INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A GAN Speech Inpainting Model for Audio Editing Software

Haixin Zhao[1]

[1]Audio Information Processing, Technical University of Munich, Germany
ge75zuk@mytum.de

Abstract

This paper proposes the generative adversarial networks (GAN) speech inpainting model consisting of the GAN magnitude inpainting network and the phase reconstruction algorithm. The GAN network with partial convolutions implements inpainting specific time-frequency (T-F) areas of spectrograms, and captures latent information of speech spectrograms and high-dimensional features using the proposed loss function, contributing to more real and speech-like results. The phase reconstruction algorithm adopts two strategies for different magnitudes, inpainting clear harmonics while reducing the buzzes in high frequency. The proposed model outperforms the conventional and the T-F mask-based deep inpainting baselines in inpainting performance of Short-Term Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ). Since it can inpaint specific T-F areas and improve the inpainting performance, the model implements the speech inpainting for audio editing software.

Index Terms: speech inpainting, audio inpainting, GAN, partial convolution, speech-VGG loss, audio editing software

1. Introduction

Speech signals inevitably suffer from undesired local distortions. These distortions could be treated as missings, such as damage after noise suppression, storage error, or packet loss during transmission. The missing information severely degrades audio signals' hearing and coherence. For speech signals, this information missing, especially harmonics missing, impairs more since the integrity of speech signals is more crucial to the quality and intelligibility of speech. Hence, speech inpainting, has wide applications in audio editing, audio enhancement [1], and audio bandwidth extension [2] [3].

In recent decades, many conventional and machine learning methods have been developed for inpainting audio signals [4]. Conventional methods mainly concentrate on directly inpainting audio using context time features. At first, D. Goodman and N. Perraudin implemented inpainting with waveform substitution according to the explicit estimation of voicing and pitch or the similarity graph of time-persistent spectral [5]. Afterward, more successful auto-regressive methods were proposed to linearly predict or interpolate the missing samples using auto-regressive coefficients, which are learned from adjacent sampling points [6]. To improve the interpolation of longer missing gaps, P. Esquef introduced the warped burg's method into the auto-regressive method [7]. Sinusoidal modeling-based long interpolation approaches using a linear prediction were attempted by M. Lagrange [8] and J. Lindblom [9]. Besides, S. Godsill introduced a statistical approach [10]. G. Chantas further proposed a variational Bayesian model with sparse signal representation [11]. Maher presented the extrapolation method by synthesizing the estimate of missing using a sinusoidal representation [12]. Moreover, I. Kauppinen proposed another Linear Predictive Coding (LPC) based extrapolation method, which extrapolates time-varying cosine waves [13] [14]. The LPC method can reconstruct longer missing gaps and is often used as a performance comparison baseline.

On the other hand, more and more deep neural network (DNN) methods were proposed. Most of these are based on the context encoder, which is responsible for extracting information from the context, then generating and inpainting missing segments with this information [15] [16]. Some GAN models were proposed for inpainting longer gaps [17] [18]. Nevertheless, these methods could only inpaint time-clip missing. Thus, they only work for some limited scenarios like packet loss. For the speech inpainting in audio editing software and other speech enhancement tasks, including noise suppression and audio bandwidth extension, T-F mask-based methods are proposed for inpainting specific T-F areas [1] [19].

However, since these methods treat T-F spectrograms as images to inpaint, they cannot capture the actual audio spectrogram features, leading to unreal inpainting results. These methods still have severe problems of over-smooth and blurry results, especially for harmonics, leading to the failure to reconstruct reasonable speech structures. Moreover, using existing phase reconstruction algorithms from other speech enhancement or synthesis tasks results in blurry inpainted harmonics and annoying buzzes. The inpainting of harmonics is paramount in speech inpainting tasks, which directly determines speech intelligibility and quality. Hence, this paper proposes the T-F mask-based GAN speech inpainting model to tackle these problems for improving the inpainting performance and further to implement the speech inpainting for audio editing software.

2. Method

The framework of the proposed T-F mask-based speech inpaint-

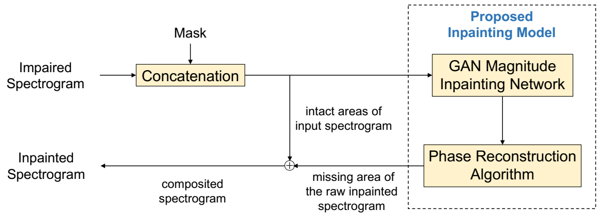

Proposed Inpainting Model
Mask
Impaired Spectrogram
Concatenation
GAN Magnitude Inpainting Network
intact areas of input spectrogram
Inpainted Spectrogram
Phase Reconstruction Algorithm
composited spectrogram
missing area of the raw inpainted spectrogram

Figure 1: *The framework of the proposed T-F mask-based speech inpainting system.*

10.21437/Interspeech.2023-904

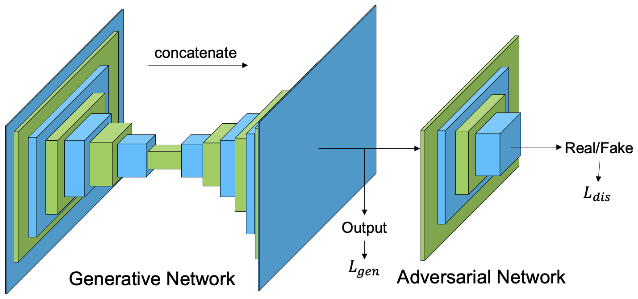


Figure 2: *The structure of the proposed GAN magnitude inpainting network.*

ing system is presented in Fig. 1. At the beginning of the system, the impaired spectrogram is concatenated with the corresponding binary masks and then fed into the inpainting model. Masks are binary and responsible to indicate the impaired areas. For audio editing software, masks can be provided by the frame or lasso from users. For other speech enhancement scenarios, masks can be estimated by methods, such as the ideal binary mask (IBM) [20] and spectral magnitude mask (SMM) [21]. The GAN network inpaints the magnitude of spectrograms [22], followed by the phase reconstruction algorithm, which rebuilds the phase from the inpainted magnitude. The network is only used to inpaint magnitude since A. Marafioti demonstrated that the magnitude DNN outperformed the complex-valued DNN in terms of signal-to-noise ratios and objective difference grades, in particular for reconstructing high-frequency content [17]. It is worth noting that the final output spectrum is composited by the unmasked part of the ground-truth input and the masked part of the raw inpainted spectrum to preserve the original intact area.

2.1. GAN Magnitude Inpainting Network

GAN architecture is employed in the proposed network since its unique adversarial training mode and joint loss function contribute to more real inpainting results. Besides, GAN architecture [22] has been demonstrated to be effective in other generative tasks including inpainting [23] [24], super-resolution [25] [26] and image generation [27] [28].

The structure of the proposed GAN magnitude inpainting network is presented in Figure 2. The network includes the generative network and the adversarial network. For the generator, the U-Net structure is used to enlarge the network's receptive field allowing for extracting more contextual information [29]. It also avoids the information loss of rim features during deconvolution and up-sampling by the skip link from the encoding block. The network consists of a seven-layer deep encoder and decoder. All convolutional layers are replaced by partial convolutional layers [30] since partial convolutions have been demonstrated to have advantages in T-F mask-based inpainting [19]. Each partial convolutional layer is followed by a BN and ReLU activation layer. For the decoder, the upsampled feature map and mask are concatenated with the corresponding ones from the encoding block. The concatenation is then inputted into the next partial convolutional layer.

The adversarial network consists of five convolutional layers. Before each convolutional layer, spectral normalization is performed to control the Lipschitz constant of the discriminator [31]. Besides, the leaky ReLU activation ($\alpha = 0.2$) is followed after each convolutional layer. Detailed network parameters of the generator and discriminator are shown in Table 1. Zero-padding is employed for each convolutional layer.

Table 1: *Network parameters (kernel size, stride and channel number).*

Layer	**Encoder**	**Decoder**	**Discriminator**
1_{st}	7×7, 2, 64	3×3, 1, 512	4×4, 2, 64
2_{nd}	5×5, 2, 128	3×3, 1, 512	4×4, 2, 128
3_{rd}	5×5, 2, 256	3×3, 1, 512	4×4, 2, 256
4_{th}	3×3, 2, 512	3×3, 1, 256	4×4, 1, 512
5_{th}	3×3, 2, 512	3×3, 1, 128	4×4, 1, 1
6_{th}	3×3, 2, 512	3×3, 1, 64	
7_{th}	3×3, 2, 512	3×3, 1, 1	

2.2. Loss Function

Based on the classic loss function of GAN models [22], The magnitude-based weight loss item is proposed and joined with VGG loss [30] and L^1 loss items to form the proposed new joint loss function. The proposed joint loss function consists of generative loss and adversarial loss. The adversarial loss of the proposed network is similar to that of the classic joint loss function [22], shown as:

$$L_{ad} = \frac{E(dis_{gt}, Real) + E(dis_{comp}, Fake)}{2} \tag{1}$$

where dis_{gt} and dis_{comp} are the outputs of the discriminator. E represents the binary cross-entropy (BCE).

For the generative network training, BCE loss L_B is still useful as computed by $E(dis_{comp}, Real)$. It is worth noting that the label is Real here, different from that of the dis_{comp} in the discriminative loss due to the adversarial training mode.

In addition to the loss item from the discriminator, more loss items from data (ground truth) are employed for the generative loss function. The basic L^1 loss items: L_h and L_v are included for the pixel-wise reconstruction accuracy as:

$$L_h = c_h(1 - \boldsymbol{M}) \odot ||\boldsymbol{S_{out}} - \boldsymbol{S_{gt}}||_1 \tag{2}$$

$$L_v = c_v \boldsymbol{M} \odot ||\boldsymbol{S_{out}} - \boldsymbol{S_{gt}}||_1 \tag{3}$$

where $\boldsymbol{S_{out}}$ and $\boldsymbol{S_{out}}$ are the magnitude spectrogram matrixes of networks's output and groundtruth. $\boldsymbol{M}$ is the binary mask matrix (1 for valid pixels) and the normalization factor c_h and c_v are $1/(1 - mean(\boldsymbol{M}))$ and $1/mean(\boldsymbol{M})$, respectively. Besides, since partial convolution operations are based on valid areas, L_v is included as well.

In addition, instead of inpainting spectrograms as images like some previous methods, we introduce two speech VGG loss items: perceptual loss L_p and style loss L_s. The VGG network pretrained by lots of speech data is used to capture the actual spectrogram latent, contributing to more real and speech-like inpainting results. Moreover, VGG loss is helpful for inpainting results to be closer to human perception, leading to better performance in subjective hearing. VGG loss items are computed based on a pre-trained deep feature extractor. After the generator's output and the ground truth are fed into the extractor, a perceptual loss can be obtained by computing the L1 distance

Table 2: *Comparative experimental results in objective metrics (PESQ & STOI)*

Inpainting Models	Loss Function	PRA	PESQ				STOI			
Missing Size (in T & F)	–	–	10%	20%	30%	40%	10%	20%	30%	40%
Before Inpainting	–	–	2.291	1.489	1.085	0.775	0.894	0.768	0.649	0.524
CNN (U-Net)	L_1	TSPRA	3.296	2.406	1.797	1.345	0.951	0.854	0.740	0.614
Proposed GAN Model	classic joint loss & L_1	TSPRA	3.281	2.590	2.108	1.736	0.952	0.879	0.799	0.709
	proposed loss function	GLA	3.299	2.424	2.108	1.572	0.954	0.872	0.792	0.706
	proposed loss function	TSPRA	**3.302**	**2.609**	**2.149**	**1.755**	**0.955**	**0.889**	**0.814**	**0.730**

of features in each pooling layer. Since our model's final output is the composition of the raw output's hole part and ground truth's valid part, we further apply the composited output rather than the raw output and update the perceptual L_p [32] as:

$$L_p = \sum_{p=0}^{P-1} \frac{||\boldsymbol{F}_p^{\boldsymbol{S}_{comp}} - \boldsymbol{F}_p^{\boldsymbol{S}_{gt}}||_1}{N_p} \tag{4}$$

where $\boldsymbol{F}_p^{\boldsymbol{S}_{gt}}$ and $\boldsymbol{F}_p^{\boldsymbol{S}_{comp}}$ are p_{th} layer feature maps of ground truth and composited output respectively. N_p is the number of pixels in p_{th} layer. Based on the perceptual loss, if each feature map is performed an auto-correlation (Gram matrix) before computing L1 distance, another helpful VGG loss item: style loss L_s to extract style information besides content [32].

Furthermore, to address the over-smoothing and blurring of inpainting results, we propose the magnitude-based weight loss item L_w as the following formula:

$$L_w = |\boldsymbol{S}_{gt}| \, ||\boldsymbol{S}_{comp} - \boldsymbol{S}_{gt}||_1 \tag{5}$$

Since the harmonics are usually the pixels with higher magnitudes in the speech spectrogram, using $\boldsymbol{S}_{gt}$ as the weight can force the network to enhance the learning of these pixels to reduce the over-smoothing and blur. As a result, harmonics with clearer structures will be inpainted accordingly.

Overall, the generative loss function of the proposed network is the combination of the above loss items as:

$$L_{gen} = 0.01L_B + L_v + 2L_h + 4L_p + 500L_s + 0.2L_w \tag{6}$$

2.3. Two-Strategy Phase Reconstruction Algorithm (TSPRA)

Previous inpainting models used some existing Phase Reconstruction Algorithms (PRA), including Griffin–Lim Algorithm (GLA) [33] and fast signal reconstruction [34], and so on. However, these methods inevitably have problems, such as annoying buzzes and over-smooth inpainting results, since they are designed for other speech enhancement and synthesis tasks. Thus, based on the Phase gradient heap integration (PGHI) [35], we employ a proper phase reconstruction algorithm for speech inpainting. The PGHI shows remarkable performance in low frequency, as spectrograms have clear-structure harmonics with high magnitude in low frequency. Nevertheless, since the spectrograms in high frequency have low magnitudes and are relatively smooth, PGHI will generate lots of annoying buzzes. Hence, we adopt two strategies for different magnitudes. The PGHI is only performed for the areas with high magnitude. For other low-magnitude areas, random phases are assigned to reduce buzzes since high frequency does not contribute much to speech intelligibility. The buzzes caused by further processings degrade the intelligibility instead.

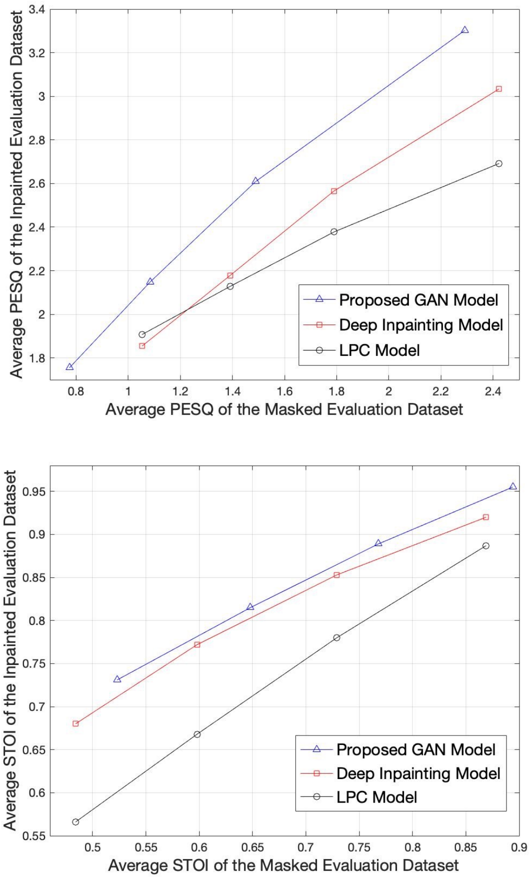


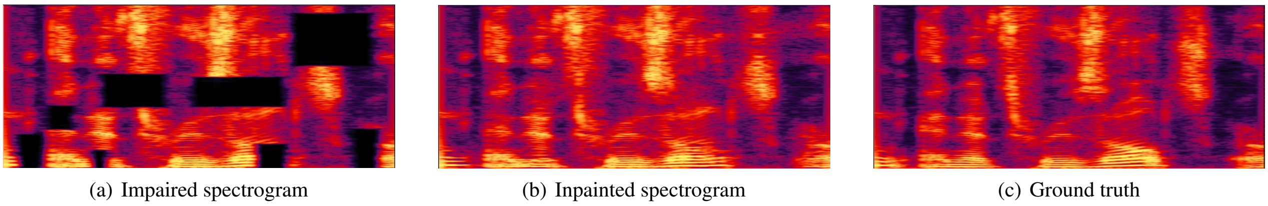
Figure 3: *Comparison with baselines in PESQ and STOI.*

3. Experimental Results

Objective experiments in speech intelligibility and quality are carried out to demonstrate the inpainting performance improvement of the proposed GAN speech inpainting model. Besides, various demos including spectrograms and audio files are generated by the proposed T-F mask-based speech inpainting system of audio editing software for presenting the improvement of inpainting and subjective hearing.

3.1. Training Settings

Train-clean-360, test-clean, and dev-clean of LibriSpeech [36] are used as training, validation, and evaluation datasets, respectively, which are consistent with the baseline [19]. All used datastes are English speech data with a sampling rate of 16


(a) Impaired spectrogram

(b) Inpainted spectrogram

(c) Ground truth

Figure 4: *An example of the proposed T-F mask-based speech inpainting system for audio editing software.*

kHz. During training and evaluation, speech time sequences are first chunked into 1-second long segments, then transformed into spectrograms by STFT. After that, normalization and transforming to log domain are performed subsequently. Then, spectrograms are multiplied by binary masks to simulate the impaired speech dataset to be inpainted. The shape of masks in the used mask dataset depends on the specific speech enhancement task since partial convolutions have been demonstrated to handle masks with any shape robustly [30]. In this paper, to better compare with baselines, the mask dataset for training is also generated with the same rules as the baseline [19]. Missing types includes both time clips and frequency bands. Besides, we use 512-sample, instead of 256-sample, STFT to increase the frequency resolution of the inpainted spectrogram.

For training, the batch size is set to 16. Adam optimizer with $\beta = 0.9$ is used for generative and adversarial networks. Initial learning rates are 0.02 and 5×10^{-5}. Step size and multiplicative factor of learning rate decay are 200 and 0.96. Each epoch consists of 9600 segments selected from the training dataset randomly, and the training is based on the Nvidia P100. Pre-trained SpeechVGG model used is the one in [37].

3.2. Comparative Experiments

Comparative experiments between a context-based CNN model and the proposed T-F mask-based GAN model are carried out. Similarly, The ablation experiments about effectiveness of the proposed loss function and the two-strategy phase reconstruction algorithm were carried. Experimental results in the PESQ and STOI are shown in Table 2. Missing sizes in the experiments are from 10% to 40% (percentage of missing areas in both the time and the frequency dimension).

As shown in Table 2, the proposed model performs best for all missing sizes in both metrics. The CNN model's scores can follow when the missing size is small, like 10 %. However, there is a noticeable performance deterioration of CNN mode for large missing such as from 20% to 40%, showing that the proposed GAN network with partial convolutions outperforms context-based models in T-F areas missing, especially for large missing. Moreover, compared to the GAN model with the classic joint loss function, the improvement of the GAN model with the proposed joint loss function is consistent, even if a little smaller. The model with TSPRA outperforms that with conventional GLA. Thus, the proposed GAN network, loss function, and TSPRA contribute to the performance improvement of the proposed model in PESQ and STOI.

3.3. Comparison with Baselines

The comparisons between the proposed GAN model and two baselines, LPC [13] and the T-F mask-based deep inpainting model [19], are also performed. These models' average scores of the impaired and inpainted datasets in PESQ and STOI are drawn as points in Figure 3. Four points of each model with missing sizes from 10% to 40% are further fitted to curves. The difference between the horizontal and vertical coordinates of points on the curve denotes the improvement by the model. In other words, the better model performs, the higher its curve is located.

Therefore, for PESQ, a remarkable improvement can be observed on the curve of the proposed GAN model compared to LPC and the deep inpainting model for all missing sizes. For STOI, the improvement of our model is consistent, even if smaller. Overall, our GAN speech inpainting model improves inpainting performance in PESQ and STOI, providing an average increase of 0.3 in PESQ and an average increase of 0.02 in STOI over the T-F mask-based deep inpainting baseline, which corresponds to an average increase of 0.45 in PESQ and an average increase of 0.09 in STOI over the LPC model.

3.4. Example Illustration

An example of the proposed system for audio editing software is illustrated in Figure 4. The impaired spectrogram is generated by randomly masking some holes. The training model used is as Sec. 3.1. As observed, harmonics are inpainted clearly, and the over-smoothing is improved. The audio file of the example and various audio demo are provided at: [1] for presenting inpainting improvement in subjective intelligibility and quality.

4. Conclusions

In this paper, the GAN speech inpainting model is proposed consisting of the GAN magnitude inpainting network using the new joint loss function, and the two-strategy phase reconstruction algorithm. The magnitude inpainting network employs the GAN structure contributing to more real inpainting results by capturing more latent information of speech spectrograms. The proposed loss function with magnitude-based weight loss, VGG loss items can enhance structures of harmonics and explore latent information of speech spectrograms and high-dimensional features. Meanwhile, the proposed network provides inpainted magnitudes with higher resolution for phase reconstruction. The phase reconstruction algorithm adopts two strategies for different magnitudes, inpainting clear harmonics while reducing the buzzes caused by the smooth and low magnitude in high frequency. The experimental results demonstrate that the proposed GAN network, loss function and TSPRA significantly improve the inpainting performance in PESQ and STOI. Moreover, comparison results show that our model with the proposed phase reconstruction algorithm outperforms the conventional and the T-F mask-based deep inpainting baselines. Furthermore, the proposed T-F mask-based speech inpainting system implements the speech inpainting for audio editing software, and is promising to be integrated into other audio enhancement and bandwidth expansion tasks.

[1] `https://github.com/HXZhao1/GSIM`

5. References

[1] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6959–6963.

[2] 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.

[3] L. Thieling and P. Jax, "Generally applicable deep speech inpainting using the example of bandwidth extension," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 451–455.

[4] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2011.

[5] D. Goodman, G. Lockhart, O. Wasem, and W.-C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1440–1448, 1986.

[6] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, 1986.

[7] P. A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen, "Interpolation of long gaps in audio signals using the warped burg's method," in *Proc. 6th Int. Conf. on Digital Audio Effects (DAFx-03)*, 2003, pp. 08–11.

[8] M. Lagrange, S. Marchand, and J.-B. Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 891–905, 2005.

[9] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in *Speech Coding, 2002, IEEE Workshop Proceedings*. IEEE, 2002, pp. 65–67.

[10] S. J. Godsill, P. J. Wolfe, and W. N. Fong, "Statistical model-based approaches to audio restoration and analysis," *Journal of New Music Research*, vol. 30, no. 4, pp. 323–338, 2001.

[11] G. Chantas, S. Nikolopoulos, and I. Kompatsiaris, "Sparse audio inpainting with variational bayesian inference," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2018, pp. 1–6.

[12] R. C. Maher, "A method for extrapolation of missing digital audio data," *Journal of the Audio Engineering Society*, vol. 42, no. 5, pp. 350–357, 1994.

[13] I. Kauppinen and K. Roth, "Audio signal extrapolation–theory and applications," in *Proc. DAFx*, 2002, pp. 105–110.

[14] I. Kauppinen and J. Kauppinen, "Reconstruction method for missing or damaged long portions in audio signal," *Journal of the Audio Engineering Society*, vol. 50, no. 7/8, pp. 594–602, 2002.

[15] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[16] A. Marafioti, N. Holighaus, P. Majdak, N. Perraudin *et al.*, "Audio inpainting of music by means of neural networks," in *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019.

[17] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "Gacela: A generative adversarial context encoder for long audio inpainting of music," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 120–131, 2020.

[18] P. P. Ebner and A. Eltelt, "Audio inpainting with generative adversarial network," *arXiv preprint arXiv:2003.07704*, 2020.

[19] M. Kegler, P. Beckmann, and M. Cernak, "Deep speech inpainting of time-frequency masks," *arXiv preprint arXiv:1910.09058*, 2019.

[20] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[23] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *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.

[25] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 185–200.

[26] J. Zhu, G. Yang, and P. Lio, "How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1669–1673.

[27] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2745–2754.

[28] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual gan," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 205–220.

[29] O. Ronneberger, P. Fischer, and T. 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.

[30] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[32] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[33] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[34] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," in *Proc. DAFx*, vol. 10, 2010, pp. 397–403.

[35] Z. Průša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from stft magnitude," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1154–1164, 2017.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[37] P. Beckmann, M. Kegler, H. Saltini, and M. Cernak, "Speech-vgg: A deep feature extractor for speech processing," *arXiv preprint arXiv:1910.09909*, 2019.