ON THE ROBUSTNESS OF NON-INTRUSIVE SPEECH QUALITY MODEL BY ADVERSARIAL EXAMPLES

Hsin-Yi Lin, Huan-Hsin Tseng, Yu Tsao

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

It has been shown recently that deep learning based models are effective on speech quality prediction and could outperform traditional metrics in various perspectives. Although network models have the potential to be a surrogate for complex human hearing perception, they may contain instabilities in predictions. This work shows that deep speech quality predictors can be vulnerable to adversarial perturbations, where the prediction can be changed drastically by unnoticeable perturbations as small as -30 dB compared with speech inputs. In addition to exposing the vulnerability of deep speech quality predictors, we further explore and confirm the viability of adversarial training for strengthening robustness of models.

***Index Terms*—** MOS, speech quality models, adversarial examples, perturbation, robustness.

1. INTRODUCTION

The need for speech quality evaluation has been raised as the increasing use of speech processing algorithms and telecommunications applications. Traditionally, the Mean Opinion Score (MOS) of a speech sample is derived from averaging subjective listening tests [1, 2, 3] of participants. Although actual human rating is considered the most faithful index to assess speech quality, listening tests are typically costly and time consuming. To reduce the last two factors, automatic speech quality predictions mimicking human perceptions have become an active research topic. Several metrics such as Perceptual Evaluation of Speech Quality (PESQ) [4] and Short-Time Objective Intelligibility (STOI) [5] were introduced as possible surrogates for speech quality. One caveat is that several viable candidates require clean references (labels) for evaluation, which are not always available in real world tasks. Among several attempts to avoid clean label requirements, deep learning is one strong candidate for such tasks, due to the complex nonlinear functionality. Indeed, several neural network based approaches have been proposed to estimate speech quality [6, 7, 8, 9, 10].

While neural network models provide a simple solution to get rid of clean reference requirement with seemingly satisfying results, stability and consistency across different data are not ensured. In fact, it has been reported in several areas that certain imperceptible perturbations on input data can drastically alter network output, so that the prediction ability is heavily questioned. Such phenomena is usually caused by the so-called *adversarial examples*, which prevail in both image and audio domain when using neural networks. Previous literature regarding audio adversarial examples was mainly focused on Automatic Speech Recognition (ASR) systems [11, 12, 13]. As the network-based speech quality prediction has gradually become a trend and has derived numerous downstream applications, it is important to carefully examine their prediction behavior in adversarial settings.

Our contribution. This work utilizes one well-known DNSMOS p.835 (non-intrusive quality) predictor to demonstrate that a deep-learning based model can be vulnerable to targeted adversarial attack. Our contribution contains two parts: (1) an approach to generate adversarial audio samples against the DNSMOS network is presented, while the adversarial attack is hardly noticeable to human ears, where the imperceptibility is supported by a human test. (2) we show that although the adversarial attack exposes the weakness of deep speech quality predictors, it can be used for model enhancement. Our experiments confirmed that the robustness can be strengthened by adversarial training.

2. ADVERSARIAL EXAMPLES CAUSING INCONSISTENT EVALUATIONS

2.1. Speech quality prediction under perturbation

This study investigates how quality predictions can be affected by small perturbations on input speech. Consider a speech quality prediction network f. An *adversarial example* $\widetilde{x}$ of f is an input data similar to another sample x under certain measurement, such that the prediction $f(\widetilde{x}) \neq f(x)$. Due to the desired property that $\widetilde{x}$ should be close to an input x, adversarial examples are naturally considered as perturbations of input samples. As such, a (small) *adversarial perturbation* δ can also be defined whenever $\widetilde{x} = x + \delta$ forms an adversarial sample.

The general description for targeted adversarial examples can be formulated as an optimization problem. Given $\epsilon \in \mathbb{R}$,

an input $x \in \mathbb{R}^T$ and a target $y \in \mathbb{R}^k$, consider:

$$\min_{\delta \in \mathbb{R}^T} L_S\left(x+\delta, x\right) + c \cdot L_T(x+\delta, y) \text{ s.t. } D(\delta) < \epsilon. \quad (1)$$

where $L_S : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ is a real-valued function measuring the *similarity* between x and the perturbed output $x+\delta$. $L_T : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ estimates the *target deviation* between output $f(x+\delta)$ from target y such that when $f(x+\delta) \to y$ one has $L_T(x+\delta, y) \to 0$. A coefficient $c \in \mathbb{R}$ is included to balance the two terms. When c is large, the optimization naturally emphasizes L_T and vice versa. $D : \mathbb{R}^T \to \mathbb{R}$ is a distortion metric for perturbations, introduced to be a constraint within tolerance ϵ allowed in a task.

In this study, we let $\mathbb{R}^T$ be the space of speech signals and choose $D = dB$ to measure the audio distortion in decibels (dB), which describes the relative loudness of a perturbation $\delta = (\delta_1, \ldots, \delta_T) \in \mathbb{R}^T$ with respect to an input $x = (x_1, \ldots, x_T)$:

$$dB_x(\delta) = 20 \log_{10} \left(\max_{t \in [0,T]} |\delta_t| / \max_{t \in [0,T]} |x_t| \right). \quad (2)$$

To confine perturbation decibel $dB_x(\delta) < \epsilon$, the perturbation form $\delta_t = A \cdot \tanh(z_t)$ is chosen with $A > 0$, $t = 1, \ldots, T$. Since $\tanh(z_t) \in (-1, 1)$ for any $z_t \in \mathbb{R}$, the perturbation amplitude is always bounded $|\delta_t| < A$.

To construct a function faithfully reflecting the similarity of two audio signals $\widetilde{x}$ and x, we consider comparisons in Fourier (spectrum) space under L^1-norm. Therefore, we define

$$L_S(\widetilde{x}, x) = \|\mathcal{F}(\widetilde{x}) - \mathcal{F}(x)\|_1 \quad (3)$$

with $\mathcal{F}$ the Short-Time Fourier Transform (STFT) onto Fourier space. The target deviation is chosen as:

$$L_T(x, y) = \|f(x) - y\|_1 \quad (4)$$

where $\|\cdot\|_1$ is the L^1-norm. With the similarity loss Eq. (3) and target loss Eq. (4) defined, together we derive the formulation for our adversarial task from Eq. (1):

$$\begin{aligned} &\min_{\delta \in \mathbb{R}^T} \|\mathcal{F}(x+\delta) - \mathcal{F}(x)\|_1 + c \cdot \|f(x+\delta) - y\|_1 \\ &\quad \text{such that} \quad dB_x(\delta) < \epsilon, \end{aligned} \quad (5)$$

This formulation is subsequently implemented to conduct adversarial training with small amplitude A.

2.2. Adversarial training to improve robustness

Although adversarial samples seem destructive to speech quality networks, there are occasions that they can be constructive. Below we explore the viability of enhancing robustness using adversarial noises.

Given a quality predictor f and an audio sample x_i from a speech corpus $\{x_i\}_{i=1}^N$, we consider the score $y_i = f(x_i)$ predicted by f as a label such that the data pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are formed. Subsequently, given a target $\widetilde{y}$, an adversarial perturbation δ_i can be derived by Eq. 5 associated to each x_i to achieve $f(x_i + \delta_i) = \widetilde{y}_i$. When $\widetilde{y}_i \neq y_i$, the network f is considered attacked. Particularly, when $\|\widetilde{y}_i - y_i\|$ is large with tiny δ_i, the network prediction is considered unstable.

To enhance a predictor with such type of weakness, we make the network be aware of adversarial examples. That is we correct "false" prediction $\widetilde{y}_i$ with the regular y_i and achieve the defense. By collecting all adversarial examples, we can teach (retrain) the predictor with these irregular data pairs, called an *adversarial dataset* $\mathcal{AD} = \{(x_i + \delta_i, y_i)\}_{i=1}^N$.

Our goal is to derive a robustness-improved model g from f by training on an adversarially-extended dataset $\mathcal{D} \cup \mathcal{AD}$, where $\mathcal{AD}$ in this case can be regarded as data augmentation to strengthen the network stability. Our loss function for training process is defined as follows:

$$\mathcal{L}(g) = \sum_i \|g\left(x_i\right) - f(x_i)\|_2^2, \quad (x_i \in \mathcal{D} \cup \mathcal{AD}). \quad (6)$$

As the training is operated on two datasets $\mathcal{D}$ and $\mathcal{AD}$, there are two types of losses involved:

$$\mathcal{L}_1(g) = \|g\left(x_i + \delta_{x_i}\right) - f(x_i)\|_2^2, \ \mathcal{L}_2(g) = \|g(x_i) - f(x_i)\|_2^2, \quad (7)$$

where $\|\cdot\|_2$ is the L^2 norm. We note that $\mathcal{L}_1$ intends to correct the adversarial perturbations with regular labels, and $\mathcal{L}_2$ serves as a forgetting loss [14], which prevents g from forgetting old knowledge inherited from f. Ideally, a new model g is free from adversarial attack so that a perturbed audio has very similar scores as the unattacked $g\left(x_i + \delta_{x_i}\right) \cong f(x_i)$. In the meantime, any unperturbed audio should maintain the same score as before $g(x_i) \cong f(x_i)$.

Recruiting adversarial data into training has been recognized as effective in defending adversarial attacks, and model robustness is indeed found improved [15, 16, 17]. Different from previous works where most of the demonstrations were in the image domain, this work devotes to speech quality assessment and intends to confirm the viability of adversarial training on speech quality models.

It should be noted that if a speech corpus has quality labels obtainable, one can always replace the surrogate index $y_i = f(x_i)$ with real (better) labels. Due to the inaccessibility in many corpus, it is our proposal to adopt $y_i = f(x_i)$ instead, which is probably more useful on numerous occasions.

3. EXPERIMENTS

The following experiments were conducted with the released **DNSMOS P.835** CNN-based model, which predicts 3 subjective scores of noisy-clips based on ITU-T P.835, speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL) [18]. The codes of this work can

be found at https://github.com/hsinyilin19/adversarial_example_speech_quality.

3.1. Adversarial examples on quality prediction model

3.1.1. Datasets

DNS-2020 is the dataset from 2020 DNS Challenge [19], containing a noise set of 65,000 clips and 150 classes, selected from Audioset and Freesound. The clean speech data has 500 hours from 60,000 clips, obtained from the public audio books dataset named *Librivox*. We adopted the resulting training dataset of 150 audio classes and 60,000 noisy clips.

TIMIT is a corpus frequently utilized in speech-related experiments. The speech data contains versatile acoustic-phonetic information including phonetically-compact sentences (SX) and phonetically-diverse sentences (SI), as well as regional diversity in dialects sentences (SA). A suggested *core test* set consists of 192 sentences from 24 speakers, where 15 speakers were randomly selected in a balanced manner to conduct the adversarial experiments.

3.1.2. Experimental setting

Under the pretrained weights of the DNSMOS network f, an adversarial perturbation $\delta_{x,\widetilde{y}}$ was sought to attain a desired target (MOS) score $\widetilde{y}$ from input x using optimization Eq. (5).

The STFT transformation $\mathcal{F}$ to measure L_1-similarity in Eq. (5) had 512 Fourier basis ($n_{\text{FFT}} = 512$) under Hann window length 512 and hop size 128, resulting in 257 STFT dimensions denoting as frequency bins. The parameter $c = 10$ was used in implementations. Input audio magnitudes were normalized, and the perturbations were set in the form $\delta = 0.03 \cdot \tanh z$ so that the resulting $dB_x(\delta) < -30dB$.

We note that a target $\widetilde{y} = (\widetilde{y}_1, \widetilde{y}_2, \widetilde{y}_3) = (\texttt{SIG}, \texttt{BAK}, \texttt{OVRL})$ can have arbitrary subscore $\widetilde{y}_i \in [1, 5]$. In our case, we intentionally consider utterly different scores to see interesting results. Particularly, we let $\widetilde{y} = (\widetilde{y}_1, \widetilde{y}_2, \widetilde{y}_3)$ to alter from the original prediction $y = (y_1, y_2, y_3)$ as follows:

$$\widetilde{y}_i = \begin{cases} 5 & \text{if } y_i \in [1, 3], \\ 1 & \text{if } y_i \in (3, 5], \end{cases} \quad (i = 1, 2, 3) \tag{8}$$

This relabelling strategy is interesting since a very clean speech with original high score $y = (5, 5, 5)$ is to be downgraded as $\widetilde{y} = (1, 1, 1)$ using adversarial perturbations. Contrarily, a very noisy audio with low predicted score $y = (1, 1, 1)$ is to be uplifted to $\widetilde{y} = (5, 5, 5)$. Another interesting case included is a mid-ranged score $y = (3.1, 2.8, 3.2)$ to be judged as $\widetilde{y} = (1, 5, 1)$, where the three MOS scores are torn to be contrasting.

3.1.3. An example of results

With limited space we demonstrate one adversarial example. In Fig. 1, an original audio (reader_01326_9_7J3kchZ5UAg) from DNS-2020 and its adversarial correspondent are shown in spectrogram, where the prediction $y = (4.06, 4.16, 3.73)$ was downgraded to $\widetilde{y} = (0.99, 1.0, 1.0)$ by a small perturbation with small distortion $dB_x(\delta) = -36.01$ dB. This audio demonstration and more examples can be found at https://hsinyilin19.github.io/Demo_adversarial_example_speech_quality/.

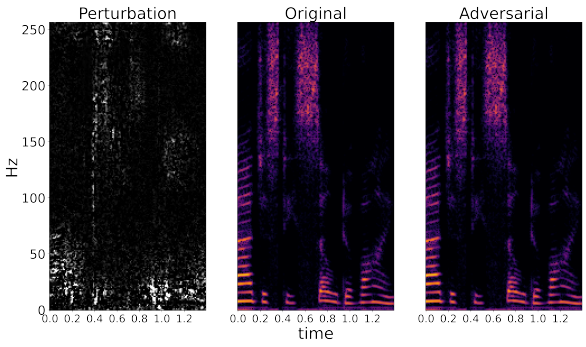

Fig. 1. Visualization of an adversarial perturbation and its corresponding audio from DNS-2020. The perturbation is observed to conceal in utterances to sneakily alter scores.

3.2. Enhancing robustness by adversarial training

3.2.1. Dataset

VCTK-DEMAND is a noisy speech corpus premixed by the Voice-Bank (VCTK) [20] with real-world noises DEMAND database [21]. VCTK-DEMAND [22] has 11,572 training samples and 824 testing samples composed by 28 speakers at 48 kHz sampling rate. The VCTK-DEMAND corpus was used to conduct this enhancing experiment owing to the suitable data size for reasonable adversarial training time.

3.2.2. Experimental setting

In this experiment, adversarial perturbations were generated for each audio sample in both training and testing dataset of VCTK-DEMAND using Eq. (5), (8). Consequently, the entire training set $\mathcal{D} = \{(x_i, f(x_i))\}_{i=1}^N$ yielded a corresponding adversarial set $\mathcal{AD} = \{(x_i + \delta_{x_i,\widetilde{y}_i}, f(x_i))\}_{i=1}^N$ with $N = 11,572$. A new network g was trained by joint data $\mathcal{D} \cup \mathcal{AD}$ with initial weights from f. The loss function was $\mathcal{L}$ in Eq. (6) and the model f was held fixed during the training process.

After training, the test set and its adversarial perturbations were used to verify the robustness of g. A model g with output (g_1, g_2, g_3) to claim an enhanced robustness should have the following property,

$$|g_j(x_i + \delta_{x_i}) - f_j(x_i)| < |f_j(x_i + \delta_{x_i}) - f_j(x_i)| \tag{9}$$

for any audio x_i along with a perturbation δ_{x_i} where (f_1, f_2, f_3) are original predictions by f. Inequality (9) simply checks whether g can better sustain adversarial perturbations than

the original f in recovering an unperturbed score. For convenience, we denote the following errors:

$$E_{g_j} = \frac{1}{N}\sum_{i=1}^{N}\left|g_j(x_i+\delta_{x_i}) - f_j(x_i)\right|$$
$$E_{f_j} = \frac{1}{N}\sum_{i=1}^{N}\left|f_j(x_i+\delta_{x_i}) - f_j(x_i)\right| \quad (10)$$
$$F_{g_j} = \frac{1}{N}\sum_{i=1}^{N}\left|g_j(x_i) - f_j(x_i)\right|$$

where E_{f_j} computes the prediction deviation of f and F_{g_j} denotes the forgetting rate to check how much knowledge of f is preserved in g.

3.2.3. *Enhancing results*

Fig. 2 shows the prediction deviation of f and g. For $j = 1, 2, 3$ (SIG, BAK, OVRL), the new deviation E_{g_j} was observed to largely reduce down to less than half of the original deviation E_{f_j}. This clearly indicates that g obtained better defense against *unseen* adversarial perturbations on the test set. In the meantime, small F_{g_j} addresses that predictions of g concurred with those of f on the unattacked test audio. As such, the robustness of g was indeed improved from f to conclude our experiment.

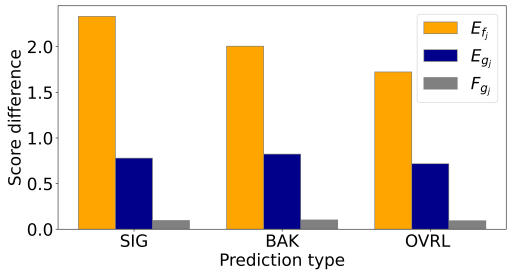

Fig. 2. The score differences before/after adversarial training.

3.3. Human Imperceptibility Evaluation

Having constructed numerous adversarial samples for the three datasets following procedure in Sec. 3.1, human evaluations were conducted to verify their imperceptibility.

In this evaluation, 35 participants were given 30 pairs of audio samples, asked to identify whether any difference might exist within each pair. The 30 pairs were composed of 10 randomly chosen pairs from each of the three datasets: DNS-2020, TIMIT, and VCTK-DEMAND. Among 10 pairs from each dataset, 7 pairs were adversarial; the other 3 were identically unperturbed ones. The participants were instructed to carefully answer either "**A:** *this pair is identical*" or "**B:** *this pair has difference*". The participants did not have a time limit and the audio can be repeated until their answers were final.

3.3.1. *Results*

The results (Fig. 3) were examined in two perspectives. First, there were at most 10 (out of 35) persons chosen "B" for each audio pair. Moreover, there were only 2 (out of 30 pairs) that received 10 B's from participants. Among the 2 pairs, one was adversarial; the other was in fact identical. In brief, no matter whether a pair is identical or adversarial, there is always more than 71.43% of the participants believing that they are identical.

Secondly, we conducted statistical hypothesis testing for each participant. Let the null hypothesis be "*the participant cannot tell the difference between identical and adversarial pairs, and so he/she was guessing*". Under the hypothesis, the z-score $= 2(x-15)/\sqrt{30}$, where x represents how many correct answers out of 30 questions each participant returned. After counting, we summarize that all participants returned between 8 and 17 correct answers. This implies that the z-score ≤ 0.73 for all participants, equivalent to a one-tailed p-value $= 23.27\%$. As the resulting p-values for all participants are fairly large, we conclude that it is very likely that all participants *cannot* tell the difference between identical and adversarial pairs, and thus the adversarial perturbations are *likely imperceptible* under the statistical sense.

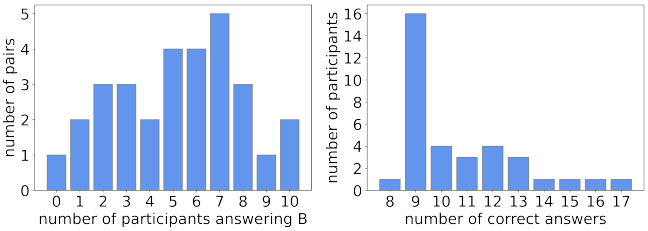

Fig. 3. [Left] the figure shows the number of pairs in terms of the numbers of B received, *e.g.* the 3^{rd} bar indicates there are 3 pairs with exactly 2 participants answered *B*. [Right] it shows the number of pairs in terms of the numbers of correct answers in each questionnaire. *e.g.*, the last bar shows there is only one participant who returned 17 correct answers.

4. CONCLUSIONS

In this work, we show that deep learning based speech quality predictors may be unstable under adversarial perturbations, where we use DNSMOS P.835 to demonstrate such vulnerability exists and may result in unreasonable quality ratings. This further suggests that a network predictor to apply to downstream tasks should be carefully examined. The study contributes to this matter further as we explore the possibility to strengthen network robustness by adversarial training. Our preliminary result on DNSMOS verifies the approach is effective for speech quality predictors and promising for future investigation.

5. REFERENCES

[1] ITUT Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[2] ITUT Rec, "P. 808, subjective evaluation of speech quality with a crowdsourcing approach," *ITU-T, Geneva*, 2018.

[3] P ITU-T, "835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," *ITU-T recommendation*, 2003.

[4] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[5] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[6] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "Mosnet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[7] Ryandhimas E. Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–17, 2022.

[8] A. Chehadi G. Mittag, B. Naderi and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *Proc. Interspeech 2021*.

[9] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP 2022*.

[10] Anderson R Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. ICASSP 2019*.

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

[12] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Pro. ICML*. PMLR, 2019.

[13] Hiromu Yakura and Jun Sakuma, "Robust audio adversarial example for a physical attack," *IJCAI-19*, pp. 5334–5341, 7 2018.

[14] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *ICLR 2015*.

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial machine learning at scale," *ICLR 2017*.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *ICLR 2018*.

[18] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, et al., "Icassp 2022 deep noise suppression challenge," in *ICASSP 2022*.

[19] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *Interspeech 2020*.

[20] Christophe Veaux, Junichi Yamagishi, and Simon King, "The Voice Bank Corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. O-COCOSDA 2013*.

[21] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA 2013*. Acoustical Society of America, 2013, vol. 19, p. 035081.

[22] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*.