National Research University Higher School of Economics

as a manuscript

Pavel Andreev

GENERATIVE MODELS FOR SPEECH ENHANCEMENT

PhD Dissertation Summary

for the purpose of obtaining academic degree

Doctor of Philosophy in Computer Science

Moscow — 2024

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Anton S. Konushin, Doctor of Philosophy in Computer Science, National Research University Higher School of Economics.

1 Introduction

1.1 Topic of the Work

Speech recordings frequently suffer from background noise, reverberation, reduced frequency bandwidth, and other distortions, diminishing both their intelligibility and the listener's aesthetic satisfaction. Speech enhancement techniques are designed to restore perceptually plausible and intelligible clean speech from such corrupted signals. Speech enhancement can lower the technical requirements for recording equipment, enabling professionals to produce studio-quality recordings without sophisticated studio equipment. These models can be used to promote communication in acoustically contaminated environments and are particularly important for hearing assistance technologies for individuals with hearing impairments. Additionally, speech enhancement models play a critical role in creating high-quality datasets for deep learning systems, which rely on extensive datasets of clean speech for effective training. As a result, even publicly available data, irrespective of the initial recording conditions, can be improved to meet the quality standards necessary for advanced speech synthesis models. Therefore, speech enhancement techniques are crucial for a wide range of applications.

A more formal definition of the speech enhancement problem could be expressed in probabilistic formulation. Let $p(y)$ be a clean speech distribution, and let $p(x|y)$ be the degradation model. The problem of speech enhancement is to retrieve a sample from the conditional distribution $p(y|x)$, where $x \sim p(x|y)$ is a speech signal corrupted by the degradation model. The degradation model distribution $p(x|y)$ can include various forms of signal transformation, including background noise injection, reduction of frequency bandwidth, codec artifacts, reverberation, etc. Depending on the available resources and the particular application, one can distinguish several formulations of the speech enhancement problem:

- **Basic Speech Enhancement** [48, 32]: This setting involves supervised speech enhancement without latency constraints. In this formulation, the degradation model distribution is known in advance, and the model is not restricted to be streaming.
- **Streaming Speech Enhancement** [19, 10]: In this scenario, it is required to build speech enhancement models with a limited algorithmic delay, i.e., causal models. Streaming speech enhancement enforces the model to use only a limited window of future information, typically ranging from 3-8 ms (low latency) to 60 ms (high latency).
- **Unsupervised Speech Enhancement** [30, 35]: This formulation does

not assume the degradation model $p(x|y)$ to be known in advance. The model is adapted to the particular degradation model only at the inference stage.

The development of efficient generative models for speech enhancement in the context of all of these formulations is **the topic of this work**.

1.2 Relevance

Historically, the field of audio processing has been dominated by methods that rely on handcrafted heuristics and statistical models, often employing unrealistic assumptions about the structure of speech and disturbances [14, 13]. However, the rise of machine learning and, more specifically, deep learning, has marked a paradigm shift towards leveraging data-driven methodologies. In contrast to traditional approaches, the data-driven paradigm learns the characteristics of the signals directly from the data. This approach has been shown to be beneficial in many domains, including speech processing.

Initial attempts to apply deep learning methods to the speech enhancement problem were based on treating this problem as a predictive problem [10, 16, 5, 19]. Following the principle of empirical risk minimization, the goal of predictive modeling is to find a model with minimal average error over the training data. Given a noisy waveform or spectrogram, these approaches try to predict the clean signal by minimizing point-wise distance in waveform or spectral domains, or jointly in both domains, thus treating this problem as a predictive task.

However, given severe degradations applied to the signal, there is an inherent uncertainty in the restoration of the speech signal (i.e., given the degraded signal, the clean signal is not restored unambiguously), which often leads to oversmoothing of the predicted speech. From the probabilistic point of view, minimization of the point-wise distance leads to an averaging effect. For example, optimization of the mean squared error between waveforms delivers the expectation of the waveform over the conditional distribution of clean speech given its degraded version. The key issue is that the expectation over this distribution is not guaranteed to lie within this distribution.

An illustrative example of this phenomenon and its impact on speech enhancement is shown in Figure 1. The model is trained to extend the frequency bandwidth of the speech signal given the signal with reduced bandwidth by minimizing the mean squared error distance between clean and generated waveforms [3]. Notably, the model is not able to restore high-frequency content while minimizing the MSE objective. Due to high uncertainty, the model oversmooths the high frequencies, being unable to restore speech content above 5

kHz. A similar effect occurs with other point-wise losses, including spectral-based losses.

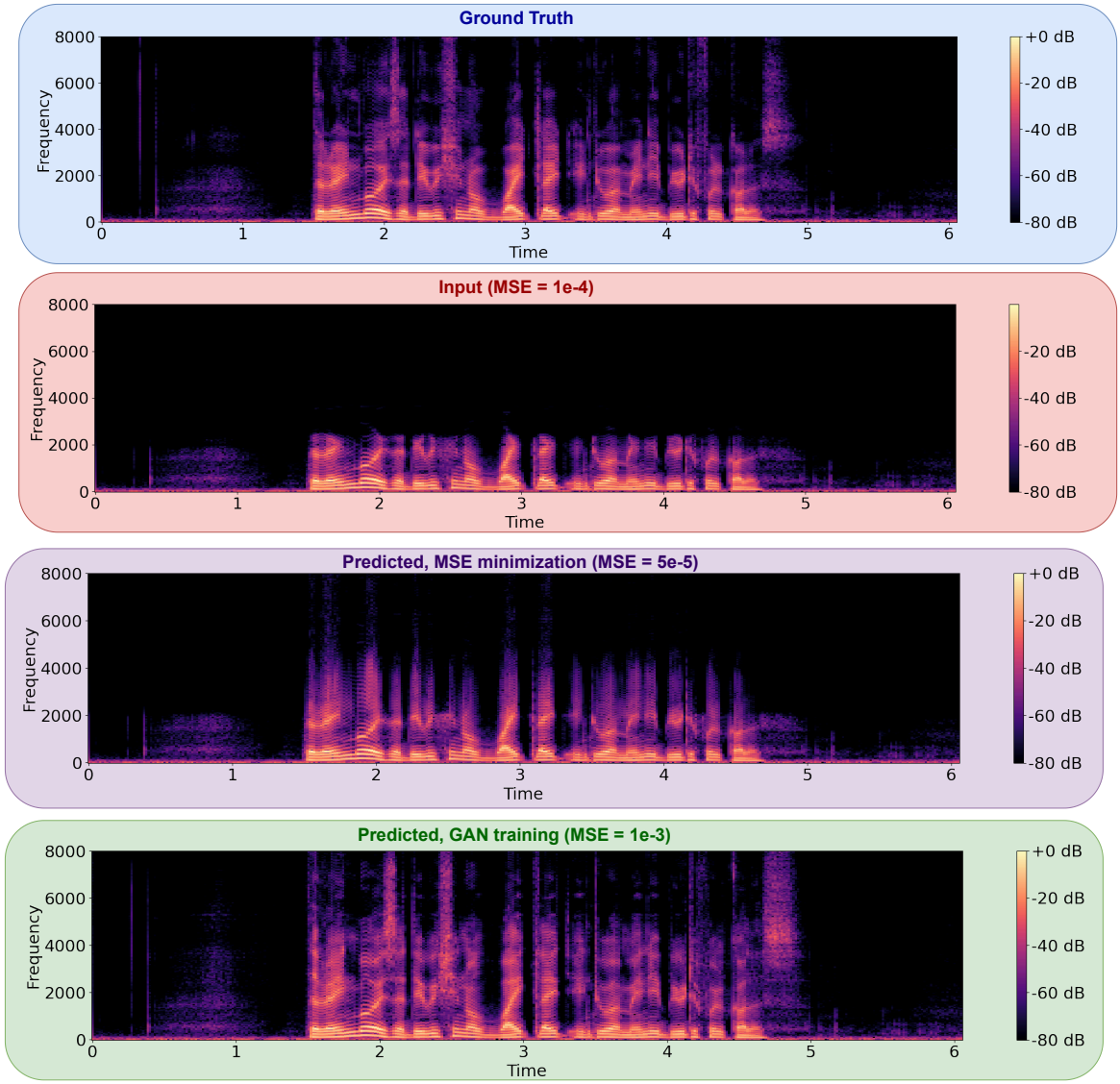

Figure 1: Example of speech signal spectrograms. Mean squared error minimization leads to oversmoothing of the predicted signal, resulting in missing high-frequency content. While the prediction provided by the generative model delivers MSE even higher than the input, the predicted signal resembles the original speech content. A similar effect is reported for image super-resolution models [27].

Unlike predictive models, generative models aim at sampling from the clean speech distribution conditioned on the degraded signal rather than minimizing point-wise loss. The advantage of this approach is that the speech enhancement model is enforced to produce a signal lying within the clean speech distribution, as illustrated in Figure 1. **This work studies generative models in**

the context of speech enhancement, proposes novel methods, and improves the efficiency and quality of existing solutions.

In the first part of this work, we focus on developing efficient solutions for basic speech enhancement [48, 32]. The basic speech enhancement formulation does not impose constraints on the available signal context, and the degradation model is assumed to be known during training. We, firstly, argue that the speech enhancement problem does not necessitate the model to learn the complete conditional distribution $p(y|x)$, instead the focus is on capturing the mode of this distribution.

We show that the GAN framework is naturally suited for this formulation of the speech enhancement problem since it tends to retrieve the main mode of the distribution—precisely what speech enhancement should typically do. Therefore, in this work, we employ the GAN framework for the basic speech enhancement formulation and design efficient architectures of generator and discriminator neural networks.

In the second part of the work, we focus on streaming speech enhancement [19, 10, 45, 51], in particular, on low-latency speech enhancement [45, 51]. Streaming models are an essential component of real-time speech enhancement tools. The streaming regime constrains speech enhancement models to use only a tiny context of future information. As a result, the low-latency streaming setup is generally considered a challenging task and has a significant negative impact on the model's quality.

However, the sequential nature of streaming generation offers a natural possibility for autoregression, that is, utilizing previous predictions while making current ones. The conventional method for training autoregressive generative models is teacher forcing, but its primary drawback lies in the training-inference mismatch that can lead to a substantial degradation in quality. We propose a straightforward yet effective alternative technique for training autoregressive low-latency speech enhancement models. We demonstrate that the proposed approach leads to stable improvement across diverse architectures and training scenarios.

Lastly, we focus on the unsupervised speech enhancement problem [30, 35]. We introduce a diffusion probabilistic model capable of solving various speech inverse tasks with unknown degradation models during training. Once trained for speech waveform generation in an unconditional manner, it can be adapted to different tasks including degradation inversion and neural vocoding.

1.3 Key Results and Conclusions

1.3.1 Contributions

The main contributions of this work can be summarized as follows:

1. We propose the HiFi++ composite generator architecture by combining the HiFi-GAN generator with three new modules: SpectralUnet, WaveUNet, and SpectralMaskNet. This new generator architecture allows building a unified framework for bandwidth extension and speech enhancement, delivering state-of-the-art results in these tasks.

2. We design a novel architecture for direct spectrogram estimation based on the fast Fourier convolution operator. The architecture allows direct manipulation with cepstrum features and further improves HiFi++ results on speech enhancement problems while being more parameter-efficient.

3. We investigate various feature extractors as backbones for speech perceptual loss and introduce criteria for selecting an extractor based on the structure of its feature space. The effectiveness of these criteria is validated by empirical results.

4. Based on these developments, we develop a novel universal speech enhancement model, FINALLY, which achieves state-of-the-art performance, outperforming all existing solutions while being more computationally efficient.

5. We propose a novel method for training autoregressive models for low-latency streaming speech enhancement. The method allows mitigating training-inference mismatch arising during training with teacher forcing. The model allows a considerable improvement in streaming speech enhancement models with autoregressive conditioning.

6. We investigate a diffusion-based technique for unsupervised speech enhancement. The proposed unconditional diffusion model can be trained for the unconditional speech generation task and then be adapted for various speech restoration tasks without additional training.

1.3.2 Theoretical and Practical Significance

This work theoretically shows that adversarial training can be used for implicit regression for the main mode of the distribution, making it a suitable tool for learning speech enhancement models. It also studies the structural properties of different speech feature extractor spaces and formulates a new perceptual loss.

Additionally, the work proposes new neural architectures, HiFi++ and FFC-SE, for deep generative models, improving the quality and computational efficiency of speech enhancement solutions. Based on these developments, the work proposes a highly efficient speech enhancement algorithm, FINALLY,

which achieves state-of-the-art quality with significantly fewer computational resources than prior methods.

The work also outlines a novel method for training autoregressive models in situations with high training-inference mismatch, significantly improving upon the conventional teacher forcing technique. The proposed iterative autoregression method holds significant practical novelty due to the widespread usage of autoregressive models nowadays.

Furthermore, the work studies the problem of unsupervised speech enhancement and proposes a novel diffusion generative model, Undiff, for unsupervised speech restoration. This work provides pioneering developments in the unsupervised speech enhancement problem.

1.3.3 Key Aspects/Ideas to be Defended

1. HiFi++ architecture for multi-domain signal processing in speech enhancement.
2. FFC-SE archietcure for direct complex spectrogram estimation.
3. FINALLY model for universal speech enhancement.
4. Iterative autoregression technique for mitigation of training-inference mismatch within autoregressive models, studied in application to low-latency speech enhancement.
5. Undiff diffusion probabilistic model for unsupervised speech enhancement.

1.3.4 Personal Contribution

The idea of HiFi++ and FFC-SE generator architectures was proposed by the author of this work. The initial implementation of the HiFi++ architecture was done by the author while Aibek Alanov and Oleg Ivanov helped to refine and prepare the codebase for experiments. The FFC-SE network was jointly developed with Ivan Shcheckotov. The experiments for validation of the networks' effectiveness were designed by the author. The implementation and paper writing were done jointly with Aibek Alanov, Oleg Ivanov, and Ivan Shcheckotov. Dmitry Vetrov provided scientific guidance for this work.

The proof of mode-seeking LS-GAN behavior and formulation of the speech enhancement problem as a mode-finding problem were developed by the author of the thesis. The FINALLY model was developed together with Kirill Tamogashev and Nicholas Babaev. The author was responsible for scientific guidance, experiment planning, and code review.

The iterative autoregression technique was proposed and theoretically studied by the author of this work. The actual implementation and experimental validation were done jointly with Nicholas Babaev. The paper was written by the author with some assistance from Nicholas Babaev and Aibek Alanov.

The Undiff generative model was designed and implemented jointly with Anastasia Yaschenko and Ivan Shcheckotov. Dmitry Vetrov provided scientific guidance for this work.

1.4 Publications and Probation of the Work

The results of this thesis are published in 3 first-tier publications and 1 second-tier publication. The PhD candidate is the main author in all of these articles[1].

1.4.1 First-Tier Publications

- **Andreev, P.***, Babaev, N.*, Saginbaev, A., Shchekotov, I., Alanov, A. (2023). Iterative autoregression: a novel trick to improve your low-latency speech enhancement model. *Proc. INTERSPEECH 2023*, 2448-2452, doi: 10.21437/Interspeech.2023-365 (Core A)

- Shchekotov, I.*, **Andreev, P.***, Ivanov, O., Alanov, A., Vetrov, D. (2022). FFC-SE: Fast Fourier Convolution for Speech Enhancement. *Proc. INTERSPEECH 2022*, 1188-1192, doi: 10.21437/Interspeech.2022-603 (Core A)

- Iashchenko, A.*, **Andreev, P.***, Shchekotov, I.*, Babaev, N., Vetrov, D. (2023). UnDiff: Unsupervised Voice Restoration with Unconditional Diffusion Model. *Proc. INTERSPEECH 2023*, 4294-4298, doi: 10.21437/Interspeech.2023-367 (Core A)

1.4.2 Second-Tier Publications

- **P. Andreev***, A. Alanov, O. Ivanov* and D. Vetrov, "HIFI++: A Unified Framework for Bandwidth Extension and Speech Enhancement," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097255. (Core B)

1.4.3 Reports at Scientific Conferences

- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 8, 2023. Topic: "HIFI++: A

[1] * indicates equal contribution

Unified Framework for Bandwidth Extension and Speech Enhancement"

- 24th INTERSPEECH Conference, Dublin, Ireland, August 22, 2023. Topic: "Iterative autoregression: a novel trick to improve your low-latency speech enhancement model"
- 24th INTERSPEECH Conference, Dublin, Ireland, August 23, 2023. Topic: "UnDiff: Unsupervised Voice Restoration with Unconditional Diffusion Model"
- 23rd INTERSPEECH Conference, Incheon, Korea, September 20, 2022. Topic: "FFC-SE: Fast Fourier Convolution for Speech Enhancement"

1.4.4 Volume and Structure of the Work

The thesis contains an introduction chapter, which formulates the topic of this work, a background chapter, which introduces the necessary context, 3 content chapters, which describe the approaches developed for each of the introduced formulations, and a conclusion chapter, which summarizes the developments of this work and concludes the study. The full volume of the thesis is 102 pages.

2 Content of the Work

In the Background chapter, we introduce basic degradation models and metrics. We also provide a literature review describing deep learning-based speech enhancement approaches from prior work.

In the three subsequent chapters, we elaborate the developments for the introduced formulations of the speech enhancement problem:

1. **Basic speech enhancement** is discussed in chapter three, "Generative Models for Basic Speech Enhancement." This chapter first refines the probabilistic formulation of the speech enhancement problem by taking a closer look at its practical goal. We argue that a speech enhancement model should retrieve the most probable reconstruction of the clean speech given the degraded version. Given the refined formulation, we show that a GAN-based training framework naturally suits this goal by encouraging the generator to retrieve the mode of the conditional distribution. The chapter then introduces practical developments for the architecture of neural networks and perceptual losses to guide training to proper solutions.

2. **Streaming speech enhancement** is discussed in the fourth chapter, "Iterative Autoregression for Streaming Speech Enhancement." In this

chapter, we argue that the streaming regime provides a natural possibility for autoregressive conditioning of speech enhancement models. We show that the conventional teacher forcing algorithm leads to high training-inference mismatch and introduce a novel training algorithm that mitigates this problem.

3. **Unsupervised speech enhancement** is discussed in the fifth chapter, "Unsupervised Speech Enhancement with Unconditional Diffusion Model." This chapter introduces UnDiff, a diffusion probabilistic model capable of solving various speech inverse tasks. Once trained for speech waveform generation in an unconditional manner, it can be adapted to different tasks of speech restoration without additional training. We first tackle the challenging problem of unconditional waveform generation by comparing different neural architectures and preconditioning domains. After that, we demonstrate how the trained unconditional diffusion model could be adapted to different tasks of speech processing by means of recent developments in post-training conditioning of diffusion models.

2.1 Generative Models for Basic Speech Enhancement

2.1.1 GANs for Speech Enhancement

We first concretize the probabilistic formulation of the speech enhancement problem by taking a closer look at its practical goal. The practical goal of a speech enhancement model is to restore the audio signal containing the speech characteristics of the original recording, including the voice, linguistic content, and prosody. Thus, the speech enhancement model should not generate new speech content but rather "refine" existing speech as if it were recorded in ideal conditions (studio-like quality). From a mathematical point of view, this means that the speech enhancement model should retrieve the most probable reconstruction of the clean speech y given the corrupted version x, i.e., $y = \arg\max_y p_{\text{clean}}(y|x)$.

Given this formulation, we argue that the framework of generative adversarial networks (GANs) is more naturally suited for the speech enhancement problem than diffusion models. We show that GAN training naturally leads to the mode-seeking behavior of the generator, which aligns with the formulation introduced above.

Let $p_g(y|x)$ be a family of waveform distributions produced by the generator $g_\theta(x)$. Mao et al. [33] showed that training with Least Squares GAN (LS-GAN) leads to the minimization of the Pearson χ^2 divergence $\chi^2_{\text{Pearson}}(p_g \| (p_{\text{clean}} + p_g)/2)$. We propose that if $p_g(y|x)$ approaches $\delta(y - g_\theta(x))$ under some parametrization, the minimization of this divergence leads to $g_\theta(x) = \arg\max_y p_{\text{clean}}(y|x)$.

This means that if the generator deterministically predicts the clean waveform from the degraded signal, the LS-GAN loss encourages the generator to predict the point of maximum $p_{\text{clean}}(y|x)$ density. We note that although prior work [29] demonstrated the mode-covering property for the optimization of Pearson χ^2 divergence, our result pertains to a deterministic generator setting, which is outside the scope of analysis provided by Li et al. [29].

Proposition 1. *Let $p_{clean}(y|x) > 0$ be a finite and Lipschitz continuous density function with a unique global maximum and $p_g^{\xi}(y|x) = \xi^n/2^n \cdot \mathbf{1}_{y-g_\theta(x)\in[-1/\xi,1/\xi]^n}$, then*

$$\lim_{\xi\to+\infty} \arg\min_{g_\theta(x)} \chi^2_{Pearson}(p_g^{\xi}||(p_{clean} + p_g^{\xi})/2) = \arg\max_y p_{clean}(y|x) \tag{1}$$

Thus, LS-GAN training under ideal conditions should lead to the solution $g_\theta(x) = \arg\max_y p_{\text{clean}}(y|x)$ for the generator. In practice, however, success is highly dependent on technicalities, such as additional losses to stabilize training and architectures of neural networks. In the following sections, we approach these problems by designing new architectures and training losses.

2.1.2 HiFi++: a Unified Framework for Bandwidth Extension and Speech Enhancement

We propose a novel HiFi++ architecture that adapts HiFi generator [21] to the speech enhancement problem by introducing new modules: SpectralUNet, WaveUNet and SpectralMaskNet (see Figure 2). The HiFi++ generator is based on the HiFi part that takes as an input the enriched mel-spectrogram representation by the SpectralUNet and its output goes through postprocessing modules: WaveUNet corrects the output waveform in time domain while SpectralMaskNet cleans up it in frequency domain. We describe the introduced modules in details in the next paragraphs.

SpectralUNet We introduce the SpectralUNet module as the initial part of the HiFi++ generator that takes the input mel-spectogram (see Figure 2). The mel-spectrogram has a two-dimensional structure and the two-dimensional convolutional blocks of the SpectralUnet model are designed to facilitate the work with this structure at the initial stage of converting the mel-spectrogram into a waveform. The idea is to simplify the task for the remaining part of the HiFi++ generator that should transform this 2d representation to the 1d sequence. We design the SpectralUNet module as UNet-like architecture with 2d convolutions. This module also can be considered as the preprocess part that prepares the input mel-spectrogram by correcting and extracting from it the essential information that is required for the desired task.

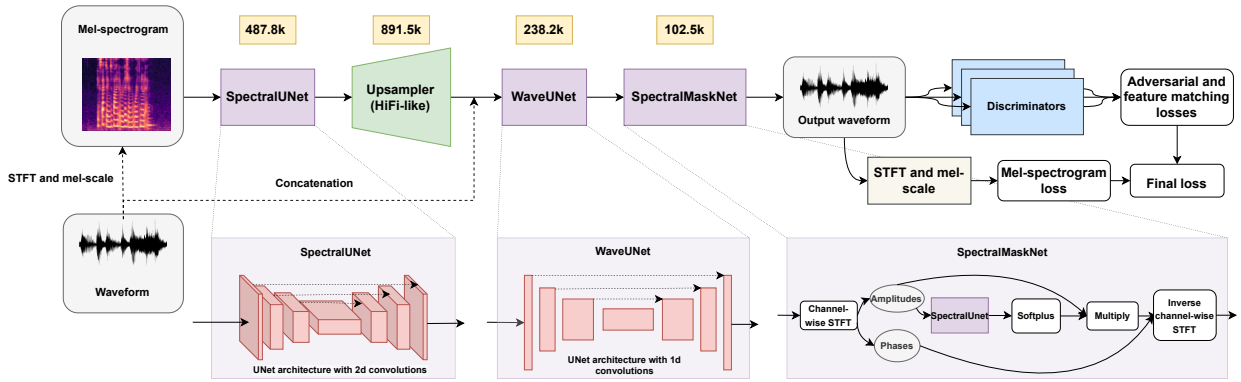

Figure 2: HiFi++ architecture and training pipeline. The HiFi++ generator consists of the HiFi-like Upsampler and three introduced modules SpectralUNet, WaveUNet and SpectralMaskNet (their sizes are in yellow boxes). The generator's architecture is identical for BWE and SE.

WaveUNet The WaveUNet module is placed after the HiFi part (Upsampler) and takes several 1d sequences concatenated with the input waveform as an input. This module operates directly on time domain and it can be considered as a time domain postprocessing mechanism that improves the output of the Upsampler and merges the predicted waveform with the source one. The WaveUNet module is an instance of the well-known architecture Wave-U-Net [42] which is a fully convolutional 1D-UNet-like neural network. This module outputs the 2d tensor which consists of m 1d sequences that will be processed and merged to the output waveform by the next SpectralMaskNet module.

SpectralMaskNet We introduce the SpectralMaskNet as the final part of the generator which is a learnable spectral masking. It takes as an input the 2d tensor of m 1d sequences and applies channel-wise short-time Fourier transform (STFT) to this 2d tensor. Further, the SpectralUNet-like network takes the amplitudes of the STFT output to predict multiplicative factors for these amplitudes. The concluding part consists of the inverse STFT of the modified spectrum (see Figure 2). Importantly, this process does not change phases. The purpose of this module is to perform frequency-domain postprocessing of the signal. We hypothesize that it is an efficient mechanism to remove artifacts and noise in frequency domain from the output waveform in a learnable way.

Training We use the multi-discriminator adversarial training framework for the time-domain models' training. It consists of three losses, namely LS-GAN loss $\mathcal{L}_{GAN}$ [33], feature matching loss $\mathcal{L}_{FM}$ [25, 24], and mel-spectrogram loss

$\mathcal{L}_{Mel}$ [21]:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \tag{2}$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \ldots, k. \tag{3}$$

where $\mathcal{L}(\theta)$ denotes loss for generator with parameters θ, $\mathcal{L}(\varphi_i)$ denotes loss for i-th discriminator with parameters φ_i (all discriminators are identical, except initialized differently).

Experiments The comparison of the HiFi++ with baselines is demonstrated in the Table 1. Our model achieves comparable performance with VoiceFixer [31] and DEMUCS [10] counterparts while being drastically smaller. Interestingly, VoiceFixer achieves high subjective quality while being inferior to other models according to objective metrics, especially to SI-SDR and STOI. Indeed, VoiceFixer doesn't use waveform information directly and takes as input only mel-spectrogram, thus, it misses parts of the input signal and is not aiming at reconstructing the original signal precisely leading to poor performance in terms of classic relative metrics such as SI-SDR, STOI, and PESQ. Our model provides decent relative quality metrics as it explicitly uses raw signal waveform as model inputs. At the same time, our model takes into account signal spectrum, which is very informative in speech enhancement as was illustrated by the success of classical spectral-based methods. It is noteworthy that we significantly outperform the SEANet [44] model, which is trained in a similar adversarial manner and has a larger number of parameters, but does not take into account spectral information.

Table 1: Speech denoising results on Voicebank-DEMAND dataset. * indicates re-implementation.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	DNSMOS	# Par (M)
Ground truth	4.60 ± 0.03	4.50	-	1.00	4.64	3.15	-
HiFi++ (ours)	$\mathbf{4.33 \pm 0.06}$	4.27	18.4	**0.95**	2.76	3.10	**1.7**
VoiceFixer	$\mathbf{4.32 \pm 0.05}$	4.14	-18.5	0.89	2.38	3.13	122.1
DEMUCS	4.22 ± 0.05	**4.37**	**18.5**	**0.95**	3.03	**3.14**	60.8
MetricGAN+	4.01 ± 0.09	3.90	8.5	0.93	**3.13**	2.95	2.7
*SEANet	3.99 ± 0.09	4.19	13.5	0.92	2.36	3.05	9.2
*SE-Conformer	3.39 ± 0.09	3.88	15.8	0.91	2.16	2.85	1.8
Input	3.36 ± 0.06	2.99	8.4	0.92	1.97	2.53	-

2.1.3 FFC-SE: Fast Fourier Convolution for Speech Enhancement

We further improve the result of HiFi++ by proposing new neural architectures based on the fast Fourier convolution (FFC) operator [7], which we adapt for speech enhancement problems. The FFC layers were originally proposed for computer vision tasks as a non-local operator replacing vanilla convolutional layers within existing neural networks. Fast Fourier convolution has a global receptive field and was shown to be helpful for the restoration of periodic backgrounds in inpainting problems [43]. These properties of FFC are especially helpful for the complex spectrum prediction. Indeed, the harmonics of spectrograms are known to form periodic structures that can be naturally handled by fast Fourier convolution (see Figure 3). Besides, we experimentally observe that a large receptive field of FFC is useful for producing coherent phases.

Based on these insights, we design new neural architectures for direct complex-valued spectrogram estimation in speech enhancement problems. The proposed models achieve state-of-the-art performance on VoiceBank-DEMAND [46] and Deep Noise Suppression [12] datasets with much fewer parameters than the baselines.

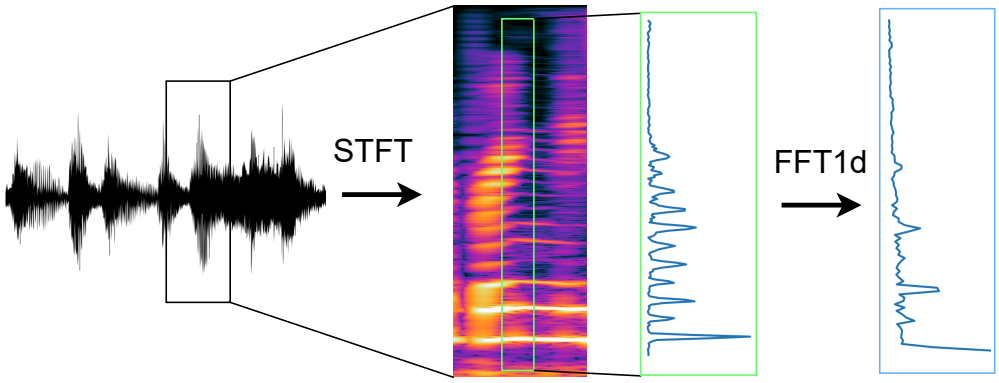

Figure 3: Harmonics of short-time Fourier transform constitute periodic structures which can be naturally processed in the Fourier domain by the global branch of fast Fourier convolution.

Fast Fourier Convolution Fast Fourier Convolution (FFC) [7] is a neural operator that allows performing non-local reasoning and generation within a neural network. FFC uses a channel-wise fast Fourier transform [36], followed by a point-wise convolution and inverse Fourier transform, thus globally affecting the input tensor across dimensions involved in the Fourier transform. FFC splits channels into local and global branches. The local branch uses conventional convolutions for local updates of feature maps, while the global branch performs a Fourier transform of the feature map and updates it in the spectral domain, affecting global context.

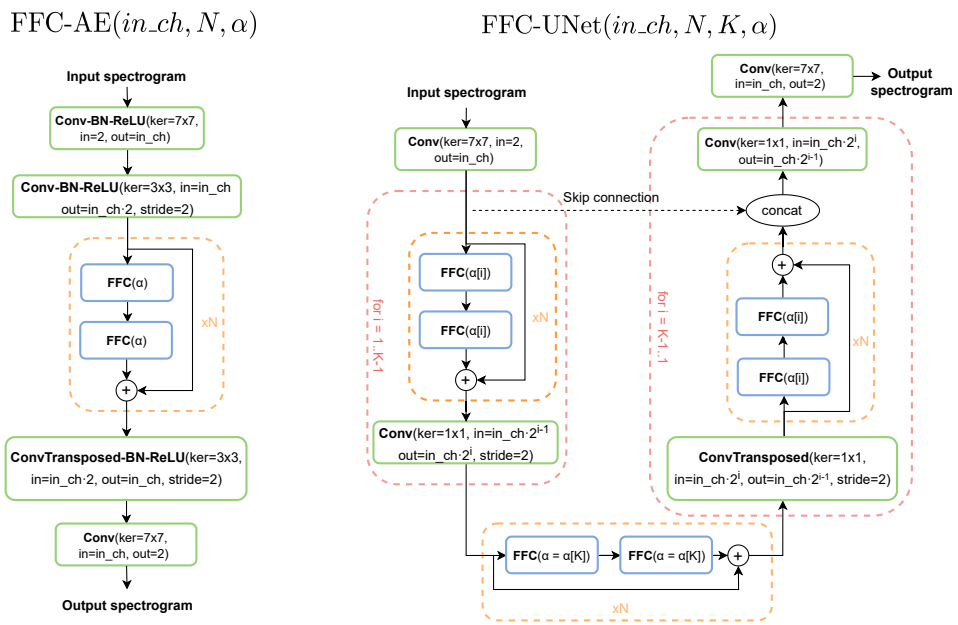

Figure 4: Proposed architectures for speech enhancement. *Left:* fast Fourier convolutional autoencoder which adopts architecture introduced in [43] for speech enhancement task. *Right:* fast Fourier convolutional U-Net.

Architectures We implement two neural network architectures for speech enhancement. The first one (FFC-AE) is inspired by [43]. This architecture consists of a convolutional encoder (strided convolution) which downsamples the input STFT representation across time and frequency dimensions by a factor of two. The encoder is followed by a series of residual blocks, each consisting of two sequential fast Fourier convolution modules. The output of the residual blocks is then upsampled by transposed convolution and used to predict the real and imaginary parts of the denoised complex-valued spectrogram. The architecture is depicted in Figure 4 (left). We call this model the fast Fourier convolutional autoencoder (FFC-AE). The second architecture is inspired by the classic work [38]. We incorporate FFC layers into the U-Net architecture as shown in Figure 4 (right). At each level of the U-Net structure, we utilize several residual FFC blocks with convolutional upsampling or downsampling.

Experiments We compare the quality of the proposed models with strong baselines. On Voicebank-DEMAND, as it can be seen from Table 2, our models significantly outperform all the baselines by MOS and give competitive results on objective metrics. Overall, we observe that neural architectures built upon fast Fourier convolution significantly outperform vanilla convolution-based architectures in terms of quality of speech enhancement, phase estimation and parameter efficiency. In general, the proposed architectures deliver state-of-art results on speech denoising benchmarks, being significantly smaller than the baselines.

Table 2: Speech denoising results on Voicebank-DEMAND dataset. Best three results are highlighted in bold.

Model	MOS	WV-MOS	SI-SDR	PESQ	# Params (M)
Ground Truth	4.46 ± 0.06	4.50	-	4.64	-
Input	3.44 ± 0.06	2.99	8.4	1.97	-
MetricGAN+ [15]	3.82 ± 0.06	3.90	8.5	**3.13**	2.7
ResUNet-Decouple+ [22]	3.94 ± 0.04	4.13	**18.4**	2.45	102.6
DEMUCS (non-caus.) [10]	4.06 ± 0.03	**4.37**	**18.5**	**3.03**	60.8
VoiceFixer [31]	4.10 ± 0.03	4.14	-18.5	2.38	122.1
HiFi++	4.15 ± 0.07	4.27	**18.4**	2.76	**1.7**
FFC-AE-V0 (ours)	$\mathbf{4.24 \pm 0.09}$	4.34	17.9	2.88	**0.42**
FFC-AE-V1 (ours)	$\mathbf{4.33 \pm 0.03}$	**4.37**	17.5	2.96	**1.7**
FFC-UNet (ours)	$\mathbf{4.28 \pm 0.03}$	**4.38**	18.1	**2.99**	7.7
FFC-AE-V1 (abl.)	3.98 ± 0.07	4.05	16.7	2.68	2.9
vanilla UNet	4.10 ± 0.07	4.11	17.2	2.73	20.7

2.1.4 FINALLY: Fast and Universal Speech enhancement with Studio-like Quality

The previously discussed approaches deliver impressive results on simulated data contaminated with additive noise. However, in practice, real-life recordings are often contaminated with several distortions at the same time. We found that models trained to eliminate additive noise generalize poorly to real data. Therefore, we revisit the HiFi++ framework for speech enhancement and demonstrate that it provides rapid and high-quality universal speech enhancement, i.e., one model could be used to reverse several degradations simultaneously. Our model outperforms both diffusion models and previous GAN-based models, achieving an unprecedented level of quality on both simulated and real-world data.

To achieve this, we investigate various feature extractors as backbones for perceptual loss and propose criteria for selecting an extractor based on the structure of its feature space. These criteria are validated by empirical results from a neural vocoding task, indicating that the convolutional features of the WavLM neural network are well-suited for perceptual loss in speech enhancement. We also develop a novel model for universal speech enhancement that integrates the proposed perceptual loss with MS-STFT discriminator training and enhances the architecture of the HiFi++ generator [1] by combining it with a self-supervised pre-trained WavLM encoder [6].

Architecture We introduce two modifications to the HiFi++ generator's architecture (see Figure 5). First, we modify the generator by incorporating WavLM-large model output (last hidden state of the transformer) as an ad-

ditional input to the Upsampler. Prior works [18, 4] have demonstrated the usefulness of Self-Supervised Learning (SSL) features for speech enhancement tasks, and we validate this by observing significant performance gains from using SSL features. Second, we introduce the Upsample WaveUNet at the end of the generator. This allows the model to output a 48 kHz signal while taking a 16 kHz signal as input.

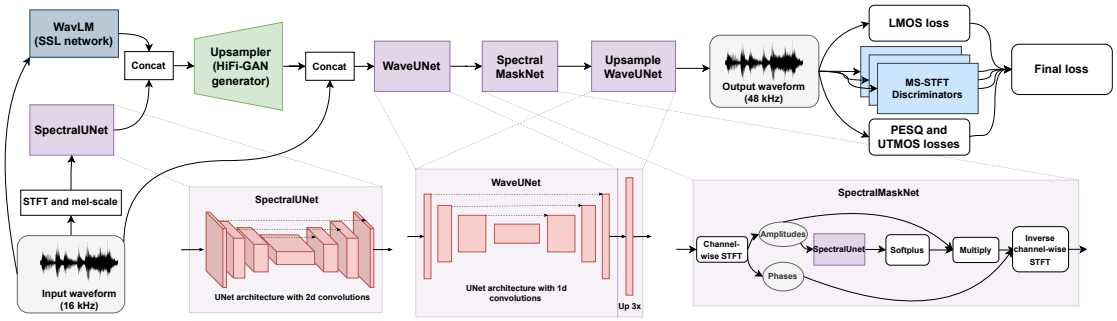

Figure 5: FINALLY model architecture.

Training During training, before mixing with noise, we convolve the speech signal with a randomly chosen microphone impulse response and apply other digital distortions. We train the model in three stages. The first two stages concentrate on restoring the original speech content, and the final stage aims to enhance the aesthetic perception of the speech. The loss functions that we use can be written as follows:

$$\mathcal{L}_{\text{LMOS}}(\theta) = \mathop{\mathbb{E}}_{x,y\sim p(x,y)} \left[100 \cdot \|\phi(y) - \phi(g_\theta(x))\|_2^2 + \||\text{STFT}(y)| - |\text{STFT}(g_\theta(x))|\|_1\right], \quad (4)$$

$$\mathcal{L}_{\text{gen}}(\theta) = \overbrace{\underbrace{\lambda_{\text{LMOS}} \cdot \mathcal{L}_{\text{LMOS}}(\theta)}_{\text{1st stage (16 kHz)}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN-gen}}(\theta) + \lambda_{\text{FM}} \cdot \mathcal{L}_{\text{FM}}(\theta)}^{\text{2nd stage (16 kHz)}} + \lambda_{\text{HF}} \cdot \mathcal{L}_{\text{HF}}(\theta), \quad (5)$$

$$\text{3rd stage (48 kHz)}$$

$$\mathcal{L}_{\text{disc}}(\varphi_i) = \mathcal{L}_{\text{GAN-disc}}(\varphi_i), \quad i = 1, \ldots, k. \quad (6)$$

Here, ϕ denotes the WavLM-conv feature mapping, $g_\theta(x)$ denotes the generator neural network with parameters θ, $\mathcal{L}_{\text{GAN-gen}}(\theta)$ denotes the LS-GAN generator loss [33], $\mathcal{L}(\theta)$ denotes the combined generator loss, $\mathcal{L}_{\text{GAN-disc}}(\varphi_i)$ denotes the LS-GAN discriminator [33] loss for the i-th discriminator with parameters φ_i, $\mathcal{L}_{\text{FM}}$ denotes the feature matching loss [24, 11], $\mathcal{L}_{\text{HF}}$ denotes the human feedback loss, and λ_* denotes the corresponding loss weights.

Table 3: Comparison with prior work.

VoxCeleb (HiFi-GAN-2 validation set, real data)						
Model	MOS	DNSMOS	UTMOS	WV-MOS	-	RTF
Input	3.46 ± 0.07	2.72 ± 0.11	2.76 ± 0.13	2.90 ± 0.16	-	-
VoiceFixer	3.41 ± 0.07	3.08 ± 0.06	2.60 ± 0.09	2.79 ± 0.09	-	0.02
DEMUCS	3.79 ± 0.07	3.27 ± 0.04	3.51 ± 0.08	3.72 ± 0.08	-	0.08
STORM	3.75 ± 0.06	3.17 ± 0.04	3.29 ± 0.08	3.54 ± 0.09	-	1.05
BBED	3.97 ± 0.06	3.23 ± 0.04	3.30 ± 0.10	3.47 ± 0.08	-	0.43
HiFi-GAN-2	4.47 ± 0.05	**3.32 ± 0.03**	3.67 ± 0.09	**3.96 ± 0.06**	-	0.50
Ours	**4.63 ± 0.04**	**3.31 ± 0.04**	**4.05 ± 0.07**	**3.98 ± 0.06**	-	**0.03**
UNIVERSE validation set (simulated data)						
Model	MOS	DNSMOS	UTMOS	WV-MOS	PhER	RTF
Input	2.87 ± 0.05	2.25 ± 0.19	2.27 ± 0.28	1.72 ± 0.61	0.31 ± 0.05	-
Ground Truth	4.39 ± 0.05	3.33 ± 0.04	4.26 ± 0.06	4.28 ± 0.06	0	-
UNIVERSE	4.10 ± 0.07	**3.23 ± 0.07**	3.89 ± 0.15	3.85 ± 0.12	0.20 ± 0.04	0.5
Ours (16 kHz)	3.99 ± 0.07	**3.24 ± 0.05**	**4.21 ± 0.08**	**4.43 ± 0.07**	**0.14 ± 0.03**	**0.03**
Ours	**4.23 ± 0.07**	**3.25 ± 0.05**	**4.21 ± 0.10**	**4.43 ± 0.08**	**0.14 ± 0.03**	**0.03**

Experiments We consider BBED [26], STORM [28], and UNIVERSE [40] diffusion models, along with Voicefixer and DEMUCS regression models, as our baselines. In addition, we consider our closest competitor, HiFi-GAN-2, as a GAN-based baseline. The data for comparison with HiFi-GAN-2 and UNIVERSE were taken from their demo pages since the authors did not release any code. We conduct comparisons with BBED, STORM, Voicefixer, DEMUCS, and HiFi-GAN-2 on real-world VoxCeleb1 samples, and the comparison with UNIVERSE on the simulated data provided by the authors of this work. We also provide results for our predictions resampled to 16 kHz, in addition to the base predictions at 48 kHz, since the UNIVERSE model outputs only 16 kHz tracks. The comparison is outlined in Table 3.

By integrating WavLM-based perceptual loss into the MS-STFT adversarial training pipeline and enhancing the HiFi++ architecture with a WavLM encoder, we develop a novel speech enhancement model, FINALLY, which achieves state-of-the-art performance, producing clear and high-quality speech at 48 kHz.

2.2 Iterative Autoregression for Streaming Speech Enhancement

The nature of streaming generation follows a sequential pattern that lends itself well to autoregression. The conventional approach for training autoregressive models is through "teacher forcing" [47], whereby the model is presented with past ground-truth samples to predict the subsequent ones during training. During the inference stage, the model utilizes its own samples for autoregressive

conditioning (free-running mode) since ground-truth is not available. Teacher forcing is an efficient means of training and convergence as it can be parallelized effectively for convolution-based networks. However, its primary limitation is the mismatch between training and inference which can result in a significant degradation in quality during the test phase. We observe that autoregressive speech enhancement models rely heavily on ground-truth conditioning and are, therefore, particularly vulnerable to training-inference mismatch. The common approaches address this mismatch by utilizing free-running mode during training [2]. However, these methods are typically used for recurrent networks operating on low resolution features as their application to convolution-based networks on high resolution features substantially slows down training, which hampers practical application.

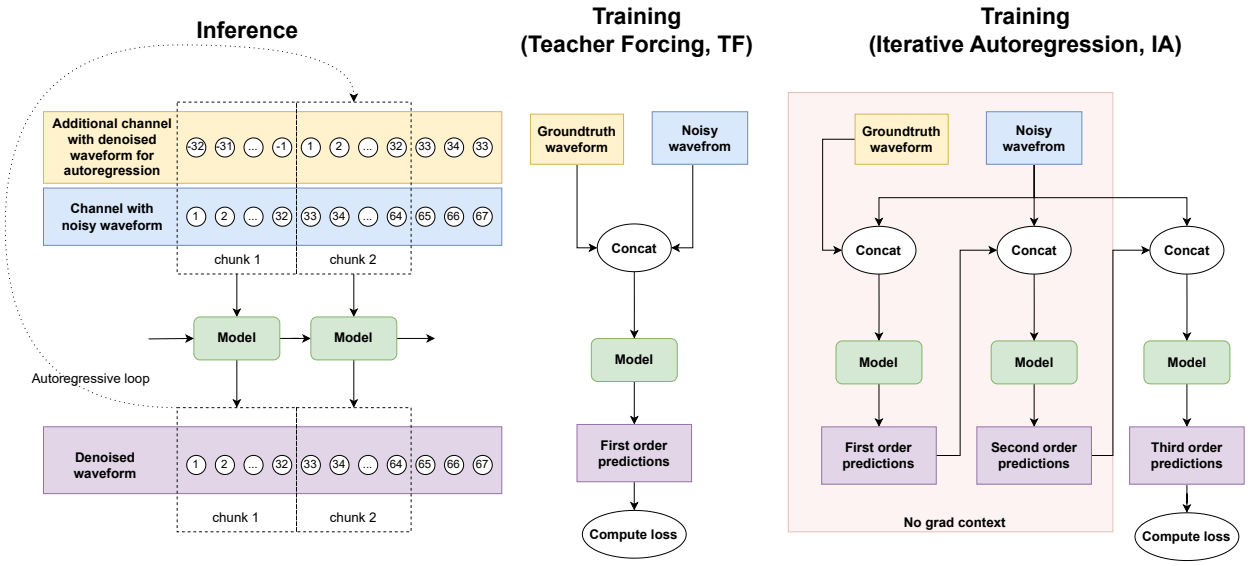

Figure 6: *Left:* illustration of autoregressive conditioning for the model with algorithmic latency 32 timesteps (2 ms at 16 kHz sampling rate). Predicted timesteps from chunk 1 are re-used then making predictions for chunk 2 during inference (free-running mode). *Middle:* Illustration of teacher forcing training. *Right:* Stage 3 of the iterative autoregression training process. The model uses its own predictions to produce predictions of higher orders. We shift ground-truth waveform and predictions before forming a channel with autoregressive conditioning to avoid leakage of future information.

Iterative autoregression We present a straightforward yet highly efficient algorithm for training autoregressive models that significantly reduces the training-inference mismatch (see Figure 6). Our approach is based on the iterative substitution of ground-truth conditioning with the model's predictions in teacher-forcing mode. Specifically, we divide the entire training process into N stages,

starting with the standard teacher forcing in the initial stage. In the second stage, the forward pass of the model comprises two steps: in the first step, model predicts conditioned on the ground-truth, in the second step, it predicts conditioned on the predictions from the first step. Similarly, at the n-th stage, the forward pass consists of n steps, and at each step, the model is conditioned on the predictions from the previous step. We refer to this algorithm as "iterative autoregression" (IA) and demonstrate that it helps to alleviate the training-inference mismatch caused by teacher forcing. Furthermore, we show that autoregressive conditioning offers significant advantages over non-autoregressive baselines across various training losses and neural architectures. Notably, our proposed IA algorithm is versatile and can potentially be applied to training autoregressive models beyond the speech enhancement domain.

Table 4: Autoregressive training improves speech enhancement quality in different scenarios. All AR models are trained with iterative autoregression if not indicated otherwise (TF).

Experiment	UTMOS	DNSMOS	SISDR	CMOS
Base configuration				
w/o AR	3.53	2.97	17.0	-
w/ AR (TF)	3.38	2.92	12.6	$-0.5^{\pm 0.08}$
w/ AR	**3.61**	**3.03**	**18.4**	$0.1^{\pm 0.05}$
Different losses				
w/o AR (adv.)	3.68	3.02	15.2	-
w/ AR (adv.)	**3.74**	**3.04**	**15.3**	$0.12^{\pm 0.04}$
w/o AR (si-snr)	3.51	2.95	17.0	-
w/ AR (si-snr)	**3.57**	**2.96**	**17.8**	$0.13^{\pm 0.05}$
DNS dataset				
w/o AR	2.42	2.98	14.5	-
w/ AR	**2.47**	**3.03**	**14.6**	$0.1^{\pm 0.05}$
ConvTasNet architecture				
w/o AR	3.08	2.86	15.3	-
w/ AR	**3.33**	**2.99**	**15.8**	$0.52^{\pm 0.06}$
Different latencies				
w/o AR (2 ms)	3.47	2.94	17.1	-
w/ AR (2 ms)	**3.55**	**2.98**	**18.3**	$0.04^{\pm 0.04}$
w/o AR (4 ms)	3.5	2.96	17.2	-
w/ AR (4 ms)	**3.59**	**3.02**	**18.6**	$0.16^{\pm 0.05}$
w/o AR (16 ms)	3.57	2.99	17.3	-
w/ AR (16 ms)	**3.64**	**3.02**	**18.6**	$0.09^{\pm 0.04}$

Experiments In all our experiments, we consider additive noise as the distortion to be removed from speech recordings. As shown in Table 4, we conduct a number of experiments to test the proficiency of iterative autoregression in different training scenarios. For each experimental setting, we train the baseline model without autoregressive conditioning (w/o AR) and the model with autoregressive conditioning (w/ AR). The training conditions and models are identical except AR models are trained with iterative autoregression if not stated otherwise. The experiments can be divided into 5 settings, depending on the training scenario employed. In each training scenario, we change only one training condition (dataset/model architecture/loss/latency) while leaving other parameters as in the base configuration described below.

The proposed method of iterative autoregressive training allows for improving the quality of streaming speech enhancement models in all studied scenarios. Furthermore, it dramatically outperforms the conventional teacher forcing method, which fails to provide any improvements over the non-autoregressive baseline due to a high training-inference mismatch. We believe that the presented technique provides a practical alternative to teacher forcing and takes an important step toward improving streaming models by means of autoregression.

2.3 Unsupervised Speech Enhancement with Unconditional Diffusion Model

In recent years, diffusion models [41, 17, 20] have gained attention due to their ability to efficiently model complex high-dimensional distributions. Diffusion models are designed to learn the underlying data distribution's implicit prior by matching the gradient of the log density. This learned prior can be useful for solving inverse problems, where the objective is to recover the input signal y from the measurements x, which are typically linked through some differentiable operator A, s.t. $x = A(y) + n$, where n is some noise. In this part of the work, we introduce UnDiff, a diffusion probabilistic model specifically designed to tackle various inverse tasks for speech processing including degradation inversion and neural vocoding.

The key advantage of UnDiff is its ability to be trained in an unconditional manner for speech waveform generation and then be adapted for the inverse problem without any additional supervised training. This is in contrast to existing approaches that utilize conditional diffusion models for waveform restoration or utilize specialized training pipelines [40, 37, 39].

Inverse problems with diffusion models The inverse problems address the task of retrieving object $\mathbf{y}$ given its partial observation $\mathbf{x}$ and the degra-

dation model $p(\mathbf{x}|\mathbf{y})$. To utilize reverse SDE for sampling from conditional distribution $p(\mathbf{y}|\mathbf{x})$, one needs to know the score function of conditional distribution $\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x})$.

One way to estimate $\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x})$ is to apply imputation guidance (data consistency) [41, 34, 8]. The idea of this method is to explicitly modify the score so that some parts of a denoised estimate $\hat{\mathbf{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}(t)}}(\mathbf{y}_t - (1-\bar{\alpha}(t))s_\theta(\mathbf{y}_t, t))$ are imputed with observations $\mathbf{x}$.

Another way to formalize the search for $\mathbf{y}$ is the usage of Bayes' rule:

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})/p(\mathbf{x}), \tag{7}$$

thus,

$$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t|\mathbf{x}) = \nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t) + \nabla_{\mathbf{y}_t} \log p_t(\mathbf{y}_t), \tag{8}$$

$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t)$ is generally intractable. However, Chung et al. [9] showed that one can make the approximation $\nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\mathbf{y}_t) \approx \nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\hat{\mathbf{y}}_0)$, where $\hat{\mathbf{y}}_0$ can be estimated from score function and $\nabla_{\mathbf{y}_t} \log p(\mathbf{x}|\hat{\mathbf{y}}_0)$ can be computed using the degradation model. Given the observation operator A and assuming Gaussian likelihood, the final approximation becomes:

$$\nabla_{\mathbf{y}_t} \log p_t(\mathbf{x}|\mathbf{y}_t) \approx -\xi(t)\nabla_{\mathbf{y}_t} \|\mathbf{x} - A(\hat{\mathbf{y}}_0)\|_2^2 \tag{9}$$

where $\xi(t)$ is a weighting coefficient which we set to be inversely proportional to the gradient norm similarly to [34]. Likewise, [34] we refer to this method as reconstruction guidance.

Bandwidth extension Frequency bandwidth extension [23, 1] (also known as audio super-resolution) can be viewed as a realistic restoration of waveform's high frequencies. The observation operator is a lowpass filter $\mathbf{x} = A(\mathbf{y}) = \mathrm{LPF}(\mathbf{y})$. Thus, imputation guidance in this case corresponds to substituting the generated estimate of low frequencies with observed low frequencies $\mathbf{x}$ at each step. More formally, this corresponds to modifying the score function during sampling as

$$\tilde{s}_\theta(\mathbf{y}_t, t) = \frac{1}{1-\bar{\alpha}(t)}(\sqrt{\bar{\alpha}(t)}\tilde{\hat{\mathbf{y}}}_0 - \mathbf{y}_t), \tag{10}$$

where $\tilde{\hat{\mathbf{y}}}_0 = \hat{\mathbf{y}}_0 - \mathrm{LPF}(\hat{\mathbf{y}}_0) + \mathbf{x}$ is imputed estimate of $\mathbf{y}_0$, and $\hat{\mathbf{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}(t)}}(\mathbf{y}_t + (1-\bar{\alpha}(t))s_\theta(\mathbf{y}_t, t))$ is estimate of $\mathbf{y}_0$ with original score function. In our bandwidth extension experiments, we use recordings with a sampling rate of 16 kHz as targets and consider two frequency bandwidths for input data: 2 kHz and 4 kHz. We artificially degrade the signal to the desired frequency bandwidth (2 kHz or 4 kHz) using polyphase filtering. The results and comparison with other techniques for the challenging 2 kHz setting are outlined in Table 5.

Table 5: Results of bandwidth extension (BWE 2 kHz) on VCTK.

Model	Supervised	WV-MOS	LSD	MOS
Ground Truth	-	4.17	0	4.09 ± 0.09
HiFi++	✓	**4.05**	1.09	$\mathbf{3.93 \pm 0.10}$
Voicefixer [31]	✓	3.67	1.08	3.64 ± 0.10
TFiLM [3]	✓	2.83	1.01	2.71 ± 0.10
UnDiff (Diffwave)	×	3.48	**0.96**	3.59 ± 0.11
UnDiff (FFC-AE)	×	3.59	1.13	3.50 ± 0.11

Decliping We consider clipping as an inverse problem with observation function defined as $A = \text{clip}(\mathbf{y}) = \frac{1}{2}(|y+c| - |y-c|)$ and apply reconstruction guidance strategy. We compare our models against popular audio declipping methods A-SPADE [49] and S-SPADE [50], as well as the general speech restoration framework Voicefixer [31] on clipped audio recordings with input SDR being equal to 3 db (see Table 6).

Table 6: Results of declipping (input SNR = 3 db) on VCTK.

Model	Supervised	WV-MOS	SI-SNR	MOS
Ground Truth	-	3.91	-	3.84 ± 0.11
A-SPADE [49]	×	2.63	8.48	2.67 ± 0.11
S-SPADE [50]	×	2.69	8.50	2.55 ± 0.11
Voicefixer [31]	✓	2.79	-22.58	2.98 ± 0.12
UnDiff (Diffwave)	×	**3.62**	**10.57**	$\mathbf{3.59 \pm 0.12}$
UnDiff (FFC-AE)	×	3.01	7.35	3.06 ± 0.12
Input	-	2.30	3.82	2.19 ± 0.09

The results show that despite UnDiff was never explicitly trained to solve any of the considered tasks, it performs comparably (though inferior) to supervised baselines for bandwidth extension and declipping. Overall, the results highlight the potential of the unconditional diffusion models to serve as general unsupervised voice restoration tools.

3 Conclusions

The main conclusions drawn from the results of this work are the following:

1. Composite multi-domain generator architectures provide a better trade-off between quality and complexity of BWE and SE models. In particular,

it is useful to enhance audio processing models with modules that perform both time-domain and spectral-domain signal correction to achieve more efficient parameter utilization and lower computational complexity, as demonstrated in the HiFi++ study.

2. Fast Fourier Convolution blocks provide an efficient architectural choice for designing spectrum-based processing modules. The global receptive field of this neural layer allows for effective phase estimation, reducing the memory consumption for the weights of the neural network.

3. The theoretical analysis reveals that LS-GAN training can be used for implicit regression for the mode of distribution, which is naturally aligned with the practical goals of the speech enhancement problem. The practical implementation of GAN-based training supports this analysis and shows that GAN-based models are able to achieve fast and high-quality speech enhancement, outperforming other types of generative models with fewer resources.

4. Autoregressive conditioning is able to improve streaming speech enhancement models by utilizing information about past predictions during inference. However, the application of the standard technique for training autoregressive models, which is teacher forcing, leads to high levels of training-inference mismatch and consequently poor enhancement quality. The developed iterative autoregression technique provides a practical alternative to teacher forcing and allows for efficient and effective training of autoregressive speech enhancement models.

5. Unsupervised speech enhancement presents a significant challenge due to the unknown degradation model during training. The problem could be tackled by the unconditional diffusion model. The unconditional diffusion model can be used to learn the prior distribution of speech signals during training and then be adapted to the particular degradation model during inference. Unfortunately, our study reveals significant challenges arising with this approach, and the resulting models tend to perform significantly worse than their supervised counterparts.

References

[1] Pavel Andreev et al. "Hifi++: a unified framework for bandwidth extension and speech enhancement". In: *arXiv preprint arXiv:2203.13086* (2022).

[2] Samy Bengio et al. "Scheduled sampling for sequence prediction with recurrent neural networks". In: *Advances in neural information processing systems* 28 (2015).

[3] Sawyer Birnbaum et al. "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations". In: *arXiv preprint arXiv:1909.06628* (2019).

[4] Jaeuk Byun et al. "An Empirical Study on Speech Restoration Guided by Self-Supervised Speech Representation". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[5] Jun Chen et al. "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7857–7861.

[6] Sanyuan Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.

[7] Lu Chi, Borui Jiang, and Yadong Mu. "Fast fourier convolution". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4479–4488.

[8] Jooyoung Choi et al. "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14367–14376.

[9] Hyungjin Chung et al. "Diffusion posterior sampling for general noisy inverse problems". In: *arXiv preprint arXiv:2209.14687* (2022).

[10] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. "Real Time Speech Enhancement in the Waveform Domain". In: *Interspeech*. 2020.

[11] Alexandre Défossez et al. "High Fidelity Neural Audio Compression". In: *Transactions on Machine Learning Research* (2023).

[12] Harishchandra Dubey et al. "Icassp 2022 deep noise suppression challenge". In: *arXiv preprint arXiv:2202.13288* (2022).

[13] Yariv Ephraim. "Statistical-model-based speech enhancement systems". In: *Proceedings of the IEEE* 80.10 (1992), pp. 1526–1555.

[14] Yariv Ephraim and David Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *IEEE Transactions on acoustics, speech, and signal processing* 32.6 (1984), pp. 1109–1121.

[15] Szu-Wei Fu et al. "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement". In: *arXiv preprint arXiv:2104.03538* (2021).

[16] Xiang Hao et al. "FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 6633–6637.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[18] Kuo-Hsuan Hung et al. "Boosting self-supervised embeddings for speech enhancement". In: *arXiv preprint arXiv:2204.03339* (2022).

[19] Umut Isik et al. "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss". In: *arXiv preprint arXiv:2008.04470* (2020).

[20] Tero Karras et al. "Elucidating the Design Space of Diffusion-Based Generative Models". In: *Advances in Neural Information Processing Systems*.

[21] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". In: *arXiv preprint arXiv:2010.05646* (2020).

[22] Qiuqiang Kong et al. "Decoupling magnitude and phase estimation with deep resunet for music source separation". In: *arXiv preprint arXiv:2109.05418* (2021).

[23] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. "Audio super resolution using neural networks". In: *arXiv preprint arXiv:1708.00853* (2017).

[24] Kundan Kumar et al. "Melgan: Generative adversarial networks for conditional waveform synthesis". In: *arXiv preprint arXiv:1910.06711* (2019).

[25] Anders Boesen Lindbo Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[26] Bunlong Lay et al. "Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement". In: *arXiv preprint arXiv:2302.14748* (2023).

[27] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[28] Jean-Marie Lemercier et al. "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[29] Cheuk Ting Li and Farzan Farnia. "Mode-seeking divergences: theory and applications to GANs". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 8321–8350.

[30] Hsin-Yi Lin et al. "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 19935–19946.

[31] Haohe Liu et al. "VoiceFixer: Toward General Speech Restoration with Neural Vocoder". In: *arXiv preprint arXiv:2109.13731* (2021).

[32] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[33] Xudong Mao et al. "Least squares generative adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2794–2802.

[34] Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. "Solving Audio Inverse Problems with a Diffusion Model". In: *arXiv preprint arXiv:2210.15228* (2022).

[35] Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. "Solving audio inverse problems with a diffusion model". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[36] Henri J Nussbaumer. "The fast Fourier transform". In: *Fast Fourier Transform and Convolution Algorithms*. Springer, 1981, pp. 80–111.

[37] Julius Richter et al. "Speech enhancement and dereverberation with diffusion-based generative models". In: *arXiv preprint arXiv:2208.05830* (2022).

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[39] Robin Scheibler et al. "Diffusion-based Generative Speech Source Separation". In: *arXiv preprint arXiv:2210.17327* (2022).

[40] Joan Serrà et al. "Universal speech enhancement with score-based diffusion". In: *arXiv preprint arXiv:2206.03065* (2022).

[41] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations*.

[42] Daniel Stoller, Sebastian Ewert, and Simon Dixon. “Wave-u-net: A multi-scale neural network for end-to-end audio source separation”. In: *arXiv preprint arXiv:1806.03185* (2018).

[43] Roman Suvorov et al. “Resolution-robust Large Mask Inpainting with Fourier Convolutions”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2149–2159.

[44] Marco Tagliasacchi et al. “SEANet: A multi-modal speech enhancement network”. In: *arXiv preprint arXiv:2009.02095* (2020).

[45] Zehai Tu et al. “A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing”. In: *Proc. Clarity* (2021), pp. 3–5.

[46] Cassia Valentini-Botinhao et al. “Noisy speech database for training speech enhancement algorithms and tts models”. In: (2017).

[47] Ronald J Williams and David Zipser. “A learning algorithm for continually running fully recurrent neural networks”. In: *Neural computation* 1.2 (1989), pp. 270–280.

[48] Yong Xu et al. “A regression approach to speech enhancement based on deep neural networks”. In: *IEEE/ACM transactions on audio, speech, and language processing* 23.1 (2014), pp. 7–19.

[49] Pavel Zaviska and Pavel Rajmic. “Analysis Social Sparsity Audio Declipper”. In: *arXiv preprint arXiv:2205.10215* (2022).

[50] Pavel Zaviska et al. “A proper version of synthesis-based sparse audio declipper”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 591–595.

[51] K Zmolikova and JH Cernock. “BUT System for the First Clarity Enhancement Challenge”. In: *Proceedings of Clarity* (2021), pp. 1–3.