Test-Time Adaptation for Automatic Pathological Speech Detection in Noisy Environments

Mahdi Amiri
Signal Processing for Communication Group
Idiap Research Institute
Martigny, Switzerland
mahdi.amiri@idiap.ch

Ina Kodrasi
Signal Processing for Communication Group
Idiap Research Institute
Martigny, Switzerland
ina.kodrasi@idiap.ch

***Abstract*—Deep learning-based pathological speech detection approaches are gaining popularity as a diagnostic tool to support time-consuming and subjective clinical assessments. While these approaches perform well in controlled environments with clean recordings, their performance significantly degrades in realistic scenarios with background noise. In this paper, we propose a test-time adaptation framework to increase the robustness of such approaches to background noise during inference. To this end, we use a voice activity detector to extract noise-only segments from the test signal. These segments are used to augment a portion of the training/validation data, which is then exploited to fine-tune the classification models. Extensive experimental results demonstrate the effectiveness of the proposed framework in increasing robustness to noise for state-of-the-art automatic pathological speech detection approaches.**

***Index Terms*—pathological speech detection, robustness, noise, data augmentation, adaptation**

I. Introduction

Pathological speech can be caused by neurological damage from diseases such as Cerebral Palsy, Amyotrophic Lateral Sclerosis, or Parkinson's disease. Traditionally, speech and language pathologists diagnose speech disorders through costly and time-consuming auditory-perceptual assessments. To address this challenge, there is a growing interest in developing automatic pathological speech detection approaches to support clinical assessments. Previously, such approaches typically relied on classical machine learning classifiers and handcrafted acoustic features [1]–[5]. With the advent of deep learning (DL) algorithms and their remarkable success in various domains such as computer vision [6], natural language processing [7], and speech processing [8], recent automatic pathological speech detection approaches focus on exploiting DL. The vast majority of DL-based approaches aim to learn pathology-discriminant cues from time-frequency input representations such as the short-time Fourier transform (STFT) [9], Mel frequency cepstral coefficients [10], [11], or Mel spectrograms [12], using architectures such as convolutional neural networks (CNNs) [9], recurrent neural networks [13], or autoencoders [12]. With the promising performance achieved by using self-supervised embeddings from transformer-based models such as wav2vec2 [14], [15] for several downstream tasks [16], researchers have also used such embeddings as input representations for automatic pathological speech detection [17], [18].

Despite DL-based approaches exhibiting a promising performance for automatic pathological speech detection using recordings collected in clean acoustic environments, their performance significantly decreases in the presence of additive noise [19]. This limitation hinders the deployment of these approaches in realistic clinical settings. While the robustness to noise has been extensively investigated in many speech applications such as automatic speech recognition [20]–[22], audio event detection [23], or speaker identification [24], the robustness to noise of pathological speech detection approaches has received significantly less attention. The investigation in [25] explores the applicability of using pitch and Mel frequency cepstral coefficients (MFCCs) in detecting dysarthria in the presence of additive Gaussian noise. Results indicate that in this artificial Gaussian noise scenario, pitch demonstrates greater resilience compared to MFCCs. In [26], domain adversarial training is used to increase the robustness of pathological speech detection approaches in scenarios where training and testing recordings come from different devices. Results show that domain adversarial training is useful in addressing minor discrepancies between training and testing data that arise from different recording devices. In [27], it is proposed to use a single-channel speech enhancement module before extracting input features from signals recorded in non-controlled noisy environments. However, specific information regarding the types of noise present in the signals or their signal-to-noise ratios (SNRs) is not provided. Furthermore, it should be noted that conventional single-channel speech enhancement methods introduce distortions to the original signals, which is problematic in the context of pathological speech detection. This is especially true as these distortions can be mistaken as pathology-discriminant cues.

This paper proposes a test-time adaptation framework aimed at increasing the robustness of pathological speech detection approaches against noise present in test signals. The proposed framework involves initially extracting noise-only segments from the test recordings using a voice activity detector (VAD) module [28]. Following this, the pathological speech detection

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

model under consideration is fine-tuned on training/validation data augmented with the extracted noise-only segments and the test recording is processed using the adapted model. It should be noted that such a framework is general and applicable to any DL-based pathological speech detection approach. In this paper, we apply the proposed framework to increase robustness of the CNN-based approach from [9] and the wav2vec2-based approach from [17]. Extensive experimental results demonstrate the advantages of the proposed framework.

II. DL-Based Pathological Speech Detection

In this paper, we consider increasing the robustness to noise of the CNN-based approach operating on STFT input representations from [9] and of the wav2vec2-based approach from [17]. In the following, these approaches are briefly reviewed.

CNN-based approach. The CNN-based approach accepts fixed-size inputs, hence, we consider fixed-size segments of speech and compute their STFT. After calculating the logarithm of the magnitude of the STFT coefficients, these representations undergo normalization before being encoded using two convolutional layers. Each layer consists of 64 channels and uses kernels sized 2×2 and 3×3, respectively. Following each convolutional layer, there's a sequence of operations: ReLU activation function, batch normalization, and max-pooling with a kernel size of 2×2. Subsequently, the second convolutional layer is followed by a dropout module with a dropout probability of 0.5. After the dropout module, a linear layer is employed for pathological speech detection, with an input size of 13376 and an output size of 2.

wav2vec2-based approach. The wav2vec2-based approach accepts variable length audio as input. Therefore, we consider full utterances as input to the wav2vec2 base model [14] and obtain their embeddings. After extracting embedings from a user-selected transformer layer, we compute their mean across time. Subsequently, we utilize two linear layers (layer 1 - input size: 768, output size: 256; layer 2 - input size: 256, output size: 2) for the classification task. The first linear layer is also followed by a dropout module with a dropout probability of 0.5. 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.

III. Test-time Adaptation Framework

Fig. 1 presents a schematic illustration of the proposed test-time adaptation framework to enhance the robustness of pathological speech detection approaches to noise. As shown, we consider a pathological speech detection model trained on the available training utterances, with hyperparameters tuned on the available validation utterances. The training and validation utterances can either be clean or augmented with noise, with the latter resulting in a more robust initial model. Given a noisy test signal, we extract the noise-only segments using a VAD. These extracted noise-only segments are then concatenated and replicated as necessary to augment a subset of the utterances from the training and/or validation sets.

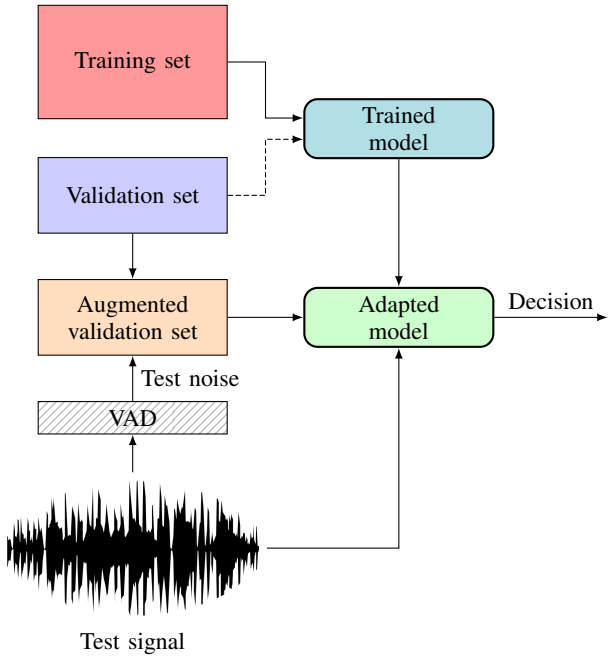

Fig. 1. Schematic illustration of the proposed test-time adaptation framework. During inference, we adapt the initial trained model for each specific test signal based on the extracted noise-only segments using a VAD.

Finally, the augmented utterances are used for fine-tuning the initial model, enhancing its robustness against the particular noise in the test signal. This model adaptation procedure is repeated for each test signal. In Section V and as depicted in Fig. 1, we augment the utterances from the validation set, such that the initial model is adapted with previously unseen data. It should be noted that during this adaptation procedure, the model learns solely from the noise characteristics extracted using the VAD, since we do not utilize labels from the test signal. The effectiveness of the proposed framework depends on the duration of the noise-only segments within the test signal, the accuracy of the VAD in isolating these segments, and the stationarity of the noise. In Section V we demonstrate that the proposed framework is highly effective for many noise types even with very short noise-only segments available, with the median length being 18.6 ms for our data and the minimum length being 8 ms. Additionally, our initial findings indicate that the overall performance of this framework is relatively unaffected by the specific VAD used. Therefore, for the results presented in Section V, we employ the VAD proposed in [28].

IV. Experimental Settings

A. Clean speech dataset

In this paper, we use the PC-GITA dataset [29] which contains clean Spanish recordings from a group of 50 patients diagnosed with Parkinson's disease and 50 neurotypical speakers. This dataset contains 25 male and 25 female speakers in each group. Each speaker is recorded with a sampling frequency of 44.1 kHz uttering 10 sentences and 1 phonetically-balanced text. Recordings are downsampled to 16 kHz prior to

using them for the considered approaches. The average length of all utterances combined for each speaker is 55.37 s.

B. Noise datasets and augmentation

To generate noisy utterances, we use the QUT-NOISE dataset (KITCHEN, LIVINGB, and CITY noises) [30] as well as the DEMAND dataset (DKITCHEN, NPARK, DLIVING, SPSQUARE, OMEETING, OOFICE, and PCAFETER noises) [31]. To generate a noisy utterance with a specific SNR, we select a random part from the noise of interest and add it to the clean utterance after scaling the noise with the coefficient α given by

$$\alpha = 10^{\frac{SNR}{20}} \times \frac{\sqrt{\sum_i x_i^2}}{\sqrt{\sum_i n_i^2}}, \tag{1}$$

where x_i denotes the i-the sample of the clean utterance and n_i denotes the i-th sample of the noise. For testing models trained on clean data, we use the QUT-NOISE dataset. To train initial robust models with noise augmentation, we use the QUT-NOISE dataset and SNR $\in \{5, 10, 15, 20\}$ dB. Since we need an additional dataset to evaluate the performance when models are trained with noise augmentation, we test these models using the DEMAND dataset, both on similar noise types as the training noises (i.e., DKITCHEN, DLIVLING, and SPSQUARE) and on different noise types from the training noises (i.e., NPARK, OMEETING, OOFICE, and PCAFETER). Please note that all tests are done for SNR $\in \{5.0, 7.5, 10.0, \ldots, 22.5\}$ dB.

C. Input Representation

As previously mentioned, the CNN-based approach accepts fixed-size segments of speech as input. For this purpose, we split each utterance into 500 ms segments with an overlap of 250 ms. The STFT of these segments is computed using a 10 ms Hanning window without overlap and the logarithm of the STFT magnitude is used as input representation.

Since transformer-based networks accept full utterances as input, full utterances are used as input to the wav2vec2-based approach. Initial investigations show that using embeddings from the 10th transformer layer of the wav2vec2 base model yields a better pathological speech detection performance than embeddings from other layers. Hence, for the results presented in the following, wav2vec2 embeddings are extracted from the 10th transformer layer.

D. Training

A 10-fold cross validation framework is used for training and evaluating the considered approaches. At each fold, we split the data into utterances from 80, 10, and 10 different speakers for training, validation, and testing, respectively. Each of the sets contains the same number of gender-matched neurotypical and pathological speakers. To train the initial CNN-based model, we use the stochastic gradient decent (SGD) optimizer with a learning rate of 0.001 and a weight decay of 5×10^{-4}. To train the inital wav2vec2-based model, we use the Adam optimiser with a learning rate of 0.1 and a weight decay of 5×10^{-4}. In addition, we use the *ReduceLROnPlateau* learning rate scheduler with $patience = 5$ and $factor = 0.5$ for all models. Training stops if the learning rate decreases below 10^{-4} of the initial value or the maximum number of epochs of 100 is reached.

Using this training procedure and the data outlined in Sections IV-A and IV-B, we train four initial models, i.e., the CNN-based model using clean data, the CNN-based model using noise augmented data, the wav2vec2-based model using clean data, and the wav2vec2-based model using noise augmented data. The models trained using noise augmented data are expected to be more robust to noise than the models trained using only clean data.

E. Adaptation

For adaptation, we fine-tune the previously trained models on the validation set augmented with the noise extracted from the test signal. It should be noted that adaptation is done for each individual test signal. The CNN-based models are adapted for 1 epoch using the SGD optimizer with a learning rate of 0.0001. The wav2ec2-based models are adapted for 1 epoch using the Adam optimiser with a learning rate of 0.01. A weight decay of 5×10^{-4} is used for all models. It should be noted that we did not optimize these hyperparameters (i.e., number of epochs, portion of the augmented data, learning rate, or weight decay) to achieve a better performance. Hence, while the proposed adaptation framework already demonstrates a considerable performance improvement (cf. Section V), one can optimise these hyperparameters to further improve the performance.

F. Performance

To assess the performance of the considered approaches, we compute speaker-level accuracy. Soft labels are first generated by passing the networks' output through a *softmax* function for all the segments/utterances belonging to each speaker. The final speaker classification decision is done through soft voting of all the segment-/utterance-level labels. To account for randomness when initializing the models during training, we train all models for 5 different random seeds. The reported speaker-accuracy values reflect the mean values obtained across these different seeds.

V. Experimental Results

In this section, we investigate the performance obtained using the initial models and the proposed adaptation framework for different scenarios. To investigate the effect of the length of the available noise-only segments on the proposed adaptation framework, we analyze two scenarios: i) the realistic scenario where the test signal is not altered and used as it is available to extract noise-only segments (yielding a median length of noisy-only segments of 18.6 ms across all signals) and ii) the non-realistic scenario where noise-only segments with a 1 s duration are appended at the beginning and end of the test signal and used for the adaptation.

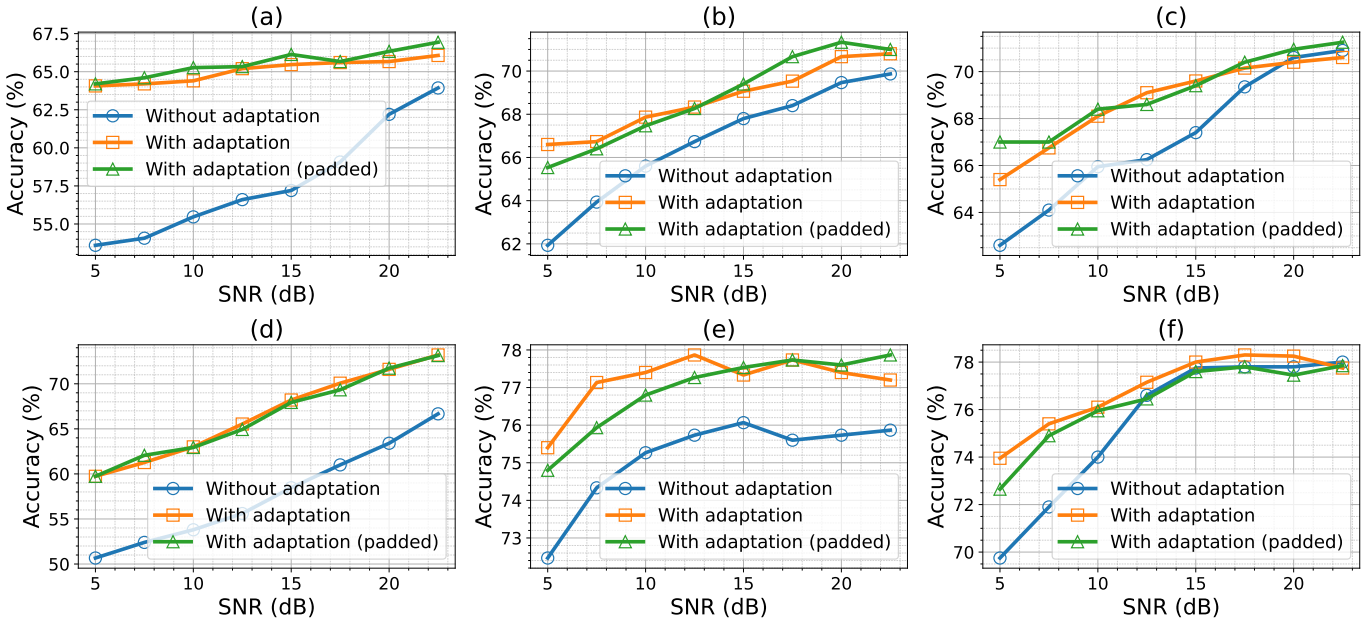

Fig. 2. Performance of the considered models without and with adaptation on noisy test signals: (a) CNN-based model trained on clean samples and tested on noisy samples (QUT-NOISE), (b) CNN-based model trained on noise augmented samples (QUT-NOISE) and tested on noisy samples with similar noise characteristics as the training samples (DEMAND), (c) CNN-based model trained on noise augmented samples (QUT-NOISE) and tested on noisy samples with different noise characteristics from training samples (DEMAND), (d) wav2vec2-based model trained on clean samples and tested on noisy samples (QUT-NOISE), (e) wav2vec2-based model trained on noise augmented samples (QUT-NOISE) and tested on noisy samples with similar noise characteristics as the training samples (DEMAND), (f) wav2vec2-based model trained on noise augmented samples (QUT-NOISE) and tested on noisy samples with different noise characteristics from training samples (DEMAND).

A. Performance for Initial Models Trained on Clean Data

In the following, we analyse the effectiveness of the proposed adaptation framework when the initial models are trained on clean data. The reported performance values are averaged across different seeds and different noise types present in the test data. Figs. 2(a) and (d) depicts the obtained performance values for different SNRs for the CNN-based and wav2vec2-based approaches, respectively. As illustrated, while the models trained on clean data have a poor performance when the testing data is noisy (particularly at low SNRs), the proposed adaptation framework improves the performance by a considerable margin. Further, it can be observed that using the available noise-only segments for the proposed adaptation framework yields a similar performance to using longer noise-only segments. This confirms the effectiveness of the proposed adaptation framework even when the available noise-only segments are short.

B. Performance for Initial Models Trained on Augmented Noisy Data

In the following, we analyse the effectiveness of the proposed adaptation framework when the initial models are trained on noisy data (and hence, are already more robust to noise in the test data). The reported performance values are averaged across different seeds and different noise types present in the test data. Figs. 2(b) and (e) depict the performance of the considered models when the noise type in the testing data is similar to the noise type in the training data (although from different databases), whereas Figs. 2(c) and (f) depict the performance when the noise type in the testing data is different from the noise type in the training data. As illustrated, while the models trained on noisy data have a better performance when the testing data is noisy (in comparison to models trained on clean data), the proposed adaptation framework still improves the performance further, particularly at low SNRs and regardless of the similarity of the noise types in the training and testing data. Similarly to before, it can be observed that using the available noise-only segments for the proposed adaptation framework yields a similar performance to using longer noise-only segments.

VI. Conclusion

In this paper, we have proposed a general adaptation framework to increase robustness to additive noise of DL-based pathological speech detection approaches. The proposed framework relies on extracting noise-only segments from the test signal using a VAD, and exploiting these segments to fine-tune models and increase their robustness to the specific noise present in the test signal. Although the proposed framework is simple, results demonstrate a considerable improvement in robsutness for pathological speech detection approaches. The proposed framework is applied to the pathological speech detection task, however, it is a general framework applicable to increase robustness to noise in other audio tasks.

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. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems*, California, USA, Dec. 2017, pp. 6000–6010.

[8] 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.

[9] 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.

[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] 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.

[13] 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.

[14] 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.

[15] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[16] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[17] 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.

[18] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[20] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, and L.-R. Dai, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3174–3178.

[21] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7097–7101.

[23] 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.

[24] 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 *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 364–368.

[25] T. Bhattacharjee, J. Mallela, Y. Belur, N. Atchayarcmf, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Effect of noise and model complexity on detection of amyotrophic lateral sclerosis and parkinson's disease using pitch and mfcc," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7313–7317.

[26] Y.-T. Hsu, Z. Zhu, C.-T. Wang, S.-H. Fang, F. Rudzicz, and Y. Tsao, "Robustness against the channel effect in pathological voice detection," 2018.

[27] 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 *Sixteenth annual conference of the international speech communication association*, 2015.

[28] H. Bredin, "pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th INTERSPEECH Conference*. ISCA, 2023, pp. 1983–1987.

[29] 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.

[30] 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.

[31] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.