MITIGATING CLOSED-MODEL ADVERSARIAL EXAMPLES WITH BAYESIAN NEURAL MODELING FOR ENHANCED END-TO-END SPEECH RECOGNITION

Chao-Han Huck Yang[1,2*] *Zeeshan Ahmed*[1] *Yile Gu*[1] *Joseph Szurley*[1]
Roger Ren[1] *Linda Liu*[1] *Andreas Stolcke*[1] *Ivan Bulyko*[1]

[1]Amazon Alexa AI, USA
[2]Georgia Institute of Technology, USA

ABSTRACT

In this work, we aim to enhance the system robustness of end-to-end automatic speech recognition (ASR) against adversarially-noisy speech examples. We focus on a rigorous and empirical "closed-model adversarial robustness" setting (e.g., on-device or cloud applications). The adversarial noise is only generated by closed-model optimization (e.g., evolutionary and zeroth-order estimation) without accessing gradient information of a targeted ASR model directly. We propose a new Bayesian neural network (BNN) based adversarial detector, which could model latent distributions against adaptive adversarial perturbation with divergence measurement. We further simulate deployment scenarios of RNN Transducer, Conformer, and wav2vec-2.0 based ASR systems with proposed adversarial detection system. Leveraging the proposed BNN based detection system, we improve detection rate by +2.77 to +5.42% (relative +3.03 to +6.26%) and reduce the word error rate by 5.02 to 7.47 % on LibriSpeech datasets compared to the current model enhancement methods against the adversarial speech examples.

Index Terms— Adversarial Robustness, Robust Speech Recognition, Speech Recognition Safety, and Sequence Modeling

1. INTRODUCTION

End-to-end automatic speech recognition [1, 2] (ASR) has many applications in human society, empowering voice-based intelligent control, spoken language understanding [3], on-device services [4], and web-based speech interactions [5]. These high-performance speech applications benefit from neural network-based ASR systems that have highly accurate on-device performance with fixed model parameters. Even when the model parameters of a simulated ASR system are secured by reliable data protection [6], encryption [7], and security measures, recent concerns [8, 9] about query-free robustness evaluation highlight the need for designing an ASR system to be robust against noisy samples generated by adversarial optimization. As shown in Fig. 1(a), query-free optimization is applied on random signals to synthesize high-confidence false examples to maliciously manipulate ASR output (e.g., "open the door"). The aforementioned noise evaluation on a simulated ASR is called closed-model adversarial robustness and could be categorized in a regime of ASR robustness against environmental noise (e.g., acoustic conditions) as a crucial challenge for ASR designs. Table 1 provides an overview of the access constraints for closed-model robustness, which is our focus in this work.

To improve robustness against adversarial examples, adversarial training-based approaches [8, 10] have been widely studied, mainly by applying augmented noisy examples with correct labels to retrain a targeted neural network model to improve its generalization. However, incorporating adversarial training [11] into a large-scale neural network is costly and unstable due to its sensitivity to high-dimensional decision boundaries. Furthermore, adversarial training is often considered to not be economical and realistic for online ASR systems, where the model parameters have to be updated frequently, and the augmented noises need to be re-generated each time.

Recently, adversarial detection mechanisms [12, 13] have emerged as new alternatives to improve system robustness against adversarial examples. Adversarial detection provides low training-cost solutions that are easily incorporated into an existing end-to-end system with an option to control the safety risk by filtering out suspicious inputs. In this work, we take advantage of randomness properties from Bayesian neural networks to design a high-performance adversarial detector empowered by statistical estimation on test samples. As shown in Fig. 1(b), the proposed method is based on an additional variational layer (Flipout [14]) for distribution estimation that could be easily integrated into end-to-end ASR architectures. Next, we will review existing methods for mitigating adversarial speech examples and demonstrate the novelty of our system design.

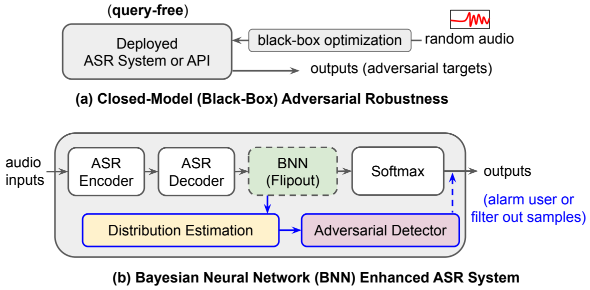
(query-free)
Deployed ASR System or API
black-box optimization
random audio
outputs (adversarial targets)
(a) Closed-Model (Black-Box) Adversarial Robustness
audio inputs
ASR Encoder
ASR Decoder
BNN (Flipout)
Softmax
outputs
Distribution Estimation
Adversarial Detector
(alarm user or filter out samples)
(b) Bayesian Neural Network (BNN) Enhanced ASR System

Fig. 1: Illustration of (a) closed-model adversarial robustness for automatic speech recognition (ASR) from query-based optimization (e.g., on-device or cloud services) and (b) proposed Bayesian neural network (BNN) enhanced model robustness by adversarial detection.

2. RELATED WORK

2.1. Adversarial Robustness for Speech Processing

To study adversarial robustness for ASR, additive noise δ is applied to the original waveform under an environmental noise ratio = 10dB reported in the previous evaluation [20, 8]. The noise is

[*] Work done as an applied scientist intern at Amazon Alexa.

Table 1: Type between closed-model ($\mathcal{A}_E$ and $\mathcal{A}_Z$) and open-model ($\hat{\mathcal{A}}$) adversarial robustness evaluation. In this work, we focuses on closed-model settings with access constraints during noise generation.

Noise Types	Para. Access	Gradient Info.	Output Access	Ref.
$\mathcal{A}_E$: Evolutionary	**No**	No	Yes	[15, 16]
$\mathcal{A}_Z$: Zeroth-order	**No**	No (estimated)	Yes	[17, 18]
$\hat{\mathcal{A}}$: Fast Gradient	Yes	Yes	Yes	[19]

trainable with an objective to degrade model performance or create malicious outputs (e.g., "open the door"). The adversarial robustness for speech modeling is first studied with a simple open-model setting (white-box), where the δ is simply generated by gradient information, such as the fast-gradient-sign method (FGSM) [21] and projected gradient descent (PGD) [22] attacks against the model's decision boundaries. However, open-model adversarial evaluations heavily rely on the hypothesis that the targeted model could be accessed for its gradient information during the noise generation process. This hypothesis is generally **not** true for real-world speech-based applications in recent discussions and evaluations [23, 9, 8]. For example, the end-point device and cloud users cannot extract gradient information from a simulated ASR model directly, where the model is prohibited from external visits and protected by different layers of information security frameworks.

By contrast, closed-model adversarial evaluations aim to study strict and empirical settings, where the additive environmental noise must be generated without knowing model parameters. Closed-model optimization techniques for the neural model have been used in these studies, such as evolutionary learning [16] and zeroth-order optimization [18, 17] from sample queries. Multiple-objective evolutionary optimization [15] has highlighted closed-model adversarial robustness challenges against popular ASR back-ends, including DeepSpeech [24] and Kaldi ASR [25]. By using only limited query information from input/output pairs, a high acoustic similarity (0.97 to 0.98) is preserved between closed-model generated audio and the original audio, while resulting in large relative word error rate (WER) degradation (up to 980%). Meanwhile, for improving adversarial robustness, adversarial training is often augmented with correctly labeled noisy samples for training. However, adversarial training requires a long training times and high model complexity. Moreover, adversarial training is costly for on-device simulation once the ASR model needs updating. Ideally, parallel model enhancement solutions, including adversarial detection and noise filtering frameworks, would provide tangible, low-complexity, and energy-efficient solutions toward safe and reliable ASR. Next, we will review the existing adversarial detection benchmarks to mitigate the closed-model perturbation for ASR.

2.2. Distribution Modeling against Closed-Model Perturbation

To concretely address empirical adversarial robustness, adaptive noise is jointly optimized with both targeted and enhancement modules during the adversarial evaluation. For example, temporal dependency [12] (TD) based detection methods have shown the best performance compared to other defensive methods. TD detector randomly divides the inputs sequences into multiple uniform segments and computes the output distribution of whole audio and segmented audio segments for abnormality detection on the adversarial input. More recently, the self-attention mechanism [8] has been incorporated into the TD properties and further denoises the adversarial features with multi-scale representation learning. Nevertheless, there is less discussion on how to design an efficient framework to improve the distributional modeling (e.g., abnormal) of the latent representations with the TD properties for the neural ASR model. In this work, we propose a novel design building upon RNN-T, conformer, and wav2vec-2.0 with Bayesian neural networks [14] to improve distributional modeling. Specifically, incorporating randomness into neural networks has recently been shown to improve smoothness [13, 26] of neural predictors, thus providing stronger robustness guarantees.

3. CLOSED-MODEL ADVERSARIAL-ROBUST NEURAL SPEECH RECOGNITION

3.1. Evaluating Closed-model Adversarial Robustness

We first consider an end-to-end classification model M, where the classifier is trained with model parameters θ to produce output prediction (e.g., words or phonemes), $M(\mathbf{x};\theta) = \mathbf{y}$, from an input $\mathbf{x}$ (e.g., acoustic features). The loss function $L_\theta(\mathbf{y}, \hat{\mathbf{y}})$ is optimized using gradient descent to minimize the prediction error between $\mathbf{y}$ and $\hat{\mathbf{y}}$. In closed-model adversarial robustness, we assume that there is no information about model parameter θ available to the adversarial noise generator. As shown in Eq. 1, untargeted closed-model adversarial noise can only be generated by queries $\mathbf{x}, \mathbf{y}$ to maximize the prediction loss with a model outcome of $\mathbf{y}_{adv}$:

$$\mathbf{x}_{adv} = \mathbf{x} + \delta;\ \underset{\delta}{\operatorname{argmax}}\, \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[L(\mathbf{y}_{adv}, \mathbf{y})\right]. \tag{1}$$

We have fixed the maximum [16] querying number (of input-output pairs $(\mathbf{x}, \mathbf{y})$) to 10k in our study, following established closed-model robustness benchmarks. We consider **adaptive adversarial noise setting** in this work, where the noise is jointly optimized against a simulated "ASR associated with an adversarial detector" as a rigorous and firm setting [23, 27] for empirical adversarial robustness evaluation with a $\delta_{\max}$=0.01 under l_∞-norm.

$\mathcal{A}_E$: Multi-objective **evolutionary adversarial perturbation**. The method initializes a population of examples around the given input example x picking random examples from a uniform distribution defined over the sphere of radius $\delta_{\max}$ centered on the original example. The algorithm computes an adversarial noisy input such that $\|\mathbf{x} - \mathbf{x}_{adv}\|_\infty \leq \delta_{\max}$. This is achieved by adding random noise in the range $(-\delta_{\max}, \delta_{\max})$ to each dimension of the input vector $\mathbf{x}$. We follow the multi-objective optimization method proposed in [15] to ensure acoustic similarity, computed as the cosine similarly ($\geq$ 0.95) of MFCCs between the noisy and original speech inputs, for the ASR robustness evaluation. Since the work of [15] does not provide a query number for study, we use the benchmark closed-model algorithm from [16] to improve the sample efficiency to generate adversarial examples.

$\mathcal{A}_Z$: Estimated gradient noises with **zeroth-order optimization**, where random additive noises are synthesized with estimated gradient information by zeroth-order optimization on open source ASR models. We follow an established zeroth order optimization benchmark proposed by AutoZoom [18] to generated distortion with an adaptive random gradient estimation strategy on the ASR system without accessing the target model parameters. AutoZoom leverages upon jointly optimizing latent embeddings (from an offline encoder) of a target input to generate query-efficient distortion by coordinate-wise gradient estimation based on random vector. We use wav2vec2.0 [28] as the offline encoder to extract latent embeddings from inputs speech. The unknown gradient information $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}$ of the i-th input speech sample is estimated by: $b \cdot \frac{f(\mathbf{x}+\beta \mathbf{u})-f(\mathbf{x})}{\beta} \cdot \mathbf{u} \approx \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}$, where $\beta > 0$ represents a smoothing parameter (e.g., coordinate-wise estimated gradient); $\mathbf{u}$ is a vector with unit-length. We set the scaling parameter $b = 1$ as in [29]. The gradient information was used to compute the standard querying-based closed-model loss function defined in [17]. We refer to [17, 18] for more details.

3.2. Bayesian Neural Network for Adversarial Detection

In previous works, BNN and other randomization methods have been used to improve robust classification accuracy, but they were not simulated to improve adversarial detection performance in ASR.

Blundell *et al.* [30] introduced an efficient algorithm to learn BNN parameters. BNN is subject to model the distribution of the hidden parameters w with the given random variables (x, y), instead of estimating the maximum likelihood values w_{MLE} for the weights. From a Bayesian perspective, each stochastic parameter is now a random variable sampling from a distribution instead of being a fixed (deterministic) parameter. Given the input x and label y, a BNN aims to estimate the posterior over the weights $p(\boldsymbol{w} \mid x, y)$ given the prior $p(\boldsymbol{w})$. The real posterior can be inferred by a parametric distribution $q_{\boldsymbol{\theta}}(\boldsymbol{w})$, where the unknown trainable parameter $\boldsymbol{\theta}$ is estimated by minimizing the KL divergence: $\mathcal{D}_{\text{KL}}\left(q_{\boldsymbol{\theta}}(\boldsymbol{w}) \| p(\boldsymbol{w} \mid x, y)\right)$ over $\boldsymbol{\theta}$. q_θ is a factorized Gaussian distribution with neural parameters, where $q_{\theta_i}\left(\boldsymbol{w_i}\right) = \mathcal{N}\left(\boldsymbol{w}_i; \mu, \sigma^2\right)$.

The objective function of the evidence lower bound for training BNNs is reformulated from the expression of KD divergence as shown in (2), which is a sum of a data-dependent part and a regularization part, for each training data pair $(x_i, y_i) \in \boldsymbol{D}$ and $(\boldsymbol{w}_i) \in \boldsymbol{W}$:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \underset{\boldsymbol{W} \sim q_\theta}{\mathbb{E}} [\log p(\boldsymbol{D} \mid \boldsymbol{W})] - \mathcal{D}_{\text{KL}}\left(q_{\boldsymbol{\theta}} \| p\right) \right\} \quad (2)$$

where $\boldsymbol{D}$ represents the data distribution; $\boldsymbol{W}$ represents a random weights distribution. In the first term of objective (2), the probability of y_i given x_i and weights is the output of the model. This part represents the classification loss. The second term of objective (2) is trying to minimize the divergence between the prior and the parametric distribution, as a form of regularization.

Based on the theoretical justification for BNN modeling [31, 32, 13], we consider $f(\boldsymbol{x}, \boldsymbol{w})$ as a model with $\boldsymbol{x} \sim \mathcal{C}_{\boldsymbol{x}}$ and $\boldsymbol{w} \sim \mathcal{C}_{\boldsymbol{w}}$, where $\mathcal{C}_{\boldsymbol{w}}$ is any distribution that is symmetric about $\boldsymbol{w}_0 = \mathbb{E}[\boldsymbol{w}]$, such as $\mathcal{N}\left(\boldsymbol{w}_0, \boldsymbol{I}\right)$. If $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{w})$ can be approximated by the first-order Taylor expansion at $\boldsymbol{w}_0$, we have:

$$\mathcal{W}(f(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{w}), f(\boldsymbol{x}, \boldsymbol{w})) \geq \mathcal{W}\left(f\left(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{w}_0\right), f\left(\boldsymbol{x}, \boldsymbol{w}_0\right)\right) \quad (3)$$

where $\boldsymbol{\delta}$ represents an adversarial perturbation and $\mathcal{W}$ represents a translation-invariant sliced-Wasserstein distance [33] measuring distribution dispersion with one-dimensional closed-form convergence [33]. The inequality performs that parameters associated with randomness will enlarge the distributional differences between normal and adversarial outputs. Thus, we can utilize the distributional differences of the BNN layer as shown in Fig. 1(b) to detect adversarial examples. We estimate the 1-Wasserstein distance between the distributions to detect adversarial examples. We show that BNNs can enlarge distributional differences with this distance metric. For a model $f(\boldsymbol{x}, \boldsymbol{w})$ with $\boldsymbol{x} \sim \mathcal{C}_{\boldsymbol{x}}$ and $\boldsymbol{w} \sim \mathcal{C}_{\boldsymbol{w}}$, where $\mathcal{C}_{\boldsymbol{w}}$ is any distribution that satisfies $\boldsymbol{w}$ is symmetric about $\boldsymbol{w}_0 = \mathbb{E}[\boldsymbol{w}]$, such as $\mathcal{N}\left(\boldsymbol{w}_0, \boldsymbol{I}\right)$. We use an efficient stochastic "Flipout" [14] method for BNN (as shown in Fig. 1(b)) to sample pseudo-independent posteriors for each input audio, which utilizes a Monte Carlo approximation of the distribution to model neural network parameters.

Distribution estimation with BNN: To interpret the detection behavior of predictions by BNNs, we visualize the distribution of different audio inputs, including the training set (blue), the test set (green), and the adversarial samples (red) during the untargeted noisy evaluation. As shown in Fig. 2, the BNN with the aforementioned stochastic modeling is accurate and robust in identifying the adversaries in the LibriSpeech dataset.

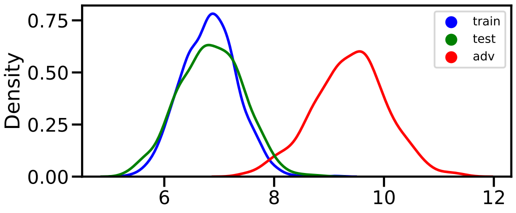

Fig. 2: Standard deviation distributions of hidden layer output of the proposed adversarial detection by Bayesian neural network.

4. EXPERIMENTS

4.1. Training Datasets and ASR Architecture

LibriSpeech dataset: The LibriSpeech [34] corpus is a collection of large-scale audiobooks. We use 960 hours of LibriSpeech for our evaluation. For the language modeling results, LibriSpeech-960 also provides the n-gram language models and the corresponding texts taken from Project Gutenberg books, comprising 803M tokens and 977K unique words. We randomly select 2,000 samples from "dev-clean" for two adversarial robustness evaluation tasks ("targeted words" or "non-targeted words" perturbation) in this study.

ASR architectures for adversarial robustness evaluation: (1) RNN Transformer: We use RNN-T ASR model with five LSTM layers as an encoder. Each LSTM layer has 1024 hidden units. The prediction network in RNN-T comes with two LSTM layers of 1024 units and an embedding layer of 512 units. The one-best sequence is produced by combining the scores of RNN-T with the language model scores. We take length-normalized probabilities from the language model (LM) network for its LM scores and determine the coefficient λ using grid search on a development data set; λ was fixed at 0.006 for all reported experiments.

(2) Conformer ASR: ConformerNet [35] has been used in streaming ASR applications. It contains two feed-forward modules sandwiching the multi-headed self-attention module and the convolution module. The first feed-forward module is built from half-step residual weights. The second part of the feed-forward module is connected to a final layer-norm module.

(3) ASR based on wav2vec-2.0 [28], a pretraining method used to provide waveform-level representations for end-to-end speech recognition. We follow the setup of "large from scratch" presented in Baevski *et al.* [28] to provide a downstream task for wav2vec-2.0 and evaluate its closed-model adversarial robustness.

4.2. Closed-model Adversarial Robustness Baseline

Task 1 ($\mathcal{T}1$) for untargeted adversarial robustness evaluation: We first evaluate all the ASR models with (1) the multi-objective evolutionary ($\mathcal{A}_E$) and (2) zeroth-order ($\mathcal{A}_Z$) estimated gradient information for untargeted noise evaluation. As shown in Eq. (1), the optimization process is based on finding a minimum noise under l_p-norm constraints (e.g., environmental noise level in our case) to let the final prediction diverge as far as possible from its original output prediction. The l_∞-norm untargeted adversarial robustness evaluation is important, owing to a recent theoretical justification [36] of its relationship to an upper bound of model robustness toward other noise (e.g., l_2 as Gaussian).

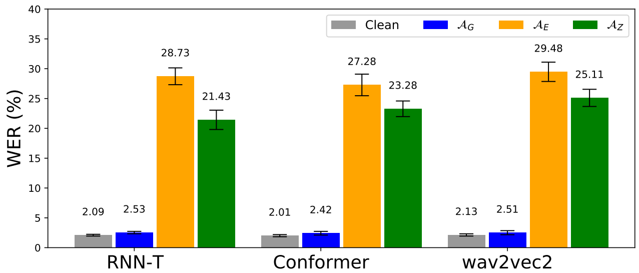

Fig. 3: Task 1 baseline under closed-model adversarial noise.

As shown in Fig. 3, RNN-T, Conformer, and wav2vec-2.0 models show relatively small variances between clean test set and additive Gaussian noise (denoted by $\mathcal{A}_G$) under an SNR of 10dB. However, all the evaluated ASR systems show major performance degradation (increased WER) by evaluating with evolutionary adversarial noise (denoted by $\mathcal{A}_E$) and zeroth-order optimized based adversarial noise (denoted by $\mathcal{A}_Z$) shared the same SNR ratio (10dB) with the additive Gaussian noise setting. While adversarial noises ($\mathcal{A}_E$ and $\mathcal{A}_Z$) causes a severe 23.11% average WER degradation, Gaussian-based perturbation only leads to a WER increase of 0.38% to 0.42%. Notably, the ConformerNet ASR system shows slightly better robustness in terms of WER compared to RNN-T and wav2vec-2.0, which degrades by 25.27% absolute under $\mathcal{A}_E$ noise and 21.27% under $\mathcal{A}_Z$ noise. The findings could be due to its scale-free feature learning during the patch-wise processes based on its enhanced convolution architecture, which echos some findings in [8].

Table 2: Task 1 with detection methods: WER (%) reduction by filtering out adversarial examples and detection performance.

ASR	RNN-T		Conformer		wav2vec-2.0	
Method	TD	BNN	TD	BNN	TD	BNN
WER	-8.41	**-13.43**	-9.64	**-15.09**	-7.45	**14.92**
AUC	82.31	**90.12**	83.42	**92.34**	80.09	**91.09**
FP	12.23	**6.23**	10.29	**5.23**	14.32	**5.78**
FN	6.43	**2.34**	7.12	**2.34**	8.23	**3.23**

Next, we integrate an additional adversarial detector into the ASR to detect and filter an **equally-mixed** of 6k noisy samples (2k for $\mathcal{A}_G$; 2k for $\mathcal{A}_E$; 2k for $\mathcal{A}_Z$). We compare the WER deduction by using a TD detector and a BNN distance-estimated detector in $\mathcal{T}1$ and report the value of WER deduction, area under the curve (AUC), false positive (FP), and false negative (FN) ratios in Table 2. We observe that Conformer-based ASR is already improved on the four evaluation metrics. The BNN-based detection and filtering system demonstrates superior results compared to the TD detection baseline [12].

Task 2 ($\mathcal{T}$2) for targeted adversarial robustness evaluation: Unlike untargeted perturbation, targeted sentence adversarial robustness evaluation aims to prevent the ASR from making misleading predictions. For example, ”open the door” is one classical malicious target used in adversarial noise evaluation. To generate noise with a targeted output, the optimization process tries to find a minimal noise signal δ that minimize prediction loss between noisy prediction $\mathbf{y}_{\text{adv}}$ and target $\mathbf{y}_{\text{target}}$. The second row of Table 3 shows that the output predictions of evaluated ASR models are not robust a mix of noisy Librispeech data. For $\mathcal{T}2$ experiments, we select (1) local smoothing [12] (LS), which uses a sliding window of fixed length for local smoothing to reduce the adversarial perturbation; (2) downsampling [12] (DS), where we downsample the original 16 kHz audio to 8 kHz, resulting in a audio file with a band-limit, that avoids sacrificing the quality of the recovered audio while reducing the adversarial noise in the reconstruction phase; and (3) TD methods from [12]. (4) For model defense based on speech enhancement (SE), we select the state-of-the-art self-attention U-Net gated speech enhancement [8] against closed-model adversarial noise. As shown in Table 3, the BNN-based detection and filtering outperforms all the existing methods and renders 91.94% of $\mathcal{A}_E$ and 94.13% of $\mathcal{A}_T$ adversarial perturbations unsuccessful—a new model robustness benchmark when evaluating with a mix of noisy Librispeech data.

Table 3: The unsuccessful rate (UR%) indicating how often adversarial noise fails to manipulate an ASR model with malicious output of “open the door”. A **higher** UR value indicates a **robust** system performance with selected defense algorithms (lines 3 to 5) discussed in Section 4.2. $\mathcal{A}_E$ represents evolutionary noise and $\mathcal{A}_Z$ represents zeroth-order optimized noise as summarized in Table 1.

Evaluation	RNN-T	Conformer	wav2vec-2.0	Avg.
$\mathcal{A}_E$ noise [15]	6.76	8.79	6.39	7.31
$\mathcal{A}_E$ + LS	10.39	10.27	10.92	10.52
$\mathcal{A}_E$ + DS	9.39	11.23	8.92	9.84
$\mathcal{A}_E$ + TD [12]	88.73	86.34	84.50	86.52
$\mathcal{A}_E$ + SE [8]	80.99	81.42	81.34	81.25
$\mathcal{A}_E$ + BNN	**91.23**	**93.34**	**91.25**	**91.94**
$\mathcal{A}_Z$ noise [18]	24.76	25.76	23.39	24.63
$\mathcal{A}_Z$ + LS	24.32	25.83	23.92	24.69
$\mathcal{A}_Z$ + DS	23.93	25.92	23.95	24.60
$\mathcal{A}_Z$ + TD [12]	90.87	92.34	89.12	90.77
$\mathcal{A}_Z$ + SE [8]	91.32	93.42	89.34	91.36
$\mathcal{A}_Z$ + BNN	**94.53**	**95.34**	**92.52**	**94.13**

5. CONCLUSIONS

We have demonstrated a new framework leveraging Bayesian neural networks for latent distribution modeling to improve adversarial robustness for end-to-end speech recognition. The newly proposed BNN method attains the best results compared to existing enhancement methods under closed-model adversarial evaluation. The experimental results suggest that recent Conformer and wav2vec-2.0 methods also suffer from the adversarial evaluation challenges for both untargeted and targeted adversarial evaluation. The proposed low-complexity detection method gives us promising new ways to design robust and reliable ASR systems.

6. ACKNOWLEDGMENT

The authors thank Ariya Rastrow, Mat Hans and Björn Hoffmeister from Alexa AI for their valuable comments and discussion.

7. REFERENCES

[1] Chin-Hui Lee, Lawrence R. Rabiner, Roberto Pieraccini, and Jay G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, vol. 4, no. 2, pp. 127–165, 1990.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[3] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system," in *Proc. IEEE Spoken Language Technology Workshop*, 2008, pp. 69–72.

[4] Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, Anirudh Raju, Denis Filimonov, and Ivan Bulyko, "Multi-task language modeling for improving speech recognition of rare words," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 1087–1093.

[5] Gideon Mendels, Erica Cooper, Victor Soto, Julia Hirschberg, Mark JF Gales, Kate M Knill, Anton Ragni, and Haipeng Wang, "Improving speech recognition and keyword search for low resource languages using web data," in *Proc. Interspeech*, 2015, pp. 829–833.

[6] Chao-Han Huck Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee, "PATE-AAE: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification," in *Proc. Interspeech*, 2021, pp. 881–885.

[7] Shi-Xiong Zhang, Yifan Gong, and Dong Yu, "Encrypted speech recognition using deep polynomial networks," in *Proc. IEEE ICASSP*, 2019, pp. 5691–5695.

[8] Chao-Han Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Chin-Hui Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 3107–3111.

[9] Piotr Żelasko, Sonal Joshi, Yiwen Shao, Jesus Villalba, Jan Trmal, Najim Dehak, and Sanjeev Khudanpur, "Adversarial attacks and defenses for speech recognition systems," *arXiv preprint arXiv:2103.17122*, 2021.

[10] Sining Sun, Ching-Feng Yeh, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie, "Training augmentation with adversarial examples for robust speech recognition," in *Proc. Interspeech*, 2018, pp. 2404–2408.

[11] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein, "Universal adversarial training," *Proc. AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 5636–5643, 2020.

[12] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song, "Characterizing audio adversarial examples using temporal dependency," in *Proc. International Conference on Learning Representations*, 2019.

[13] Yao Li, Tongyi Tang, Cho-Jui Hsieh, and Thomas Lee, "Detecting adversarial examples with bayesian neural network," *arXiv preprint arXiv:2105.08620*, 2021.

[14] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," in *Proc. International Conference on Learning Representations*, 2018.

[15] Shreya Khare, Rahul Aralikatte, and Senthil Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," *Proc. Interspeech*, pp. 3208–3212, 2019.

[16] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proc. Genetic and Evolutionary Computation Conference*, 2019, pp. 1111–1119.

[17] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[18] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng, "Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks," in *Proc. AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 742–749.

[19] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 1–7.

[20] Hiromu Yakura and Jun Sakuma, "Robust audio adversarial example for a physical attack," in *Proc. 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 5334–5341.

[21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *Proc. International Conference on Learning Representations*, 2015.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. International Conference on Learning Representations*, 2018.

[23] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry, "On adaptive attacks to adversarial example defenses," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[24] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., "Deep Speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[26] Arturs Bekasovs and Iain Murray, "Bayesian adversarial spheres: Bayesian inference and adversarial examples in a noiseless setting," in *Proc. Third Eorkshop on Bayesian Deep Learning*, 2018.

[27] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[28] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[29] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[30] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[31] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh, "Adv-BNN: Improved adversarial defense through robust bayesian neural network," in *Proc. International Conference on Learning Representations*, 2018.

[32] Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner, "Bayesian layers: A module for neural network uncertainty," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14660–14672, 2019.

[33] Soheil Kolouri, Yang Zou, and Gustavo K Rohde, "Sliced Wasserstein kernels for probability distributions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5258–5267.

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[35] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[36] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *Proc. International Conference on Learning Representations*, 2018.