ARTIFICIAL BANDWIDTH EXTENSION USING A CONDITIONAL GENERATIVE ADVERSARIAL NETWORK WITH DISCRIMINATIVE TRAINING

Jonas Sautter, Friedrich Faubel, Markus Buck

Nuance Communications,
Speech Signal Enhancement,
89077 Ulm, Germany
firstname.lastname@nuance.com

Gerhard Schmidt

Kiel University,
Digital Signal Processing and System Theory,
24143 Kiel, Germany
gus@tf.uni-kiel.de

ABSTRACT

The aim of artificial bandwidth extension is to recreate wideband speech (0 - 8 kHz) from a narrowband speech signal (0 - 4 kHz). State-of-the-art approaches use neural networks for this task. As a loss function during training, they employ the mean squared error between true and estimated wideband spectra. This, however, comes with the drawback of over-smoothing, which expresses itself in strongly underestimated dynamics of the upper frequency band. We previously proposed to tackle this problem by discriminative training, i.e., a modification of the loss function that is designed to improve the separation between fricatives and vowels. Other authors instead took a generative adversarial network (GAN) approach. This was motivated by the fact that GANs demonstrated big reductions of over-smoothing in speech synthesis. In this work, we combine these two approaches. In particular, we show that conditional GANs improve the speech quality by a CMOS score of 0.28 compared to GANs while the combined approach yields an improvement of 0.84.

Index Terms— artificial bandwidth extension, generative adversarial networks, discriminative training

1. INTRODUCTION

The aim of artificial bandwidth extension (BWE) is to improve the quality of narrowband (NB) telephony speech (0 - 4 kHz) by artificially extending the signal to wideband (WB), i.e., a bandwidth of 0 - 8 kHz. This is achieved by estimating the missing upper band (UB) between 4 and 8 kHz based on the NB spectrum. In most approaches, the signal is first transformed to the frequency domain and then decomposed into a spectral envelope and an excitation signal. Both these parts are then extended separately, which simplifies the estimation problem. The extension of the excitation signal is typically achieved with rather simple methods like spectral shifting. This is perfectly sufficient, as the introduced degradation of the speech quality is rather low [1, 2]. In early BWE approaches, the estimation of WB features was often performed with codebooks, Gaussian mixture models (GMM), or hidden Markov models (HMM).

More recent publications have shown that the performance can be improved by using deep neural networks (DNN) [3, 4]. In most of these approaches, the DNN is used to estimate the spectral envelope. Other approaches use the DNN to estimate the entire magnitude spectrum [5] or even the complete time-domain waveform [6]. Besides different DNN training targets, various DNN structures have been employed. Basic feed-forward neural networks with fully connected layers were used first [7, 8, 9]. The mean squared error (MSE) loss function used in these approaches, however, leads to over-smoothing [10, 11]. This degrades the speech quality because of low dynamics of the inserted energy. Different types of recurrent neural networks (RNN) have been implemented [12, 13] in an effort to model the time dependencies and dynamics more accurately. Convolutional neural networks (CNN) have been examined in a direct waveform modeling approach to BWE [14] as well as in a combined CNN / RNN approach for BWE envelope estimation [15].

In 2014, Goodfellow et al. proposed a network architecture that may be more suitable for this task: the generative adversarial network (GAN) [16]. It employs a generator network that learns to generate data and a discriminator network that learns to distinguish between real and generated data. These two networks with opposite objectives are trained in tandem. Hence, over the training iterations, the generator network generates data with its current model. The discriminator network learns to identify the "mistakes" the generator network makes during this process. The identified mistakes are propagated back to the generator network in the form of a gradient, such that it can continuously improve its model. Convergence is reached when the output becomes indistinguishable from real data. Generative adversarial networks have been applied successfully to different fields of research, ranging from image processing [17, 18] to speech enhancement [19, 20, 21]. In particular, it has been shown that they can solve the over-smoothing problem in statistical parametric speech synthesis [22, 23]. This suggests that they may also be beneficial for BWE. In 2018, Li et al. picked up on this idea and showed that GANs outperform codebook and HMM approaches on the BWE task [4]. However, they did not directly compare its performance to other DNN approaches.

In this contribution, we propose to replace the GAN in [4] by a conditional generative adversarial network (CGAN). This is motivated by the fact that the original GAN approach [16] was designed to generate data from statistically independent random noise. In BWE, however, the estimation problem is an input-dependent task, in which NB features are mapped to WB features. CGANs [24] take this explicitly into account by training the discriminator with both WB and NB features as an input. This enables the discriminator to learn the conditional classification task whether its input is real or generated WB data, based on the given NB data. Next to replacing the GAN from [4] by a CGAN, we add a discriminative term to the cost function in order to improve the separation between fricatives and vowels [11]. This has been found to reduce the over-smoothing problem of BWE compared to plain-vanilla MSE training [11]. The proposed combination of CGAN BWE with the discriminative term gives a total CMOS improvement of 1.7 over NB speech, an improvement of 0.84 over plain GAN training and an improvement of 0.56 compared to CGAN training without discriminative term.

The remainder of this paper is organized as follows: In Section 2, we briefly introduce the architecture and the training algorithm for DNN, GAN, and CGAN models and we describe discriminative training. The BWE algorithm using DNNs is explained in Sec. 3. In the evaluation (Sec. 4), the BWE approaches using different DNNs are compared and the results are discussed. The findings are summarized in Sec. 5.

2. DNN TRAINING METHODS

This section briefly describes the neural network architectures and training methods used in this paper, including basic mean square error training, adversarial training, and discriminative training.

2.1. Mean Square Error Training

The baseline approach for BWE uses a simple regression DNN R with two fully connected hidden layers. The basic setup is described in more detail in [11]. As in most regression tasks, the MSE between real and estimated output features, y and $\hat{y} = R(x)$, is used as loss function that is to be minimized during training:

$$\mathcal{L}_{\text{MSE}}(R) = \mathbb{E}_{x,y}\left[\|R(x) - y\|_2^2\right]. \tag{1}$$

For BWE, the input features x are narrowband (NB) feature vectors. The output features are wideband (WB) feature vectors. Initialization of the weight matrices of the DNN is performed by pre-training with a stacked auto-encoder.

2.2. Generative Adversarial Network Training

For GAN training, the network architecture needs to be modified. More precisely, the regression network R is replaced by a generator network G, and an auxiliary discriminator network D is added during the training stage (see Fig. 1a). These two networks have opposite tasks. While the generator network learns to generate data $\hat{y}$ that can hardly be distinguished from real data y, the discriminator network learns to distinguish real data from generated data. In other words: the generator and discriminator networks are "adversaries" during the training process. Hence, the name GAN.

In the original GAN approach [16], G gets random noise z as input. For BWE, the random noise is replaced by a NB feature vector x [4]. The output of G is the corresponding WB feature vector $\hat{y} = G(x)$, as in the simple regression DNN from above. The discriminator network D classifies if its input is real or generated by G, i.e., it is trained to produce $D(y) = 1$ and $D(\hat{y}) = 0$, respectively. This objective can be formulated as finding the maximum of the following function with respect to D [17]:

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_y\left[\log\left(D(y)\right)\right] + \mathbb{E}_x\left[\log\left(1 - D(G(x))\right)\right]. \tag{2}$$

The generator network G, on the other hand, is trained to minimize (2), i.e., it learns to trick D into believing that the data it generated is real data. It is important to note that the training of G and D is done alternately, i.e., the weights of D are not updated while training G and vice versa. In recent approaches [4, 17, 26], it has been found to be beneficial to mix the pure GAN loss function $\mathcal{L}_{\text{GAN}}(G, D)$ with the minimum mean square error loss function from (1). This can be interpreted as not allowing G to deviate too strongly from the target features y while becoming indistinguishable for D from real data. Following these approaches, we use the following combined objective for GAN training [17]:

$$\min_G \max_D \mathcal{L}_{\text{MSE}}(G) + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}}(G, D). \tag{3}$$

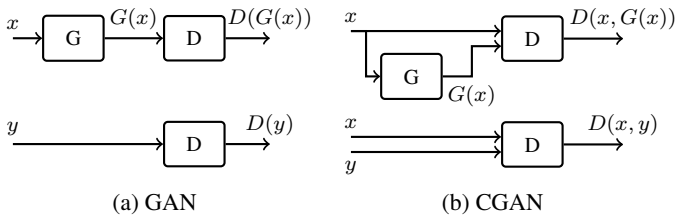


Fig. 1. Differences between GAN and CGAN training. The upper block diagrams show the path for generated data with a generator network G. The lower diagrams show the path for real data. D is trained using the upper and the lower method simultaneously on tuples of x and y.

After optimal convergence, G should estimate the WB features so accurately that D cannot distinguish the output of G from real WB features.

2.3. Conditional GAN Training

In standard GANs, the discriminator network D gets real or estimated WB features, y or $\hat{y} = G(x)$, as an input. The idea behind conditional GANs (cGANs) [24, 17] is to improve the discrimination by also providing the corresponding NB input features x belonging to y or $\hat{y} = G(x)$. More specifically, $D(y)$ and $D(G(x))$ are replaced by $D(x, y)$ and $D(x, G(x))$, as shown in Fig. 1b. This enables cGANs to decide whether the estimated output features are a good fit for the given input features. A standard GAN, in contrast, can just evaluate if the estimated output features seem realistic in general. Apart from this difference, the training procedure is identical. Just the loss function (2) needs to be modified as follows [17]:

$$\begin{aligned}\mathcal{L}_{\text{CGAN}}(G, D) =& \mathbb{E}_{x,y}\left[\log\left(D(x, y)\right)\right] + \\ & \mathbb{E}_x\left[\log\left(1 - D\left(x, G(x)\right)\right)\right].\end{aligned} \tag{4}$$

2.4. Discriminative Training

GAN and CGAN training are suitable alternatives to MSE training if a regression task is subject to the over-smoothing problem [22]. Unfortunately, they do not completely solve the problem when combined with the MSE like in our case. Hence, we add a "discriminative" term to the loss function that forces the network to better preserve the differences between sharp fricatives and vowels [11]:

$$\mathcal{L}_{\text{d}}(G) = \left|\frac{SFPR(G(x)) - SFPR(y)}{SFPR(y)}\right|^2. \tag{5}$$

In this euqtaion, SFPR stands for the sharp fricative power ratio [11], i.e., the ratio between the UB energy of sharp fricatives (*s* and *z*) and the UB energy of all other phonemes. The idea behind this loss function is to punish deviations of the SFPR between real WB features y and estimated WB features $\hat{y} = G(x)$, such that the network "learns" to preserve the energy difference of the classes.

3. PROPOSED BWE ALGORITHM

The BWE algorithm used in this work is shown as a block diagram in Fig. 2. It works entirely in the frequency domain. For the extension, the NB spectrum S_{NB} is first decomposed into a spectral envelope

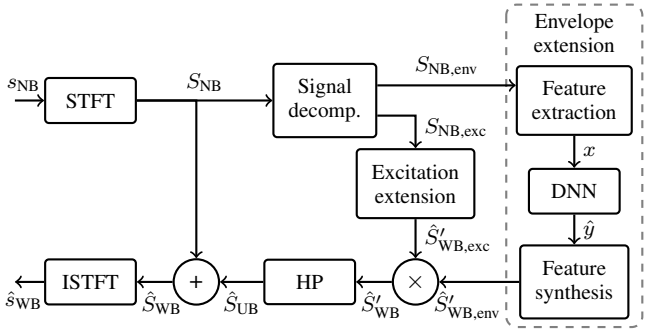


Fig. 2. Block diagram of BWE using a DNN where HP denotes a high-pass at 4 kHz, STFT denotes the short-time Fourier transform and ISTFT the inverse STFT.

$S_{\text{NB,env}}$ and an excitation signal $S_{\text{NB,exc}}$. These parts are then extended separately to a bandwidth of 8 kHz. While the NB excitation is extended with multiple spectral shifting [2], the spectral envelope is extended with a neural network. This is achieved by extracting an input feature vector x for each frame, based on the NB spectrum, and then mapping these NB features to a WB feature vector y with DNN regression. After this step, the estimated WB envelope $\hat{S}'_{\text{WB,env}}$ is reconstructed from the WB feature vector and subsequently multiplied with the estimated WB excitation $\hat{S}'_{\text{WB,exc}}$ [11]. To avoid the introduction of artifacts, the NB part of the estimated WB spectrum is replaced with the original NB spectrum.

In this work, the input feature vector x of dimension 135 consists of 30 NB Mel-frequency Cepstral coefficients (MFCCs), their first and second derivatives in time (Delta features) as well as several other features. A more detailed description of the feature set is found in [27]. The output feature vector y of dimension 30 consists of 30 WB MFCC coefficients. The DNNs are trained offline, on a training data set. For the baseline regression DNN, the mean square error $\mathcal{L}_{\text{MSE}}$ is used as a loss function. For GAN training, the combined objective from (3) is used. The weight λ_{GAN} is set to 0.1, as this was found to give good results in preliminary experiments. The CGAN is trained analogously, just using the CGAN loss function (4) for the discriminator network instead of (2).

This paper proposes to combine CGAN training (Sec. 2.3) with discriminative training (Sec. 2.4). This is motivated by the fact that both approaches tackle the over-smoothing problem in a different way. Hence, they may profit from each other. The combination is achieved by simply adding the discriminative loss term $\mathcal{L}_{\text{d}}(G)$ from (5) to the combined objective (3) of GAN training with the CGAN loss function (4):

$$\min_G \max_D \mathcal{L}_{\text{MSE}}(G) + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{CGAN}}(G, D) + \lambda_{\text{d}} \cdot \mathcal{L}_{\text{d}}(G), \quad (6)$$

where the weight λ_{d} was set to $2.5 \cdot 10^{-3}$. For all GAN / CGAN trainings, the learning rate of the discriminator training was set to two times the learning rate of the generator training. To avoid convergence to a local minimum, G was trained independently of D for 2000 steps when the training process started. Subsequently, D and G were trained alternately, with the weights of D staying fixed while G was trained and vice versa. In the following evaluation, G has exactly the same architecture as the regression DNN R from Sec. 2.1, with the only difference that no pre-training is applied. This makes the different approaches comparable.

4. EVALUATION

Finding objective measures for the evaluation of a BWE algorithm is a difficult task. Especially the well known measures PESQ and POLQA are not reliably predicting the overall speech quality of a BWE signal [28]. Therefore, the perceived speech quality assessment is done with subjective listening tests. However, some objective measures are inspected to find out why listeners prefer a BWE method to another one. The evaluation setup is similar to that in [11]. The database used for training, validation and testing was the TIMIT corpus of American English speech [29]. DNN training was performed with the Adam optimizer [30] and an initial learning rate of $1 \cdot 10^{-5}$. L2-regularization and dropouts [31] were applied in all cases. The regression network R and the GAN networks G and D consist of two hidden layers with 128 nodes each. Rectified linear units (ReLU) [25] were used as activation function. In the following subsections, the listening test setup is described. The results of the listening tests are shown and discussed. And the objective measures are introduced and evaluated.

4.1. Listening Test Setup

We conducted subjective listening tests to evaluate the perceived speech quality of the presented BWE methods. The data set that was used consisted of 8 short German sentences, of which 4 were spoken by a male and 4 by a female speaker. Six conditions were tested in a comparison category rating (CCR) [32] test:

- **WB**: Clean speech with a bandwidth of 8 kHz
- **NB**: WB filtered with a low-pass at 4 kHz
- **DNN**: BWE using a regression DNN (see Sec. 2.1)
- **GAN**: BWE using a modified GAN (see Sec. 2.2)
- **CGAN**: BWE using a CGAN (see Sec. 2.3)
- **CGAND**: BWE using a CGAN with discriminative training (see Sec. 2.4)

All BWE approaches were applied to the generated NB files. Following the ITU-T recommendation [32], we mapped the seven possible answers from the CCR scale (*Much worse* to *Much better*) to integer values from -3 to 3. CGAND was compared to all other conditions once for each sound file, as CGAND is the proposed method in this paper. The test consists of direct comparisons of two conditions A and B. The order of the comparisons and the assignment of A and B was randomized. Additionally, the reliability of listeners was tested by adding an anchor point that presents the same CGAND sample as both A and B, for each sound file. This results in 6 comparisons per sound file or 48 comparisons per participant. 24 listeners, aged between 25 and 59 years, participated in the test (21 male, 3 female). 21 of them were experienced listeners in the field of speech enhancement. All participants were German speakers. This is important because the listeners have to know how the words should be pronounced.

4.2. Listening Test Results

The results of the CCR listening test are presented in Fig. 3 in terms of comparison mean opinion score (CMOS) values. The rank order shows that every step towards more complexity improved the perceived speech quality. None of the 95% confidence intervals overlap. This indicates that all the differences are statistically significant. Remarkable is the high quality improvement from NB to CGAND of 1.7 CMOS points. This is higher than the degradation of CGAND

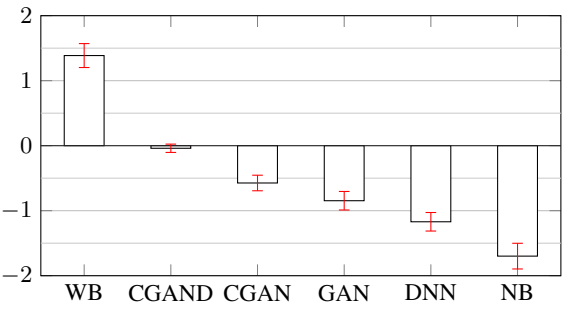


Fig. 3. Mean CMOS ratings (bars) and 95% confidence intervals CI_{95} (red error plots). Note that the limits of the y-axis do not reflect the whole range of possible values which is $[-3..3]$ for CMOS.

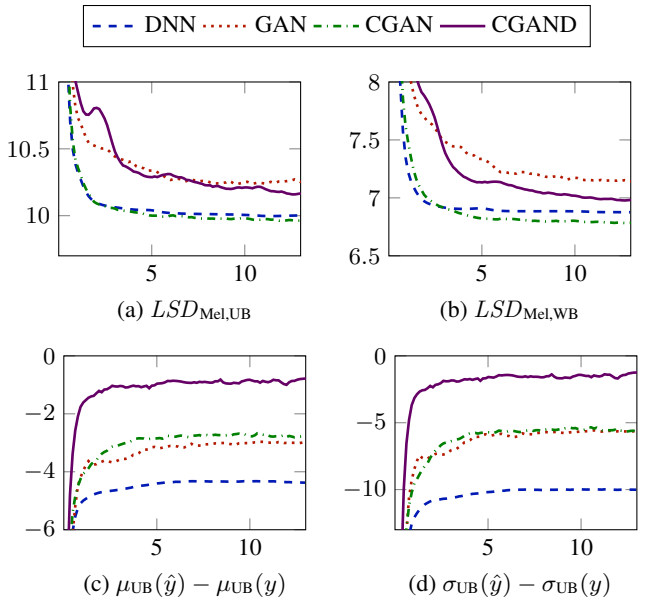
compared to WB (1.39 CMOS points). It is also way higher than the improvement that was measured between NB and a basic DNN with discriminative training in comparable listening tests (1.24 CMOS points) [11]. Many listeners gave the feedback that they were not sure whether they preferred a higher speech bandwidth or slight artifacts. Especially some expert listeners rated the NB speech quality very high. A study with only non-experts might have shown an even higher quality improvement between NB and BWE. The GAN training is roughly comparable to the training proposed by Li et al. [4]. The listening tests show that the performance could be improved significantly by the proposed CGAND training (1.14 points on the CMOS scale). It was noticed in a small pre-study that the effects of applying discriminative training and changing the GAN to a CGAN seem to be orthogonal. The objective results in Sec. 4.3 also show that CGAN training tends to minimize the distance measure while discriminative training optimizes the statistical distribution.

4.3. Objective Quality Measures

In informal listening tests and interviews with participants of the subjective evaluation, we observed some frequent causes for a low subjective speech signal quality. These are similar to the main challenges for BWE that were often formulated in previous studies: A first effect that leads to low perceived speech quality is a lisping sound. It occurs if the introduced UB energy for frames with sharp fricatives like s and z is not high enough. A second effect that leads to a perceived signal degradation is the insertion of high UB energy for frames with vowels or even silence. These two effects show that the perceived speech quality is low if the introduced energy is too evenly distributed. The energy distribution is assessed in this study by inspecting the mean μ_{UB} and the standard deviation σ_{UB} of the frame-wise UB energy over time p_{UB} [11]. The deviation of these measures, which is introduced by UB estimation, is plotted in Fig. 4c and 4d. Both deviations are relatively large for the basic DNN, which indicates some amount of over-smoothing. While GAN and CGAN lead to more or less the same improvement regarding the two distribution measures, the deviations are highly reduced by discriminative training (CGAND).

The presented measures do not take into account how the UB energy is distributed over frequencies or whether the right frames are chosen for a strong extension. These two aspects can be monitored with log-spectral distance (LSD) measures that are calculated based on just the UB or the entire WB Mel spectrum. The LSD based on the WB Mel spectrum can be written as:

$$LSD_{\text{Mel,WB}}(y, \hat{y}) = \mathbb{E}_{y,\hat{y}}\left[\|m_{\text{WB}}(\hat{y}) - m_{\text{WB}}(y)\|_2\right]. \quad (7)$$



(a) $LSD_{\text{Mel,UB}}$ (b) $LSD_{\text{Mel,WB}}$

(c) $\mu_{\text{UB}}(\hat{y}) - \mu_{\text{UB}}(y)$ (d) $\sigma_{\text{UB}}(\hat{y}) - \sigma_{\text{UB}}(y)$

Fig. 4. Performance measures, plotted while training four different DNN models (DNN, GAN, conditional GAN (CGAN) and CGAN with discriminative training (CGAND)). The x-axes denote the training steps in units of 10^5 while the values on the y-axes are given in dB. The plots in the first row show log-spectral distance measures, based on either the UB Mel spectrum (a) or the WB Mel spectrum (b). The measures in the second row are based on the UB energy per frame, namely the deviation of the mean μ_{UB} (c) and the deviation of the standard deviation σ_{UB} (d).

Here, $m_{\text{WB}}(y)$ is a function that converts the MFCC coefficients y to a logarithmic Mel spectrum by inverse discrete cosine transform (IDCT). $LSD_{\text{Mel,UB}}$ is calculated analogously with the only difference that m_{UB} just returns the UB Mel frequencies [11]. Both distance measures are low for the baseline DNN that underestimates the UB dynamics. While the GAN achieves a better energy distribution over time, the distance measures increase. This increased distance can be recovered by using a CGAN, without degrading performance on the distribution measures σ_{UB} and μ_{UB}. When adding discriminative training, the distance measures increase again. So, a trade-off seems to be made between correct distribution and low spectral distance to maximize the perceived speech quality.

5. CONCLUSION

In this paper, we apply a conditional GAN with discriminative regression training to BWE. The objective function for a basic GAN does not take the NB input features into account. The conditional training ensures that the CGAN can learn the dependencies between NB and WB features by feeding the NB inputs to the discriminator part. Using cGANs instead of GANs yielded a significant improvement in subjective listening tests. The results of our BWE using a CGAN are, however, still subject to the over-smoothing problem. Hence, in this work, we propose to combine CGAN training with a discriminative loss term that preserves the energy differences between different phoneme classes in the UB. This leads to a big improvement of the perceived speech quality and results in a CMOS score of 1.7 when compared to NB.

6. REFERENCES

[1] J. Abel, M. Kaniewska, C. Guillaume, et al., "A subjective listening test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean," in *Proc. of ICASSP*, Shanghai, China, March 2016, pp. 5915–5919.

[2] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Evaluation of different excitation generation algorithms for artificial bandwidth extension," in *Proc. of Elektronische Sprachsignalverarbeitung (ESSV)*, Ulm, Germany, 2018.

[3] J. Abel and T. 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, Jan 2018.

[4] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proc. of ICASSP*, April 2018, pp. 5029–5033.

[5] K. Li and C. H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. of ICASSP*, April 2015, pp. 4395–4399.

[6] Z. Ling, Y. Ai, Y. Gu, and L. R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, May 2018.

[7] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. of Interspeech*, 2003.

[8] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, March 2007.

[9] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, Sept 2011.

[10] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *Proc. of ChinaSIP*, July 2014, pp. 71–75.

[11] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Discriminative training of deep regression networks for artificial bandwidth extension," in *Proc. of IWAENC*, Sept 2018.

[12] Y. Wang, S. Zhao, J. Li, and J. Kuang, "Speech bandwidth extension using recurrent temporal restricted boltzmann machines," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1877–1881, Dec. 2016.

[13] Y. Gu, Z. H. Ling, and L. R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Proc. of Interspeech*, 2016.

[14] Y. Gu and Z. H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," in *Proc. of Interspeech*, 08 2017, pp. 1123–1127.

[15] K. Schmidt and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in *Proc. of ICASSP*, April 2018, pp. 5444–5448.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," in *Proc. of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA, 2014, NIPS'14, pp. 2672–2680, MIT Press.

[17] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, 2016.

[18] C. Ledig, L. Theis, F. Huszar, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[19] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *CoRR*, 2017.

[20] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," *CoRR*, 2017.

[21] K. Wang, J. Zhang, S. Sun, et al., "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *CoRR*, 2018.

[22] T. Kaneko, H. Kameoka, N. Hojo, et al., "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. of ICASSP*, March 2017, pp. 4910–4914.

[23] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan 2018.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, 2014.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of International Conference on Artificial Intelligence and Statistics*, Geoffrey Gordon, David Dunson, and Miroslav Dudk, Eds., Fort Lauderdale, FL, USA, 11–13 Apr 2011, vol. 15 of *Proceedings of Machine Learning Research*, pp. 315–323, PMLR.

[26] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *CoRR*, 2016.

[27] J. Sautter, F. Faubel, and G. Schmidt, "Feature selection for DNN-based bandwidth extension," in *Proc. of Jahrestagung für Akustik (DAGA)*, Munich, Germany, 2018.

[28] J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 384–396, Feb 2017.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, et al., "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

[31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[32] "Methods for subjective determination of transmission quality," ITU-T Recommendation P. 800, 1996.