GENERALIZABILITY OF PREDICTIVE AND GENERATIVE SPEECH ENHANCEMENT MODELS TO PATHOLOGICAL SPEAKERS

Mingchi Hou[1,2], *Ante Jukić*[3], *Ina Kodrasi*[1]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
[3]NVIDIA, USA

ABSTRACT

State-of-the-art speech enhancement (SE) models achieve strong performance on neurotypical speech, but their effectiveness is substantially reduced for pathological speech. In this paper, we investigate strategies to address this gap for both predictive and generative SE models, including (*i*) training models from scratch using pathological data, (*ii*) fine-tuning models pre-trained on neurotypical speech with additional data from pathological speakers, and (*iii*) speaker-specific personalization using only data from the individual pathological test speaker. Our results show that, despite the limited size of pathological speech datasets, SE models can be successfully trained or fine-tuned on such data. Fine-tuning models with data from several pathological speakers yields the largest performance improvements, while speaker-specific personalization is less effective, likely due to the small amount of data available per speaker. These findings highlight the challenges and potential strategies for improving SE performance for pathological speakers.

Index Terms— noise reduction, pathological speakers, Parkinson's disease

1. INTRODUCTION

Speech communication is essential for human interactions, facilitating information exchanges across diverse environments. However, in real-world settings, background noise often degrades speech quality and intelligibility. Speech enhancement (SE) techniques aim to address this challenge by estimating clean speech from noisy recordings. The widespread adoption of voice-based technologies, such as mobile communication, hearing aids, automatic speech recognition, and teleconferencing systems, alongside the need to alleviate cognitive burden and ease listener's fatigue, has increased the demand for robust and generalizable SE solutions [1]. This need drives efforts like the URGENT challenge [2] for universal SE robustness across languages, noises, and domains.

Traditional SE approaches, such as spectral subtraction [3] or Wiener filtering [4], rely on assumed statistical properties of the clean speech and noise signals. While effective in stationary noise environments, these methods often struggle with non-stationary interferences. The advent of deep learning has revolutionized the field, with approaches employing data-driven solutions to model complex signal patterns from training data [5, 6]. Data-driven solutions have traditionally relied on predictive models that aim to learn a single deterministic mapping from noisy to clean speech, for example by estimating spectral masks [7, 8] or spectral coefficients [9]. Such approaches, however, encounter challenges to generalize to unseen noise environments [10]. More recently, the emergence of generative models has become an increasingly popular area of research in various speech related tasks, including SE [11, 12]. State-of-the-art (SOTA) generative SE approaches are typically based on diffusion models, operating by progressively adding noise to a clean speech signal in a forward diffusion process [13]. The reverse process, which can be guided by a neural network, learns to recover clean signals from noisy inputs. Recently, a Schrödinger bridge-based generative model was proposed for SE [14]. Differently from typical forward diffusion, this model results in exact interpolation between the clean and noisy speech spectral coefficients.

Despite beneficial enhancement performance, the vast majority of SE research is using recordings from neurotypical speakers exhibiting no speech disorders from datasets such as Wall Street Journal [14–16] or VoiceBank [16]. Pathological speech, produced by individuals with hearing impairments, head and neck cancers, or neurological conditions such as Parkinson's disease (PD), has received far less attention. Such speech often exhibits atypical acoustic characteristics [17], leading to reduced intelligibility and increased vulnerability to noise. Research has further shown that pathological speech differs markedly in its statistical distribution compared to neurotypical speech [18, 19], which implies that SE models trained exclusively on neurotypical data are likely to generalize poorly to pathological conditions. Although these conditions are widespread, with approximately 360 million people experiencing hearing impairment [20], 650 thousand new cases of head and neck cancers diagnosed annually [21], and over 1 billion individuals worldwide affected by neurological disorders [22], the effectiveness of SOTA SE models for pathological speakers remains largely unexplored. To the best of our knowledge, SE performance for pathological speech has only been explicitly studied in [23], where a hybrid variational autoencoder (VAE)–non-negative matrix factorization (NMF) model was evaluated. While results showed a marked deterioration in enhancement quality for pathological speakers, this finding is limited in scope since the VAE–NMF model neither reflects the architectures nor the performance levels of current SOTA systems [24]. Consequently, the true capability of modern SE models to handle pathological speech remains unknown. Given the global prevalence of pathological speech, systematically assessing the performance of SOTA SE models and developing tailored solutions for handling pathological speech is therefore a critical research need.

This paper systematically investigates the performance of SOTA predictive and generative SE models on pathological speech and proposes strategies to improve their effectiveness. Specifically, we consider: (*i*) training models from scratch using pathological data, (*ii*) fine-tuning models pre-trained on neurotypical speech with additional pathological data, and (*iii*) speaker-specific personalization

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

where models pre-trained on large neurotypical speech corpora are fine-tuned using only data from the pathological speaker of interest.

2. SPEECH ENHANCEMENT

Consider the noisy microphone signal $y(\tau)$ at time τ, i.e.,

$$y(\tau) = x(\tau) + n(\tau), \tag{1}$$

with $x(\tau)$ denoting the clean speech signal and $n(\tau)$ denoting the additive noise signal. In the short-time Fourier transform (STFT) domain, the signal model in (1) is given by

$$Y(j,k) = X(j,k) + N(j,k), \tag{2}$$

where $Y(j,k)$, $X(j,k)$ and $N(j,k)$ are the complex-valued STFT coefficients of noisy, clean speech, and noise signals, j represents the frequency bin index, and k represents the time frame index. The objective of single-channel SE is to estimate the clean speech signal given the noisy microphone recording. In the following, several enhancement approaches considered in this paper are briefly reviewed. We consider predictive models, which learn a single mapping between noisy and clean speech, as well as generative models, which learn the distribution of clean speech.

2.1. Predictive models

Magnitude spectrogram-based masking (MM): Mask-based models estimate the clean speech signal by selectively masking unwanted (i.e., noise-dominated) time-frequency components [25]. The enhanced signal is computed as $\hat{X}(j,k) = M(j,k)Y(j,k)$, where $M \in [0,1]$ is the time-frequency mask. In this paper, we use a deep neural network (DNN) model trained to estimate the traditional ideal ratio mask, which is defined as the ratio between the spectral magnitudes of the clean and noisy speech [8]. While various training losses have been investigated for mask-based models for SE [26], in this paper we use the widely adopted scale-invariant signal-to-distortion ratio (SI-SDR) [27] between the time-domain predicted signal $\hat{x}(\tau)$, computed using the inverse STFT of $\hat{X}(j,k)$, and the clean reference signal $x(\tau)$.

Complex-valued spectrogram-based regression (CR): Instead of enhancing only the noisy speech magnitude and retaining the noisy phase as in the magnitude spectrogram-based masking, approaches that jointly enhance the magnitude and the phase component of the noisy signal have also become popular. For this category of approaches, we use a DNN model trained to estimate the real and imaginary part of the clean STFT coefficients from the noisy STFT coefficients [28]. Since the mean square error (MSE) is a standard and straightforward choice in regression problems, the training loss used for this approach is the MSE between the time-domain predicted signal $\hat{x}(\tau)$ and the clean reference signal $x(\tau)$.

2.2. Generative models

Score-based diffusion model (SGMSE+): Diffusion models typically operate by progressively adding noise to a clean speech signal in a forward diffusion process [15, 28–30]. A score-model can be trained to guide the reverse process, and hence, recover clean signals from noisy inputs by removing the noise at each reverse step [15,30]. The diffusion process is defined by a forward stochastic differential equation (SDE) [28, 29]. The corresponding reverse SDE can be expressed in terms of the forward SDE parameters with an additional term based on the score function $\nabla_x \log p_t(x)$ of the marginal distribution p_t at process time t. To enable inference using the reverse SDE, a DNN is trained to estimate the score [15] and an iterative sampler is used to obtain the clean speech estimate. In this work, an SGMSE+ model incorporating an affine drift term with a predictor-corrector sampler is used. For additional details, the reader is referred to [15].

Schrödinger Bridge-based model (SB): Schrödinger Bridge is a generative model that aims to find an optimal transport path that minimizes the discrepancy between noisy and clean distributions [14, 31, 32]. As opposed to typical forward diffusion [15], where clean data is transformed into a sample from a broad distribution centered around the noisy observation, the SB results in exact interpolation between the clean speech and the observed noisy speech coefficients [14]. Thus, a weighted data prediction loss is used. In this work, a SB model with SDE sampler is used. For additional details, the reader is referred to [14].

3. PATHOLOGICAL SE STRATEGIES

It is well established that SE models exhibit variable performance due to mismatches in speaker characteristics between training and testing sets [15, 33]. This issue is expected to be even more pronounced for pathological speakers, whose acoustic characteristics differ substantially from neurotypical speakers [18, 19]. To improve the performance of SOTA SE models for pathological speakers, we investigate three strategies. First, we consider the feasibility of training models from scratch on a (small) dataset that includes both neurotypical and pathological speech. Second, we explore fine-tuning models pre-trained on large neurotypical datasets on smaller datasets containing pathological speech. Based on our previous work with the hybrid VAE–NMF model [23], fine-tuning with pathological data is expected to improve performance for pathological speakers. However, a performance gap relative to neurotypical speakers may persist due to the high variability in pathological speech characteristics. Finally, we consider speaker-specific personalization, where the parameters of models pre-trained on large nurotypical speech datasets are adapted to individual test speakers. While this approach can provide some gains, its effectiveness is limited in our experiments, likely due to the small amount of data available per speaker.

4. EXPERIMENTAL SETTINGS

4.1. Neurotypical and pathological datasets

Clean speech datasets: For the results presented in the following section, we use the Spanish CROWD [34] and PC-GITA [35] clean speech datasets. CROWD is a large dataset of neurotypical speech used to pretrain the considered SE models. It contains 37.8 hours of recordings from 174 healthy speakers sampled at 48 kHz. To reflect the structure of standard SE datasets in terms of duration and number of speakers (cf. e.g., [14, 15]), recordings are downsampled to 16 kHz, and we use 23 h, 2.2 h, and 1.5 h of the data for training, validation, and testing respectively. PC-GITA contains 2.8 hours of recordings sampled at 44.1 kHz from 50 patients diagnosed with PD and 50 neurotypical controls, reflecting the small size of pathological speech datasets that are typically available. Each speaker utters 12 utterances (10 sentences, 1 read text, 1 monologue). Recordings are downsampled to 16 kHz. Given the small dataset size, we adopt a 10-fold speaker-independent cross-validation strategy when training or fine-tuning models using the PC-GITA dataset, with 80%, 10%, and 10% of the data used for training, validation, and testing respectively in each fold.

Noisy mixtures: To generate noisy mixtures, we use the CHiME3 dataset [36]. All noise signals are first downsampled

to 16 kHz. For each clean utterance, we randomly select a noise file recorded on a bus, in a cafe, in a pedestrian area, or at a street junction, and add it at an SNR uniformly sampled between -6 dB and 14 dB for training and validation. For testing, we use fixed SNRs of -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB to provide a consistent evaluation of model performance.

4.2. Training Configuration

As in [16], signals are transformed to the STFT domain using a window size of 510 samples and a hop size of 128 samples. Further, the hyperparameters used to compress the dynamic range of the spectrogram are $\alpha = 0.5$ and $\beta = 0.33$ [14].

The *MM model* is trained using a 5 layer Bidirectional Long Short-Term Memory network [8]. The *CR model* follows the NCSN+ architecture in [28], i.e., a multi-resolution U-Net architecture with skip connections, incorporating 3 ResNet blocks with 2D convolutions, group normalization, and both upsampling and downsampling layers, further modified in [16] to take complex valued input. The *SGMSE+* model is trained via denoising score matching, with hyperparameters $\sigma_{min} = 0.05$, $\sigma_{max} = 0.5$, and $\gamma = 1.5$. The predictor-corrector sampler with 30 time steps (60 DNN evaluations) is used during inference [15]. Under the scope of Ornstein-Uhlenbeck SDE, the *SB model* uses the Variance Exploding (VE) schedule with $\sigma_{min} = 0.7$ and $\sigma_{max} = 1.82$. For inference, SDE sampler with 50 time steps (50 DNN evaluations) is used. Both SGMSE+ and SB models utilize NCSN+ backbones with extra noise scheduling layers. Further, exponential moving average with a weight decay of 0.999 is used for both models [14].

In terms of model complexity, the number of trainable parameters is 7.6M for the MM model, 22.1M for the CR model, 25.2M for the SGMSE+ model, and 25.2M for the SB model.

Training is carried out using a batch size of 8 and a total of 1000 epochs, with early stopping if the validation loss does not decrease for 20 consecutive epochs. Furthermore, we use the Adam optimizer and a learning rate of 10^{-4}. Training the CR, SGMSE+ and SB using the CROWD dataset was conducted on an NVIDIA H100 GPU, while all other models were trained on an RTX 3090 GPU.

4.3. Evaluation

The performance is evaluated through four objective measures, i.e., the extended short-time objective intelligibility (ESTOI)) [37], the wideband perceptual evaluation of speech quality (PESQ) [38], the frequency-weighted segmental SNR (fwSSNR) [39], and the SI-SDR [27]. For all these measures, the clean speech signal is used as the reference signal, and higher values indicate better performance. To facilitate comparison among different datasets, we report the difference of the metric values between the enhanced signals and the noisy mixtures.

5. RESULTS AND DISCUSSION

5.1. Performance of SE models on the CROWD dataset

Since SE models are traditionally benchmarked on English data, we first evaluate performance on the neurotypical Spanish CROWD dataset to establish a baseline. This step is important because our subsequent analyses focus on Spanish, given that the available pathological speech dataset (PC-GITA) is in Spanish. Table 1 presents the performance of the considered models when trained and tested on the neurotypical CROWD dataset. As expected from the SOTA literature, the CR and SB models generally achieve the best performance, followed by the SGMSE+ model, while the MM model typically performs the worst. More specifically, the CR model yields the highest ΔE-STOI, ΔPESQ, and ΔSI-SDR, whereas the SB model yields the highest ΔfwSSNR. The CR and SB models will therefore be used in our subsequent analysis as representatives of SOTA predictive and generative SE models.

Table 1. Performance of SE models trained and tested on the neurotypical Spanish CROWD dataset. Values in bold indicate the highest performance improvement obtained for each measure.

Model	ΔE-STOI	ΔPESQ	ΔfwSSNR	ΔSI-SDR
MM	0.12 ± 0.00	1.19 ± 0.01	2.55 ± 0.04	9.35 ± 0.08
CR	$\mathbf{0.16 \pm 0.00}$	$\mathbf{1.40 \pm 0.01}$	4.13 ± 0.04	$\mathbf{11.60 \pm 0.09}$
SGMSE+	0.11 ± 0.00	0.75 ± 0.01	3.71 ± 0.04	6.33 ± 0.06
SB	0.15 ± 0.00	1.36 ± 0.01	$\mathbf{5.19 \pm 0.04}$	8.29 ± 0.09

5.2. Performance of SE models on pathological speech

In this section, we evaluate the CR and SB models trained on the neurotypical CROWD dataset from Section 5.1 on both neurotypical and pathological speech from the PC-GITA dataset. Results are shown in Table 2. Comparing the results in Tables 1 and 2, it can be observed that the performance on neurotypical PC-GITA speakers is lower than on neurotypical CROWD speakers, showing that SOTA SE models still face challenges with cross-database generalization even within the same speaker group. More importantly, Table 2 shows that for both considered SE models across all evaluation metrics, performance drops even further for pathological speakers compared to neurotypical speakers. This gap is expected due to the different statistical distributions of neurotypical and pathological speech [18, 19] and highlights a critical limitation of current SOTA models that are trained exclusively on neurotypical speech.

5.3. Pathological SE models

In this section, we explore the following strategies to improve the SE performance of the CR and SB models for pathological speakers:

1. Training models from scratch. First, we investigate whether current SOTA SE enhancement models can be trained from scratch on small clean speech databases such as PC-GITA. This contrasts with the conventional practice of training SE models on much larger corpora such as CROWD, consisting of only neurotypical speech.

2. *Fine-tuning pre-trained models.* We further explore the potential of fine-tuning SE models that have been pre-trained on the larger CROWD dataset using the PC-GITA dataset.

3. Speaker-specific models. We finally assess the performance of speaker-specific SE models, where the models pre-trained on the CROWD dataset are fine-tuned using data from each individual speaker from the PC-GITA dataset. To this end, we use a subset of a speaker's recordings as fine-tuning/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 fine-tuning/validation/testing, we report the average performance of two speaker-specific models, one using the monologue recording for fine-tuning/validation and the sentences and read text recordings for testing, and another vice versa. 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.

Table 2. Performance of the CR and SB models trained on the neurotypical CROWD dataset and tested on neurotypical and pathological speakers from the PC-GITA dataset. Values in bold indicate the speaker group for which the highest performance improvement is obtained for each model and each measure.

	CR model				SB model			
Speaker	ΔE-STOI	ΔPESQ	ΔfwSSNR	ΔSI-SDR	ΔE-STOI	ΔPESQ	ΔfwSSNR	ΔSI-SDR
Neurotypical	$\mathbf{0.09 \pm 0.00}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{3.57 \pm 0.08}$	$\mathbf{4.22 \pm 0.19}$	$\mathbf{0.06 \pm 0.00}$	$\mathbf{0.52 \pm 0.02}$	$\mathbf{3.10 \pm 0.09}$	$\mathbf{1.40 \pm 0.18}$
Pathological	0.05 ± 0.00	0.63 ± 0.02	2.78 ± 0.09	2.81 ± 0.20	0.01 ± 0.00	0.31 ± 0.02	2.24 ± 0.10	0.36 ± 0.19

Table 3. Performance of the CR and SB models trained using various strategies outlined in Section 3 and tested on neurotypical and pathological speakers from the PC-GITA dataset (using a stratified cross-validation framework).

	CR model				SB model			
Speaker	ΔE-STOI	ΔPESQ	ΔfwSSNR	ΔSI-SDR	ΔE-STOI	ΔPESQ	ΔfwSSNR	ΔSI-SDR
	Models trained from scratch on PC-GITA							
Neurotypical	0.15 ± 0.00	1.21 ± 0.02	5.38 ± 0.07	8.19 ± 0.12	0.17 ± 0.00	1.39 ± 0.02	6.13 ± 0.08	8.00 ± 0.14
Pathological	0.13 ± 0.00	1.11 ± 0.02	4.94 ± 0.08	7.75 ± 0.14	0.15 ± 0.00	1.22 ± 0.02	5.47 ± 0.09	7.49 ± 0.14
	Models trained on CROWD and fine-tuned on PC-GITA							
Neurotypical	0.18 ± 0.00	1.40 ± 0.02	6.03 ± 0.07	8.99 ± 0.12	0.20 ± 0.00	1.53 ± 0.02	6.73 ± 0.07	8.48 ± 0.13
Pathological	0.15 ± 0.00	1.25 ± 0.02	5.37 ± 0.08	8.29 ± 0.13	0.17 ± 0.00	1.31 ± 0.02	5.94 ± 0.09	7.66 ± 0.14
	Models trained on CROWD and fine-tuned with speaker-specific PC-GITA data							
Neurotypical	0.13 ± 0.00	1.11 ± 0.02	4.71 ± 0.10	6.63 ± 0.21	0.14 ± 0.00	0.70 ± 0.02	5.17 ± 0.08	6.20 ± 0.13
Pathological	0.10 ± 0.00	0.88 ± 0.02	3.86 ± 0.10	6.02 ± 0.18	0.11 ± 0.00	0.55 ± 0.02	4.32 ± 0.08	5.31 ± 0.13

The performances obtained using these different strategies are shown in Table 3 and can be summarized as follows:

Training from scratch: Although the CR and SB models are relatively large, they can both be successfully trained on the relatively small PC-GITA dataset. This is confirmed by the fact that performance on neurotypical speakers is comparable to that of the same models trained on the much larger CROWD corpus (cf. Table 1). Most importantly, it can be observed that including pathological data in training considerably improves performance for pathological speakers. However, a consistent gap remains between neurotypical and pathological speakers for both models. Between the two models, SB generally outperforms CR across most metrics, except for SI-SDR improvement where CR shows a slight advantage.

Fine-tuning pre-trained models: Fine-tuning models trained on CROWD with pathological data yields the best overall performance, outperforming training from scratch across all metrics for both considered models. This suggests that large neurotypical corpora are valuable for learning general clean speech characteristics, while fine-tuning with relatively little pathological data enables models to capture disorder-specific traits. As with training from scratch, SB typically surpasses CR except in SI-SDR improvement. The gap between neurotypical and pathological speakers, however, persists.

Speaker-specific fine-tuning: Speaker-specific models achieve the lowest performance among all strategies, across both speaker groups and all metrics. Additional experiments (not shown here due to space constraints) suggest this is due to limited amount of data available per speaker. With only $\approx$ 50 s of data per speaker, the models cannot learn robust representations, leading to poorer generalization compared to strategies that leverage more data from multiple speakers. In summary, the most effective strategy for improving SE performance on pathological speech is to pretrain SOTA models on large neurotypical corpora and fine-tune them on smaller pathological datasets. Across these strategies, and among the considered exemplary models, SB proves to be the most effective model overall. Importantly, a performance gap between neurotypical and pathological speakers remains. We suspect this is due to the high variability in pathological speech characteristics such that training on data from other pathological speakers may not sufficiently capture the acoustic traits of a given test speaker. Future work should explore pathology-aware fine-tuning strategies that explicitly integrate domain knowledge about speech disorders and variability across individuals. Additionally, listening tests should be conducted to confirm the conclusions derived from objective metrics.

6. CONCLUSION

This paper systematically analyzed the performance of SOTA predictive and generative SE models for pathological speakers. While these models perform well for neurotypical speakers, their effectiveness on pathological speech remains limited. To address this, we explored three strategies, i.e., training from scratch on pathological datasets, fine-tuning neurotypical-pretrained models with pathological data, and adapting neurotypical-pretrained models with speaker-specific data. Results showed that fine-tuning neurotypical-pretrained models with pathological data consistently yields the best performance for pathological speakers across both the predictive CR and generative SB models, with the SB model generally providing the strongest performance. Nevertheless, a clear performance gap between neurotypical and pathological speakers persists, likely due to the high variability across pathological speech. These findings highlight the importance of developing pathology-aware fine-tuning strategies that explicitly incorporate domain knowledge about speech disorders, enabling SE systems to be more inclusive and effective.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in [34, 35]. Ethical approval was not required as confirmed by the license attached with the data.

8. REFERENCES

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

[2] W. Zhang et al., "URGENT Challenge: Universality, Robustness, and Generalizability For Speech Enhancement," in *Proc. Interspeech*, 2024, pp. 4868–4872.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[4] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[5] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[6] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[7] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement," in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 764–768.

[8] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7009–7013.

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

[10] 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, Language Process.*, vol. 31, p. 3390–3403, Sep. 2023.

[11] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech Enhancement and Dereverberation With Diffusion-Based Generative Models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, 2023.

[12] Y.-J. Lu et al., "Conditional Diffusion Probabilistic Model for Speech Enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7402–7406.

[13] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A Study on Speech Enhancement Based on Diffusion Probabilistic Model," in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021.

[14] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proc. Interspeech*, 2024, pp. 1175–1179.

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

[16] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2724–2737, 2023.

[17] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Hearing Research*, vol. 12, no. 2, pp. 246–269, Jun. 1969.

[18] 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.

[19] ——, "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1210–1222, 2020.

[20] WHO, "Multi-country assessment of national capacity to provide hearing care," 2013.

[21] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, Nov. 2018.

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

[23] M. Hou and I. Kodrasi, "Variational autoencoder for personalized pathological speech enhancement," in *Proc. Eur. Signal Process. Conf.*, 2025, pp. 116–120.

[24] J. Richter, D. de Oliveira, and T. Gerkmann, "Investigating training objectives for generative speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 1–5.

[25] T. Gerkmann and E. Vincent, "Spectral Masking and Filtering," in *Audio Source Separation and Speech Enhancement*, 1st ed. Wiley, Sep. 2018, pp. 65–85.

[26] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. Int. Conf. Telecommun. Signal Process.*, 2020.

[27] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[28] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Neural Inf. Process. Syst.*, 2019.

[29] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2021.

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

[31] E. Schrödinger, "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique," in *Annales de l'institut Henri Poincaré*, vol. 2, no. 4, 1932, pp. 269–310.

[32] Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, "Schrodinger Bridges Beat Diffusion Models on Text-to-Speech Synthesis," *arXiv preprint arXiv:2312.03491*, Dec. 2023.

[33] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Diffusion-based speech enhancement in matched and mismatched conditions using a heun-based sampler," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 10 431–10 435.

[34] A. Guevara-Rukoz et al., "Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech," in *Proc. Int. Conf. Lang. Resour. Eval*, 2020, pp. 6504–6513.

[35] 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. Lang. Resour. Eval*, 2014, pp. 342–347.

[36] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[38] 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. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[39] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.