# BANDWIDTH EXTENSION IS ALL YOU NEED

*Jiaqi Su*[1], *Yunyun Wang*[1], *Adam Finkelstein*[1], *Zeyu Jin*[2]

[1]Princeton University, USA
[2]Adobe Research, USA

## ABSTRACT

Speech generation and enhancement have seen recent breakthroughs in quality thanks to deep learning. These methods typically operate at a limited sampling rate of 16-22kHz due to computational complexity and available datasets. This limitation imposes a gap between the output of such methods and that of high-fidelity (≥44kHz) real-world audio applications. This paper proposes a new bandwidth extension (BWE) method that expands 8-16kHz speech signals to 48kHz. The method is based on a feed-forward WaveNet architecture trained with a GAN-based deep feature loss. A mean-opinion-score (MOS) experiment shows significant improvement in quality over state-of-the-art BWE methods. An AB test reveals that our 16-to-48kHz BWE is able to achieve fidelity that is typically indistinguishable from real high-fidelity recordings. We use our method to enhance the output of recent speech generation and denoising methods, and experiments demonstrate significant improvement in sound quality over these baselines. We propose this as a general approach to narrow the gap between generated speech and recorded speech, without the need to adapt such methods to higher sampling rates.

***Index Terms*—** bandwidth extension, audio super resolution, generative adversarial networks, deep features, speech enhancement

## 1. INTRODUCTION

A variety of speech generation and processing applications target 16kHz audio signals, including vocoders for text-to-speech (TTS) synthesis [1], voice conversion [2], source separation [3], and speech denoising and enhancement [4, 5]. This 16kHz sampling rate constitutes a "sweet spot" in the trade-off between intelligibility and computational cost: speech content is fully encompassed within the corresponding frequency range, while audio processing is not too expensive. However, the resulting sound quality remains unsatisfactory for some user listening experiences, as a sense of presence and environment is lost. Of course it is possible that these methods could be adapted to wider bandwidth in the future when greater computational resources and more high-quality speech data are available. Instead, this paper argues that: **(Claim 1)** bandwidth extension from 16kHz to 48kHz can yield convincing results, and therefore **(Claim 2)** there is no need to adapt existing methods to higher sampling rates, as we can simply use methods native to 16kHz and then expand the output to 48kHz. In short, *bandwidth extension is all you need.*

The first claim is not obvious because traditional bandwidth extension (BWE) research has focused on lifting narrow-band signals to 16kHz (from 4-8kHz), primarily for telephony. As far as we are aware, the only previous work that extends to as high as 44kHz (with moderate success) is that of Feng et al. [6].

This paper introduces a new BWE method capable of extending recorded speech from 16kHz to 48kHz, such that the result is typically indistinguishable from real full bandwidth recordings. Moreover, even when the method extends audio from 8kHz to 48kHz, it significantly improves quality, and outperforms baseline methods. The key ideas of the new approach are adapted from the HiFi-GAN method of Su et al. [5], which employs adversarial training together with deep feature matching in multi-domain and multi-scale discriminators. HiFi-GAN was designed and evaluated for speech enhancement (denoising, dereverb, and equalization correction), but here we adapt it to the BWE setting.

Many deep learning based audio generation methods operate at moderate bandwidth. For example, the state-of-the-art voice conversion methods such as the zero-shot AUTO-VC [2] as well as speech denoising methods such as HiFi-GAN [5], DEMUCS [7] and Deep-MMSE [8], all generate audio at 16kHz. Likewise, many source separation methods such as deep clustering [3] and Conv-TasNet [9] work with audio at an even lower rates like 8kHz. This limitation is partly due to efficiency concerns, since doubling the sampling rate will double (or worse) the computational cost. Higher sampling rates also present more modeling challenges relating to longer time series and complex high-frequency structures. For example, modifying vocoders based on WaveNet [1], Parallel WaveNet [10], or MelGAN [11] for higher sampling rates would vastly increase their model sizes in order to achieve a sufficient receptive field. The recurrent neural network (RNN) architecture limits the temporal span of the "memory" in WaveRNN [12], and thus the use of higher sampling rate would likely result in lower quality. In addition to the computation challenges associated with such large models, many speech datasets on which those data-driven methods rely are themselves limited to 16-22kHz. Therefore, rather than adapting such methods to higher sampling rates, we propose to achieve higher temporal resolution simply by applying BWE as a post-process.

To support our claims, we conduct two sets of experiments. First, subjective tests show that the proposed BWE method outperforms several baselines, and achieves perceptual quality that is close to that of real full bandwidth recordings. Next, subjective experiments show that our bandwidth extension method consistently offers significant perceptual quality improvement to the results of speech denoising systems including HiFi-GAN [5], DEMUCS [7] and DeepMMSE [8]. It also improves the quality of vocoders including WaveNet [1], WaveRNN [12] and HiNet [13] which could potentially be applied to TTS as well.

## 2. RELATED WORK

Bandwidth extension aims to estimate the missing high-frequency content, or in other words, to increase the resolution of speech signal, usually from 4-8kHz to 16kHz. The early works estimate the wideband spectral parameters, such as its spectral envelope and gain, from those of the narrowband. They utilize techniques including non-negative matrix factorization [14], linear predictive coding [15], hidden Markov models [16] and Gaussian mixture models [17]. The use of deep learning has significantly improved performance over the traditional methods by enabling greater modeling power.

Li et al. [18] proposed to use deep neural network for estimation of the log-power spectrogram (LPS) of the upperband from that of

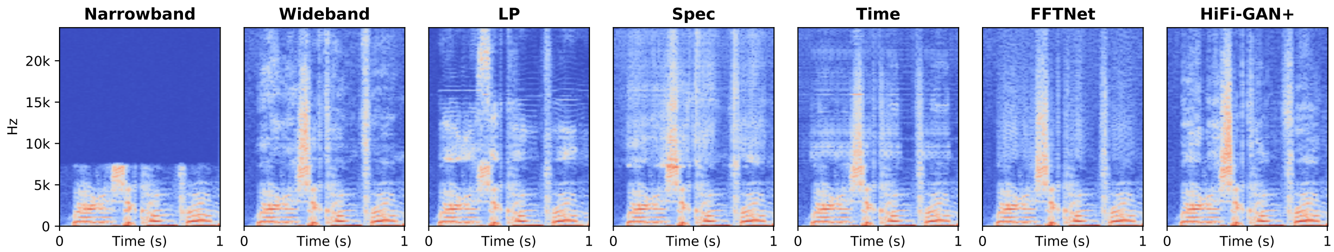

**Fig. 1**. Comparing the log spectrograms of various bandwidth extension methods. **Narrowband** is the 16kHz input; **Wideband** is the 48kHz target; **LP**, **Spec**, **Time** and **FFTNet** are baselines; **HiFi-GAN+** is our proposed approach. It can be observed that our method generates the most plausible details for the missing frequencies, while the others have either blurred energy or artifacts.

the narrowband. Various network architectures have been explored, such as variational auto-encoders [19], U-Nets [20] and recurrent neural networks [21]. While the spectral methods are good at compensating energy for the missing frequencies, the estimated spectrogram usually lacks details due to the smoothing effects of the commonly used MSE and MAE objective functions. They do not eliminate noise or artifacts not represented in the time-frequency domain either. To reconstruct waveform from the estimation, the wideband phase is approximated by repeatedly flipping the narrowband phase, but this process often introduces artifacts.

The recent advances in network architectures have enabled audio processing directly on waveform. Kuleshov et al. [22] used a convolutional encoder-decoder network inspired by image super resolution. WaveNet [23] and its variants for BWE [24, 25] use dilated convolutions to enable large receptive field while preserving the original resolution. Feng et al. [6] used FFTNet [26] which resembles the classical FFT process. Ling et al. [27] proposed a hierarchical RNN to utilize the waveform structures. Several other efforts incorporated time-frequency information while still operating in the time domain. Li et al. [28] adapted an EnvNet structure to approximate the spectral feature extraction from waveform. Time-Frequency Networks [29] use dual branches with a spectral fusion layer to combine the information. In addition, time-frequency losses have been widely adopted by time-domain methods to encourage matching of the upperband spectral energy [28, 30].

Generative Adversarial Networks (GAN) have recently been explored in audio processing to improve authenticity of speech. The generator is driven to approximate the real data distribution via the dynamic competition with the discriminator. For BWE, its variable discrimination helps to refine details in the high frequencies. The previous GAN works in BWE [31, 20, 19] typically follow simple designs, using a discriminator of a few fully connected layers or convolutional layers on the spectral features, while the discrimination directly on waveform has rarely been employed.

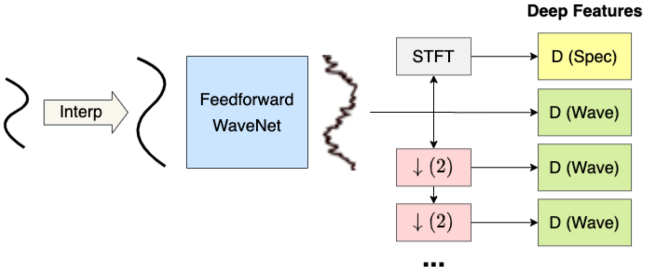

**Fig. 2**. Networks. The feed-forward WaveNet super-resolves the interpolated narrowband signal. Adversarial training with deep feature matching involves a spectrogram discriminator and multiple waveform discriminators for the signal at different resolutions.

The usage of GAN for sound quality has been more thoroughly explored in other domains such as speech synthesis [32, 11] and speech enhancement [33]. HiFi-GAN [5] shows that discrimination in both the time domain and the time-frequency domain is necessary to achieve the best sound quality. MelGAN [11] proposed to use the learnt feature space of the discriminator as a distance metric as it dynamically picks up the noticeable differences between the generated audio and the real audio. This feature matching loss stabilizes GAN training and avoids the notorious mode collapse issue by forcing content consistency. Similar ideas and properties of adversarial training can be transferred to the BWE problem.

## 3. METHOD

The bandwidth extension problem can be viewed as a special case of the speech enhancement problem, since both require transforming degraded audio signal to its high-quality form. Therefore, we adopt the HiFi-GAN approach [5] from speech enhancement, which is shown successful for obtaining clean high-fidelity audio recordings from noisy reverberant conditions. It uses an end-to-end feed-forward WaveNet structure together with deep feature matching in multi-scale multi-domain discriminators to identify artifacts from various aspects and resolutions.

This WaveNet takes in narrowband signal (resampled to the same length as the fullband signal) and outputs fullband signal. It uses non-casual dilated convolutions with exponentially increasing dilation rates to achieve sufficient temporal context for estimating the high frequency structures. We use a power of three (1 to 2187) as dilation rate as we are working with 3× and 6× upsampling. In our experiments, two WaveNet stacks with channel size 128 are used. We did not use the postnet module in the original HiFi-GAN, because extra convolution layers can smooth the output signal, reducing high frequency resolution. We use weight normalization [34] across all the networks to speed up convergence.

### 3.1. Perceptually-motivated loss

To regularize the network and speed up early convergence, we put loss on both the waveform and spectrograms. The waveform loss is the absolute difference between prediction and target waveform. It helps to match the overall shape and the phase, but it can hinder further optimization once the output signal is close to the ground truth. This is due to the fact that noise is unpredictable: when the ground truth contains high-frequency noise, minimizing L1 or L2 distance will result in predicting the average of noise, which is 0, and thus a loss of high frequency content. Therefore, we also use spectrogram loss, defined as L1 distance of log spectrograms with different FFT window sizes (i.e 512, 1024, 2048, and 4096 for 48kHz fullband signal, each with one-fourth as its hop size).

In addition, we compute the L1 log mel-spectrogram loss using 128 coefficients for the upperband to focus on the missing high frequencies. These spectrogram losses help to match high frequency components especially noises (presented as predictable constant in spectrogram). However, the use of L1 or L2 distance may still introduce over-smoothing effects that cause new artifacts. This is where adversarial training helps.

### 3.2. Adversarial training

Our adversarial training takes the similar design as HiFi-GAN. Discrimination in spectral domain encourages generation of details for the missing bands. We use a discriminator on the full-band 128-coefficient log mel-spectrogram. It consists of four stacks of 2D convolution layer, batch normalization and Gated Linear Unit (GLU), and lastly a convolution layer followed by global average pooling, similar to the one used in StarGAN-VC [35]. It uses kernel sizes of (7, 7), (4, 4), (4, 4), (4, 4) and stride sizes of (1, 2), (1, 2), (1, 2), (1, 2) for the stacks, and the last convolution layer uses a kernel size of (15, 5). The channel sizes is 32 across all the layers.

Meanwhile, we use a set of waveform discriminators, respectively operating at the fullband signal down-sampled by different ratios as a power of two, following the design in MelGAN [11]. Thus, each waveform discriminator learns features for a different frequency range. The number of waveform discriminators is determined based on the up-sampling scale from the narrowband signal to the fullband signal in the task. For example, for BWE from 8kHz to 48kHz, we used four waveform discriminators operating at 48kHz, 24kHz, 12kHz and 6kHz sampled versions of the fullband signal. Each waveform discriminator is composed of a set of grouped convolutions and global average pooling at the end, with Leaky Relu used between the layers. Specifically, the kernel sizes are 15, 41, 41, 41, 41, 5, 3; stride sizes 1, 4, 4, 4, 4, 1, 1; channel sizes 16, 64, 256, 1024, 1024, 1024, 1; and group sizes 1, 4, 16, 64, 256, 1, 1.

In experiments, we found that the waveform discriminators contribute significantly to the perceptual qualities of the resulting fullband audio by removing commonly observed metallic artifacts and improving naturalness of unvoiced sound.

As is identified previously, the common metric functions are not able to evaluate the overall perceptual quality of the generated fullband signal. However, our discriminators essentially learn a representation space for real fullband audio while trying to discriminate whether a provided audio falls in the same representation space as the real ones. Thus, we also impose the feature matching loss from each discriminator to the generator, which calculates the L1 distance between the deep features of the generated audio and those of the corresponding real fullband audio.

### 3.3. Noise augmentation

When used as a post-processing step for other audio applications, the BWE model needs to be robust to various artifacts in the input narrowband signal. For example, the recordings from denoising algorithms may contain residuals of noise and reverberation, and the synthesized speech from vocoders may contain robotic sound. Therefore, to match the test-time conditions, we add 15-25dB noise randomly drawn from the DNS Challenge Dataset [36] to the input narrowband signal during training.

## 4. EXPERIMENTS

Throughout the experiments, we used the architecture described in Section 3. Training happens in two stages. First we train WaveNet for 1000k steps with learning rate 0.001, using the waveform loss and the spectrogram losses. Then we train the generator at learning rate 0.00001 with the randomly initialized discriminators at learning rate 0.001 for 100k steps, including adversarial losses, deep feature matching losses and the previously used losses. We update the discriminators twice for every step of the generator with Adam optimizers. A batch size of 4 is used on a Tesla V100 with each input of 48k samples (i.e. 1 second).

We evaluated our model on three different tasks: (1) BWEs from 8kHz to 48kHz and from 16kHz to 48kHz on clean speech to compare our method with baselines, (2) applying our BWE method to the results of speech denoising algorithms, and (3) applying our BWE method to the results of speech generation algorithms. The experiments are to demonstrate that our proposed BWE method achieves high quality results comparable to real fullband audio and it can be applied as a post-processing step for various audio applications. Audio samples for the experiments are available at: `https://daps.cs.princeton.edu/projects/Su2020BWE/`

### 4.1. Comparison study

In this study, we compare three variants of our proposed BWE method, including the base feed-forward WaveNet (**Base**), the use of the spectrogram discriminator alone (**SpecGAN**), and the full model (**HiFi-GAN+**), i.e. using both the spectrogram discriminator and the waveform discriminators. We also compare to four other state-of-the-art baselines: a traditional method using linear prediction based analysis synthesis [15] (**LP**), a spectral-domain method using 1D convolutional U-Net with GAN [20] (**Spec**), a time-domain method using EnvNet structure with GAN [28] (**Time**), and FFTNet variant for BWE [6] (**FFTNet**). Note that all the baseline methods except for **FFTNet** were originally designed for BWE up to 16kHz, and therefore, we adapted their methods to this new BWE setting by either adjusting the filter designs (i.e. for **LP**), or changing the ratios of up-sampling layers (i.e. for **Time**).

We use the VCTK dataset [37] to train models for both 8kHz to 48kHz and 16kHz to 48kHz extensions, following the same split as Kuleshov et al. [22]: the first 99 speakers for training and the remaining 9 speakers for validation. The evaluation is then conducted on a separate dataset: the Device and Produced Speech (DAPS) Dataset's clean set [38] using the last four male and four female voices. Note that the DAPS dataset has different recording conditions from the VCTK dataset in that VCTK contains slight background noise in the recordings while DAPS is made in studio and has been professionally treated. This examines whether our approach generalizes well across different datasets and conditions.

Table 1(a) shows the scores of PSNR (Peak Signal-to-Noise Ratio) and LSD (Log-Spectral Distance) on the VCTK test set and the DAPS test set. While the two metrics have been commonly used for BWE evaluations, we observe that they do not correlate with perceptual quality in this fullband setting. Any processing to the narrowband input simply lowers its PSNR. Also GAN-based methods learn to generate plausible high frequencies rather than the exact same as ground truth, and thus their objective scores tend to be lower. In contrast, the methods trained with just spectrogram loss (**FFTNet**, **Base**) achieve smallest LSD but their results contain noticeable artifacts, possibly due to over-fitting to matching the spectrogram which introduces over-smoothing effects.

Therefore, we also conduct a subjective evaluation using Amazon Mechanical Turk. A subject needs to first pass a pre-test to identify 44kHz recordings out of 5 recordings (the other 4 are 16kHz or less). This is to make sure the subjects are using headphones and can hear high frequencies. The pre-test is followed by a series of Mean Opinion Score (MOS) tests, where a subject is asked to rate the

**Table 1**. Objective measures.

| **Method** | **PSNR↑** | **LSD↓** | **PSNR** | **LSD** | **PSNR** | **LSD** | **PSNR** | **LSD** |
|---|---|---|---|---|---|---|---|---|
| **Input SR** | 8k | | 16k | | 8k | | 16k | |
| | **VCTK Dataset** | | | | **DAPS Dataset** | | | |
| **NB Input** | **38.56** | 15.81 | **44.40** | 14.84 | 35.95 | 12.87 | **41.98** | 11.50 |
| **LP** | 15.74 | 4.06 | 15.74 | 3.83 | 15.78 | 5.00 | 13.73 | 4.61 |
| **Spec** | 26.19 | 2.42 | 35.74 | 2.06 | **36.26** | 3.06 | 40.65 | 2.58 |
| **Time** | 22.99 | 2.03 | 29.90 | 1.92 | 31.60 | 2.82 | 31.07 | 3.10 |
| **FFTNet** | 36.33 | **2.00** | 40.59 | **1.67** | 35.38 | 2.80 | 39.62 | 2.44 |
| **Base** | 31.70 | 2.26 | 32.40 | 2.03 | 29.26 | **2.67** | 30.08 | **2.34** |
| **SpecGAN** | 12.75 | 2.15 | 31.78 | 1.95 | 10.57 | 2.85 | 26.56 | 2.45 |
| **HiFi-GAN+** | 33.53 | 2.13 | 32.16 | 1.83 | 30.60 | 2.80 | 29.28 | 2.35 |

(a) BWE on VCTK and DAPS datasets.

| | **PSNR↑** | **LSD↓** | | **PSNR** | **LSD** | **PSNR** | **LSD** |
|---|---|---|---|---|---|---|---|
| **Method** | | | **Method** | | | **With BWE** | |
| **Noisy-16k** | 27.69 | 12.04 | **DeepMMSE** | 35.29 | 13.28 | 29.24 | 3.04 |
| **Clean-8k** | 37.57 | 13.95 | **DEMUCS** | 34.76 | 12.77 | 29.37 | 2.81 |
| **Clean-16k** | 43.93 | 12.78 | **HiFi-GAN** | 28.98 | 13.17 | 26.81 | 2.86 |

(b) Denoising with BWE on DEMAND dataset.

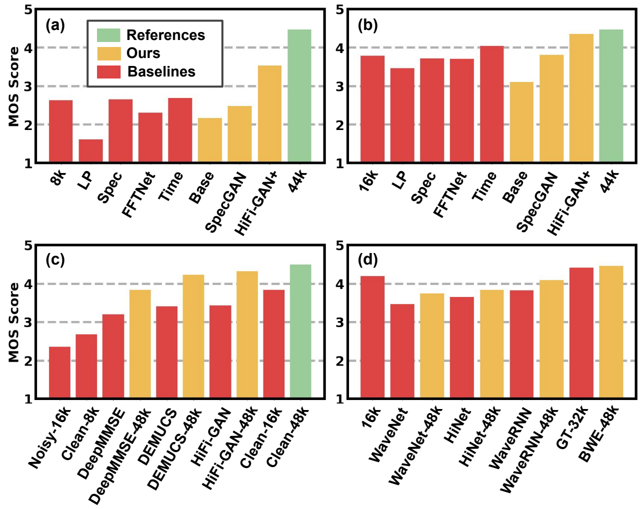

**Fig. 3**. MOS Scores for BWE experiments. (a) 8kHz to 48kHz; (b) 16kHz to 48kHz; (c) 16kHz to 48kHz for the outputs of denoising methods; (d) 16kHz to 48kHz for the outputs of vocoders. Green is the fullband reference; yellow is our methods; red is the baselines.

sound quality of an audio recording on a scale of 1 to 5, with 1=*Bad*, 5=*Excellent*. The audio recordings are randomly picked from the results of the seven methods (three ours and four baselines), as well as 8kHz, 16kHz and 48kHz versions of the clean recordings. We also include 4 validation tests to exclude workers who are not paying attention. 382 unique workers participated in this experiment, and we collected 23,400 ratings in total.

The MOS scores are shown in Figure 3. Figure 3(a) shows BWE from 8kHz to 48kHz, in which our full method significantly outperforms all baselines by a large margin. Figure 3(b) shows BWE from 16kHz to 48kHz. Visually, all BWE methods perform well while our full method **HiFi-GAN+** stands out with the highest MOS (4.35). The second best is **Time** with MOS (4.04). It is also worth noting that the full bandwidth recording has a MOS score of 4.48 and it is not statistically significant enough to say that our method is inferior to real 48kHz samples. Therefore we conducted additional pairwise comparison study (AB test) to reveal the actual gap between our method and real 48kHz samples.

In this AB test, a subject is presented with two test audio clips, the real 48kHz recording and the 16kHz recording expanded to 48kHz using our **HiFi-GAN+**, with a reference clip being the real 48kHz recording. The task is to select the sample that sounds closest to the reference. The test used the same pre-test and validation strategy presented before. We collected 2,675 answers from 200 subjects, in which 1,139 prefers our method and 1,536 prefers the real samples. This means in 42.6% cases (7.4% above random) people prefer our method; or in average 85.2% of the subjects have no preference and thus answer randomly. Though there is still a small gap between our method and real 48kHz samples, it is fair to say that our method is able to improve fidelity of 16kHz audio to 48kHz that is typically indistinguishable from real 48kHz samples.

### 4.2. Bandwidth extension for speech denoising

We apply our full BWE method to three state-of-the-art speech denoising algorithms, including HiFi-GAN [5], DeepMMSE [8] and DEMUCS [7], and use the benchmark DEMAND Dataset's test set [39] for evaluation. Since the VCTK dataset we previously trained on overlaps with the DEMAND dataset and contains background noise, we specifically train a new **HiFi-GAN+** model on DAPS's clean set for this task. We generated the denoised audio samples using DEMUCS's and DeepMMSE's official implementations and pre-trained models, and also took the denoised audio samples from HiFi-GAN's project website. We evaluate the same set of objective measures (PSNR, LSD) on the results of denoising with and without BWE (Table 1(b)). As the reasons mentioned previously, the scores do not correspond with human perception well. The same MOS test is conducted as in Section 4.1 for 16kHz noisy recordings, clean recordings sampled at 8kHz, 16kHz and 48kHz, the denoised samples from the methods, and the bandwidth-extended denoised samples from our BWE model. We collected 8,589 answers from 180 workers. The result is shown in Figure 3(c) where our method improves all speech enhancement quality by a large margin. HiFi-GAN-48k (4.32) and DEMUCS-48k (4.23) perform the best and receive the most quality boost; both are comparable to the real 48kHz samples (4.50). This experiment shows our BWE method is an effective post-processing technique for speech denoising and enhancement.

### 4.3. Bandwidth extension for waveform generation

We use the same trained model as in the denoising task in Section 4.2, and apply it to the outputs of three vocoding algorithms, including WaveNet [1], WaveRNN [12] and HiNet [13]. We took their audio samples from HiNet's project website. We also included the 16kHz input data (16k) to these networks and 32kHz raw audio (GT-32k) from the CMU Arctic Dataset [40]. We conducted the same MOS test as in Section 4.1. The result (Figure 3(d)) shows that our BWE method improves quality across all waveform generation methods. We also found that the improvement is stronger if the input generated waveform has higher sound quality. It is because any artifact that resides in the lower frequencies will be carried over by BWE, causing lower MOS ratings.

## 5. CONCLUSION

In this paper, we presented a novel bandwidth extension method based on WaveNet and adversarial training with deep feature matching. We conducted extensive experiments to show that the proposed method outperforms other state-of-the-art approaches in 8k/16kHz to 48kHz bandwidth extension tasks. We also showed via pairwise comparison that our 16-to-48kHz BWE generates audio that's comparable to real 48kHz recordings in fidelity. Thus, we propose our method as a general tool to enhance the output of speech enhancement and generation algorithms. We demonstrated fidelity improvement in these tasks via subjective evaluations.

## 6. REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *ICASSP 2018*.

[2] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP 2020*.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP 2016*.

[4] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP 2018*.

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

[6] B. Feng, Z. Jin, J. Su, and A. Finkelstein, "Learning bandwidth expansion using perceptually-motivated loss," *ICASSP 2019*.

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

[8] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: a deep learning approach to mmse-based noise power spectral density estimation," *TASLP*, vol. 28, 2020.

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *TASLP*, vol. 27, no. 8, 2019.

[10] A. Oord, Y. Li, I. Babuschkin, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018.

[11] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS 2019*.

[12] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv:1802.08435*, 2018.

[13] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *TASLP*, vol. 28, 2020.

[14] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in *European Conference on Speech Com. and Tech.*, 2005.

[15] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," in *ICASSP 2018*.

[16] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," in *ICASSP 2003*.

[17] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *ICASSP 2014*.

[18] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *ICASSP 2015*.

[19] P. Bachhav, M. Todisco, and N. Evans, "Artificial bandwidth extension using conditional variational auto-encoders and adversarial learning," in *ICASSP 2020*.

[20] S. E. Eskimez and K. Koishida, "Speech super resolution generative adversarial network," in *ICASSP 2019*.

[21] K. Schmidt and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural nets," *ICASSP 2018*.

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

[23] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., "Wavenet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[24] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, "Speech bandwidth extension with wavenet," *WASPAA 2019*.

[25] M. Wang, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, D. Yu, and H. Meng, "Speech super-resolution using parallel wavenet," in *ISCSLP 2018*.

[26] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: a real-time speaker-dependent neural vocoder," in *ICASSP 2018*.

[27] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *TASLP*, vol. 26, no. 5, 2018.

[28] X. Li, V. Chebiyyam, et al., "Speech audio super-resolution for speech recognition.," *Interspeech 2019*.

[29] T. Y. Lim, R. A. Yeh, Y. Xu, et al., "Time-frequency networks for audio super-resolution," in *ICASSP 2018*.

[30] H. Wang and D. Wang, "Time-frequency loss for cnn based speech super-resolution," in *ICASSP 2020*.

[31] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *ICASSP 2018*.

[32] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, et al., "High fidelity speech synthesis with adversarial networks," *arXiv:1909.11646*, 2019.

[33] S. Pascual, A. Bonafonte, et al., "Segan: Speech enhancement generative adversarial network," *arXiv:1703.09452*, 2017.

[34] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *NeurIPS 2016*.

[35] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv:1806.02169*, 2018.

[36] C. K. Reddy, E. Beyrami, H. Dubey, V. Gopal, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," *arXiv:2001.08662*, 2020.

[37] C. Veaux, J. Yamagishi, K. MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[38] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments...," *IEEE Signal Proc. Letters*, vol. 22, no. 8, 2015.

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

[40] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *SSW*, 2004, pp. 223–224.