On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression

Jun Qi, *Student Member, IEEE*, Jun Du, *Member, IEEE*, Sabato Marco Siniscalchi, *Senior Member, IEEE*, Xiaoli Ma, *Fellow, IEEE*, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—In this paper, we exploit the properties of mean absolute error (MAE) as a loss function for the deep neural network (DNN) based vector-to-vector regression. The goal of this work is two-fold: (i) presenting performance bounds of MAE, and (ii) demonstrating new properties of MAE that make it more appropriate than mean squared error (MSE) as a loss function for DNN based vector-to-vector regression. First, we show that a generalized upper-bound for DNN-based vector-to-vector regression can be ensured by leveraging the known Lipschitz continuity property of MAE. Next, we derive a new generalized upper bound in the presence of additive noise. Finally, in contrast to conventional MSE commonly adopted to approximate Gaussian errors for regression, we show that MAE can be interpreted as an error modeled by Laplacian distribution. Speech enhancement experiments are conducted to corroborate our proposed theorems and validate the performance advantages of MAE over MSE for DNN based regression.**

***Index Terms*—Deep neural network, vector-to-vector regression, vector-to-vector regression.**

I. Introduction

MEAN absolute error (MAE), originated from a measure of average error [1], is often employed in assessing vector-to-vector (a.k.a. multivariate) regression models [2]. Another form of average error is a root-mean-squared error (RMSE), but MAE was shown to outperform RMSE for measuring an average model accuracy in most situations except the Gaussian noisy scenarios [3]–[5]. An exception occurs when the expected error satisfies Gaussian-distributed and enough training samples are available [3]. Besides, mean squared error (MSE) is the squared form of RMSE and is commonly adopted as a regression loss function [6]–[9].

In the literature, there are some discussions on the relationship between MSE and MAE. Berger [10] presented pros and cons of squared and absolute errors from an estimation point of view. In [11], a better solution to support vector machines could be obtained based on a loss function of an absolute difference instead of the quadratic error. Li *et al.* [12] discussed the effectiveness of MAE and its variations when training a deep model for energy load forecasting; Imani *et al.* [13] investigated distributional losses, including both MAE and MSE, for regression problems from the perspective of efficient optimization. Pandey and Wang [14] exploited the MAE and MSE loss functions for generative adversarial nets (GANs). However, a comparison between MAE and MSE in terms of generalization capabilities [15]–[17] is still missing in theory. Thus, this paper aims at bridging this gap. In particular, we investigate MAE and MSE in terms of performance error bounds and robustness against various noises in the context of the deep neural network (DNN) based vector-to-vector regression, since DNNs offer better representation power and generalization capability in large-scale regression problems, such as those addressed in [18]–[21].

In this paper, we prove that the Lipschitz continuity property [22], [23], which holds for MAE but not for MSE, is a necessary condition to derive the upper bound on the Rademacher complexity [24], [25] of DNN based vector-to-vector regression functions, as we have demonstrated in [26]. Next, we show that the MAE Lipschitz continuity property can also result in a new upper bound on the generalization capability of DNN-based vector-to-vector regression in the presence of additive noise [27]–[29]. Moreover, another contribution of this work is that we establish a connection between the MAE loss function and Laplacian distribution [30], which is in contrast to the MSE loss function associated with Gaussian distribution [31]. In doing so, we can highlight the key advantages of MAE over MSE by comparing the characteristics of those two distributions.

Our experiments of speech enhancement are used as the regression task to assess our theoretical derivations and empirically verify the effectiveness of MAE over MSE. We choose regression-based speech enhancement because it is an unbounded mapping from $\mathbb{R}^d \to \mathbb{R}^q$, where enhanced speech features are expected to closely approximate the clean speech features in regression.

The remainder of this paper is presented as follows: Section II introduces some necessary math notations and theorems. Sections III, and IV highlight some key properties of the MAE loss function for DNN based vector-to-vector regression. Section V associates the MAE loss function with the Laplacian distribution. The related experiments of speech enhancement are given in Section VI, and Section VII concludes this work.

Manuscript received May 6, 2020; revised July 5, 2020; accepted August 11, 2020. Date of publication August 14, 2020; date of current version September 3, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniele Giacobello. *(Corresponding author: Jun Qi.)*

Jun Qi, Xiaoli Ma, and Chin-Hui Lee are with the School of Electrical, and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: jqi41@gatech.edu; xiaoli@gatech.edu; chl@ece.gatech.edu).

Jun Du is with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: jundu@ustc.edu.cn).

Sabato Marco Siniscalchi is with the Faculty of Architecture and Engineering, University of Enna "Kore," 94100 Enna, Italy, and also with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: marco.siniscalchi@unikore.it).

Digital Object Identifier 10.1109/LSP.2020.3016837

II. Preliminaries

1) **Notations**

- $f \circ g$: The composition of functions f and g.
- $||\mathbf{x}||_p$: L_p norm of the vector $\mathbf{x}$.
- $\mathbb{R}^d$: d-dimensional real coordinate space.
- $[n]$: An integer set $\{1, 2, \ldots, n\}$.
- $\mathbf{1}$: Vector of all ones.

2) **Lipschitz continuity**

Definition 1: A function f is β-Lipschitz continuous if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, for an integer $p \geq 1$,

$$||f(\mathbf{x}) - f(\mathbf{y})||_p \leq \beta ||\mathbf{x} - \mathbf{y}||_p. \tag{1}$$

3) **Mean Absolute Error (MAE)**

Definition 2: MAE measures the average magnitude of absolute differences between N predicted vectors $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $S^* = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, the corresponding loss function is defined as:

$$\mathcal{L}_{MAE}(S, S^*) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}_i||_1, \tag{2}$$

where $|| \cdot ||_1$ denotes L_1 norm.

4) **Mean Squared Error (MSE)**

Definition 3: MSE denotes a quadratic scoring rule that measures the average magnitude of N predicted vectors $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and N actual observations $S^* = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, the corresponding loss function is shown as:

$$\mathcal{L}_{MSE}(S, S^*) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}_i||_2^2, \tag{3}$$

where $|| \cdot ||_2$ denotes L_2 norm.

5) **Empirical Rademacher Complexity**

Definition 4: The empirical Rademacher complexity of a hypothesis space $\mathbb{H}$ of functions $h : \mathbb{R}^n \rightarrow \mathbb{R}$ with respect to N samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is:

$$\hat{\mathcal{R}}_S(\mathbb{H}) := \mathbb{E}_{\sigma_1,\ldots,\sigma_N}\left[\sup_{h \in \mathbb{H}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i h(\mathbf{x}_i)\right]. \tag{4}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_N$ are the Rademacher random variables, which are defined by the uniform distribution as:

$$\sigma_i = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2}. \end{cases} \tag{5}$$

In [32]–[34], it was shown that a function class with larger empirical Rademacher complexity is more likely to be overfit to the training data.

III. MAE Loss Function for Upper Bounding Empirical Rademacher Complexity

The Lipschitz continuity property is fundamental to derive an upper bound of the estimated regression error. In the following in Lemma 1, we show that the MAE loss function can ensure the Lipschitz continuity property. In Lemma 2, we instead show that the property does not hold for MSE.

Lemma 1: The MAE loss function is 1-Lipschitz continuous.

Proof: For two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^q$, and a target vector $\mathbf{x} \in \mathbb{R}^q$, the MAE loss difference is

$$\begin{aligned} &|\mathcal{L}_{MAE}(\mathbf{x}_1, \mathbf{x}) - \mathcal{L}_{MAE}(\mathbf{x}_2, \mathbf{x})| \\ &\quad = |||\mathbf{x}_1 - \mathbf{x}||_1 - ||\mathbf{x}_2 - \mathbf{x}||_1| \\ &\quad \leq ||\mathbf{x}_1 - \mathbf{x}_2||_1 \qquad \text{(triangle inequality)} \\ &quad = \mathcal{L}_{MAE}(\mathbf{x}_1, \mathbf{x}_2). \end{aligned} \tag{6}$$

■

Lemma 2: The MSE loss function cannot lead to the Lipschitz continuity property.

Proof: $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^q$, and $||\mathbf{x}_2||_2^2 > ||\mathbf{x}_1||_2^2$, there is

$$||\mathbf{x}_1 - \mathbf{x}_2||_2^2 = ||\mathbf{x}_1||_2^2 + ||\mathbf{x}_2||_2^2 - 2\mathbf{x}_1^T\mathbf{x}_2. \tag{7}$$

Next, we assume $\mathbf{x} = 2\mathbf{x}_2$, we have that

$$\begin{aligned} &||\mathbf{x}_1 - \mathbf{x}||_2^2 - ||\mathbf{x}_2 - \mathbf{x}||_2^2 \\ &\quad = ||\mathbf{x}_1||_2^2 - 2\mathbf{x}_1^T\mathbf{x} - ||\mathbf{x}_2||_2^2 + 2\mathbf{x}_2^T\mathbf{x} \\ &\quad = ||\mathbf{x}_1||_2^2 - 4\mathbf{x}_1^T\mathbf{x}_2 - ||\mathbf{x}_2||_2^2 + 4||\mathbf{x}_2||_2^2 \\ &\quad = ||\mathbf{x}_1||_2^2 - 4\mathbf{x}_1^T\mathbf{x}_2 + 3||\mathbf{x}_2||_2^2. \end{aligned} \tag{8}$$

By reducing Eq. (7) from Eq. (8),

$$\begin{aligned} &||\mathbf{x}_1 - \mathbf{x}||_2^2 - ||\mathbf{x}_2 - \mathbf{x}||_2^2 - ||\mathbf{x}_1 - \mathbf{x}_2||_2^2 \\ &\quad = 2||\mathbf{x}_2||_2^2 - 2\mathbf{x}_1^T\mathbf{x}_2 \\ &\quad > ||\mathbf{x}_2||_2^2 + ||\mathbf{x}_1||_2^2 - 2\mathbf{x}_1^T\mathbf{x}_2 \\ &\quad = ||\mathbf{x}_1 - \mathbf{x}_2||_2^2 \\ &\quad > 0, \end{aligned} \tag{9}$$

we derive that

$$\left|||\mathbf{x}_1 - \mathbf{x}||_2^2 - ||\mathbf{x}_2 - \mathbf{x}||_2^2\right| > ||\mathbf{x}_1 - \mathbf{x}_2||_2^2, \tag{10}$$

which contradicts the property of Lipschitz continuity. Thus, the MSE loss function is not Lipschitz continuous. ■

We now discuss the characteristic of Lipschitz continuity derived from the MAE loss function for upper bounding the estimation error $\mathcal{T}$, which is associated with the generalization capability and defined as:

$$\mathcal{T} = \sup_{f_v \in \mathbb{F}} \left|\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)\right| \leq \hat{\mathcal{R}}_S(\mathbb{L}). \tag{11}$$

where $\mathbb{F} = \{f_v : \mathbb{R}^d \rightarrow \mathbb{R}^q\}$ is a family of DNN based vector-to-vector functions and $\mathbb{L} = \{\mathcal{L}(f_v, f_v^*) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}, f_v \in \mathbb{F}\}$ denotes the family of generalized MAE loss functions. In [26], we have shown that the estimation error $\mathcal{T}$ can be upper bounded by the empirical Rademacher complexity $\hat{\mathcal{R}}_S(\mathbb{L})$.

In [26], we have also shown that the estimation error $\mathcal{T}$ can be further upper-bounded as:

$$\mathcal{T} = \sup_{f_v \in \mathbb{F}} \left|\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)\right| \leq \hat{\mathcal{R}}_S(\mathbb{L}) \leq \hat{\mathcal{R}}_S(\mathbb{F}), \tag{12}$$

where $\hat{\mathcal{R}}_S(\mathbb{F})$ is defined as:

$$\hat{\mathcal{R}}_S(\mathbb{F}) = \frac{1}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_v \in \mathbb{F}} \sum_{i=1}^{N} (\sigma_i \mathbf{1})^T f_v(\mathbf{x}_i)\right], \tag{13}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ denotes a set of Rademacher random variables as shown in Definition 4.

IV. Mae Loss Function for Dnn Robustness Against Additive Noises

We now show that the MAE loss function can give an upper bound for regression errors to ensure DNN robustness against additive noises.

Theorem 1: For an objective function $h = \mathcal{L} \circ f_v : \mathbb{R}^d \to \mathbb{R}$ with the MAE loss function $\mathcal{L} : \mathbb{R}^q \to \mathbb{R}$ and a vector-to-vector regression function $f_v : \mathbb{R}^d \to \mathbb{R}^q$, the difference of the objectives for adding noise $\boldsymbol{\eta}$ to signal $\mathbf{x}$ is bounded as:

$$|h(\mathbf{x} + \boldsymbol{\eta}) - h(\mathbf{x})| \leq L_2 ||\boldsymbol{\eta}||_2, \tag{14}$$

where $L_2 = \sum_{i=1}^{q} L_{2,i}$ is the Lipschitz constant for DNN based vector-to-vector regression, and each $L_{2,i}$ is shown as:

$$L_{2,i} = \sup\{||\nabla f_i(\mathbf{x})||_2 : \mathbf{x} \in \mathbb{R}^d\}. \tag{15}$$

Proof: To prove Theorem 1, we first introduce Lemma 3, which is achieved by the modification of Theorem 1 in [35].

Lemma 3: For a vector-to-vector regression function $f : \mathbb{R}^d \to \mathbb{R}^q$ with the property of Lipschitz continuity, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, there exists an inequality as:

$$||f(\mathbf{x}) - f(\mathbf{y})||_1 \leq L_p ||\mathbf{x} - \mathbf{y}||_q, \tag{16}$$

where $L_p = \sup\{||\nabla f(\mathbf{x})||_p : \mathbf{x} \in \mathbb{R}^d\}$ is a Lipschitz constant, and $\frac{1}{p} + \frac{1}{q} = 1, p, q \geq 1$.

We employ the fact that DNNs with the ReLU activation function are Lipschitz continuous [36]. Then, based on both triangle inequality and Lemma 3, we can upper bound the difference of objective functions with and without the additive noise $\boldsymbol{\eta}$ as:

$$\begin{aligned} |h(\mathbf{x} + \boldsymbol{\eta}) - h(\mathbf{x})| &= |||f(\mathbf{x} + \boldsymbol{\eta})||_1 - ||f(\mathbf{x})||_1| \\ &\leq ||f(\mathbf{x} + \boldsymbol{\eta}) - f(\mathbf{x})||_1 \qquad \text{(triangle ineq.)} \\ &= L_2 ||\boldsymbol{\eta}||_2 \qquad \text{(Lemma 2)} \end{aligned}$$

which completes the proof. ■

Theorem 1 holds for the MAE loss function but is not valid for MSE loss because it is not Lipschitz continuous. In other words, the difference of additive noises imposed upon the DNN based vector-to-vector function is unbounded on the MSE loss function but the MAE can guarantee an upper bound.

The upper bound makes more sense when the additive noise is small because the upper bound suggests that the imposed noise cannot lead to significant performance degradation.

V. Connection of Mae Loss Function to Laplacian Distribution

We now separately link the MAE and MSE loss functions to Laplacian distribution (LD) and Gaussian distribution (GD) based loss functions as defined in [37]. Both LD and GD based losses for DNN-based multivariate regression were experimentally compared and contrasted in [37], and it was shown that the LD loss can attain better vector-to-vector regression accuracies than those obtained optimizing GD losses.

For N input samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and N target vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, assuming $f : \mathbb{R}^d \to \mathbb{R}^q$ is a vector-to-vector regression function, we change the MAE loss function as:

$$\begin{aligned} \mathcal{L}_{MAE}(S, S^*) &= \frac{1}{N} \sum_{i=1}^{N} ||f(\mathbf{x}_n) - \mathbf{y}_n||_1 \\ &= \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{d} |f_m(\mathbf{x}_n) - y_{n,m}| \\ &= \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{d} \frac{|\hat{f}_m(\mathbf{x}_n) - \hat{y}_{n,m}|}{\alpha_m}, \end{aligned} \tag{17}$$

where $\hat{f}_m(\mathbf{x}_n) = \alpha_m f_m(\mathbf{x}_n)$, $\hat{y}_{n,m} = \alpha_m y_{n,m}$, and α_m is the variance of dimension m.

To link the LD based loss function $\mathcal{L}_{LD}(S, S^*)$ in [37], an additional term $N \sum_{m=1}^{d} \ln \alpha_m$ is added to $\mathcal{L}_{MAE}(S, S^*)$, and we obtain

$$\mathcal{L}_{LD}(S, S^*) = \mathcal{L}_{MAE}(S, S^*) + N \sum_{m=1}^{d} \ln \alpha_m. \tag{18}$$

Moreover, an MSE based loss function can be modified as:

$$\mathcal{L}_{MSE}(S, S^*) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{d} \frac{|\hat{f}_m(\mathbf{x}_n) - \hat{y}_{n,m}|^2}{\alpha_m^2}. \tag{19}$$

Then, the GD based loss $\mathcal{L}_{GD}(S, S^*)$ can be derived by adding the term $N \sum_{m=1}^{d} \ln \alpha_m$ to the MSE loss $\mathcal{L}_{MSE}(S, S^*)$,

$$\mathcal{L}_{GD}(S, S^*) = \mathcal{L}_{MSE}(S, S^*) + N \sum_{m=1}^{d} \ln \alpha_m. \tag{20}$$

We can observe that $\mathcal{L}_{MAE}(S, S^*)$ and $\mathcal{L}_{MSE}(S, S^*)$ are special cases of $\mathcal{L}_{LD}(S, S^*)$ and $\mathcal{L}_{GD}(S, S^*)$ without concerning the variance terms. When $\forall m \in [d]$, the variance α_m is a constant, $\mathcal{L}_{LD}(S, S^*)$ and $\mathcal{L}_{GD}(S, S^*)$ exactly correspond to $\mathcal{L}_{MAE}(S, S^*)$ and $\mathcal{L}_{MSE}(S, S^*)$, respectively.

Since the work [37] suggests that the LD based loss function can achieve better regression performance than the GD based one, we show that the MAE based loss function can also keep the advantage over the MSE when the variance related terms are the same. Our experiments of speech enhancement in Section VI, where both MAE and MSE loss functions are involved, are used to verify that.

VI. Experiments

This section presents our speech enhancement experiments to corroborate the aforementioned theorems. The goal of the experiments is to verify that MAE can achieve better regression performance than MSE under various noisy conditions because of the ensured upper bounds on the MAE loss functions for DNN-based vector-to-vector regression.

A. Data Preparation

Our experiments were conducted on the Edinburgh noisy speech database, where there were a total 23075 and 824 clean utterances for training and testing, respectively. The noisy training dataset at four SNR levels (15 dB, 10 dB, 5 dB, 0 dB), was obtained using the following noises: a domestic noise (inside a kitchen), an office noise (in a meeting room), three public space noises (cafeteria, restaurant, subway station), two transportation noises (car and metro) and a street noise (busy traffic intersection). In sum, we had 40 different noisy types to synthesize 23075 noisy training speech utterances. As for the noisy test set, the noisy conditions include a domestic (living room), an office noise (office space), one transport (bus) and two street noises (open area cafeteria and a public square) at four SNR values (17.5 dB, 12.5 dB, 7.5 dB, 2.5 dB). Thus, there were 20 various noisy conditions for generating totally 824 noisy test speech utterances. The Edinburgh noisy speech corpus provides a more challenging speech scenario, which allows us to better support our Theorems.

B. Experimental Setup

In this work, DNN based vector-to-vector regression models followed feed-forward architectures, where the inputs were normalized log-power spectral (LPS) feature vectors of noisy speech [38], [39], and the outputs were LPS features of either clean or enhanced speech. At training time, clean LPS vectors were assigned to the top layer of DNN to function as targets. At test time, the top layer of DNN generated enhanced LPS vectors. The architecture of DNN had the structure 771-800-800-800-800-800-1600-257, which corresponds to Input$-$Hidden$-$Output. The ReLU activation function was employed in the hidden neurons, and the top layer was connected to a linear function for vector-to-vector regression. The enhanced waveforms were reconstructed based on the overlap-add method as shown in [20]. The technique of global variance equalization [40] was utilized to improve the subjective perception of speech enhancement. At training time, the standard back-propagation (BP) algorithm was adopted to update the model parameters. The MAE and MSE loss functions were separately used to measure the difference between normalized LPS features and the reference ones. The stochastic gradient descent (SGD) based optimizer with a learning rate of 1×10^{-3} and a momentum rate of 0.4 was set up for the BP algorithm. Moreover, noise-aware training (NAT) [41] was also used to enable non-stationary noise awareness. Context information was accounted at the input by using 3 LPS vectors by concatenating frames within a sliding window [42]–[44]. During the training time, the maximum 20 epochs were set, and one-tenth of training data were randomly split into a validation set. If the performance of the model on the validation dataset started to degrade, the training process was stopped.

The evaluation metrics were based on three types of criteria: MAE, MSE, perceptual evaluation of speech quality (PESQ) [45], and short-time objective intelligibility (STOI) [46]. PESQ, which ranges from -0.5 to 4.5, is an indirect evaluation which is highly correlated with speech quality. A higher PESQ score corresponds to a better perception quality. Similarly, the STOI score, which ranges from 0 to 1, also refers to a measurement of predicting the intelligibility of noisy or enhanced speech. A higher STOI score corresponds to a better speech intelligibility.

TABLE I
THE MAE AND MSE VALUES ON EDINBURGH SPEECH CORPUS.

Models	MAE	MSE
DNN-MAE	0.7812	0.7954
DNN-MSE	0.8278	0.8371

TABLE II
THE PESQ AND STOI SCORES ON EDINBURGH SPEECH CORPUS

Models	PESQ	STOI
DNN-MAE	2.93	0.8509
DNN-MSE	2.85	0.8317

C. Evaluation Results

Using the DNN models trained with the MAE criterion (DNN-MAE) and the MSE criterion (DNN-MSE), Table I lists the MAE values for speech enhancement experiments with test data. The MAE values evaluated with DNN-MAE in the top row are always lower than those in the bottom row evaluated with DNN-MSE under the same noisy condition in each column. More specifically, MAE scores by DNN-MAE achieves a lower value than DNN-MSE (0.7812 vs. 0.8278). Similarly, DNN-MAE achieves a lower MSE score than DNN-MSE (0.7954 vs. 0.8371). Besides, the MAE scores for both DNN-MAE and DNN-MSE are consistently lower than the MSE values.

Moreover, Table II shows PESQ and STOI scores obtained with the DNN-MAE and DNN-MSE models. It can be seen that the DNN model trained with the MAE criterion consistently outperforms models trained with the MSE criterion (2.93 vs. 2.85 for PESQ, and 0.8509 vs. 0.8317 for STOI), which further confirms that MAE is a good objective function to optimize when training DNNs for speech enhancement.

Furthermore, the performance advantages of DNN-MAE over DNN-MSE corresponds to the aforementioned theorems: (1) the upper bound in Eq. (14) ensures more robust performance against the additive noise; (2) the performance gain is consistent with the connection between MAE loss function and the Laplacian distribution.

VII. CONCLUSION

This work investigates the advantages of the MAE loss function for DNN based vector-to-vector regression. On one hand, we emphasize that the Lipschitz continuity property can not only ensure a performance upper bound on DNN-based vector-to-vector regression but also give an upper bound to predict the robustness against additive noises. On the other hand, we associate the MAE loss function with Laplacian distribution. Our experiments show that DNN based regression optimized with the MAE loss function can achieve lower loss values than those obtained with the MSE counterpart. Moreover, the MAE loss function can also lead to better-enhanced speech quality in terms of the PESQ and STOI scores. Our empirical results are in line with the proposed theorems for MAE and indirectly reflect that the MAE loss functions can benefit from its related upper bounds as shown in this study.

REFERENCES

[1] C. Willmott *et al.*, "Statistics for the evaluation of model performance," *J. Geophys. Res*, vol. 90, no. C5, pp. 8995–9005, 1985.

[2] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Math. Methods Appl. Sci.*, vol. 5, no. 5, pp. 216–233, 2015.

[3] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[4] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[5] C. J. Willmott, K. Matsuura, and S. M. Robeson, "Ambiguities inherent in sums-of-squares-based error statistics," *Atmos. Environ.*, vol. 43, no. 3, pp. 749–752, 2009.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Berlin, Germany: Springer, 2001.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.

[9] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U. K.: Cambridge Univ. Press, 2014.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Berlin, Germany: Springer, 2013.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[12] N. Li, L. Wang, X. Li, and Q. Zhu, "An effective deep learning neural network model for shortterm load forecasting," *Concurrency Comput.: Pract. Experience*, vol. 32, no. 7, Jan. 2020, Paper e5595.

[13] E. Imani and M. White, "Improving regression performance with distributional losses," in *Prof. Int. Conf. Mach. Learn.*, 2018, pp. 2157–2166

[14] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5414–5418.

[15] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probability Appl.*, vol. 16, no. 2, pp. 264–280, 2018.

[16] Z. Charles and D. Papailiopoulos, "Stability and generalization of learning algorithms that converge to global optima," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 745–754.

[17] J. Qi, J. Du, S. M. Siniscalchi, and C.-H. Lee, "A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1932–1943, Dec. 2019.

[18] A. Lorencs, I. Mednieks, and J. Sinica-Sinavskis, "Biomedical image processing based on regression models," in *Proc. 14th Nordic-Baltic Conf. Biomed. Eng. Med. Phys.*, 2008, pp. 536–539.

[19] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 349–366, Feb. 2007.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[21] J. Qi, H. Hu, Y. Wang, C. H. Yang, S. Marco Siniscalchi, and C. Lee, "Tensor-to-vector regression for multi-channel speech enhancement based on tensor-train network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7504–7508.

[22] O. L. Mangasarian and T.-H. Shiau, "Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems," *SIAM J. Control Optim.*, vol. 25, no. 3, pp. 583–595, 1987.

[23] M. O'Searcoid, *Metric Spaces*. Berlin, Germany: Springer, 2006.

[24] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2002.

[25] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Annu. Statist.*, vol. 33, no. 4, pp. 1497–1537, Aug. 2005.

[26] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "Analyzing upper bounds on mean absolute errors for deep neural network based vector-to-vector regression," *IEEE Trans. Signal Process.*, vol. 68, pp. 3411–3422, 2020.

[27] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[28] C. Yang, J. Qi, P. Chen, X. Ma, and C. Lee, "Characterizing speech adversarial examples using self-attention U-Net enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3107–3111.

[29] T.-W. Weng *et al.* "Evaluating the robustness of neural networks: An extreme value theory approach," in *Proc. Int. Conf. Representation Learn.*, 2018, *arXiv:1801.10578*.

[30] S. Kotz, T. Kozubowski, and K. Podgorski, *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance*. Berlin, Germany: Springer, 2012.

[31] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 152–177, 1963.

[32] J. Fan, C. Ma, and Y. Zhong, "A selective overview of deep learning," *Statistical Sci.*, 2020, *arXiv:1904.05526*.

[33] J. Zhu, B. R. Gibson, and T. T. Rogers, "Human rademacher complexity," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 2322–2330.

[34] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[35] R. Paulavičius and J. Žilinskas, "Analysis of different norms and corresponding Lipschitz constants for global optimization," *Technological Econ. Develop. Economy*, vol. 12, no. 4, pp. 301–306, 2006.

[36] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas, "Efficient and accurate estimation of lipschitz constants for deep neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 11 423–11 434.

[37] 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, 2019.

[38] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[39] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on Gammatone filters for robust speech recognition," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2013, pp. 305–308.

[40] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech*, 2012, pp. 1436–1439.

[41] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. Interspeech*, 2014, pp. 2670–2674.

[42] J. Qi, D. Wang, and J. Tejedor Noguerales, "Subspace models for bottleneck features," in *Proc. Interspeech*, 2013, pp. 1746–1750.

[43] J. Qi, D. Wang, J. Xu, and J. Tejedor Noguerales, "Bottleneck features based on Gammatone frequency cepstral coefficients," in *Proc. Interspeech*, 2013, pp. 1751–1755.

[44] J. Qi and J. Tejedor, "Robust submodular data partitioning for distributed speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 2254–2258.

[45] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. 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.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.