Interspeech 2024
1-5 September 2024, Kos, Greece

Adversarial Robustness Analysis in Automatic Pathological Speech Detection Approaches

Mahdi Amiri[1,2], *Ina Kodrasi*[1]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
{mahdi.amiri,ina.kodrasi}@idiap.ch

Abstract

Automatic pathological speech detection relies on deep learning (DL), showing promising performance for various pathologies. Despite the critical importance of robustness in healthcare applications like pathological speech detection, the sensitivity of DL-based pathological speech detection approaches to adversarial attacks remains unexplored. This paper explores the impact of acoustically imperceptible adversarial perturbations on DL-based pathological speech detection. Imperceptibility of perturbations, generated using the projected gradient descent algorithm, is evaluated using speech enhancement metrics. Results reveal a high vulnerability of DL-based pathological speech detection to adversarial perturbations, with adversarial training ineffective in enhancing robustness. Analysis of the perturbations provide insights into the speech components that the approaches attend to. These findings highlight the need for research in robust pathological speech detection.
Index Terms: deep learning, pathological speech detection, adversarial attacks, robustness

1. Introduction

The increasing aging population has led to a rise in pathological speech conditions such as dysarthria or apraxia of speech. These conditions are associated with neurological disorders such as Parkinson's disease, Amyotrophic Lateral Sclerosis, or stroke. Diagnosing pathological speech in clinical practice involves time-consuming and expensive auditory-perceptual assessments by speech and language pathologists. To help alleviate this burden on the healthcare system, efforts within the research community are focused on developing automatic pathological speech detection approaches. Previous research relied on handcrafted acoustic features combined with classical machine learning algorithms for this task [1–5]. With the emergence of deep learning (DL) and its success in many fields [6,7], there is now a substantial number of studies aiming to leverage DL for automatic pathological speech detection [8–15]. These methods accept various input speech representations and utilize different architectures.

Despite the remarkable success of DL algorithms, they are vulnerable to adversarial attacks where crafted imperceptible perturbations can mislead a network into making incorrect predictions [16]. Significant research has been conducted to develop effective adversarial attacks and defense algorithms, with adversarial training being among the first successful defense strategies [17]. Adversarial robustness has been extensively studied in computer vision [18], yet it has received considerably less attention in the realm of audio applications. In the audio domain, adversarial robustness has primarily been explored in the context of audio classification tasks [19–22], with limited research focusing on the adversarial robustness of speech recognition systems [23, 24]. Although the robustness of healthcare applications such as pathological speech detection is of crucial importance, to the best of our knowledge, the adversarial robustness of DL-based pathological speech detection approaches has never been investigated.

The objective of this paper is to investigate adversarial robustness of automatic DL-based pathological speech detection approaches. Two exemplary approaches are considered, i.e., the convolutional neural network (CNN)-based approach proposed in [8] and the wav2vec2-based approach in [10]. Adversarial attacks are generated using the projected gradient descent (PGD) algorithm [17], with the broadband signal-to-noise-ratio (SNR) used to bound perturbations as in [22].

Results reveal that although the considered approaches yield a reasonable performance for clean speech samples, their performance is significantly affected by adversarial perturbations, even at high SNR budgets. Because the broadband SNR may not align well with how perturbations are subjectively perceived, we propose to use speech enhancement metrics to evaluate the quality of the perturbed samples. The used metrics are the short-time objective intelligibility (STOI) measure [25], perceptual evaluation of speech quality (PESQ) [26], frequency-weighted segmental SNR (fwSSNR) [27], and cepstral distance (CD) [28]. All metrics confirm that at high SNR budgets, perturbations are imperceptible, while the performance of the considered approaches is significantly affected. Analysis of the perturbations provide insights into the speech components that the approaches attend to. In addition, we show that although adversarial training has been successful in improving the robustness of DL-based classifiers [17], its effectiveness at improving the robustness of DL-based automatic pathological speech detection approaches is limited. Finally, we show that DL-based automatic pathological speech detection approaches are highly vulnerable to adversarial attacks even when noisy data augmentation is used for training. These results highlight a crucial need for developing robust automatic pathological speech detection approaches if they are to be deployed in practical clinical settings.

2. DL-Based Pathological Speech Detection

DL-based automatic pathological speech detection approaches can be broadly grouped into two categories, i.e., i) architectures operating on time-frequency input representations and ii) architectures operating on embeddings learned in a self-supervised manner. Approaches in the first category exploit architectures such as CNNs [8], recurrent neural networks [14], or autoencoders [11], to learn pathology-discriminant features from input representations such as the short-time Fourier trans-

10.21437/Interspeech.2024-1075

form (STFT) [8], Mel frequency cepstral coefficients [12, 13], or Mel spectrograms [11]. Approaches in the second category generally exploit linear layers to learn pathology-discriminant features from transformer-based feature extractors such as wav2vec2 [10]. To investigate adversarial robustness in automatic pathological speech detection, we consider one exemplary approach from each of the previously described categories, i.e., the CNN-based approach operating on STFT input representations from [8] and the wav2vec2-based approach from [10].

CNN-based approach. Since the CNN-based approach accepts only fixed-size inputs, we consider fixed-size segments of speech and compute their STFT. After computing the logarithm of the magnitude of the STFT coefficients, these representations are passed through a normalization layer to set their mean and standard deviation to 0 and 1 respectively. They are then encoded through two convolutional layers with 64 channels, having 2×2 and 3×3 kernel respectively. Each convolutional layer is followed by a ReLU activation function, batch normalization, and max-pooling with a 2×2 kernel. The second convolutional layer is followed by a dropout module ($p = 0.5$). After the dropout module, a linear layer (input size: 13376, output size: 2) is used for classification.

wav2vec2-based approach. wav2vec2-based approach accepts variable length audio as their inputs. Therefore, we consider full utterances as input to the wav2vec2 base model [29] and obtain their embeddings.

The average of the embeddings accross time after a user-selected transformer layer (cf. Section 4.3) is then obtained for each utterance and used as input to a model with two linear layers (layer 1 - input size: 768, output size: 256; layer 2 - input size: 256, output size: 2) for classification. A dropout module ($p = 0.3$) is also used after the first linear layer. It should be noted that the wav2vec2 base model is frozen and not trained/fine-tuned, with only the linear layers trained for pathological speech detection.

3. Adversarial Attacks

The objective of adversarial attacks is to obtain imperceptible perturbations which mislead the network. Such perturbations are found through maximizing the loss of the network's output for a given input. To generate imperceptible perturbations, we bound their norm to be below an upper bound ϵ. The optimization problem we solve to obtain these perturbations is

$$\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f(x+\delta;\theta), y)), \tag{1}$$

with x and y denoting the input (i.e., raw waveform) and its corresponding label (i.e., neurotypical or pathological), $f(\cdot;\theta)$ denoting the neural network with parameters θ, $\mathcal{L}$ denoting the loss function, δ denoting the perturbations, $\|\cdot\|_p$ denoting the p-norm (with $p = 2$ or $p = \infty$ commonly used).

There are two different settings to solve (1): i) the white-box setting where the adversary has access to the model's architecture and parameters and ii) the black-box setting where the adversary has access only to the network's output. Since the adversary has full access and knowledge of the model in the white-box setting, it can attack the network more effectively. In this paper, we consider the white-box setting and solve (1) iteratively using the PGD algorithm [17].

To obtain the p-norm bound, we bound perturbations through the broadband SNR, which is a more natural measure of perturbations for audio data [22]. The broadband SNR is given by

$$\text{SNR}(\delta) = 20\log\frac{\sqrt{\sum_{n=0}^{N-1} x_n^2}}{\sqrt{\sum_{n=0}^{N-1} \delta_n^2}}, \tag{2}$$

with $\{\cdot\}_n$ denoting the signal sample at time index n and N denoting the total number of samples in the waveform. As shown in [22], using an ∞-norm bound with $\epsilon = \frac{1}{N}10^{-\frac{\alpha}{20}}$ guarantees a minimum of SNR of α. Using such an ϵ, one can generate adversarial attacks for a given SNR threshold. However, it is known that the broadband SNR in (2) does not necessarily correlate with how perturbations are subjectively perceived. To resolve this issue, we propose to evaluate the perturbed samples for a given SNR bound through commonly used speech enhancement metrics such as STOI [25], PESQ [26], fwSSNR [27], and CD [28].

4. Experimental Settings

In this section we describe the experimental settings used for our analysis.

4.1. Database

We use Spanish recordings from gender-balanced groups of 50 patients diagnosed with Parkinson's disease and of 50 neurotypical speakers from the PC-GITA database [30]. Each speaker utters 10 sentences and a phonetically-balanced text recorded at a sampling frequency of 44.1 kHz. Recordings are downsampled to 16 kHz. The average length of the considered speech material for each speaker is 55.4 s.

4.2. Input representations

CNN-based approach. Available utterances are divided into 500 ms segments with an overlap of 250 ms. The STFT of these segments is computed using a weighted overlapp-add framework with a Hanning analysis window without overlap and a frame size of 10 ms.

wav2vec2-based approach. As described in Section 2, full utterances are used as input for the wav2vec2-based approach.

4.3. Evaluation and training

Evaluation is done in a speaker-independent stratified 10-fold cross-validation framework. At each fold, 80%, 10%, and 10% of the data is used for training, validation, and testing, respectively. To account for the effect of random initialization of the models when training, we train each model with 5 different seeds and report the average and standard deviation of the performance across these different seeds. The performance is evaluated in terms of speaker-level accuracy, which is computed through soft voting of the probability of decisions for all segments/utterances belonging to each speaker.

CNN-based approach. The CNN-based approach is trained using the stochastic gradient decent optimizer with a learning rate of 0.001 and a weight decay of 5×10^{-4}. We use a learning rate scheduler, such that if the validation loss does not decrease for 5 consecutive epochs, the learning rate is decreased by a factor of 0.5. Training stops if the learning rate decreases beyond 10^{-4} times the initial learning rate or if the maximum number of epochs is reached, which is set to 100 in our experiments.

wav2vec2-based approach. For the wav2vec2-based approach, we freeze the wav2vec2 model and train the linear layers. Training is done using the Adam optimizer with a learning

rate of 0.1 and a weight decay of 5×10^{-4}. The previously described learning rate scheduler is also used for the wav2vec2-based approach. It should be noted that the linear layers are trained using embeddings from the 10th layer of the wav2vec2 model. This decision was made based on our initial investigations, which revealed that these embeddings result in the best performance for pathological speech detection on clean samples (i.e., without adversarial perturbations). Nevertheless, the derived conclusions regarding the adversarial robustness when using wav2vec2 embeddings are applicable to embeddings from any of the wav2vec2 layers.

Adversarial perturbations. For generating adversarial perturbations, we use the PGD algorithm with 10 iterations (PGD-10) without random initialization and the same settings as in [22]. The ∞-norm is used to bound the perturbations based on the broadband SNR bound. The considered SNR bounds are $\text{SNR} = \{0\text{ dB}, 10\text{ dB}, \ldots, 100\text{ dB}\}$.

5. Experimental Results

In this section, results and insights are provided to investigate adversarial robustness of the considered DL-based automatic pathological speech detection approaches.

5.1. Adversarial robustness of approaches trained on clean speech samples

In the following, the performance of the considered approaches trained on clean data is analyzed in the presence of adversarial perturbations in the test data for different SNR budgets. The quality of the perturbations is evaluated through different speech enhancement metrics. For reference, Table 1 presents the accuracy of the considered approaches trained and tested on clean data, as is typically done in the state-of-the-art literature. It can be observed that both approaches result in an advantageous performance when the test data are not perturbed, with the wav2vec2-based approach outperforming the CNN-based approach.

Figure 1 presents the accuracy of the considered approaches (trained on clean data) in the presence of adversarial perturbations in the test data for different SNR budgets. For ease of comparison, the accuracy of the considered approaches for clean test data is also presented (i.e., the same results as in Table 1). The shaded area illustrates the standard deviation of the performance across the different seeds when training. As can be observed, the presence of adversarial perturbations at test time highly affects the performance of both approaches. Although the wav2vec2-based approach is more resilient to adversarial attacks than the CNN-based approach, both approaches fail already at very high SNR budgets of 60 dB, with the accuracy being below chance-level.

Since one cannot directly evaluate whether adversarial perturbations at a given broadband SNR are imperceptible, we evaluate the quality of the perturbed test signals with various speech enhancement metrics. Figure 2 presents the obtained

Table 1: *Speaker-level accuracy of the considered approaches when training and testing on clean data.*

Approach	Accuracy (%)
CNN-based	$76.6 \pm 2.57\%$
wav2vec2-based	$82.6 \pm 3.00\%$

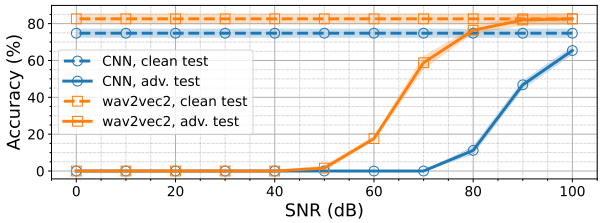

Figure 1: *Speaker-level accuracy of the considered approaches (trained on clean samples) for clean and adversarially perturbed test samples with different SNR bounds.*

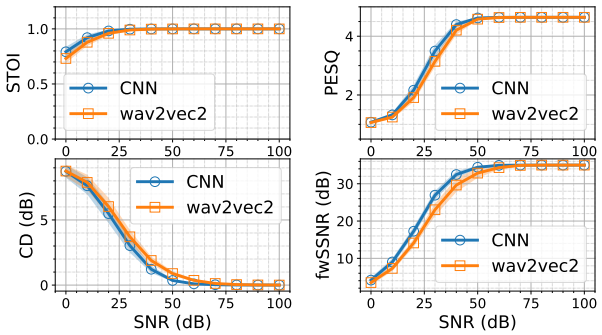

Figure 2: *STOI, PESQ, CD, and fwSSNR values of the adversarially perturbed test samples with different SNR bounds.*

STOI, PESQ, CD, and fwSSNR values of the perturbed test signals using their clean counterparts as reference. It should be noted that in the used implementation of these metrics[1], fully imperceptibe differences between the perturbed and clean samples result in a STOI, PESQ, CD, and fwSSNR value of 1, 4.6, 0 dB, and 35 dB, respectively. All metrics show that the generated adversarial perturbations are imperceptible at an SNR budget of 60 dB or higher. As previously discussed, the accuracy of the considered approaches at the SNR budget of 60 dB is below chance-level, confirming that DL-based automatic pathological speech detection approaches are highly vulnerable to imperceptible adversarial attacks.

5.2. Perturbation analysis

To gain further insight on the vulnerability of automatic pathological speech detection approaches to adversarial attacks, in this section we analyse the generated adversarial perturbations. To this end, we compute the average power spectral density (PSD) of the generated adversarial perturbations for the exemplary SNR budget of 60 dB. It should be noted that an iterative gradient-based algorithm is used to generate adversarial perturbations by solving (1). Hence, the PSDs of the generated perturbations provide insights into the frequency components that the approaches are attending to.

Figure 3 depicts the average PSD of the adversarial perturbations for both considered approaches. For reference, we also present the average PSD of the clean samples. The PSD of the adversarial perturbations for both approaches exhibits higher values at higher frequencies compared to lower frequencies, despite speech signals containing more information in the lower frequency components. This analysis suggests that the consid-

[1] https://github.com/schmiph2/pysepm

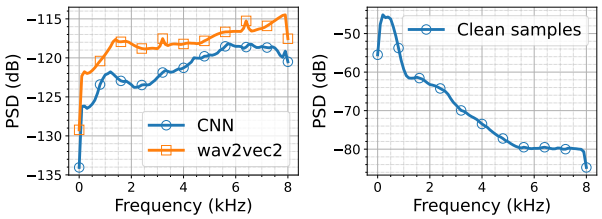

Figure 3: *PSD of the generated adversarial perturbations and of the clean samples for an exemplary SNR budget of* 60 *dB.*

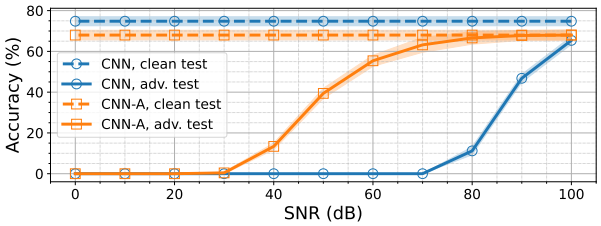

Figure 4: *Clean and adversarial accuracies for the CNN-based approach: model trained on clean samples (CNN) and adversarially trained model with an SNR budget of* 60 *dB (CNN-A).*

ered approaches are highly susceptible to input changes in the higher frequency components, indicating their lack of robustness and tendency to learn irrelevant features from the higher frequency components of speech. Future research will focus on methods to mitigate this sensitivity to high frequency components.

5.3. Adversarial training of the CNN-based approach

As mentioned in Section 1, adversarial training is one of the most effective defense strategies for adversarial robustness [17]. In this section, we investigate whether adversarial training is effective at increasing the adversarial robustness of pathological speech detection. As adversarial training of the wav2vec2-based approach demands computational resources beyond our current capabilities, the results presented in the following are focused on the CNN-based approach.

For adversarial training, we use an exemplary SNR budget of 60 dB. Perturbations generated for this SNR budget are then incorporated into the training of the CNN-based approach. Once the model is trained, we attack it and evaluate its performance on adversarially perturbed test data. Figure 4 depicts the performance of the adversarially trained model (denoted by CNN-A) for newly generated perturbations with different SNR bounds. The performance on clean test samples is also depicted. For ease of comparison, the performance of the non-adversarially trained CNN-based approach (denoted by CNN) from Section 5.1 is presented again. It can be observed that the performance of CNN-A on clean samples is lower than the performance of CNN. This is to be expected, since adversarially perturbed samples are used for training CNN-A. Most importantly, it can be observed that although the performance of CNN-A is overall higher than the performance of CNN in the presence of adversarial perturbations, the accuracy remains nevertheless very low. These results demonstrate that adversarial training is not sufficient on its own at increasing the robustness of pathological speech detection. This could be attributed to label noise in the data, the use of inadequate or underpowered models, and reliance on the typically small datasets available in this field.

5.4. Adversarial robustness of the CNN-based approach trained on noisy speech samples

Data augmentation has been shown to benefit generalization of DL-based classifiers. In this section, we analyse the adversarial robustness of pathological speech detection when training data is augmented with noisy speech samples, as opposed to Section 5.1 where clean speech samples are used for training. Similarly to Section 5.3, these analyses are presented only for the CNN-based approach.

To introduce noise in the training data, we use 6 different noise types (cf. Table 2) from the QUT-Noise database [31]. To examine whether the noise type used for augmenting the data affects the adversarial robustness, the CNN-based approach is trained on speech samples augmented with each individual noise type at an exemplary broadband SNR of 15 dB. Adversarial perturbations for these noisy samples are then generated using an exemplary SNR budget of 60 dB, which is 4 times higher than the SNR of the noisy samples, resulting in perturbations that are fully imperceptible.

Table 2: *Speaker-level accuracy of the CNN-based approach (trained on noisy samples) for noisy and adversarially perturbed noisy test samples with an SNR budget of* 60 *dB.*

Noise type	Noisy accuracy	Adversarial accuracy
CAFE	$67.2 \pm 5.0\%$	$1.8 \pm 1.2\%$
FOODCOURT	$65.2 \pm 3.2\%$	$2.0 \pm 1.7\%$
KITCHEN	$67.4 \pm 3.1\%$	$3.4 \pm 1.0\%$
LIVINGB	$70.6 \pm 2.7\%$	$0.2 \pm 0.4\%$
CITY	$70.0 \pm 3.3\%$	$1.4 \pm 2.6\%$
KG	$67.2 \pm 4.7\%$	$0.8 \pm 0.7\%$

Table 2 presents the accuracy obtained by the CNN-based approach (trained on noisy samples) for both noisy test samples and adversarially perturbed noisy test samples. It can be observed that the performance on noisy test samples considerably decreases when compared to the performance on clean test samples (cf. Table 1), although the same noise type and SNR is used for training and testing. Most importantly, it can be observed that independently of the noise type used for augmenting the data, the adversarial accuracy is remarkably low. These results indicate that the pathology-discriminant features that the CNN-based approach learns even in the presence of noise augmentation are highly susceptible to small changes in the input.

5.5. Conclusion

In this paper we have investigated adversarial robustness of two state-of-the-art pathological speech detection approaches, i.e., the CNN-based and wav2vec2-based approaches. Perturbations, generated using the PGD algorithm, have been evaluated using speech enhancement metrics. Results have shown a high vulnerability of these approaches to imperceptible perturbations, with no considerable improvement achieved through adversarial training. Analysis of the perturbations suggested that these approaches may be learning irrelevant features from high frequency components of speech. These findings emphasize the critical need for robust automatic pathological speech detection methods suitable for clinical deployment.

6. Acknowledgments

This work was supported by the Swiss National Science Foundation project no CRSII5_202228 on "Characterisation of motor speech disorders and processes".

7. References

[1] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, Nov. 2016.

[2] I. Kodrasi, M. Pernon, M. Laganaro, and H. Bourlard, "Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features." in *Proc. Annual Conference of the International Speech Communication Association*, Oct. 2020, pp. 4991–4995.

[3] I. Kodrasi and H. Bourlard, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 6, pp. 1210–1222, June 2020.

[4] N. P. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. Annual Conference of the International Speech Communication Association*, Hyederabad, India, Sept. 2018, pp. 3403–3407.

[5] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based learning for automatic dysarthric speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 96–100, Jan. 2020.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, Nevada, USA, Dec. 2012, pp. 1097–1105.

[7] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 6645–6649.

[8] P. Janbakhshi and I. Kodrasi, "Experimental investigation on STFT phase representations for deep learning-based dysarthric speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, USA, May 2022, pp. 6477–6481.

[9] ——, "Supervised speech representation learning for Parkinson's disease classification," in *Proc. ITG Conference on Speech Communication*, Kiel, Germany, Sept. 2021, pp. 154–158.

[10] D. Wagner, I. Baumann, F. Braun, S. P. Bayerl, E. Nöth, K. Riedhammer, and T. Bocklet, "Multi-class detection of pathological speech with latent features: How does it perform on unseen data?" in *Proc. Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2023, pp. 2318–2322.

[11] P. Janbakhshi and I. Kodrasi, "Adversarial-free speaker identity-invariant representation learning for automatic dysarthric speech classification," in *Proc. Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sept. 2022, pp. 2138–2142.

[12] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, Jan. 2016.

[13] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023.

[14] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019, pp. 5831–5835.

[15] D. Escobar-Grisales, T. Arias-Vergara, C. D. Ríos-Urrego, E. Nöth, A. M. García, and J. R. Orozco-Arroyave, "An automatic multimodal approach to analyze linguistic and acoustic cues on Parkinson's disease patients," in *Proc. Annual Conference of the International Speech Communication Association*, Dublin, Ireland, Sept. 2023, pp. 1703–1707.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. International Conference on Learning Representations*, Banff, Canada, Apr. 2014.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. International Conference on Learning Representations*, Vancouver, Canada, May 2018.

[18] R. Rade and S. Moosavi-Dezfooli, "Reducing excessive margin to achieve a better accuracy vs. robustness trade-off," in *Proc. International Conference on Learning Representations*, Virtual, Apr. 2022.

[19] V. Subramanian, E. Benetos, and M. B. Sandler, "Robustness of adversarial attacks in sound event classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events*, New York, USA, Oct. 2019, pp. 239–243.

[20] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "A robust approach for securing audio classification against adversarial attacks," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 4, pp. 2147–2159, Apr. 2020.

[21] R. Sallo, M. Esmaeilpour, and P. Cardinal, "Adversarially training for audio classifiers," in *Proc. International Conference on Pattern Recognition*, California, USA, Jan. 2021, pp. 9569–9576.

[22] K. Lu, M. C. Nguyen, X. Xu, and C. S. Foo, "On adversarial robustness of audio classifiers," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, May 2023.

[23] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," in *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sept. 2018, pp. 2404–2408.

[24] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, July 2011.

[26] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Dec. 2008.

[28] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. New Jersey, USA: Prentice-Hall, 1988.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annual Conference on Neural Information Processing Systems*, Virtual, Dec. 2020, pp. 12 449–12 460.

[30] J. R. Orozco, J. D. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Noeth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May. 2014, pp. 342–347.

[31] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms," in *Proc. Annual Conference of the International Speech Communication Association*, Makuhari, Japan, Sept. 2010, pp. 3110–3113.