representations*, 2022.

[62] A. Stirn and D. A. Knowles, "Variational variance: Simple, reliable, calibrated heteroscedastic noise variance parameterization," 2020, *arXiv:2006.04910*.

[63] N. Skafte, M. Jørgensen, and S. Hauberg, "Reliable training and estimation of variance networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6326–6336.

[64] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[65] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[66] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6359–6363.

[67] R. M. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, Univ. of Toronto, Toronto, ON, Canada, 1995.

[68] C. K. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[69] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A Pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. Interspeech*, 2011, pp. 1509–1512.

[70] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[71] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[72] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 23–27.

[73] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[74] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6924–6932.

[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014.

[76] A. S. Wannenwetsch, M. Keuper, and S. Roth, "Probflow: Joint optical flow and uncertainty estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 1173–1182.

[77] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[78] J. Jensen and C.H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

Huajian Fang (Student Member, IEEE) is currently working toward the Ph.D. degree. Since 2019, he has been with the Signal Processing Group and the Knowledge Technology Group, University of Hamburg, Hamburg, Germany. His research interests include speech signal processing and machine learning.
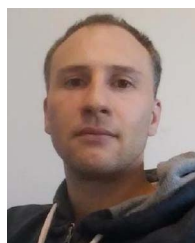
Dennis Becker received the Ph.D. degree from the University of Lüneburg, Lower Saxony, Germany, is currently a Postdoctoral Researcher with the University of Hamburg, Hamburg, Germany. At the University of Lüneburg, he researched the use of mobile phone data for prediction of onset and recurrence of symptoms related to mental disorders. His main research interests include neural networks, data analysis, and human-robot interaction.

Stefan Wermter (Member, IEEE) is currently a Full Professor with the University of Hamburg, Hamburg, Germany, leading the Knowledge Technology Research Group with the Department of Informatics. He held positions with the University of Dortmund, Dortmund, Germany, University of Massachusetts, Amherst, MA, USA, the International Computer Science Institute in Berkeley and the University of Sunderland, Sunderland, U.K. His main research interests include neural networks, hybrid knowledge technology, neuroscience-inspired computing, cognitive robotics, natural language processing and human-robot interaction. He has been an Associate Editor of the journal IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and he is an Associate Editor of *Connection Science* and *International Journal for Hybrid Intelligent Systems* and he is on the Editorial Board of the journals *Cognitive Computation* and *International Journal of Computational Intelligence*. He is the Coordinator of the international doctoral network TRAIL, co-Coordinator of the international collaborative research centre on Crossmodal Learning (TRR-169) and he is the President of the European Neural Network Society 2023-2025.

Timo Gerkmann (Senior Member, IEEE) is currently a professor with the University of Hamburg, Hamburg, Germany, where he is the head of the Signal Processing Research Group, and a deputy head of the Department of Informatics. He has previously held positions with Technicolor Research & Innovation in Hanover, Germany, the Carl-von-Ossietzky Universität Oldenburg, Oldenburg, Germany, KTH Royal Institute of Technology, Stockholm, Sweden, Ruhr-Universität Bochum, Bochum, Germany, and Siemens Corporate Research in Princeton, NJ, USA. His main research interests include statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann is currently a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and a Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.