SUPPRESSING NOISE DISPARITY IN TRAINING DATA FOR AUTOMATIC PATHOLOGICAL SPEECH DETECTION

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

Although automatic pathological speech detection approaches show promising results when clean recordings are available, they are vulnerable to additive noise. Recently it has been shown that databases commonly used to develop and evaluate such approaches are noisy, with the noise characteristics between healthy and pathological recordings being different. Consequently, automatic approaches trained on these databases often learn to discriminate noise rather than speech pathology. This paper introduces a method to mitigate this noise disparity in training data. Using noise estimates from recordings from one group of speakers to augment recordings from the other group, the noise characteristics become consistent across all recordings. Experimental results demonstrate the efficacy of this approach in mitigating noise disparity in training data, thereby enabling automatic pathological speech detection to focus on pathology-discriminant cues rather than noise-discriminant ones.

***Index Terms*—** pathological speech detection, noise disparity, data augmentation, TORGO, UA-Speech

1. INTRODUCTION

Early identification of pathological speech conditions such as dysarthria or apraxia of speech may serve as an early indicator of neurological disorders like Parkinson's disease, highlighting the critical need for swift diagnosis of pathological speech. Traditionally, speech and language pathologists rely on auditory-perceptual tests to identify pathological speech. However, these tests are costly, subjective, and time-consuming. As a result, there is a growing interest in the research community to explore methods for automatically detecting pathological speech. Previously, researchers predominantly relied on handcrafted acoustic features coupled with traditional machine learning methods to detect pathological speech [1–5]. More recently, the success of deep learning algorithms in many fields has prompted the adoption of deep learning approaches for automatic pathological speech detection [6–12]. These approaches leverage a variety of speech representations, such as e.g., the short-time Fourier transform (STFT) [6], Mel frequency cepstral coefficients [10, 11], Mel spectrograms [9], or self-supervised embeddings like wav2vec2 [8], integrated with diverse architectures to achieve automatic pathological speech detection. For example, the approach proposed in [6] has shown promising performance using a convolutional neural network (CNN) to extract pathology-discriminant cues from STFT input representations. Similarly, in [8], a high performance is achieved using linear layers to extract pathology-discriminant cues from wav2vec2 embeddings.

Despite the reported success of automatic pathological speech detection approaches, state-of-the-art literature typically assumes that recordings from both healthy and pathological speakers are acquired in identical, noise-free environments using the same recording setup. However, obtaining high-quality, clean recordings from pathological speakers, such as e.g., by recording all speakers in an anechoic chamber, poses a challenge. In [11] it has been shown that widely used databases in the research community, such as UA-Speech [13] and TORGO [14], are noisy, with the noise characteristics in healthy recordings being distinctly different from the noise characteristics in pathological recordings. Given the disparity in noise characteristics between these speaker groups, classifiers trained on such data for pathological speech detection capture noise-discriminant cues rather than pathology-discriminant ones [11]. Consequently, the assessment of pathological speech detection approaches using such databases remains inconclusive.

Although the robustness to noise has been extensively investigated in many speech applications such as automatic speech recognition [15], audio event detection [16], or speaker identification [17], the robustness to noise of pathological speech detection approaches has received significantly less attention, particularly for recordings with noise disparity among the two groups of speakers. In [18], it has been proposed to use a single-channel speech enhancement module

This work was supported by the Swiss National Science Foundation project CRSII5_202228 on "Characterisation of motor speech disorders and processes"

prior to training a pathological speech detection approach. However, employing traditional single-channel speech enhancement on recordings with varying noise characteristics leads to signal distortions dependent on these characteristics. Hence, instead of capturing pathology-discriminant cues, automatic pathological speech detection models trained on enhanced recordings will learn cues associated with the introduced distortions.

To address the challenge posed by noise disparity in training data for pathological speech detection approaches, one can either develop approaches that are robust to noise or suppress this noise disparity. In this paper, we propose a method to suppress the noise disparity in training data through noise augmentation. Given the available noisy recordings, we suggest using a voice activity detection (VAD) module [19] to extract an estimate of the noise present in each recording. Subsequently, the estimated noise signals from recordings of one speaker group are incorporated into the recordings of the other group, with appropriate scaling factors computed to maintain consistent signal-to-noise ratios (SNRs). This procedure ensures that the noise characteristics (in terms of the noise type and SNR) are similar between the two speaker groups. Therefore, with noise-discriminant cues suppressed in the data, automatic pathological speech detection approaches prioritize learning pathology-discriminant cues over noise-discrminant ones.

2. PROBLEM STATEMENT AND PROPOSED METHOD

We consider a noisy signal y at time index k given by

$$y_k = s_k + n_k, \tag{1}$$

with s denoting the clean speech signal and n denoting the noise signal. The power of the noisy signal y can be computed as

$$P_y = \frac{1}{K}\sum_k y_k^2, \tag{2}$$

with K denoting the signal length. Further, we define the SNR of the noisy signal y with respect to the noise n as

$$\text{SNR}_y^n = 20\log_{10}\frac{\sqrt{P_s}}{\sqrt{P_n}}, \tag{3}$$

with P_s and P_n being the clean speech and noise powers defined similarly to (2). For conciseness, the time index k is omitted in the remainder of this section unless explicitly required.

Without loss of generality, we consider one noisy utterance y_h from the healthy group and one noisy utterance y_p from the pathological group. These utterances (of possibly different length) are given by

$$y_h = s_h + n_h, \quad y_p = s_p + n_p, \tag{4}$$

with s_h and s_p denoting the respective clean signals and n_h and n_p denoting the respective noise signals. Since $n_h \neq n_p$ and $\text{SNR}_{y_h}^{n_h} \neq \text{SNR}_{y_p}^{n_p}$, a pathological speech detection classifier trained using y_h and y_p can easily learn the noise differences between the two signals rather than pathology-discriminant cues in s_h and s_p. To suppress the noise disparity between the two utterances, we propose to use noise augmentation and add (an estimate of) the noise present in one utterance to the other utterance as

$$\hat{y}_h = \underbrace{s_h + n_h}_{y_h} + \alpha_h \hat{n}_p, \quad \hat{y}_p = \underbrace{s_p + n_p}_{y_p} + \alpha_p \hat{n}_h, \tag{5}$$

with $\hat{n}_p$ an estimate of n_p, $\hat{n}_h$ an estimate of n_h, and the scalars α_h and α_p computed as described in the following. Although both noises are present in both signals in (5), there remains a noise disparity between $\hat{y}_h$ and $\hat{y}_p$ since their SNRs might still differ. For $\hat{y}_h$ and $\hat{y}_p$ to contain similar noise characteristics such that a network trained on these utterances fails to learn noise-discriminant cues, it is required that

- the SNR of both noisy signals with respect to each of the noises is the same, and
- the SNR of both noisy signals with respect to the total noise is the same.

More precisely, these conditions can be expressed as

$$\text{SNR}_{\hat{y}_h}^{n_h} = \text{SNR}_{\hat{y}_p}^{\alpha_p \hat{n}_h} \tag{6}$$

$$\text{SNR}_{\hat{y}_h}^{\alpha_h \hat{n}_p} = \text{SNR}_{\hat{y}_p}^{n_p} \tag{7}$$

$$\text{SNR}_{\hat{y}_h}^{n_h + \alpha_h \hat{n}_p} = \text{SNR}_{\hat{y}_p}^{n_p + \alpha_p \hat{n}_h}, \tag{8}$$

with the different SNRs computed similarly to (3). The system of equations in (6)-(8) is an over-determined system. However, if $\hat{n}_p$ and $\hat{n}_h$ are good estimates of n_p and n_h, one can assume that

$$P_{n_h} = P_{\hat{n}_h}, \quad P_{n_p} = P_{\hat{n}_p}, \tag{9}$$

with $P_{\{\cdot\}}$ denoting the power of the different noise signals defined similarly to (2). If the assumption in (9) holds, scalars α_h and α_p that satisfy (6)-(8) can be computed using the power of clean speech signals P_{s_h} and P_{s_p} as

$$\alpha_h = \sqrt{\frac{P_{s_h}}{P_{s_p}}}, \quad \alpha_p = \sqrt{\frac{P_{s_p}}{P_{s_h}}}. \tag{10}$$

Using these scalars in (5) and using $\hat{y}_h$ and $\hat{y}_p$ instead of y_h and y_p to train a pathological speech detection system results in a system that learns pathology-discriminant cues rather than noise-discriminant ones.

The proposed approach requires noise estimates $\hat{n}_p$ and $\hat{n}_h$ (cf. (5)) as well as estimates of the clean speech powers P_{s_p} and P_{s_h} (cf. (10)). To obtain these estimates, we propose to extract the noise-only segments from each utterance

using a VAD and assume that they yield a good approximation of the true noise. Such an assumption holds for relatively stationary noise, which is a reasonable assumption in clinical settings where these recordings are typically collected. The extracted noise-only segments are then concatenated and repeated as necessary to obtain noise signals $\hat{n}_p$ and $\hat{n}_h$ of appropriate lengths required in (5). Furthermore, assuming that the speech and noise signals are uncorrelated, the clean speech powers are computed as

$$P_{s_h} = P_{y_h} - P_{n_h}, \;\; P_{s_p} = P_{y_p} - P_{n_p}, \tag{11}$$

with P_{n_h} and P_{n_p} computed using the extracted noise-only segments. Although we have made several assumptions and approximations which may not hold in practice, experimental results in Section 4 show that these assumptions and approximations are effective in practice to (partially) suppress the noise disparity in the training data.

3. EXPERIMENTAL SETTINGS

For the experimental results, we adopt a methodology similar to that of [11]. We consider noisy healthy and pathological speech recordings, with the noise characteristics differing among the two groups. Further, we consider two state-of-the-art automatic pathological speech detection approaches, i.e., the CNN-based approach from [6, 20] and the wav2vec2-based approach from [8, 20]. To evaluate the impact of noise disparity on automatic pathological speech detection approaches, these approaches are trained using noisy utterances and their performance on noisy and clean test utterances is compared. To evaluate the impact of suppressing the noise disparity using the proposed method, these approaches are trained using augmented utterances and their performance on noisy and clean test utterances is compared. Obtaining a better classification performance on noisy test utterances than on clean test utterances confirms that noise-discriminant cues are being learnt instead of pathology-discriminant cues. Obtaining a similar classification performance on both noisy and clean test utterances confirms that pathology-discriminant cues are being learnt instead of noise-discriminant cues.

In the following, we describe the experimental settings used in our evaluation.

3.1. Databases

Since we do not have access to the clean speech signals in the UA-Speech and TORGO databases, we construct synthetic noisy databases for our analysis.

Clean speech. We consider clean recordings of healthy and pathological speakers from the PC-GITA database [21]. This database contains Spanish recordings from gender-balanced groups of 50 healthy speakers and 50 patients suffering from Parkinson's disease. Recordings are down-sampled to 16 kHz prior to using them for our experiments.

Noise. To generate noisy recordings, we augment the clean recordings with different noise types from the DEMAND database [22]. Three different noise types are used, i.e., D-type (DKITCHEN and DLIVING), N-type (NPARK and NRIVER), and O-type (OOFFICE and OMEETING). To simulate noise disparity in the healthy and pathological recordings, one group of speakers is augmented with one specific noise from a given type (e.g., DKITCHEN used for healthy speakers), whereas the other group of speakers is augmented with the other noise from the same type (e.g., DLIVING used for pathological speakers). For each of the three considered noise types, we generate data for three different SNR settings, i.e.,

A. healthy and pathological recordings have the same SNR of 20 dB,
B. healthy and pathological recordings have the same SNR of 40 dB, and
C. healthy recordings have an SNR of 20 dB whereas pathological recordings have an SNR of 40 dB.

The choice of such relatively high SNR values is done to reflect the (more recent) reality where care is taken to have better quality pathological recordings as well as to show that even an SNR of 40 dB is problematic for state-of-the-art pathological speech detection approaches.

3.2. Model Architecture

CNN-based approach. The CNN-based approach operates on fixed-size segments of speech. To compute inputs to this approach, we segment utterances into 500 ms long segments (with an overlap of 250 ms) and compute their STFT. The used STFT parameters and the architecture of the CNN-based approach is the same as in [20].

wav2vec2-based approach. Differently from the CNN-based approach, the wav2vec2-based approach operates on full utterances of variable length. As in [20], the wav2vec2-base model is frozen and used to extract features from full utterances. These features are then used as input to a linear layer (input size 768, output size 256). After a ReLU activation function, batch normalization, and dropout ($p = 0.3$), a final linear layer (input size 256, output size 2) is used for pathological speech detection.

3.3. Training and Evaluation

A stratified 10-fold cross validation framework is used to evaluate the performance of the proposed method, with no overlap between speakers in the training, validation, and test sets.

As in [20], the CNN-based approach is trained using the SGD optimizer with a learning rate of 0.001, whereas the wav2vec2-based approach is trained using the Adam optimizer with a learning rate of 0.01. For both approaches, we

Table 1. Mean and standard deviation of the classification accuracy (%) of the CNN-based and wav2vec2-based approaches using different training procedures for different SNR settings.

SNR SETTING	TEST	CNN-BASED			WAV2VEC2-BASED		
		STANDARD	ORACLE	PRACTICAL	STANDARD	ORACLE	PRACTICAL
A	NOISY	99.9 ± 0.1	70.5 ± 4.8	76.3 ± 3.2	98.1 ± 0.6	78.0 ± 1.7	86.3 ± 2.7
	CLEAN	54.1 ± 5.8	68.3 ± 1.7	67.4 ± 2.0	70.3 ± 4.6	77.5 ± 1.6	77.3 ± 1.2
B	NOISY	84.8 ± 1.7	75.9 ± 1.3	70.8 ± 0.8	82.9 ± 1.6	81.1 ± 0.9	82.5 ± 0.2
	CLEAN	77.2 ± 2.3	74.6 ± 1.1	69.7 ± 0.4	81.5 ± 1.6	81.2 ± 0.3	82.4 ± 0.6
C	NOISY	99.5 ± 0.5	71.5 ± 6.5	76.7 ± 3.8	95.5 ± 2.8	78.5 ± 2.0	86.2 ± 3.0
	CLEAN	51.3 ± 1.9	70.3 ± 1.9	68.2 ± 1.7	53.6 ± 5.9	80.0 ± 2.1	77.6 ± 5.0

use a weight decay of 5×10^{-4} and the learning rate scheduler *ReduceLROnPlateau* with $patience = 5$ and $factor = 0.5$. Training is stopped if the learning rate decreases beyond 10^{-4} of the initial learning rate or if the maximum number of epochs of 100 is reached.

To implement the method proposed in Section 2, noise is added to each utterance from each group of speakers according to (5). This noise is extracted from a randomly selected utterance from the other group of speakers. This procedure is repeated in each epoch.

The performance is evaluated using the speaker-level accuracy computed similarly as in [20]. To account for the effect of randomness in initializing networks, we analyze the mean performance for networks trained with 5 different seeds.

4. EXPERIMENTAL RESULTS

In this section, we present several results to validate the effectiveness of the proposed method. To demonstrate the validity of the proposed method and to decouple it from potential estimation errors in practice, we consider the oracle scenario where one has access to the noise signals required for augmentation in (5) and to the clean speech powers required for computing appropriate scalars in (10). To demonstrate the effectiveness of the proposed method in practice, we also consider the practical scenario where the different required quantities are estimated through the VAD as described in Section 2. The performance obtained when using the proposed method in these scenarios is compared to the performance when the standard training procedure is used, i.e., without accounting for the existing noise disparity in the training data. All results are given in Table 1, where the mean and standard deviation of the performance across the different considered noise types is presented.

Standard training. Table 1 shows that in all SNR settings, using standard training for the CNN-based approach yields a very high performance on noisy test utterances and a considerably lower performance on clean test utterances. Although the wav2vec2-based approach is expected to be more robust to noise, using standard training for this approach also results in a considerably higher performance on noisy test utterances than on clean test utterances for SNR settings A and C. These results confirm that in the presence of noise disparity in the data, these approaches learn noise-discriminant cues rather than pathology-discriminant cues.

Oracle scenario. Table 1 shows that when using the proposed method with oracle quantities, the performance of both approaches on noisy and clean test utterances is very similar for all SNR settings. These results confirm the validity of the proposed method, with noise augmentation being effective at suppressing noise disparity in the data and allowing pathological speech detection approaches to learn pathology-discriminant cues instead of noise-discriminant cues.

Practical scenario. Table 1 shows that in comparison to the standard training procedure, the proposed method with realistically estimated quantities increases the performance on clean test utterances and decreases the performance on noisy test utterances as desired. Nevertheless, there remains a gap between the effectiveness of the proposed method in practice in comparison to the oracle implementation for SNR settings A and C. Additional investigations confirm that the estimated clean speech powers are very similar to the oracle speech powers, and hence, this remaining performance gap can be attributed to the approximation of the noise by concatenating and repeating noise-only segments extracted through the VAD. In the future, alternative methods to extract the noise-only signals will be investigated.

5. CONCLUSION

In this paper we have proposed a method to suppress the noise disparity in training data for automatic pathological speech detection. The goal is to enable these approaches to prioritize learning pathology-discriminant cues over noise-discriminant ones. The proposed method involves extracting noise from recordings using a VAD and then augmenting recordings from one speaker group with noise extracted from another speaker group. Extensive experimental results have demonstrated the effectiveness of the proposed method.

6. 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*, Shanghai, China, 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] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. R. Gunderson, T. S. Huang, K. L. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sept. 2008, pp. 1741–1744.

[14] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[15] Q.-S. Zhu, L. Zhou, J. Zhang, S.-J. Liu, Y.-C. Hu, and L.-R. Dai, "Robust data2vec: Noise-robust speech representation learning for ASR by combining regression and improved contrastive learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, June 2023.

[16] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[17] M. MohammadAmini, D. Matrouf, J.-F. Bonatsre, S. Dowerah, R. Serizel, and D. Jouvet, "A comprehensive exploration of noise robustness and noise compensation in ResNet and tdnn-based speaker recognition systems," in *Proc. European Signal Processing Conference*, Belgrade, Serbia, 2022, pp. 364–368.

[18] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, and E. Nöth, "Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions," in *Proc. Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sept. 2015.

[19] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe," in *Proc. Annual Conference of the International Speech Communication Associa- tion*, Dublin, Ireland, Aug. 2023, pp. 1983–1987.

[20] M. Amiri and I. Kodrasi, "Test-time adaptation for automatic pathological speech detection in noisy environments," in *Proc. European Signal Processing Conference*, Lyon, France, Aug. 2024.

[21] 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.

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. International Congress on Acoustics*, Montreal, Canada, June 2013.