SPEECH SUPER RESOLUTION GENERATIVE ADVERSARIAL NETWORK

Sefik Emre Eskimez[*] *, Kazuhito Koishida*[†]

[*]University of Rochester, 500 Wilson Blvd, Rochester, NY 14627, USA
[†]Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

The goal of speech super-resolution (SSR) or speech bandwidth expansion is to generate the missing high-frequency components for a given low-resolution speech signal. It has the potential to improve the quality of telecommunications. We propose a new method for SSR that leverages the generative adversarial networks (GANs) and a regularization method for stabilizing the GAN training. The generator network is a convolutional autoencoder with 1D convolution kernels, operating along time-axis and generating the high-frequency log-power spectra from the low-frequency log-power spectra input. We employ two recent deep neural network (DNN) based approaches to compare them with our proposed method, including both objective speech quality metrics and subjective perceptual tests. We show that our proposed method outperforms the baseline methods in terms of both objective and subjective evaluations.

Index Terms— generative adversarial networks, speech super-resolution, artificial speech bandwidth extension

1. INTRODUCTION

Speech enhancement is one of the most studied problems in the speech processing field. The primary goal of speech enhancement is to increase the quality and intelligibility of the input speech signal. The majority of work in this field focuses on removing the background noise or reverberation, where some of these works focus on generating the missing high-frequency content to increase the resolution of the speech signal, which is called *artificial speech bandwidth expansion* or *speech super-resolution (SSR)* in the literature. In the rest of this paper, we refer to this problem as SSR.

SSR has applications in many practical scenarios and has a potential for improving the quality of life for people. A typical example is the public switched telephone network (PSTN), where the bandwidth is still limited to a narrowband (300-3400 Hz). In a study [1], it was shown that the users favor high-resolution speech signal in telephony compared to narrowband. Kepler et al. [2] pointed out that the narrowband range is challenging for the hearing impaired population when communicating through telephony. In another study, Liu et al. [3] showed that artificially increasing the resolution from narrowband to wideband (up to 8 kHz) improves the speech recognition rates for cochlear implant users.

In this work, we introduce a novel speech super-resolution neural network that employs adversarial training and a regularization method to stabilize the adversarial training. We are motivated by the success of adversarial training for the single image and video super-resolution. The generator is a sequence-to-sequence convolutional autoencoder network that accepts log power spectrogram (LPS) as input and generates the corresponding high-frequency range LPS. The filters in the convolutional layers are 1D, and they operate along the time axis of the spectrogram. By adopting 1D kernels, we reduce the computational complexity for both training and inference. The resulting system is light-weight with a real-time processing capability on mobile devices and consumer level CPUs. The training procedure is as follows: First, we initialize the generator network by training it with only a reconstruction loss for a few epochs. Then, we train the framework using the adversarial loss in addition to weighted reconstruction loss. During GAN training, we add a weighted gradient penalty to discriminator loss in order to stabilize the process. We use the Centre for Speech Technology Research (CSTR) Voice Cloning Toolkit (VCTK) Corpus [4] for training our system. To confirm the robustness against unseen speakers and recording conditions, we evaluate our system using a completely different dataset from our training set, namely the Wall Street Journal (WSJ0) corpus [5]. We compare our method with baselines [6, 7]. We show that our method outperforms the baseline methods in terms of objective and subjective evaluations. A set of examples is publicly available[1].

The rest of the paper is organized as follows: Section 2 describes the related works. Section 3 outlines the system overview, the neural network framework. In Section 4, we describe the experiment details and present the objective and subjective evaluation results. Section 5 concludes the paper.

2. RELATED WORK

The early works focused on estimating the spectral envelope of the speech signal and model the mapping from narrowband to wideband signals. These works relied on Gaussian mixture models (GMMs) [8–10], hidden Markov models (HMMs) [11–14], and neural networks (NNs) [6, 7, 15–17] to be able to learn the transfer function between the narrowband and wideband signals. Recently, the deep learning based-methods [6, 7] outperformed these approaches.

Li et al. [6] proposed a DNN to predict the log-power spectrogram (LPS) of the wideband from the LPS of the narrowband. To artificially create the missing phase information, they flipped the phase of the low-frequency band as that of the high-frequency band to reconstruct the time domain signal. They showed that their method outperforms the GMM-based methods. Kuleshov et al. [7] proposed to use the raw waveforms directly and introduced an end-to-end network. They used a convolutional autoencoder network with mean-squared error (MSE) objective function. Compared to signal processing based methods, this method is more straightforward regarding implementation since there is no pre-processing. However, it is computationally expensive and might not be suitable for running on the edge devices.

Generative adversarial networks (GANs) [18] are shown to be powerful in image, video and speech generation tasks. In essence,

The work was done when SEE was an intern at Microsoft Research.

[1] http://www.ece.rochester.edu/projects/air/projects/SSRGANc.html

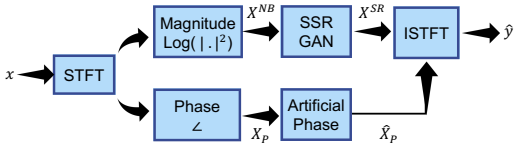

Fig. 1: The proposed speech super-resolution (SSR) system overview during the test time. The short-time Fourier transform (STFT) is applied to time domain signal x to obtain the log-power spectra (LPS) X^{NB} and the phase spectrogram X_P. The narrowband (NB) LPS X^{NB} is fed to SSR-GAN in order to obtain the estimated high-frequency (HF) range LPS, and it is concatenated to the NB LPS to obtain the wideband (WB) LPS $\hat{X}^{SR}$. The phase of the HF range is artificially produced by flipping and repeating the NB phase X_P and adding a negative sign. Finally, the estimated WB LPS and artificial phase are used to reconstruct the time-domain signal $\hat{y}$ by inverse STFT (ISTFT) and overlap-add.

GANs are a zero-sum game that contains multiple neural networks, usually a generator, and a discriminator. The generator tries to deceive the discriminator by generating fake but realistic data, while the discriminator tries to distinguish between the real and fake data. Although GANs yield impressive and realistic results, they suffer from instabilities during training [19]. The researchers focused on stabilizing the GAN frameworks by introducing regularizations [19–23]. Some of these regularization methods add a penalty to the norm of the gradients in order to stabilize the training [19, 21, 23].

GANs have been successfully applied to image and video super-resolution [24, 25]. Since spectrograms are similar to images or video frames, these studies encouraged us to investigate the adversarial networks in the context of speech super-resolution.

Li et al. [26] proposed a speech bandwidth expansion method using adversarial training recently. Their neural network (NN) predicts the Line Spectral Frequencies (LSF) and speech energy of the high band (HB) from LSF, delta LSF and speech energy of the low band signal. The generator and discriminator are fully connected neural networks with four layers. The high-resolution speech signals were synthesized by the EVRC-WB framework [27] and a synthesis filterbank using the predicted speech parameters. Our method and [26] both use GAN framework for SSR. However, our method directly generates the speech spectrograms and employs a regularization method to stabilize GAN training, while [26] synthesizes speech with a synthesis framework using the estimated LSF and energy parameters.

3. PROPOSED METHOD

In the following, we describe how our system works during inference. Let x be the time domain waveform of the narrowband speech. First, the short-time Fourier transform (STFT) is applied to x. Then, the log-power spectrogram (LPS) X^{NB} and the phase spectrogram X_P are computed from X. The high-frequency range LPS, $\hat{X}^{WB}$ is estimated from the X^{NB} using the proposed generator network. The original narrowband and the predicted high-frequency range LPSs are concatenated to get the estimated wideband LPS X^{SR}. We also predict the highest C frequency bins of the narrowband spectrogram, where C is called the *offset* parameter. During concatenation, the narrowband spectrogram less than C frequency bins is concatenated with the predicted high-frequency range. This way, we avoid discontinuities at the concatenation [6]. We follow Li et al. [6] to create an artificial phase by flipping the narrowband phase and reverting the sign. For the 2x super-resolution version, we concatenate this flipped phase with the narrowband phase to obtain an artificial phase $\hat{X}_P$ of the entire wideband signal. For the 4x super-resolution version, we repeat the flipped phase three times. In the last step, we reconstruct the time domain signal using the overlap-add method from the inverse STFT of estimated wideband LPS X^{SR} and artificial phase $\hat{X}_P$. The system overview is shown in Figure 1.

Table 1: Detailed parameters of the proposed network architectures. K and N are the narrowband and the high-frequency range LPS dimensions along the frequency axis, respectively. K is 129 and 65 for 2x and 4x super-resolution scales, respectively. N is 141 and 199 for 2x and 4x super-resolution scales, respectively.

Layers	Activation	Filter No	Filter Size	Strides	Output Shape
The Generator Network					
Input	-	-	-	-	32 × K
Conv	LeakyReLU	256	(7, 1)	(2, 1)	16 × 256
Conv	LeakyReLU	512	(5, 1)	(2, 1)	8 × 512
Conv	LeakyReLU	512	(3, 1)	(2, 1)	4 × 512
Conv	LeakyReLU	1024	(3, 1)	(2, 1)	2 × 1024
Conv	LeakyReLU	512	(3, 1)	(1, 1)	4 × 512
Conv	LeakyReLU	512	(5, 1)	(1, 1)	8 × 512
Conv	LeakyReLU	256	(7, 1)	(1, 1)	16 × 256
Conv	LeakyReLU	N	(7, 1)	(1, 1)	32 × N
Conv	LeakyReLU	N	(9, 1)	(1, 1)	32 × N
The Discriminator Network					
Input	-	-	-	-	32 × (K+N)
Conv	LeakyReLU	1024	(7, 1)	(2, 1)	16 × 1024
Conv	LeakyReLU	1024	(5, 1)	(2, 1)	8 × 1024
Conv	LeakyReLU	1024	(3, 1)	(2, 1)	4 × 1024
Flatten					4096
FC	LeakyReLU	2048			2048
FC	Sigmoid	1			1

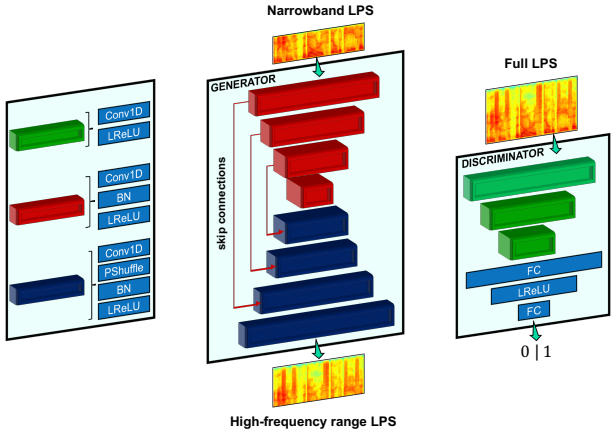

Fig. 2: The proposed network architectures for the generator (middle) and the discriminator (right). Each rectangular block is a convolutional layer with structures color coded and detailed on the left subfigure. Notations: *BN* - batch normalization layer, *FC* - fully connected layer, *LReLU* - LeakyReLU activation, and *PShuffle* - pixel shuffle or sub-pixel layer, *LPS* - log-power spectrogram.

3.1. Network Architecture

The generator is a sequence-to-sequence model, which accepts the narrowband LPS with T time steps and outputs the high-frequency range LPS with T time steps. We use a common bottleneck autoencoder architecture described in [7]. The convolutional kernels are 1D, which operates on the time axis of the LPSs. Compared to 2D kernels, the computational cost is much lower, permitting real-time processing of the network on CPUs and mobile devices. We use batch normalization (BN) layers after the convolutional layer followed by leaky rectified linear units (LeakyReLU) activations with a slope of 0.2, except for the output layer, where we use linear ac-

tivation and do not use a BN layer. We employ sub-pixel (or pixel shuffle) layers introduced in [28] for upsampling, which is shown useful for image and video super-resolution.

The discriminator includes three convolutional layers that are followed by two fully connected (FC) layers. We use LeakyReLU activation with a slope of 0.2 in all layers, except for the output layer, where we use a linear activation function. Since the BN layers lead to instabilities during training in the discriminator network, which is especially true if the discriminator loss is regularized [19, 23], we do not use BN layers. The discriminator network receives the concatenated narrowband and high-frequency range LPSs as input. The high-frequency range LPS could be coming directly from the data distribution or generated by the generator network. The details of both network architectures are shown in Table 1.

3.2. Training Objective Functions

First, we train the generator network with only a reconstruction loss for several epochs to initialize. The generator is typically trained to produce the overly smooth results after this initialization. To obtain sharper and more detailed LPSs, we switch to using an adversarial loss (GAN loss) in addition to the reconstruction loss. We use log-spectral distance (LSD) (or log-spectral distortion) function as our training objective. The LSD measures the distance between two spectrograms in decibels, and it is mathematically defined as follows:

$$l_{LSD} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}[X^{HR}(l,k) - X^{SR}(l,k)]^2}, \quad (1)$$

where K is the number of frequency bins, and X^{HR} and X^{SR} are the ground truth and estimated LPSs, respectively.

The original generative adversarial network (GAN) is a two player, zero-sum (minimax) game between a generator and a discriminator. We formulate this problem in the context of SSR, which can be defined as follows:

$$\begin{gathered}\min_{\theta}\max_{\psi}\mathbb{E}_{\mathbb{P}}[\log D_{\psi}(X^{HR})] + \mathbb{E}_{\mathbb{Q}}[\log(1 - D_{\psi}(G_{\theta}(X^{NB})))], \\ \mathbb{P} : X^{HR} \sim p(X^{HR}), \\ \mathbb{Q} : X^{NB} \sim p(X^{NB}),\end{gathered} \quad (2)$$

where X^{HR} is the high-resolution data (real data), X^{NB} is the narrowband data. $G_{\theta}(\cdot)$ is the generator and $D_{\psi}(\cdot)$ is the discriminator, where θ and ψ are the trainable parameters. $\mathbb{P}$ is the distribution of real data and $\mathbb{Q}$ is the distribution of the narrowband data. The generator ($G_{\theta}(\cdot)$) handles the concatenation of narrowband and high-band spectrograms. This notation can be simplified as follows:

$$\min_{\theta}\max_{\psi}\mathbb{E}_{\mathbb{P}}[\log \varphi_R] + \mathbb{E}_{\mathbb{Q}}[\log(1 - \varphi_F)], \quad (3)$$

where φ_R and φ_F are the discriminator output for real and fake data, respectively.

We add a penalty on the weighted gradient-norms of the discriminator as described in [23] to stabilize the GAN training. The regularization term is described as:

$$\Omega = \mathbb{E}_{\mathbb{P}}[(1 - \varphi_R)^2\|\nabla\phi_R\|^2] + \mathbb{E}_{\mathbb{Q}}[\varphi_F^2\|\nabla\phi_F\|^2], \quad (4)$$

where $\phi = \sigma^{-1}(\varphi)$, and σ is the sigmoid activation used in generating the output of the discriminator. We add this term to the objective function of the discriminator as follows:

$$l_{DIS} = \mathbb{E}_{\mathbb{P}}[\log \varphi_R] + \mathbb{E}_{\mathbb{Q}}[\log(1 - \varphi_F)] - \frac{\gamma}{2}\Omega, \quad (5)$$

where γ is the weight for the regularization term.

The generator loss is the weighted sum of the reconstruction loss and the GAN loss, and defined as follows:

$$l_{GEN} = \mathbb{E}_{\mathbb{Q}}[-\log(D_{\psi}(G_{\theta}(X^{NB})))] + \lambda l_{LSD}, \quad (6)$$

where l_{LSD} is the objective function described in Equation 1 and λ is the weighting parameter for the LSD loss.

4. EXPERIMENTS

We used the CSTR Voice Cloning Toolkit Corpus (VCTK) to train our network, which is initially designed for training text-to-speech (TTS) synthesis systems. The recordings are 16-bit WAV files with 48 kHz sampling rate and contain clean speech. There are a total of 109 English speakers with different accents, where each speaker utters 400 sentences. We used the utterances from six random speakers as a validation set and used the rest for training the network. To create our training pairs, we applied the band-limited sinc interpolation method described in [29] to high-resolution signal in order to obtain the downsampled version.

In order to evaluate the generalization ability of our network, we employed another dataset for evaluation that has different speakers and recording conditions than the VCTK corpus, namely the Wall Street Journal (WSJ0) corpus. The sampling rate of the recordings is 16 kHz, where they contain natural background noise. In our objective evaluations, we use a random subset with 5000 samples (around 12 hours) from this dataset.

Our network was trained for 50 epochs using only the LSD loss (Equation 1) with a learning rate of 10^{-4}, and it was trained for another 100 epochs using GAN plus LSD loss (Equation 6) with a learning rate of 10^{-5}. We determined the number of epochs experimentally. The number of time-steps of our input and output spectrograms were set to 32. We used Adam optimizer to train the generator network and RMSProp optimizer to train the discriminator network with a mini-batch size of 64. The input and output LPSs were normalized to have zero mean and unit variance; We calculated these statistics from the training data and applied them during inference. The K variable shown in Table 1 was 129 for 2x experiments and 65 for 4x experiments. The frequency offset value was calculated according to the following formula:

$$C = floor(\frac{K}{10}) + 1, \quad (7)$$

where K is the number of frequency bins in the input spectrogram. The N variable shown in Table 1 was set to 141 and 199 for 2x and 4x super-resolution scales, respectively. We set the γ variable shown in Equation 5 to 2.

We employed two baseline methods from existing works described in Section 2. The first baseline is an STFT-based method [6], which we name as *BL1* through the rest of the paper. Since this work only considers 2x SSR, we did not implement 4x SSR version. The second baseline is raw waveform-based method [7], which we name as *BL2* through the rest of the paper. We adopted the code provided by the authors to reproduce the results for 2x and 4x SSR. We name our proposed method as *SSR-GAN*.

4.1. Objective Metrics

We employed the LSD defined by Equation 1, segmental signal-to-noise ratio (SegSNR) [30], and perceptual evaluation of speech quality (PESQ) [31] objective metrics in order to evaluate and

Table 2: The objective evaluation results for 2x and 4x SSR experiments. Our method (SSR-GAN) outperforms the baselines for all metrics. *LSD HF* shows the LSD value calculated only for the high-frequency range, where *LSD Full* shows the LSD value calculated for the whole spectrogram.

Scale	Method	LSD HF (dB)	LSD Full (dB)	SegSNR (dB)	PESQ
2x	BL1 [6]	9.32	7.06	15.73	4.21
	BL2 [7]	10.56	7.64	14.96	4.19
	SSR-GAN	**8.20**	**5.95**	**19.64**	**4.32**
4x	BL2 [7]	16.20	14.96	8.24	2.89
	SSR-GAN	**12.90**	**10.24**	**13.01**	**3.40**

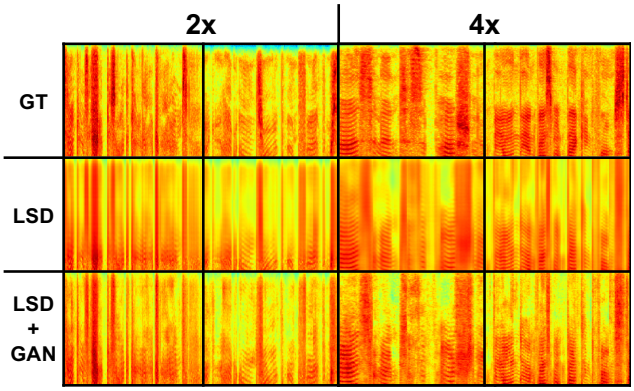

Fig. 3: Spectrogram examples for 2x and 4x are shown. The samples are randomly selected from the WSJ0 corpus (unseen speakers). The first row shows the ground truth high-frequency range spectrograms. The second and third rows show the generated high-frequency range spectrograms of the proposed network trained with only the LSD loss (second rows) and with both LSD and GAN losses (third rows).

compare our method with the baseline methods. These metrics are widely used in speech enhancement, and SSR works. PESQ measures speech quality, and it is standardized by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T). SegSNR is the signal-to-noise (SNR) ratio, averaged over segments of audio samples, and defined as follows:

$$SegSNR = \frac{1}{L}\sum_{l=1}^{L} 10log\frac{\sum_{n=1}^{N}[x(l,n)]^2}{\sum_{n=1}^{N}[x(l,n)-\hat{x}(l,n)]^2}, \quad (8)$$

where L is the number of segments, and N is the number of data points in the utterance. For SegSNR and PESQ, the higher value is better; for LSD, the lower value is better.

4.2. Results

The objective evaluation results are shown in Table 2. Our method outperforms both baselines in 2x and 4x SSR tasks with a good margin in terms of all of the three objective evaluation metrics. LSD values are improved by around 1.1 dB compared to *BL1*. For SegSNR, the improvement is around 3.9 dB. There is a slight PESQ improvement, which is around 0.1. The improvement of our method, compared to BL2, is more noticeable in the 4x setting. The LSD improvements are around 3.3 dB and 4.7 dB for high-frequency range and whole spectrum, respectively. The SegSNR is improved by around 4.7 dB. Compared to the 2x scale, PESQ is improved significantly, which is around 0.5.

Figure 3 shows the example spectrograms, where the first row is the ground truth high-frequency range spectrogram, the second row is the high-frequency range spectrograms obtained from a neural network trained only with LSD loss, and the third row shows SSR-GAN results, for 2x and 4x, respectively. Note that the LPSs on the second rows are overly smooth. After the GAN training (third row), the results are sharper, containing fine details and more energy.

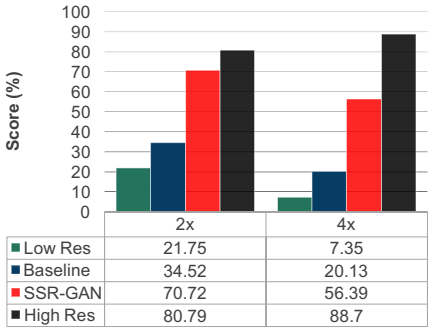

Fig. 4: The subjective test results for 2x and 4x scales are shown.

4.3. Subjective Evaluations

We carried out subjective evaluations to test how our method compares to baselines and ground-truth data in terms of human perception. We generated two test sets where each of them contained 40 utterances, for 2x and 4x scales. The sets included the narrowband signal, ground-truth high-resolution signal, predicted super-resolution signals of our method and the baselines. We wanted to limit the test time for each subject within 30 minutes; therefore we used only the samples from one of the baseline methods for each resolution scale, we employed [6] and [7] for 2x and 4x, respectively. There were a total of 20 volunteers, where each of them evaluated 80 samples. Each volunteer was trained by listening to the 5 pairs of low and ground-truth high-resolution utterances. The testing samples were randomly presented to the volunteers, and they assigned a score between 0 and 100 for each sample, where 0 corresponds to the low-resolution signal and 100 corresponds to the high-resolution signal.

The 2x and 4x scale experiment results are shown in Figure 4. The ground truth high-resolution speech has a score of 80.79%, which is followed by our method with a score of 70.72%. The low-resolution signal and *BL1* has lower scores, which are 21.75% and 34.52%, respectively. Since SSR-GAN score is close to the high-resolution signal, we can conclude that for 2x scale, SSR-GAN can convince the listeners in terms of speech quality and can outperform the baseline method. The 4x experiments are more challenging, and the missing phase information is more apparent compared to 2x experiments. The gap between the high-resolution score and SSR-GAN is around 32%. SSR-GAN can still outperform the baseline method and has more than 50% score.

5. CONCLUSION

In this work, we presented a novel method that leverages adversarial training for the speech super-resolution. Through objective and subjective evaluations, we showed that our method outperforms the DNN based baseline methods. The subjective evaluations revealed that for a 2x resolution scale, our method could score close to the ground-truth high-resolution signal, and could obtain a decent performance for a 4x resolution scale. Our method is light-weight in terms of computational complexity and capable of running in real-time on edge devices. Our future work includes estimating the phase information along with the spectrograms.

6. REFERENCES

[1] ITU, "Paired comparison test of wideband and narrowband telephony," in *Tech. Rep. COM 12-9-E*. Mar. 1993.

[2] Laura Jennings Kepler, Mark Terry, and Richard H Sweetman, "Telephone usage in the hearing-impaired population.," *Ear and hearing*, vol. 13, no. 5, pp. 311–319, 1992.

[3] Chuping Liu, Qian-Jie Fu, and Shrikanth S Narayanan, "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. EL77–EL83, 2009.

[4] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2016.

[5] John Garofalo, David Graff, Doug Paul, and David Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[6] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4395–4399.

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

[8] Kun-Youl Park, "Narrowband to wideband conversion of speech using gmm based transformation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2000, pp. 1843–1846.

[9] Samir Chennoukh, A Gerrits, G Miet, and R Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 1, pp. 665–668.

[10] Hyunson Seo, Hong-Goo Kang, and Frank Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6087–6091.

[11] Peter Jax and Peter Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2003, vol. 1, pp. I–I.

[12] Guo Chen and Vijay Parsa, "Hmm-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–709.

[13] Patrick Bauer and Tim Fingscheidt, "An hmm-based artificial bandwidth extension evaluated by cross-language training and test," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4589–4592.

[14] Geun-Bae Song and Pavel Martynovich, "A study of hmm-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[15] Bernd Iser and Gerhard Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[16] Juho Kontio, Laura Laaksonen, and Paavo Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 873–881, 2007.

[17] Johannes Abel and Tim Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[19] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin, "Which training methods for gans do actually converge?," in *International Conference on Machine Learning*, 2018, pp. 3478–3487.

[20] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[22] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016.

[23] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Advances in Neural Information Processing Systems*, 2017, pp. 2018–2028.

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network.," in *CVPR*, 2017, vol. 2, p. 4.

[25] Alice Lucas, Santiago Lopez Tapia, Rafael Molina, and Aggelos K Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *arXiv preprint arXiv:1806.05764*, 2018.

[26] Sen Li, Stéphane Villette, Pravin Ramadas, and Daniel J Sinder, "Speech bandwidth extension using generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5029–5033.

[27] 3GPP2 C.S0014-C v1.0, "Enhanced variable rate codec, speech service option 3, 68 and 70 for wideband spread spectrum digital systems," .

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[29] Julius O. Smith, "Digital audio resampling home page center for computer research in music and acoustics (ccrma)," .

[30] Paul Mermelstein, "Evaluation of a segmental snr measure as an indicator of the quality of adpcm coded speech," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1664–1667, 1979.

[31] AW Rix, J Beerends, M Hollier, and A Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, vol. 862, 2001.