Interspeech 2022
18-22 September 2022, Incheon, Korea

VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration

Haohe Liu[1,*,†], *Xubo Liu*[1,*], *Qiuqiang Kong*[2], *Qiao Tian*[2], *Yan Zhao*[2],
DeLiang Wang[3] *Chuanzeng Huang*[2], *Yuxuan Wang*[2]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, China
[3]Department of Computer Science and Engineering, The Ohio State University, USA

haohe.liu@surrey.ac.uk, kongqiuqiang@bytedance.com, dwang@cse.ohio-state.edu

Abstract

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on a single type of distortion, such as speech denoising or dereverberation. However, speech signals can be degraded by several different distortions simultaneously in the real world. It is thus important to extend speech restoration models to deal with multiple distortions. In this paper, we introduce VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and can expand degraded speech (e.g., noisy speech) with a low bandwidth to 44.1 kHz full-bandwidth high-fidelity speech. We design VoiceFixer based on (1) an analysis stage that predicts intermediate-level features from the degraded speech, and (2) a synthesis stage that generates waveform using a neural vocoder. Both objective and subjective evaluations show that VoiceFixer is effective on severely degraded speech, such as real-world historical speech recordings. Samples of VoiceFixer are available at https://haoheliu.github.io/voicefixer.

Index Terms: speech restoration, speech super-resolution, neural vocoder, speech synthesis, deep learning

1. Introduction

Human speech often suffers from distortions such as background noise, room reverberations, or clipping from low-quality devices. Those distortions degrade the perceptual quality of human listeners. Speech restoration is a task to restore degraded speech to high-quality speech, which is useful in a wide range of applications such as online meeting [1] and hearing aids [2].

Previous speech restoration methods mainly focus on a single type of distortion, such as speech denoising [3], dereverberation [4], super-resolution [5], and declipping [6]. However, in the real world, speech signals can be degraded by several different distortions simultaneously. These mismatches limit the performance of these systems. Several works have explored restoring speech with multiple distortions, such as noise and reverberation [7, 8]. But other distortions such as low-resolution and clipping receive less attention, despite their significant impacts on speech perceptual quality.

Speech fidelity is important to perceptual quality. However, existing methods show limited performance on high-fidelity speech restoration. For example, for a noisy speech with low bandwidth, although the speech denoising method could remove noises, the restored speech would be still in low fidelity. One way to address this issue is to concatenate speech restoration methods (e.g., denoising) with the speech super-resolution method. However, this approach has limitations such as increasing computational cost and accumulating the artifacts introduced by each speech restoration model. To our knowledge, restoring low-bandwidth speech with multiple distortions has not been studied in the literature.

This paper introduces VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and could expand distorted speech with a low bandwidth between 1 kHz and 22.05 kHz to a full-bandwidth high-fidelity speech signal. We design VoiceFixer based on a two-stage strategy: (1) an analysis stage that performs mel spectrogram estimation; (2) a synthesis stage that generates the speech signal from the estimated mel spectrogram. Compared to the conventional speech restoration methods that operate on spectrogram or waveform, VoiceFixer uses the low dimensional mel spectrogram as the intermediate-level feature, which alleviates the difficulties of restoring multiple distortions simultaneously. In addition, neural vocoders [9] are usually trained on large-scale speech datasets. This provides prior knowledge on synthesizing waveform from low-dimensional mel spectrogram. The contributions of this paper are listed as follows:

- We present VoiceFixer, a unified framework for 44.1 kHz high-fidelity speech restoration. VoiceFixer can restore degraded speech from multiple distortions (e.g., noise, reverberation, clipping, and low-bandwidth).
- Evaluation result shows the effectiveness of VoiceFixer, which achieves a 0.256 higher mean opinion scores (MOS) than the baseline method.
- We release the pre-trained model and source code[1] of VoiceFixer to encourage future research.

The rest of this paper is organized as follows. Section 2 introduces the formulations of speech distortions we addressed. Section 3 describes the architecture of our proposed VoiceFixer. Experiments are presented in Section 4. In Section 5, we summarize this study and discuss our future directions.

2. Problem Formulation

We denote a segment of a speech signal as $s \in \mathbb{R}^L$, where L is the number of samples in the segment. We model the speech distortion process as function $d(\cdot)$. The degraded speech $x \in \mathbb{R}^L$ thus can be written as $x = d(s)$. Speech restoration aims to restore high-quality speech $\hat{s}$ from x by $\hat{s} = f(x)$, where $f(\cdot)$ is the restoration function and can be viewed as an inverse approximation of $d(\cdot)$. The target of the restoration function is to estimate s by restoring $\hat{s}$ from the degraded speech x.

* The first two authors contributed equally to this work.
† Part of this work was done during the internship at ByteDance

[1]https://github.com/haoheliu/voicefixer_main

10.21437/Interspeech.2022-11026

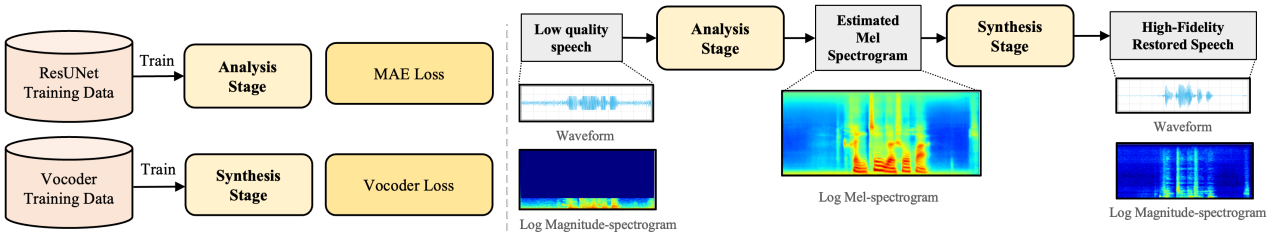

Figure 1: *Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.*

Distortion modeling is an important step to simulate training data when building speech restoration systems. Previous works model distortions in a sequential order [10, 11]. Similarly, we model the distortion $d(\cdot)$ as a composite function:

$$d(x) = d_1 \circ d_2 \circ ...d_Q(x), d_q \in \mathbb{D}, q = 1, 2, ..., Q, \quad (1)$$

where $\circ$ stands for function composition and Q is the number of distortions consisted in $d(\cdot)$. $\mathbb{D} = \{d_v(\cdot)\}_{v=1}^{V}$ is the set of distortion types, where V is the total number of types. Equation 1 describes the procedure of compounding different distortions from $\mathbb{D}$ in a sequential order. The four types of speech distortions we addressed in this work are introduced as follows.

Additive noise is one of the most common distortion and can be modeled by the addition between speech s and noise $n \in \mathbb{R}^L$:

$$d_{\text{noise}}(s) = s + n. \quad (2)$$

Reverberation is caused by the reflections of signal within a space. Reverberation makes speech signals sound distant and blurred. It can be modeled by convolving speech signals with a room impulse response filter (RIR) r:

$$d_{\text{rev}}(s) = s * r, \quad (3)$$

where $*$ stands for convolution operation.

Clipping distortion refers to the clipped amplitude of audio signals when their amplitude exceeds the maximum level. Clipping can be modeled by restricting signal amplitudes within a range $[-\eta, +\eta]$:

$$d_{\text{clip}}(s) = \max(\min(s, \eta), -\eta), \eta \in [0, 1]. \quad (4)$$

In the frequency domain, the clipping effect produces harmonic components in the high-frequency part and degrades speech intelligibility accordingly.

Low-bandwidth distortion refers to the limited bandwidth in the audio recordings caused by low sampling rate or defects in the recording device. We follow the description in [12] to produce low-bandwidth distortions but add more filter types [13]. After designing a low pass filter h, we first convolve it with $\boldsymbol{s}$ to avoid the aliasing phenomenon. Then we perform resampling on the filtered result from the original bandwidth o to a lower bandwidth u:

$$d_{\text{low\_bw}}(s) = \text{Resample}(s * h, o, u). \quad (5)$$

3. Approach

The two-stage strategy of VoiceFixer is formulated as follows:

$$f : x \mapsto z, \quad (6)$$

$$g : z \mapsto \hat{s}. \quad (7)$$

Equation 6 denotes the analysis stage of VoiceFixer where a distorted speech x is mapped into a intermediate-level feature z. Equation 7 denotes the synthesis stage of VoiceFixer, which synthesizes z to the restored speech $\hat{s}$. The overview of VoiceFixer framework is depicted in Figure 1.

3.1. Analysis stage

The goal of the analysis stage is to predict the intermediate representation $\boldsymbol{z}$, which can be used later to recover the speech signal. In our study, we choose the mel spectrogram as the intermediate representation. Mel spectrogram has been widely used in tasks such as speech enhancement [14] and audio synthesis [15, 16]. The frequency dimension of the mel spectrogram is usually much smaller than the magnitude spectrogram calculated using short-time-fourier-transform (STFT), thus working on mel-scale can reduce the dimension of feature space and offer a more tractable restoration process. The objective of the analysis stage is to restore the mel spectrogram of the target signals, which can be written as follows:

$$\hat{S}_{\text{mel}} = f_{\text{mel}}(X_{\text{mel}}; \alpha) \odot (X_{\text{mel}} + \epsilon), \quad (8)$$

where X_{mel} is the mel spectrogram of x. It is calculated by $X_{\text{mel}} = |X| W$, where $|X|$ is the magnitude spectrogram of x and W is a set of mel filter banks. The columns of W are not divided by the numbers of mel bands, because this will make the restoration model difficult to recover the high-frequency part. The mapping function $f_{\text{mel}}(\cdot; \alpha)$ is the mel-restoration mask-estimation model parameterized by α. X_{mel} is added with a minimum value ϵ before multipling with the output of f_{mel}. ϵ is set to 1×10^{-8} in this work to avoid zero values in X_{mel}.

We use ResUNet [17, 18] to model the analysis stage. ResUNet consists of six encoder and six decoder blocks. There are skip connections between encoder and decoder blocks at the same level. Both encoder and decoder blocks have a similar structure of four residual convolutions. Each residual convolutional consists of a batch normalization, a leakyReLU activation, and a two-dimensional convolutional operation. We utilize average pooling and transpose convolution for the upsampling and downsampling in the encoder and decoder blocks. We will refer to ResUNet as UNet in the remaining parts. We optimize the model in the analysis stage using the MAE loss between the estimated and the target mel spectrogram, $\hat{S}_{\text{mel}}$ and S_{mel}:

$$\mathcal{L}_{\text{MAE}} = \left\| \hat{S}_{\text{mel}} - S_{\text{mel}} \right\|_1. \quad (9)$$

3.2. Synthesis stage

We realize the synthesis stage with a neural vocoder, which synthesizes the mel spectrogram into waveform, as denoted in the following Equation 10:

$$\hat{s} = g(X_{\text{mel}}; \beta), \quad (10)$$

where $g(\cdot;\beta)$ stands for the vocoder model parameterized by β. The number of speakers used for the training of vocoder is much larger than that used in the analysis stage, which increases the robustness of VoiceFixer when generalizing to unseen speakers. We employ a pre-trained[2] time and frequency domain-based generative adversarial network (TFGAN) [19] as a vocoder. TFGAN achieves strong performance on 44.1 kHz speaker-independent speech vocoding, which will be discussed in detail in Section 4.

4. Experiments

We conduct two types of experiments to evaluate the performance of VoiceFixer: (1) High-fidelity speech restoration from simultaneously appearing noise, reverberation, clipping, and low-bandwidth distortions; (2) Single type restoration from speech with only one type of distortion (e.g., denoising). In the following sections, we first describe the experimental data preparation, then present the results of these two experiments. The test sets used in this section are publicly available[3].

4.1. Experimental data preparation

Training a speech restoration system relies on pairs of distorted speech and clean speech. In the high-fidelity speech restoration task, we simulate speech with multiple distortions. As introduced in Section 2, we simulate four types of speech distortion: additive noise, reverberation, clipping, and low-bandwidth. Three types of datasets are used for the simulation, including clean speech, noise data, and room impulse response (RIR). Note that clipping and low-resolution distortion only need the clean speech dataset for simulation and do not depend on other datasets. We introduce the three types of datasets as follows.

Clean speech we used is based on VCTK [20], which is a multi-speaker English corpus that consists of 110 speakers with different accents. The version of VCTK we used is 0.92. Following the setups in other studies [21], speakers p280 and p315 are omitted for the technical issues. The remaining part is split into a training set VCTK-Train with 98 speakers and a testing set VCTK-Test with the last 8 speakers. The remaining 2 speakers are omitted as they appear in the test set for denoising.

Noise data we used is based on two datasets. The first one is VCTK-Demand (VD) [22]. VD contains a training part VD-Train and a testing part VD-Test. Both parts contain clean speech and noisy speech data. To obtain the noise data from VD, we minus each noisy data in VD-Train with its corresponding clean part to get the noise dataset VD-Noise for training. The second noise dataset we use is the TUT urban acoustic scenes 2018 dataset [23], which contains 89 hours of high-quality recording from 10 acoustic scenes (e.g., airport). This dataset contains a development part and an evaluation part. We only use the evaluation part (DCASE-Eval) for the simulation of the test set for high-fidelity speech restoration.

Room impulse response is randomly simulated to add reverberation effect on 44.1 kHz speech. Simulation is performed using an open-source tool[4]. All the related parameters are randomized, including the size of the room, the placement of the microphone and the sound source, the RT60 value, and the pickup pattern of the microphone. In total, 43239 filters are simulated, in which we randomly split out 5000 filters as the test set RIR-Test and named the other 38239 filters as RIR-Train.

[2] https://github.com/haoheliu/voicefixer

[3] https://zenodo.org/record/5528144

[4] https://github.com/sunits/rir_simulator_python

4.2. High-Fidelity speech restoration

4.2.1. Data sets and distortion modeling

In this task, we simulate low-quality speech with four distortions in the training and test set, including noise, reverberation, clipping, and low bandwidth. Training data is simulated on the fly based on the speech data in VCTK-Train, the noise in VD-Noise, and RIR in RIR-Train. We set up the parameters of each distortion to be completely random to better cover the real-world cases. The test set we used in high-fidelity speech restoration, HiFi-Res, is constructed based on the clean speech in VCTK-Test, the noise in DCASE-Eval, and RIR in RIR-Test. HiFi-Res consists of 501 three seconds utterances with similar random distortions simulated as the training process. We first generate the distortions following specific order: reverberation, noise, and clipping. Then the degraded speech is low-pass filtered and down-sampled to an arbitrary low sampling rate between 2 kHz to 44.1 kHz. Details of the distortion modeling in this work are made available on GitHub[5].

4.2.2. Experiment details

All the audio files in our datasets are resampled to 44.1 kHz sampling rate. We calculate STFT using the Hanning window with a window length of 2048 and a hop size of 441. The mel filterbank we used consists of 128 filters. For training, We use Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.999$, an initial learning rate of 3×10^{-4} and a batch size of 24. The first 1000 steps are warmup steps, during which the learning rate grows linearly from 0 to 3×10^{-4}. The learning rate is scheduled for decay by 0.9 every 400 hours of training data. We trained our model using four Nvidia-V100-32GB GPUs for two days.

4.2.3. Baseline systems

We mainly use four baseline systems in the experiment. We implemented an UNet-based system (Baseline-UNet) for the high-fidelity speech restoration task, which structure is similar to the analysis module of VoiceFixer. It performs restoration by estimating STFT of the high-quality speech and reusing the phase of the degraded speech, which is a common approach in previous speech restoration systems [24]. As for the Oracle-Mel system, we directly use the target mel spectrogram as input to the vocoder to simulate the case when the analysis module works ideally. So, Oracle-Mel marks the theoretical upper bound of the VoiceFixer performance. For the Target system, scores are calculated using the ground truth clean speech. Conversely, the Unprocessed system evaluated directly on the distorted speech.

4.2.4. Evaluation metrics

We use both objective and subjective evaluation metrics. The objective metrics including log-spectral distance (LSD) [25], scale-invariant signal-to-noise ratio (SISNR) [26], wideband perceptual evaluation of speech quality (PESQ-wb) [27], and structural similarity (SSIM) [28]. Since neural vocoders generate waveforms directly from mel spectrograms, even with the same perceptual quality, the generated waveforms may not align with the target waveform in the time domain. This mis-alignment can considerably degrade the objective metrics, as is often the case in generative model [29]. Nevertheless, we report the model performance on these objective metrics for reference.

We use mean opinion scores (MOS) as the subjective eval-

[5] https://github.com/haoheliu/voicefixer_main

Table 1: *Evaluation result on the high-fidelity speech restoration test set HiFi-Res. Higher PESQ-wb, SSIM, MOS value indicates better performance, while LSD is the opposite. The best value for each metric is shown in bold.*

Models	PESQ-wb	LSD	SSIM	MOS
Unprocessed	1.94	2.00	0.64	2.38
Oracle-Mel	2.52	0.91	0.74	3.74
Target	4.64	0.01	1.00	3.95
Baseline-UNet	**2.67**	1.01	**0.79**	3.37
VoiceFixer	2.05	**1.01**	0.71	**3.62**

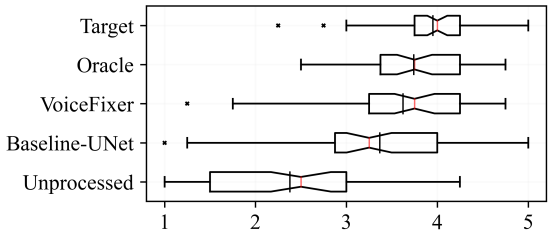

Figure 2: *Box plot of the MOS scores on HiFi-Res test set. Red and black vertical lines represent median and mean values.*

uation metric and invite eight internal language experts in ByteDance to perform evaluation. Their task is to rate the overall speech quality of an audio clip with a score between 1 (bad) to 5 (excellent). Each system has 38 samples for evaluation. We average the MOS values across all language experts as the final result.

4.2.5. Evaluation results

Table 1 shows the experimental results and Figure 2 depicts the box plot of the MOS scores. The Oracle-Mel system achieves a MOS score of 3.74, which is close to the Target MOS of 3.95, indicating that the vocoder performs well in the synthesis stage. We observe that VoiceFixer obtains 0.256 higher MOS score than that of Baseline-UNet and is only 0.11 lower than the Oracle-Mel, demonstrating its good performance for high-fidelity speech restoration. Although VoiceFixer performs worse on PESQ-wb and SSIM metrics, it has a much better MOS score than Baseline-UNet. This result shows that the improvement in subjective metrics in VoiceFixer is not always consistent with objective evaluations.

4.3. Single type restoration

To further demonstrate the effectiveness of VoiceFixer, we conduct two benchmark speech restoration experiments: speech denoising and speech declipping.

4.3.1. Denoising

For speech denoising, we evaluate the model performance on VD-Test (as described in Section 4.1). VD-Test contains 824 utterances from a female speaker and a male speaker. The test set is simulated at four SNR levels, which are 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. The original data is sampled at 48 kHz. We downsample it to 44.1 kHz to fit our experiments. We adopt three recent methods SEGAN [30] , WaveUNet [31], and the model trained with weakly labeled data [32] (referred to WL-Model) as baseline methods.

Experimental results are shown in Table 2. The PESQ-wb score of VoiceFixer reaches 2.43, higher than SEGAN, WaveUNet, and WL-Model. The MOS evaluations demonstrate that VoiceFixer outperforms the baseline speech denoising model Baseline-UNet. In addition, we observe that VoiceFixer even outperforms Oracle-Mel and achieves the same level as Target on the MOS scores. This is because the restored results of the VoiceFixer contain more energy in the high-frequency part, which potentially leads to a better perceptual quality for the listener. The SISNR of Oracle-Mel and VoiceFixer is significantly lower than Baseline-UNet because of the alignment issue mentioned in Section 4.2.4.

Table 2: *Evaluation result on the VD-Test test set. Superscript $*$ indicates the model is only trained on a single restoration task.*

Models	SISNR	PESQ-wb	MOS
Unprocessed	8.40	1.97	3.20
Oracle-Mel	-17.52	2.85	3.64
Target	/	4.50	3.69
*SEGAN [30]	/	2.16	/
*WaveUNet [31]	/	2.40	/
*WL-Model [32]	/	2.28	/
Baseline-UNet	**17.58**	**2.82**	3.64
VoiceFixer	-16.23	2.43	**3.69**

Table 3: *Evaluation result on the declipping test set DECLI.*

Clipping Level	0.25		0.1	
Models	STOI	MOS	STOI	MOS
Unprocessed	0.95	2.56	0.89	2.72
Oracle-Mel	0.81	3.44	0.81	3.42
Target	1.00	3.42	1.00	3.49
*SSPADE [33]	**0.98**	3.34	0.92	2.63
Baseline-UNet	0.97	3.38	**0.96**	3.23
VoiceFixer	0.82	**3.38**	0.80	**3.38**

4.3.2. Declipping

For the declipping task, we compare VoiceFixer with a state-of-the-art synthesis-based method SSPADE [33]. To evaluate the model performance, we create a test set DECLI based on VCTK-Test (as described in Section 4.1). DECLI is constructed by first normalizing the amplitude of VCTK-Test into $[-1, 1]$, and then simulating clipping on each audio with two clipping levels 0.25 and 0.1. This resulted in two declipping test sets, each containing 2937 clipped and clean speech audios.

We adopt MOS as the subjective metric and STOI [34] as the objective metric. A higher STOI value indicates better performance. Experimental results are shown in Table 3. VoiceFixer outperforms SSPADE on MOS by 0.04 and 1.25 in 0.25 and 0.1 clipping levels, respectively. The higher performance on MOS demonstrates a better perceptual quality restoration offered by VoiceFixer on speech declipping.

5. Conclusion

In this study, we propose VoiceFixer, an effective approach for high-fidelity speech restoration. VoiceFixer consists of an analysis stage modeled by a ResUNet and a synthesis stage using a TFGAN. The two stages can also be replaced by other deep learning models. The subjective evaluation results show that VoiceFixer achieves superior performance on high-fidelity speech restoration from distortions such as noise, reverberation, clipping, and low bandwidth. In the future, VoiceFixer will be extended to more types of distortions.

6. References

[1] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv:2006.12847*, 2020.

[2] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

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

[4] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted magnitude-phase loss for speech dereverberation," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 5794–5798.

[5] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv:1708.00853*, 2017.

[6] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 5–24, 2020.

[7] Y. Ai, H. Li, X. Wang, J. Yamagishi, and Z. Ling, "Denoising-and-dereverberation hierarchical neural vocoder for robust waveform generation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 477–484.

[8] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv:2006.05694*, 2020.

[9] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv:2010.05646*, 2020.

[10] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.

[11] X. Shu, Y. Zhu, Y. Chen, L. Chen, H. Liu, C. Huang, and Y. Wang, "Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation," *arXiv:2107.09298*, 2021.

[12] H. Wang and D. Wang, "Towards robust speech super-resolution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021.

[13] S. Sulun and M. E. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 132–142, 2020.

[14] S. Maiti and M. I. Mandel, "Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2020, pp. 206–210.

[15] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *arXiv:1910.06711*, 2019.

[16] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2021, pp. 1–6.

[17] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation." in *The International Society for Music Information Retrieval*, 2021.

[18] H. Liu, Q. Kong, and J. Liu, "CWS-PResUNet: Music source separation with channel-wise subband phase-aware resunet," *arXiv:2112.04685*, 2021.

[19] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, "TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis," *arXiv:2011.12206*, 2020.

[20] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2019.

[21] J. Lee and S. Han, "NU-wave: A diffusion probabilistic model for neural audio upsampling," *arXiv:2104.02321*, 2021.

[22] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv:1807.09840*, 2018.

[24] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with U-Net," *arXiv:2006.00687*, 2020.

[25] A. Erell and M. Weintraub, "Estimation using log-spectral-distance criterion for noise-robust speech recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 853–856.

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 749–752.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "NU-GAN: High resolution neural upsampling with GAN," *arXiv:2010.11362*, 2020.

[30] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[31] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," *arXiv:1811.11307*, 2018.

[32] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from AudioSet," in *INTERSPEECH*, 2021.

[33] P. Záviška, P. Rajmic, O. Mokrỳ, and Z. Průša, "A proper version of synthesis-based sparse audio declipper," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2019, pp. 591–595.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.