Speech Denoising: An Inception-Enhanced Wave-U-Net Approach

NG Hoi Kan
The Hong Kong University of Science and Technology
hkngae@connect.ust.hk
Submission for COMP5214 Final Report

May 15, 2025

1 Abstract

Speech denoising is vital for robust audio applications. Traditional methods often struggle with non-stationary noise and introduce phase artifacts. Deep learning models processing raw waveforms offer an end-to-end solution, avoiding these issues. Building on the U-Net architecture of Wave-U-Net, this work proposes an enhanced model for speech denoising by integrating Inception-inspired multi-scale convolutional blocks into the encoder. This design enables the model to capture acoustic features across varied temporal scales, addressing the limitations of fixed-kernel architectures in modeling diverse noise patterns. The model is trained end-to-end on combined datasets covering a wide range of SNR conditions (-5 dB to 20 dB) and noise types, predicting both clean speech and noise components using a masked L2 loss. Experimental evaluation on the adopted test set demonstrates that the proposed Inception-enhanced architecture achieves superior objective performance (higher SNR, PESQ, STOI) compared to baseline Wave-U-Net configurations. This study highlights the effectiveness of multi-scale feature extraction for robust end-to-end waveform speech denoising.

2 Introduction

Effective speech denoising is crucial for the performance of numerous audio applications, including communication systems and speech recognition. The presence of background noise significantly degrades speech quality and intelligibility. While traditional signal processing and early deep learning methods have addressed this task, many operate in the spectral domain (e.g., STFT), requiring phase reconstruction prone to artifacts, or rely on restrictive assumptions like noise stationarity.

End-to-end deep learning models operating directly on raw audio waveforms offer a promising alternative, inherently preserving temporal correlations and circumventing phase reconstruction challenges. Architectures like SEGAN [3] and Wave-U-Net [6] demonstrate the viability of this approach. However, standard convolutional layers used in these models employ fixed kernel sizes, potentially limiting their capacity to effectively capture features across the diverse temporal scales of speech and varied noise characteristics.

Motivated by this limitation, this work proposes an enhanced deep learning architecture for time-domain speech denoising. The key innovation is the integration of Inception-inspired multi-scale convolutional blocks into the encoder of a U-Net-like structure. This allows the model to process inputs concurrently at different temporal resolutions, capturing both fine details and broad context. The model predicts both clean speech and noise components, optimized via a masked L2 loss.

The main contributions of this report are:

- An Inception-enhanced Wave-U-Net architecture for end-to-end speech denoising.
- Empirical validation of the benefit of multi-

scale convolutions for improved denoising performance compared to fixed-kernel baselines.

- Evaluation using diverse datasets spanning a wide range of SNRs and noise types.

The remainder of this report details the proposed method, experimental setup, results, and conclusions.

3 Related Work

Recent advances in deep learning have significantly advanced speech enhancement, particularly through architectures operating directly on raw waveforms. This section reviews foundational works that inform the approach presented in this report.

3.1 SEGAN

Pascual et al.[3] introduced SEGAN, a generative adversarial network (GAN) framework for speech enhancement. SEGAN employs a fully convolutional architecture designed to preserve temporal correlations in raw waveforms while reducing computational complexity by eliminating dense layers. Importantly, skip connections connect the generator's encoder and decoder to mitigate information loss during downsampling, thereby ensuring retention of low-level acoustic details, such as phoneme structure and harmonicity. The adversarial loss, implemented via least squares GAN (LSGAN), is augmented with an L1-norm reconstruction loss between the generator's output and clean speech. This hybrid loss enhances output fidelity by suppressing artifacts common in purely GAN-based systems. While effective, SEGAN's reliance on fixed kernel sizes limits its capacity to capture multi-scale noise patterns, a limitation addressed by the method proposed herein.

3.2 Wave-U-Net

Stoller et al.[6] proposed Wave-U-Net, a U-Net [5] based model primarily developed for end-to-end audio source separation in the time domain. By iteratively downsampling and upsampling waveforms using strided convolutions, Wave-U-Net achieves a large receptive field, which is crucial for modeling long-range dependencies in audio signals (e.g., a 2-second clip at 16 kHz spans 32,000 samples). Skip connections propagate localized features from the encoder to the decoder, enabling precise reconstruction of target sources. Despite its success in audio tasks, Wave-U-Net's uniform convolutional kernel sizes constrain multi-scale feature extraction, potentially overlooking transient noise events or broad spectral trends relevant to the denoising task.

3.3 Proposed Advancements

Both SEGAN and Wave-U-Net avoid short-time Fourier transform (STFT) representations, thereby circumventing phase estimation errors inherent to spectrogram-based methods. Building on these foundations, the present work integrates Inception-inspired multi-scale convolutional blocks into the downsampling, bottleneck, and upsampling blocks of the Wave-U-Net architecture.. By employing parallel convolutional layers with varying kernel sizes, this architecture is designed to concurrently capture fine-grained temporal details and broader noise patterns, directly addressing a key limitation observed in the aforementioned fixed-kernel architectures.

4 Data

The performance of speech enhancement models is highly dependent on the diversity and representativeness of the training data, particularly given the complex and non-stationary nature of real-world acoustic environments. To address this requirement, two complementary datasets were utilized for training and evaluation of the proposed model. This section details their composition, sources, preprocessing procedures, and rationale for selection.

4.1 VoiceBank-DEMAND-16k

The primary dataset employed was VoiceBank-DEMAND-16k [8]. This corpus consists of 12,396 paired mono-channel clean and noisy speech samples, sampled at 16 kHz. The signal-to-noise ratios (SNRs)

range from 0 dB to 15 dB. Approximately 87% of the samples exhibit durations under 4 seconds, reflecting characteristics of typical conversational speech. The dataset is divided into predefined training (11,572 samples) and test (824 samples) partitions, designed to ensure speaker disjointness between the subsets.

Prior to use, all waveforms from this dataset were normalized to the range [-1,1] using peak amplitude normalization to ensure consistent scaling across samples. No further preprocessing steps were applied to this dataset.

4.2 PTDB-TUG and AudioSet Synthesized Data

To augment the training data and enhance model generalizability, a secondary dataset was synthesized. This involved combining clean speech samples from the PTDB-TUG corpus [4] with noise excerpts obtained from the AudioSet database [1]. The PTDB-TUG corpus comprises recordings of 20 native English speakers (10 male, 10 female) captured in an anechoic chamber, providing high-fidelity clean speech. Noise samples were curated from AudioSet's YouTube-sourced clips, encompassing diverse categories including environmental sounds (e.g., rainfall, wind), mechanical sounds (e.g., keyboard typing), animal sounds (e.g., duck quacking), and others (e.g., church bells, waterfalls, white noise).

Synthetic noisy speech was generated by programmatically mixing clean speech utterances and noise excerpts at target SNRs of 20, 15, 10, 5, 0, and -5 dB, specifically chosen to simulate challenging low-SNR scenarios. The mixing procedure involved several steps. First, both clean speech and noise waveforms were normalized to the range [-1,1] using peak amplitude normalization. Noise segments were then dynamically truncated to match the duration of the corresponding clean speech utterance. Subsequently, the noise segments were scaled to achieve the desired target SNRs and combined additively with the clean speech. To prevent sudden onset artifacts and better mimic the characteristics of natural noise exposure, a 50-ms linear fade-in was applied to the beginning of each noise excerpt prior to its addition. This technique helps to smooth the noise introduction and enhances the perceptual realism of the synthesized noisy speech samples.

This augmentation strategy resulted in the generation of an additional 8,000 mono-channel samples. These synthesized samples were partitioned into training, validation, and test sets with a 60%, 20%, and 20% split, respectively. The incorporation of a wider SNR range, greater noise diversity, and the application of fade-in processing were specifically implemented to contribute to mitigating overfitting and improve the model's robustness to unseen acoustic conditions.

5 Methods

5.1 Problem Formulation

The speech denoising task is formally defined using the additive noise model:

$$x(t) = s(t) + n(t) \tag{1}$$

where $x(t)$, $s(t)$, and $n(t)$ represent the time-domain noisy speech waveform, the clean speech target waveform, and the additive noise waveform, respectively, at time index t. The objective is to train a function $\mathcal{F}_\theta$ parameterized by θ that estimates the clean speech signal $\hat{s}(t)$ from the noisy input $x(t)$. This estimation is performed under variable signal-to-noise ratio (SNR) conditions, specifically ranging from -5 dB to 20 dB, and importantly, without making explicit assumptions regarding the stationarity of the noise component $n(t)$.

5.2 Architecture Overview

The proposed system architecture integrates three key principles to address the speech denoising problem in the time domain: 1) direct processing of raw waveforms to circumvent phase artifacts associated with spectral representations, 2) utilization of an Inception-enhanced U-Net structure to enable multi-scale temporal modeling, and 3) adoption of a fully convolutional design to preserve temporal resolution while maintaining parameter efficiency. By employing kernels of different sizes, these blocks are

designed to capture both fine-grained local speech patterns and broader global audio structures concurrently. The upsampling path, leveraging transposed convolutions and skip connections, facilitates the reconstruction of high-frequency details lost during downsampling.

5.3 Key Design Choices

Direct processing of raw waveforms was selected primarily to avoid the phase reconstruction errors inherent in magnitude-only STFT-based methods. This approach negates the requirement for manual tuning of spectrogram parameters and intrinsically preserves temporal relationships across all relevant time scales. The integration of Inception modules facilitates the concurrent extraction of features at multiple temporal scales by combining outputs from parallel convolutional layers with different kernel sizes through concatenation. This enables the model to capture both fine-grained phonetic details and broader, potentially non-stationary noise patterns simultaneously. For the final output layer, a fully convolutional architecture replaces dense layers with 1D convolutions. This design choice reduces the total number of parameters compared to fully connected alternatives while preserving the temporal resolution of the output. Notably, the final 1D convolutional layer is configured with an output channel size of two. These two channels are intended to correspond to estimates of the clean speech component and the additive noise component, respectively.

5.4 Training Protocol

The model was trained using a masked L2 loss function, defined in Equation 2. This loss function computes the L2 distance between the model's estimated outputs and their respective targets, weighted by a binary mask. Specifically, the model produces two outputs, corresponding to the estimated clean speech signal ($\hat{s}$) and the estimated noise signal ($\hat{n}$). The overall loss is an aggregation of the L2 losses computed for both estimations against their ground truth targets (s and n). Optimization was performed using the Adam optimizer with an initial learning rate of 1×10^{-4}. Training incorporated a learning rate scheduler to dynamically adjust the learning rate during the optimization process. If the validation loss did not show improvement compared to the lowest recorded validation loss for 5 consecutive epochs, the learning rate was reduced by a factor of 0.2. This reduction strategy continued until a minimum learning rate of 1×10^{-7} was reached.

$$\mathcal{L}_{\text{total}} = \frac{\sum_i \left(\|\hat{s}_i - s_i\|_2^2 \cdot \text{mask}_i + \|\hat{n}_i - n_i\|_2^2 \cdot \text{mask}_i \right)}{\sum_i \text{mask}_i + \epsilon} \tag{2}$$

where:

- $\hat{s}_i$ and $\hat{n}_i$: The model's estimated clean speech and noise components at time index i.
- s_i and n_i: The ground truth clean speech and noise components at time index i.
- $\| \cdot \|_2^2$: The squared L2 norm (or squared Euclidean distance) for the scalar elements.
- mask_i: A binary weight (0 or 1) for the i-th element, typically derived from speech activity detection to emphasize speech-dominant regions.
- ϵ: A small constant (e.g., 10^{-8}) added to the denominator to prevent division by zero.

The final denoised output waveform is typically obtained by using the estimated clean speech component, i.e.,

$$\hat{s}_{\text{denoised}}(t) = \hat{s}(t) \tag{3}$$

5.5 Rejected Alternatives

Several alternative design choices were intentionally avoided in the proposed architecture based on their potential theoretical limitations within the context of end-to-end speech denoising:

- **VAE Bottlenecks:** While Variational Autoencoders (VAEs) [2] offer potential for disentangling speech and noise through a probabilistic latent space, their stochastic nature may introduce undesirable interference with the deterministic feature propagation facilitated by U-Net's

skip connections. Skip connections are designed to transmit precise, unaltered encoder features, whereas VAE bottlenecks involve sampling from a learned distribution.

- **Dilated Convolutions:** Although dilated convolutions are effective for expanding the receptive field without increasing computational cost, their application in 1D waveforms can lead to sparse sampling patterns that risk introducing temporal gridding artifacts. In this specific application, the potential for such artifacts was deemed to outweigh the benefits of an increased receptive field achieved through this method.
- **STFT-Based Processing:** While processing in the spectrogram domain simplifies frequency-domain modeling, it fundamentally separates magnitude and phase. The reliance on phase reconstruction from the estimated magnitude spectrum often introduces irreversible artifacts, particularly at low SNRs. Since phase estimation typically remains decoupled from magnitude processing in most architectures, this necessitates that networks either disregard crucial phase information or attempt to approximate it[7].

Table 1 presents a comparison of the proposed model against other relevant speech enhancement methods.

6 Experiments

This section details the experimental methodology employed to evaluate the performance of both the baseline Wave-U-Net architecture and the proposed enhanced model for speech denoising. The experimental procedure commenced with an initial phase focused on establishing a functional denoising baseline using the standard Wave-U-Net.

6.1 Initial Model Evaluation and Parameter Tuning

Given that the Wave-U-Net architecture served as the foundational element for this study, the initial step of the experimental process involved assessing its fundamental capability to perform audio denoising. This was conducted using a standard training/validation setup with the primary objective of noise reduction from speech signals.

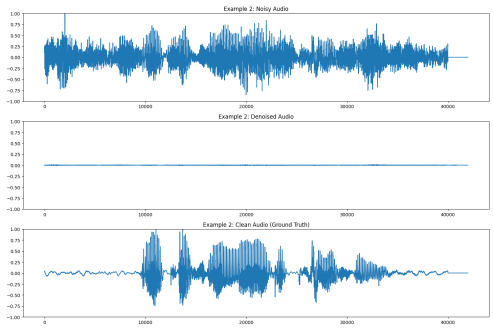

Figure 1: Initial Wave-U-Net Output After Training with Suboptimal Parameters

Initial trials conducted with default or slightly varied model parameters yielded suboptimal denoising results, suggesting the model had not effectively learned to separate speech from noise in its initial configuration. As depicted in Figure 1, the denoised output waveform is nearly indistinguishable from silence, consistently exhibiting amplitudes close to zero. This outcome indicated a failure mode in these preliminary training stages where the model suppressed both noise and speech signals.

To address this deficiency, an iterative process of hyperparameter tuning was conducted. Through empirical investigation, the impact of modifying key architectural hyperparameters on achieving a functional denoising capability was explored. Specifically, tuning focused on the following parameters:

6.1.1 Channel Configuration

This parameter, represented as a list of integers, defines the number of convolutional channels at each level of the U-Net structure. The length of this list determines the total number of downsampling/upsampling levels (U-Net depth).

Table 1: Comparison of Speech Enhancement Methods

Method	Key Features	Input Domain	Potential Limitations
Wiener Filter	• Statistical signal estimation • Linear filtering • Frequency-domain processing	STFT	• Assumes signal/noise stationarity • Phase estimation challenges • Typically targets Gaussian noise • Can introduce musical noise artifacts
Wave-U-Net	• U-Net architecture • Skip connections • End-to-end waveform processing • Fully convolutional	Waveform	• Single-scale convolutional paths • Fixed receptive field per layer type • Struggle with highly non-stationary noise
SEGAN	• Generative Adversarial Network • Adversarial loss + L1 loss • End-to-end waveform processing	Waveform	• Training instability can be challenging • No explicit noise modeling component • Perceptual quality hard to evaluate
VAE-Based	• VAE component • Probabilistic latent space • Generative modeling capabilities	Waveform	• May cause over-smoothing of denoised audio • Reconstruction quality can vary
Proposed Model	• Fully Convolutional U-Net • Inception modules • Two distinct outputs • Masked aggregated L2 loss	Waveform	• Higher computational complexity • Increased number of hyperparameters

6.1.2 Block Depth

This parameter controls the number of sequential 1D convolutional layers contained within each downsampling, upsampling, and bottleneck block at every level of the network. Adjusting this value impacts the representational capacity and complexity of the operations performed at each network stage. Empirical results indicated that a block depth of 2 yielded optimal performance for the speech denoising task. A depth of 1 resulted in insufficient model expressiveness, while a depth of 3 led to unstable training convergence.

6.1.3 Sources

This parameter specifies the number of distinct audio sources the model is trained to predict. Initial attempts predicting only a single source ('speech') did not yield effective learning for noise reduction. Predicting two sources ('speech' and 'noise') was found to be optimal for the speech denoising task, allowing the network to explicitly model and separate the noise component.

6.1.4 Target Output Size

This parameter determines the desired length of the output audio segments produced by the model. This choice also influences the required input segment length. An output size of approximately 64,000 samples (corresponding to 4 seconds at a 16 kHz sampling rate) was found to provide adequate temporal context for the model and was subsequently adopted.

6.2 Experimental Setup

Following preliminary hyperparameter tuning to identify suitable configurations for the speech denoising task, comprehensive training was initiated using the complete set of prepared training data described in Section 2. All training runs were conducted with a

batch size of 32, leveraging data parallelism across 6 NVIDIA RTX 3090 GPUs to accelerate computation.

6.2.1 Baseline Wave-U-Net Experiments

Wave-U-Net, as illustrated in Figure 2, was adapted for the speech denoising task using several distinct model configurations. Table 2 presents the performance metrics for various tested configurations, two of which are particularly noteworthy for comparison. The first configuration utilized a channel structure defined by [24, 48, 72, 96], while the second employed a configuration of [24, 96, 218, 384]. Both models processed input segments of length 64995 samples and produced output segments of length 64001 samples. Other hyperparameters remained consistent across these two variants. The second configuration, with approximately 15.7 million parameters, was significantly larger than the first, which had approximately 1.6 million parameters. Despite this substantial difference in model capacity, the observed denoising performance, as measured by objective metrics, remained comparable between the two Wave-U-Net variants. This result suggested a potential limitation inherent to the standard Wave-U-Net architecture's ability to leverage increased parameters effectively for this specific denoising task, potentially due to its fixed-scale convolutions.

6.2.2 Proposed Architecture Experiments

Motivated by the limitations observed in the standard Wave-U-Net baseline, the 1D convolutional layers within the downsampling blocks, bottleneck, and upsampling blocks were replaced with Inception-inspired modules. The number of parallel branches and their respective kernel sizes within these Inception modules were treated as adjustable hyperparameters to capture features at multiple scales. Four variants of the proposed architecture incorporating these Inception modules were trained and evaluated. These experiments indicated that the integration of multi-scale feature extraction via Inception modules indeed yielded an improvement in objective denoising performance compared to the baseline Wave-U-Net models, as detailed in the following results section.

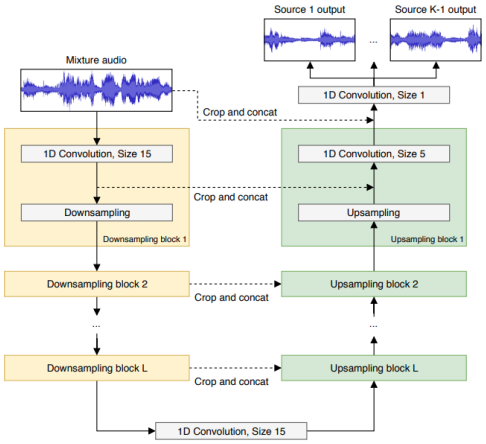

Figure 2: Original Wave-U-Net Architecture Overview [6]

6.3 Results

Table 2 presents a comprehensive overview of the performance metrics obtained for all tested model configurations, including the baseline Wave-U-Net variants and the proposed architectures. The metrics reported include SNR, PESQ, and STOI.

As indicated by the objective metrics in Table 2, Proposed Model 3 achieved the highest overall performance. It demonstrated the best SNR, PESQ, and STOI scores among all tested models.

Subjective evaluations were also conducted using custom speech samples to qualitatively assess the performance of Proposed Model 2 (illustrated by the output waveform in Figure 3) and Proposed Model 3 (illustrated in Figure 4). In both cases, background guitar noise was effectively suppressed from the noisy input. However, in this specific subjective test case, Proposed Model 2 was perceived to perform marginally better than Proposed Model 3. This qualitative observation suggested that while Model 3 achieved superior objective scores, it might have introduced minor, perceptually noticeable suppression of speech components in certain complex scenarios, a phenomenon less apparent in Model 2.

Table 2: Model Performance on VoiceBank-DEMAND Test Set

Method	Model Configurations	Parameters (Million)	SNR (dB)	PESQ	STOI
Noisy Baseline	N/A	N/A	9.55 ±0.06	1.72 ±0.01	0.88 ±0.01
Wave-U-Net 1	channels=[24, 48, 72, 96]	1.6	10.78	1.69	0.82
Wave-U-Net 2	channels=[24, 96, 218, 384]	15.7	10.71	1.69	0.82
Proposed Model 1	channels=[24, 48, 72, 96]; kernel=[1, 3, 5, 7, 11]	**1.0**	11.00	1.62	0.83
Proposed Model 2	channels=[24, 48, 72, 96]; kernel=[3, 7, 31, 127]	3.3	13.68	1.91	0.89
Proposed Model 3	channels=[24, 48, 72, 96]; kernel=[3, 7, 31, 127, 255]	5.8	**15.51**	**2.16**	**0.91**
Proposed Model 4	channels=[24, 48, 72, 96]; kernel=[3, 7, 31, 127, 255, 511]	10.3	10.60	1.74	0.86

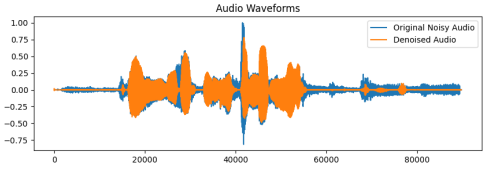

Figure 3: Output Waveform Example from Proposed Model 2

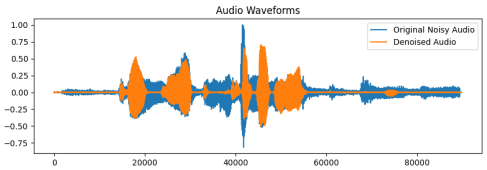

Figure 4: Output Waveform Example from Proposed Model 3

7 Conclusion

This report presented an Inception-enhanced Wave-U-Net architecture for end-to-end raw waveform speech denoising. The integration of Inception-inspired multi-scale convolutional blocks in the encoder aimed to improve the model's ability to capture features across different temporal scales, addressing a limitation of fixed-kernel models in handling complex noise.

Trained on diverse datasets and utilizing a masked L2 loss for explicit speech and noise prediction, the proposed model demonstrated better objective performance on the adopted test set compared to baseline Wave-U-Net configurations. Specifically, the variant with multi-scale kernels achieved the highest SNR, PESQ, and STOI scores. This finding empirically supports the hypothesis that multi-scale feature extraction is advantageous for robust time-domain speech denoising. Preliminary subjective testing indicated potential trade-offs in specific scenarios, suggesting a balance between noise reduction and speech preservation warrants further investigation.

In conclusion, this work developed and evaluated a multi-scale convolutional architecture for speech denoising within an end-to-end waveform framework.

Future research could explore alternative multi-scale designs, investigate the impact of different loss functions or training strategies to balance speech fidelity and noise reduction, conduct large-scale formal subjective evaluations, or adapt the architecture for diverse real-world noise scenarios.

References

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 3

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 4

[3] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network, 2017. 1, 2

[4] Gerhard Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf. Pitch tracking database from graz university of technology (ptdb-tug). Graz University of Technology, Institute of Broadband Communications, 2011. Database includes microphone/laryngograph signals and pitch trajectories from 20 native English speakers (2342 TIMIT sentences, 4720 recordings). 3

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs]. 2

[6] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation, 2018. 1, 2, 7

[7] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng. Joint time-frequency and time domain learning for speech enhancement. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3816–3822. International Joint Conferences on Artificial Intelligence Organization. 5

[8] Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and TTS models, 2016. [Sound]. 2