Variational Autoencoder for Personalized Pathological Speech Enhancement

Mingchi Hou[1,2], Ina Kodrasi[1]
[1]*Idiap Research Institute, Switzerland*
[2]*École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
{mingchi.hou,ina.kodrasi}@idiap.ch

***Abstract*—The generalizability of speech enhancement (SE) models across speaker conditions remains largely unexplored, despite its critical importance for broader applicability. This paper investigates the performance of the hybrid variational autoencoder (VAE)-non-negative matrix factorization (NMF) model for SE, focusing primarily on its generalizability to pathological speakers with Parkinson's disease. We show that VAE models trained on large neurotypical datasets perform poorly on pathological speech. While fine-tuning these pre-trained models with pathological speech improves performance, a performance gap remains between neurotypical and pathological speakers. To address this gap, we propose using personalized SE models derived from fine-tuning pre-trained models with only a few seconds of clean data from each speaker. Our results demonstrate that personalized models considerably enhance performance for all speakers, achieving comparable results for both neurotypical and pathological speakers.**

***Index Terms*—speech enhancement, variational autoencoder, generalizability, Parkinson's disease, personalization**

I. Introduction

Speech communication is essential for conveying information, ideas, and emotions. However, noise from sources like traffic, machinery, or crowded environments impairs speech intelligibility and quality, posing challenges for many voice-based applications such as hearing aids and speech recognition systems [1]. To mitigate these challenges, speech enhancement (SE) approaches focusing on suppressing undesired interferences have become indispensable [2].

In recent years, deep learning has revolutionized SE, leading to data-driven techniques that leverage large datasets and powerful learning algorithms to improve speech quality [3]–[5]. However, such approaches typically require pairs of clean and noisy data for training, and their performance heavily depends on the quantity and diversity of the training samples. As a result, they often struggle to generalize in noisy environments that are not encountered during training [6]. To achieve robustness in various noisy environments, generative models have recently become more prevalent [7]–[16]. These models can be broadly categorized into three categories, i.e., i) diffusion-based [7]–[9], ii) Schrödinger bridge-based [10], and iii) hybrid variational autoencoder (VAE) and non-negative matrix factorization (NMF)-based [11]–[16] models. Diffusion-based models operate by progressively adding noise to a clean speech signal in a forward process. In the reverse process, typically guided by a neural network, these models learn to reconstruct the clean signal from the noisy input. The more recent Schrödinger bridge-based generative model, in contrast, enables optimal interpolation between the clean and noisy speech spectral components, moving beyond the conventional forward-backward noise process seen in diffusion models. Lastly, hybrid VAE-NMF-based models leverage a probabilistic latent space to model the clean speech prior through a VAE and an NMF to model the signal's structure dynamically, separating speech from noise in a more interpretable and structured way. Although diffusion-based and Schrödinger bridge-based models have generally shown better SE performance [8], [10], VAE-NMF-based models remain relevant and advantageous as they only require clean signals for training and typically have smaller model sizes.

Despite the beneficial enhancement performance, a vast majority of SE research is done using English data and recordings from neurotypical speakers exhibiting no speech disorders. The generalizability to other languages has received seldom attention [17], as outlined in the recent URGENT challenge [18]. More importantly, to the best of our knowledge, the performance of SE models has never been investigated for pathological speakers with speech disorders. Pathological speech, produced by individuals with neurological conditions such as Parkinson's disease (PD) or Amyotrophic Lateral Sclerosis, exhibits irregularities in speech characteristics [19], which leads to reduced intelligibility and increased susceptibility to noise. Research has shown that there is a considerable difference in the statistical distribution of pathological speech compared to neurotypical speech [20], [21]. Consequently, the performance of SE models trained using neurotypical speech recordings is expected to degrade when applied to pathological speech. Given the widespread prevalence of neurological disorders, affecting more than 1 billion individuals [22], analyzing the performance of SE models for pathological speakers and developing SE solutions tailored to such speakers is crucial.

In this paper, we investigate the SE performance of the hybrid VAE-NMF model [15], [16] across different languages and across speakers. We analyze cross-language generalizability using English and Spanish datasets, showing that the hybrid VAE-NMF model is sensitive to domain shifts caused by language differences. Additionally, we compare the SE performance on both neurotypical and pathological speak-

This work was supported by the Swiss National Science Foundation project 200021_215187 on "Pathological Speech Enhancement".

ers using VAE models trained on a large dataset of only neurotypical speech as well as VAE models trained on a considerably smaller dataset that includes both neurotypical and pathological speech. Unsurprisingly, we observe a performance gap between neurotypical and pathological speakers, with poorer SE performance for pathological speakers. Finally, we propose fine-tuning and personalization strategies, with personalization yielding not only the best overall performance for both neurotypical and pathological speakers, but also a comparable performance for both groups of speakers.

II. VAE-NMF for Speech Enhancement

Mixture model. In the short-time Fourier transform (STFT) domain, the complex noisy mixture y_{ft} at frequency bin index f and time frame index t is given by

$$y_{ft} = \sqrt{g_t} s_{ft} + n_{ft}, \tag{1}$$

with s_{ft} the clean speech, n_{ft} the additive noise, and $g_t \in \mathbb{R}_+$ a frequency-independent time-varying gain introduced to provide robustness to the varying loudness level of different speech signals typically used for training [15]. Given the noisy mixture and assuming that the speech and noise spectral coefficients are uncorrelated and follow a circularly symmetric Gaussian distribution, the minimum mean square error estimator of the clean speech coefficients is given by the Wiener filter [15]

$$\hat{s}_{ft} = \frac{\hat{g}_t \hat{\sigma}^2_{s,ft}}{\hat{g}_t \hat{\sigma}^2_{s,ft} + \hat{\sigma}^2_{n,ft}} y_{ft}, \tag{2}$$

with $\hat{g}_t$, $\hat{\sigma}^2_{s,ft}$ and $\hat{\sigma}^2_{n,ft}$ the estimated gain, clean speech variance, and noise variance, respectively. As briefly outlined below, the speech and noise variances can be estimated using a Monte Carlo expectation maximization (MCEM) algorithm combining a VAE which learns the prior distribution of clean speech and an untrained NMF-based noise model [15].

Clean speech prior. The VAE consists of an encoder network $\mathbf{e}_\phi$, parametrized by ϕ, and a decoder network $\mathbf{d}_\theta$, parameterized by θ, operating on the squared magnitude of the STFT coefficients, i.e., the F-dimensional vector of speech power coefficients $|\mathbf{s}_t|^2$ at time frame index t, with F being the total number of frequency bins. The encoder maps $|\mathbf{s}_t|^2$ to the parameters of the distribution over the latent variable $\mathbf{z}_t \in \mathbb{R}^D$, where D is the dimensionality of the latent space. The decoder then reconstructs the clean speech power coefficients from the latent space. The likelihood of the clean speech coefficients given the latent variables is modeled as

$$p_\theta(\mathbf{s}_t|\mathbf{z}_t) = \mathcal{N}_\mathbb{C}(\mathbf{0}, \mathrm{diag}(\mathbf{d}_\theta(\mathbf{z}_t))), \quad \text{with} \quad \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

The network parameters θ and ϕ are optimized by maximizing the evidence lower bound on the clean speech log-likelihood. For additional details, the interested reader is referred to [15].

Noise model. NMF is used to model the noise variance as

$$\sigma^2_{n,ft} = \{\mathbf{WH}\}_{ft}, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times T}$ are non-negative matrices representing the temporal activations and noise basis, with K being the NMF rank.

Noisy mixture model. Using the previously described clean speech and noise models, the distribution of the noisy mixture coefficients given the latent variable $\mathbf{z}_t$ follows

$$y_{ft}|\mathbf{z}_t \sim \mathcal{N}_\mathbb{C}(0, g_t\{\mathbf{d}_\theta(\mathbf{z}_t)\}_f + \{\mathbf{WH}\}_{ft}), \tag{5}$$

with the parameters g_t, $\mathbf{W}$, and $\mathbf{H}$ the unknown parameters to be estimated through MCEM [15].

SE with the hybrid VAE-NMF model. Given the noisy mixture signal, the gain g_t and the speech and noise spectral variances are estimated as previously described. These estimated parameters are then used to compute the Wiener gain in (2) and enhance the noisy signal. Such a hybrid VAE-NMF approach to SE has been shown to achieve advantageous performance for neurotypical speech recordings [15]. Given the different distributions of neurotypical and pathological speech spectral coefficients [20], [21], it can be expected that the clean speech prior learned through a VAE trained on a neurotypical speech dataset does not generalize well to pathological speech.

III. Personalized Pathological Speech Enhancement

It is already known that SE models exhibit highly variable performance due to mismatches in speaker characteristics between training and testing sets [23], [24], although this has not been explored for pathological speakers or in the context of the hybrid VAE-NMF model. Personalized SE models, which optimize performance for a single speaker within a specific acoustic environment, have been shown to outperform general-purpose SE models for that particular speaker and environment [25], [26]. Several approaches have been proposed to adapt general SE models to test speakers, including incorporating speaker embeddings into models like Deep Complex Convolution Recurrent Network and Deep Convolution Attention U-Net [27], training speaker encoder networks [28], and using internal embeddings to capture speaker profiles [29]. Another approach involves fine-tuning pre-trained SE models on a pseudo-SE task, using only noisy signals from the test speaker of interest [23].

In exploring strategies to enhance the hybrid VAE-NMF model's performance on pathological speech, we consider two approaches. First, we investigate the impact of including pathological speech data during VAE training. Instead of random initialization, we fine-tune a pre-trained VAE model (originally trained on a large dataset of neurotypical speakers) on a smaller dataset containing both neurotypical and pathological speech. As shown in Section V-C, while fine-tuning the VAE on a pathological dataset improves the SE performance of the hybrid VAE-NMF model for pathological speakers, a performance gap between neurotypical and pathological speaker remains. We argue that the characteristics of pathological speech vary widely between speakers, and including pathological speech in the VAE training set does not necessarily improve generalization to unseen pathological test speakers. To further improve performance, our second approach involves using a personalized SE model for each speaker. Given the

TABLE I
WSJ0 (ENGLISH) AND CROWD (SPANISH) DATASET CHARACTERISTICS.

	WSJ0 [30]	CROWD [31]
Training set	101 speakers, 24.9 hours	101 speakers, 23.0 hours
Validation set	10 speakers, 2.2 hours	10 speakers, 2.2 hours
Test set	8 speakers, 1.5 hours	8 speakers, 1.5 hours

challenges of obtaining robust embeddings for pathological speakers, we follow a personalization approach similar to [23], where signals from the test speaker are used to fine-tune the pre-trained SE model. Differently from [23] where fine-tuning is done using noisy signals, we assume availability of clean signals from the test speakers (since the VAE is trained on clean speech).

IV. EXPERIMENTAL SETTINGS

A. Datasets

Clean speech signals are used for training and validation of VAE models, while noisy mixtures are used for testing.

Clean signals. To assess cross-language generalizability, we use the English Wall Street Journal (WSJ0) [30] and the Spanish Crowdsourced Latin American Spanish Corpora (CROWD) [31] datasets, which contain recordings from neurotypical speakers. The characteristics of these datasets, including the number of speakers and total duration of the training, validation, and test sets, are presented in Table I. As shown, the datasets are closely matched in terms of speaker count and total duration, ensuring a systematic evaluation of cross-language generalizability. To assess cross-speaker generalizability, fine-tune models, and develop personalized models, we use the Spanish PC-GITA dataset [32], comprising recordings from 50 neurotypical and 50 PD speakers. PC-GITA includes recordings of 10 sentences, one text passage, and a monologue per speaker, and has a total duration of 2.8 hours. Due to this dataset's limited size (which is a common characteristic of pathological speech datasets) compared to those used in SE literature (cf. Table I), we use a 10-fold speaker-independent cross-validation evaluation framework, partitioning the data within each fold into 80% training, 10% validation, and 10% testing. Personalized models are trained, validated, and tested per speaker independently (cf. Section IV-C).

Noisy mixtures. Noisy mixtures required for testing the VAE-NMF model are generated using the previously described clean signals and noise recordings from the QUT dataset [33]. The noise recordings are randomly selected from the set {"cafe", "car", "home", "street"}, with the signal-to-noise ratio (SNR) randomly selected from the set {-5, 0, 5} dB [12].

B. Training

All signals are resampled to 16 kHz if necessary and are transformed to the STFT domain using a 64 ms Hann window with a hop size of 16 ms, resulting in 513 unique frequency bins. As in [11], [13], the VAE latent dimension is $D = 16$ and the NMF rank is $K = 8$. The settings of the MCEM follow [15]. Additionally, we use a batch size of 128, the Adam optimizer with standard settings, and an initial learning rate of 10^{-4}. The learning rate scheduler halves the learning rate if the validation loss plateaus, with a patience of 10 epochs. The maximum number of epochs is set to 500, with early stopping if the validation loss does not decrease for 20 consecutive epochs.

C. Fine-tuning and personalization

Given the small size of the PC-GITA dataset, training a VAE model from scratch (i.e., with random weight initialization) on this dataset is expected to yield suboptimal results. To improve performance, we explore the possibility of leveraging the pre-trained VAE model from the CROWD dataset and fine-tuning it on the PC-GITA dataset. After initializing the model weights, fine-tuning is done following the same training procedure outlined in Section IV-B.

As outlined in Section III, we also train personalized models for each speaker in the PC-GITA dataset. To this end, we use a subset of a speaker's recordings as training/validation data, while the remaining recordings are used as testing data. To ensure that the conclusions we draw from these models are not influenced by the specific subset of data used for training/validation/testing, we report the average performance of two personalized models: one using the monologue recording for training/validation and the sentences and read text recordings for testing, and another using the sentences and read text recordings for training/validation and the monologue recording for testing. It should be noted that these subsets of data have a similar duration, i.e., the average duration of the monologue recording across all speakers is 47.1 s whereas the average duration of the sentences and read text recording across all speakers is 55.3 s. Personalized models are initialized with the weights of the pre-trained VAE model from the CROWD dataset. The training procedure is the same as outlined in Section IV-B.

D. Evaluation

Performance is evaluated using the wideband perceptual evaluation of speech quality (PESQ) measure [34], the frequency-weighted segmental SNR (fwSSNR) [35], and the scale-invariant signal-to-distortion ratio (SI-SDR) [36]. The clean speech signal is used as a reference signal for computing these measures. The improvement in these instrumental measures, i.e., ΔPESQ, ΔfwSSNR, and ΔSI-SDR, is computed as the difference between the PESQ, fwSSNR, and SI-SDR values of the enhanced signal and the noisy microphone signal.

V. RESULTS AND DISCUSSION

A. Cross-lingual generalization

Table II presents the performance of the hybrid VAE-NMF model with the VAE trained on the English WSJ0 and Spanish CROWD datasets, evaluated under both within-database and cross-lingual testing conditions. It should be noted that the noise types and SNRs remain consistent across all evaluations. However, since the VAE is trained on the different considered

TABLE II

WITHIN-DATABASE AND CROSS-DATABASE PERFORMANCE OF THE VAE-NMF SE MODELS. VAE (WSJ0) DENOTES MODELS TRAINED ON THE ENGLISH WSJ0 DATASET, WHEREAS VAE (CROWD) DENOTES MODELS TRAINED ON THE SPANISH CROWD DATASET. MODELS ARE TESTED ON ENGLISH (WSJ0-QUT) AND SPANISH (CROWD-QUT) TEST SETS. VALUES IN BOLD INDICATE THE TEST SET FOR WHICH THE HIGHEST PERFORMANCE IMPROVEMENT IS OBTAINED FOR EACH MODEL AND EACH MEASURE.

Model	English (WSJ0-QUT) test set			Spanish (CROWD-QUT) test set		
	ΔPESQ	ΔfwSSNR	ΔSI-SDR	ΔPESQ	ΔfwSSNR	ΔSI-SDR
VAE (WSJ0) - English train set	$\mathbf{0.21 \pm 0.02}$	$\mathbf{0.87 \pm 0.13}$	$\mathbf{5.77 \pm 0.28}$	0.13 ± 0.01	-0.91 ± 0.10	4.53 ± 0.25
VAE (CROWD) - Spanish train set	0.13 ± 0.01	0.69 ± 0.13	4.65 ± 0.26	$\mathbf{0.19 \pm 0.01}$	$\mathbf{0.78 \pm 0.08}$	$\mathbf{6.25 \pm 0.22}$

TABLE III

PERFORMANCE OF THE VAE (CROWD) MODEL ON THE PC-GITA-QUT TEST SET OF NEUROTYPICAL AND PATHOLOGICAL SPEAKERS. VALUES IN BOLD INDICATE THE SPEAKER GROUP FOR WHICH THE HIGHEST PERFORMANCE IMPROVEMENT IS OBTAINED FOR EACH MEASURE.

Test Speakers	ΔPESQ	ΔfwSSNR	ΔSI-SDR
Neurotypical	$\mathbf{0.10 \pm 0.01}$	$\mathbf{2.22 \pm 0.14}$	$\mathbf{4.72 \pm 0.26}$
Pathological	0.05 ± 0.01	1.51 ± 0.17	3.76 ± 0.30

TABLE IV

PERFORMANCE OF DIFFERENT MODELS ON THE PC-GITA-QUT TEST SET OF NEUROTYPICAL AND PATHOLOGICAL SPEAKERS. VALUES IN BOLD INDICATE THE SPEAKER GROUP FOR WHICH THE HIGHEST PERFORMANCE IMPROVEMENT IS OBTAINED FOR EACH MEASURE.

Model	Test Speakers	ΔPESQ	ΔfwSSNR	ΔSI-SDR
M_S	Neurotypical	0.13 ± 0.01	1.96 ± 0.19	4.48 ± 0.35
	Pathological	0.07 ± 0.01	1.08 ± 0.20	3.22 ± 0.37
M_F	Neurotypical	0.13 ± 0.01	2.53 ± 0.16	5.02 ± 0.29
	Pathological	0.08 ± 0.01	1.79 ± 0.17	4.10 ± 0.32
M_P	Neurotypical	$\mathbf{0.19 \pm 0.02}$	$\mathbf{2.74 \pm 0.16}$	$\mathbf{7.87 \pm 0.34}$
	Pathological	0.18 ± 0.02	2.40 ± 0.16	7.75 ± 0.37

clean speech datasets, domain shifts are introduced due to different languages (and potentially different recording setups). As expected, the VAE achieves a better performance in within-database testing (i.e., when the training and testing datasets are in the same language). In contrast, performance degrades in cross-lingual settings due to the domain shift introduced by the different language.

B. Cross-speaker generalization

In this section, we assess the performance of the VAE-NMF model on both neurotypical and pathological speakers from the PC-GITA database. Since PC-GITA consists of Spanish recordings, our analysis focuses on the VAE (CROWD) model, i.e., the model trained on the Spanish CROWD dataset in Section V-A. The results, presented in Table III, reveal that the model performs consistently better on neurotypical speakers than on pathological speakers across all evaluation metrics. This degradation is expected, as VAE (CROWD) is trained exclusively on neurotypical speech, with pathological speech characteristics exhibiting substantial deviations from it [21].

C. Fine-tuning and personalization

In this section, we explore various strategies to improve the SE performance of the VAE-NMF model for pathological speakers. Specifically, we investigate the performance of the following models:

- *Training a VAE from scratch on PC-GITA (model M_S).* We examine the feasibility of training a VAE model from scratch on a small, clean speech dataset, such as PC-GITA, which includes both neurotypical and pathological speakers. This contrasts with the conventional practice of training VAEs on much larger corpora, such as WSJ0 or CROWD, which consist of only neurotypical speech.
- *Fine-tuning a pre-trained VAE (model M_F).* We explore the potential of fine-tuning a VAE model that has been pre-trained on the larger CROWD dataset, using the PC-GITA dataset, as described in Section IV-C.
- *Personalized VAE model (model M_P).* We assess the performance of personalized VAE models, where the model is trained for each individual speaker as outlined in Section IV-C.

The SE performance of the different models for both neurotypical and pathological speakers is summarized in Table IV. Results show that training the VAE solely on the PC-GITA dataset leads to a poorer performance in terms of fwSSNR, and SI-SDR for both neurotypical and pathological speakers, compared to training the VAE on the larger CROWD dataset (cf. Table III). This highlights the importance of using a sufficiently large and diverse dataset when training VAEs. Rather than training the VAE from scratch on PC-GITA, Table IV demonstrates that fine-tuning the VAE model pre-trained on CROWD using the PC-GITA dataset leads to improvements across all metrics for both groups of speakers. However, a persistent performance gap remains between neurotypical and pathological speakers. We hypothesize that the high variability in pathological speech across individuals restricts the model's ability to generalize effectively to new pathological speakers, even when pathological speech is included in the training data. Finally, Table IV highlights the value of using personalized models. Taking advantage of the knowledge embedded in the pre-trained VAE (CROWD), incorporating only a few seconds (≈ 50 s) of clean speech from the test speaker can boost the SE performance. While this personalization strategy proves effective for both groups of speakers, the improvement in SE performance is particularly notable for pathological speakers.

VI. CONCLUSION

This paper evaluates the generalizability of the hybrid VAE-NMF model for SE across languages (i.e., English and

Spanish) and speaker conditions (i.e., neurotypical speakers and pathological speakers with PD). The results show that, as expected, the hybrid VAE-NMF performs best when trained and tested within the same database. However, cross-lingual and cross-speaker testing leads to performance degradation, with a larger drop observed in the latter case. To improve performance for pathological speech, we have proposed fine-tuning pre-trained models by incorporating pathological data into the training set, as well as training personalized models for each speaker using only a few seconds of clean data. Results have shown that personalized SE models significantly improve performance for all speakers, achieving comparable results for both neurotypical and pathological speakers.

REFERENCES

[1] D. O'Shaughnessy, "Speech Enhancement—A Review of Modern Methods," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, Feb. 2024.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*. Routledge & CRC Press, 2013.

[3] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[4] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*. ISCA, 2018, pp. 3229–3233.

[5] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "An integrated cnn-gru framework for complex ratio mask estimation in speech enhancement," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 764–768.

[6] P. Gonzalez, T. S. Alstrøm, and T. May, "Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, p. 3390–3403, Sep. 2023.

[7] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional Diffusion Probabilistic Model for Speech Enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7402–7406.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 31, pp. 2351–2364, 2023.

[9] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration: A review," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2024.

[10] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proc. Interspeech*. ISCA, 2024.

[11] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A Recurrent Variational Autoencoder for Speech Enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 371–375.

[12] J. Richter, G. Carbajal, and T. Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks," in *Proc. Interspeech*. ISCA, Oct. 2020, pp. 4516–4520.

[13] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 676–680.

[14] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 2993–3007, Sep. 2022.

[15] S. Leglaive, L. Girin, and R. Horaud, "A Variance modeling framework based on Variational Autoencoders for Speech Enhancement," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2018, pp. 1–6.

[16] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 716–720.

[17] G. Close, T. Hain, and S. Goetze, "The Effect of Spoken Language on Speech Enhancement using Self-Supervised Speech Representation Loss Functions," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[18] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, A. Kumar, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "URGENT Challenge: Universality, Robustness, and Generalizability For Speech Enhancement," in *Proc. Interspeech 2024*, Sep. 2024, pp. 4868–4872.

[19] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.

[20] I. Kodrasi and H. Bourlard, "Statistical modeling of speech spectral coefficients in patients with parkinson's disease," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[21] ——, "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1210–1222, 2020.

[22] WHO, "Neurological disorders: public health challenges," 2006.

[23] A. Sivaraman, S. Kim, and M. Kim, "Personalized speech enhancement through self-supervised data augmentation and purification," in *Proc. Interspeech*. ISCA, 2021, pp. 2208–2212.

[24] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Interspeech*. ISCA, 2014, pp. 2685–2689.

[25] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[26] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 176–180.

[27] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: new models and comprehensive evaluation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 356–360.

[28] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, Y. Wang, L. Xie, and S. Shang, "Tea-pse: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2022 dns challenge," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9291–9295.

[29] T. Pärnamaa and A. Saabas, "Personalized Speech Enhancement Without a Separate Speaker Embedding Model," in *Proc. Interspeech*. ISCA, Sep. 2024, pp. 4863–4867.

[30] Garofolo, John S., Graff, David, Paul, Doug, and Pallett, David, "CSR-I (WSJ0) Complete," May 2007.

[31] A. Guevara-Rukoz, I. Demirsahin, F. He, S.-H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, and O. Kjartansson, "Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech," in *Proc. Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 6504–6513.

[32] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, May 2014, pp. 342–347.

[33] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*. ISCA, Sep. 2010, pp. 3110–3113.

[34] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[35] 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, 2008.

[36] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 626–630.