Speech Audio Super-Resolution For Speech Recognition

Xinyu Li, Venkata Chebiyyam, Katrin Kirchhoff

Amazon AI

{xxnl,chebiyya,katrinki}@amazon.com

Abstract

Automatic bandwidth extension (restoring high-frequency information from low sample rate audio) has a number of applications in speech processing. We introduce an end-to-end deep learning based system for speech bandwidth extension for use in a downstream automatic speech recognition (ASR) system. Specifically we propose a conditional generative adversarial network enriched with ASR-specific loss functions designed to upsample the speech audio while maintaining good ASR performance. Evaluations on the speech commands dataset and the LibriSpeech corpus show that our approach outperforms a number of traditional bandwidth extension methods with respect to word error rate.

Index Terms: generative adversarial network, super-resolution, bandwidth extension, speech recognition

1. Introduction

Automatic bandwidth extension (BWE) has been used successfully in many applications, such as speech compression, emotion recognition, speech enhancement, and speaker identification [1, 2, 3]. Upsampling narrow-band (NB) speech for use with a wide-band (WB) ASR model has number of potential applications, such as single model ASR and ASR with limited transmission bandwidth. However, there are few studies that specifically focus on developing bandwidth extension methods optimized for ASR performance and that evaluate word error rate rather than perceptual quality. In this paper we introduce an end-to-end audio super-resolution approach for converting NB to WB speech for the purpose of speech recognition. Unlike previous BWE methods that predict high-band (HB) parameters using low band (LB) features, we treat the NB audio as WB audio mixed with noise that cancels the HB spectrum. We propose a conditional GAN (cGAN) based approach that removes the noise and recovers the WB audio. The cGAN consists of a generator network that produces WB audio while a discriminator network learns to distinguish between artificially generated and true WB audio. We enrich the standard cGAN approach by integrating ASR-specific loss functions and discriminator training. We evaluate our approach on a speech command data set and the LibriSpeech data set [4] against traditional bandwidth extension methods as well as previously proposed super-resolution techniques adapted to this problem from image and audio processing. Our results outperform all baseline methods and demonstrate that our proposed loss functions are critical for improved ASR performance using speech super-resolution.

2. Related Work

Bandwidth extension of speech is the problem of recovering missing high-frequency information by exploiting the redundancies in the human speech production system [5]. It has been approached from two directions: (a) digital signal processing based signal restoration and (b) machine learning based prediction of missing information. The digital signal processing based techniques typically involve a LB speech coder and may either contain explicit HB information transmitted from an encoder (guided BWE) [1, 6] or perform "blind" BWE without this information [3, 7]. The commonly used guided BWE method includes Spectral Band Replication [6], time-domain BWE [1] and methods that transmit information about the HB fine spectral structure [1, 6]. The commonly used blind BWE techniques include predicting HB parameters and fine spectral structure [3] and predicting the HB line spectral frequencies (LSFs) [8]. The machine learning based approaches typically predict HB features using LB features. A DNN was applied to predict the HB cepstrum [9] and log power spectrum [10]. Although GANs were recently used for speech BWE [7], it was not implemented end-to-end to directly produce WB speech signal, instead they worked with predicting key HB parameters, which needs additional intelligence to produce HB waveforms. There are end-to-end systems which use GANs in related applications including audio synthesis, speech enhancement, speech/music super-resolution [11, 12, 13] etc. However, these studies rely on the vanilla GAN structure using adversarial loss and focus on evaluating perceptual quality rather than generating a signal that can be used for ASR directly. One novelty of our work compared to previous research is that we incorporate several task-specific loss functions, including those to equivalent to "perceptual loss" (distance between representations extracted by a pre-trained learner from real vs. generated input). This is widely considered the most important loss component in image super-resolution work using GANs [14] and we demonstrate that it is critical for speech applications as well.

3. Methodology

The goal of BWE is to recover the WB signal from a given NB signal. Previous bandwidth extension (BWE) methods have attempted to generate WB audio $\tilde{x}$ from NB audio x by predicting the missing HB information (Figure 1, top). Typically, BWE focuses on minimizing the distance between estimated HB parameters and actual HB parameters extracted from transformation of the true WB audio. This does not guarantee improved ASR performance on the generated audio signal. For example, when mel-frequency cepstral coefficients (MFCCs) are used for ASR, using oracle LSFs [1] does not necessarily guarantee that MFCCs extracted from the generated signal will match the oracle MFCC features produced from true WB speech. We consider the NB signal x as the WB signal with zero spectral content in the HB portion $\bar{x}$, which can be further considered as the original WB signal $\tilde{x}$ mixed with noise n that cancels all the HB energy in phase as:

$$x \equiv \bar{x} = \tilde{x} + n \tag{1}$$

If we apply the short-time Fourier transform (STFT) transform to both $\tilde{x}$ and n, the noise happens to cancel the HB information

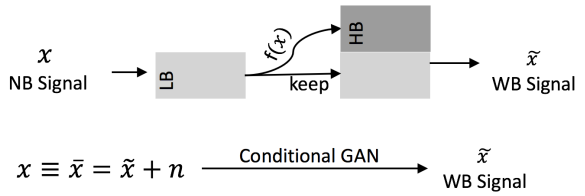

Figure 1: *Top, traditional bandwidth extension that predict HB feature based on LB feature and then concatenate them together to recover the WB audio. Bottom, our end-to-end approach using cGAN.*

from $\tilde{x}$ as: $STFT(n)_{HB} = -STFT(\tilde{x})_{HB}$ and maintain the LB information as: $STFT(n)_{LB} = 0$. Given NB audio x is equivalent to the $\bar{x}$ that has same number of samples as WB signal $\tilde{x}$, we propose to recover the WB signal $\tilde{x}$ by using a cGAN that removes noise n from $\bar{x}$. This is similar to pixel to pixel mapping in computer vision where related problems (e.g. converting a sketch to a photo-like image) have been successfully addressed with GANs [15]. The conditional GAN [16] consists of a generator and a discriminator. Unlike the vanilla GAN that generates samples from noise, the conditional GAN generates samples that are subject to task-specific constraints (e.g., generates an image that has the same shape but different texture as input image). In our application, the generator tries to generate WB audio that shares similar acoustic features as input NB audio. The discriminator is indeed a classifier that tries to distinguish the generated WB audio and real WB audio [16].

3.1. Generator

We design an end-to-end generator that takes the LB 1D audio signal x as input and generates WB 1D audio signal $\tilde{x}$. This has the advantage that: 1. the system can be trained end-to-end which avoids the information loss in input parameter estimation as well as lack of information for audio reconstruction; 2. the generated signal can be used as input to any ASR system directly; and 3. the proposed system is able to generate both frequency and phase information at once. We propose a generator based on a conv-deconv structure that first decomposes the signal into a low dimensional feature representation and then reconstructs the features in the transfer domain [17] (Figure 2, generator). We use the EnvNet structure that first learns log-mel features from the raw audio with 1D convolutions and then finds spatio-temporal associations with 2D convolution [18]. The deconvolution module has the reverse structure. To maintain a tight association between input and output signal, we use u-net connections [19], which append the features learned by convolution layers to associated de-convolution layers.

3.2. Discriminator

The discriminator serves the purpose of distinguishing the generated audio from true WB audio. In most cases, the discriminator takes the generated samples from the generator to calculate the adversarial loss. We designed our system to calculate adversarial loss based on distance in MFCC feature space rather than based on the waveform. This design helps the generator produce audio that will result in MFCCs comparable to those extracted from WB audio which are the features typically used for ASR purposes. To this end we include an MFCC extraction layer into the discriminator that first converts the 1D audio into a 2D MFCC representation. The MFCCs are then passed into a 2D CNN, and a fully-connected layer with the sigmoid activation function is used for the discriminator's decision making (Figure 2, discriminator).

3.3. Losses

Standard adversarial loss in a GAN is defined as

$$L_a = log(1 - D(G(x))) \tag{2}$$

where D denotes the discriminator and G denotes the generator. The adversarial loss expresses the similarity between generated samples and real samples estimated by the discriminator. This is usually combined with additional loss functions designed for task-specific purposes [15]. We introduce three loss functions in addition to adversarial loss that cover both feature-wise differences and global waveform differences between real WB audio and generated WB audio. The overall loss is defined as:

$$L = L_a + \alpha L_p + \beta L_{MFCC} + \gamma L_{MAE} \tag{3}$$

L_p is a perceptual loss that expresses the distance between features learned by a pre-trained ASR network:

$$L_{p_i} = (\sum_{j=1}^{N_i} H_{ij} - \tilde{H}_{ij})/N_i \tag{4}$$

where L_{p_i} denotes the perceptual loss calculated based on the i^{th} intermediate layer of an ASR system's acoustic model with N_i neurons (assume the acoustic model typically uses LSTM structure). $H_{ij} - \tilde{H}_{ij}$ denotes the difference between the same neuron's output with real WB audio as input and using our generated audio as input. The perceptual loss further encourages the generator to generate WB audio that has a similar feature representation in an ASR network compared with real WB audio, which is likely to lead to improved ASR performance. The drawback is that calculating perceptual loss requires a pre-trained ASR network, which might not be a convenient resource to get access to. As an alternative, we propose the MFCC loss (L_{MFCC}), which is not dependent on such a resource and simply computes the distance between the MFCC feature representations typically used in ASR systems:

$$L_{MFCC} = ||MFCC(G(x)) - MFCC(\tilde{x})|| \tag{5}$$

where $MFCC(x)$ denotes the associated MFCC map extracted from signal x. $G(x)$ denotes the generated WB audio. Finally, we add the mean absolute error between the actual WB signal and the generated WB signal:

$$L_{MAE} = abs(G(x) - \tilde{x}) \tag{6}$$

The MAE penalizes the per-sample difference between generated audio and real audio waveform thus incorporating a phase loss of the dominant energy bins in the STFT implicitly. Typically, the low-frequency components are dominant in energy. Thus, MAE ensures that the waveform of the LB frequencies of the generated signal is close to that of the LB frequencies in the original NB signal. The different loss functions are weighted by coefficients α, β, γ – in our implementation we use values of 1.0 for α and β and a much smaller weight (0.2) for γ, to highlight the feature-wise losses.

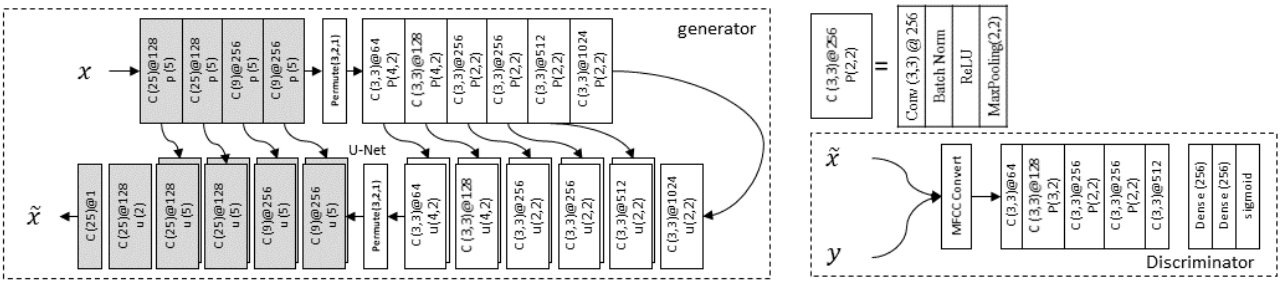

Figure 2: *Overview of the cGAN structure. x denotes the input NB audio, the $\tilde{x}$ denotes the generated WB audo and y is the real WB signal.*

3.4. Implementation

We implemented our model with Keras using a Tensorflow backend. The MFCC extraction functions were wrapped as Tensorflow layers using Keras Lambda function. We used the routines in the Tensorflow contrib library to calculate MFCCs, based on STFTs calculated with a window of size 512. We used the 25ms Hamming window with appropriate zero padding to achieve the STFT window size of 512 and further applied 80 Mel filters to the STFT bins between 80Hz to 8kHz. The generator and the discriminator were trained for 3000 epochs with the ADAM optimizer and auto-decay set at 1e-7.

Methods	Speech Commands	LibriSpeech
Interpolation	8.00%	12.66%
Noise Filling	6.82%	13.99%
LB Duplication	7.35%	13.31%
DNN [10]	7.13%	12.35%
GAN [7]	7.15%	12.34%
Real 16 kHz (WB) audio	4.54%	10.81%
Our approach (without perceptual loss)	6.27%	12.30%
Our approach (with perceptual loss)	5.80%	12.17%

Table 1: *Comparison of ASR word error rates for different audio super-resolution methods.*

4. Experimental Results

4.1. Datasets

Our goal is to test audio superresolution as an alternative to baseline BWE methods, with single WB-model based ASR as a backend application. We tested our system on the:

Speech Commands Dataset [20]: which has 65,000 one-second long utterances of 30 short words recorded by thousands of speakers. We randomly selected 70%, 10% and 20% of the data in each class for training, validation, and testing respectively. We down-sampled the acoustic data from 44.1k to 8k as NB audio and to 16K as WB ground truth. The same training and testing split was used for training both the speech recognition model and our cGAN model. The speech recognizer is a 4-layer bi-directional LSTM network trained on the 16k audio. Since the Speech Commands Dataset only contains single-command utterances, no language model is used. For preprocessing we used the same MFCC configuration as was used for calculating the MFCC loss (see above). The recognition network is also used for perceptual loss calculation, with representations being extracted from the second and third layers.

LibriSpeech dataset [4]: is a corpus of approximately 1000 hours of 16kHz (WB) read English speech, with an official training and testing split. Using the official clean-training subset of LibriSpeech, we trained an acoustic model for conversational speech recognition based on a hybrid LSTM/HMM acoustic model and a 4-gram language model. This trained acoustic model was also used to calculate the perceptual loss. We down-sampled the official clean testing data to 8 kHz (NB) for testing purposes.

4.2. Results and Comparison

Our baselines include: 1. Interpolation using sox, using target sample rate of 16000 and 16-bit quantization. 2. Noise filling with spectrally shaped and scaled random noise. Spectral shaping is done to match HB and LB tilts and energies around 4kHz. 3. LB Duplication, which is the same as Noise Filling except for a spectrally flattened LB instead of Random Noise. 4. The DNN and GAN based methods in [10, 7]. A comparison of results (Table 1) shows that (a) Our model significantly outperforms the baseline approaches ($\sim$20% relative lower WER). (b) The perceptual loss results in slightly lower WER compared to the MFCC loss (but comes at the higher cost of requiring a trained ASR system). (c) Our approach performs better than previous approaches using either DNNs or GANs for HB feature prediction [10, 7], confirming the importance of the task-specific loss functions. (d) One argument against the speech super-resolution for ASR is that one could train a separate ASR model for NB speech. The problem is that the NB training data is often hard to collect. Our preliminary results on English, Portuguese and Italian language audio shows that it takes several hundreds hours of NB audio data to produce a well-trained NB ASR system that has a performance comparable to using a WB ASR system using NB speech audio in conjunction with bandwidth extension by our proposed method.

4.3. Analysis

We further tested the impact of different design choices in our system (Table 2). The results show that: 1. The GAN with only adversarial loss and MAE loss has significantly higher WER (Table 2, first row), which demonstrates that the perceptual loss and MFCC-based loss are critical for better ASR performance. 2. The discriminator with the MFCC conversion layer has lower WER (Table 2, second and fifth row), which shows that calculating adversarial loss based on MFCCs instead of the waveform helps with ASR performance. 3. We also implemented a 1D convolution based generator similar to the one used in [11, 12].

The proposed 1D and 2D conv-deconv generator outperformed the simple 1D convolutional generator (Table 2, third and fifth row), which indicates that the EnvNet structure works well as a generator. 4. We also evaluated stack GANs [21] on our task but did not observe any performance increase (Table 2, fourth and fifth row) from stacking multiple generators. 5. Finally, we tried to replace the MFCC loss with a simpler STFT spectrogram loss (Table 2), fifth and seventh row); however this resulted in higher WERs.

Model Config	Speech Commands Dataset
$L_a + L_{MAE}$	11.72%
Discriminator without MFCC convert layer	6.43%
$L_a + L_{MAE} + L_{MFCC}$ with 1D generator	6.95%
$L_a + L_{MAW} + L_{MFCC}$ stacked GAN	6.31%
$L_a + L_{MAE} + L_{MFCC}$	6.27%
$L_a + L_{MAE} + L_p$	5.80%
$L_a + L_{MAE} + L_{STFT}$	6.47%

Table 2: *Comparison of ASR word error rates between different audio super-resolution methods.*

5. Discussion

We visualize and compare the spectrograms extracted from waveforms generated by different methods (Figure 3). The results show that the proposed cGAN is able to better recover the detailed high-frequency components compared with other baseline methods. Note that the energy dip in the spectrogram around 4kHz is caused in baselines because of wider rolloff of the low-pass filter while generating the LB from the original WB audio. Such gaps can be filled up by using prior information about the cutoff frequency or estimate the cutoff frequency in conjunction with methods such as noise filling or frequency shifting etc. Our proposed method can fill up such gaps without the explicit need to rely upon any prior information. This makes the proposed method more robust.

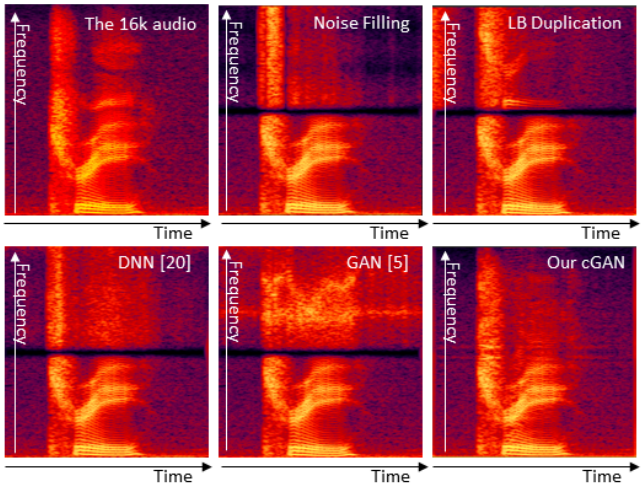

Figure 3: *Comparison of STFT spectrogram generated by different methods.*

6. Conclusion

We have introduced a conditional GAN for audio super-resolution with ASR-specific loss functions. Evaluation of this method in an ASR system for speech commands demonstrated that our proposed cGAN is able to generate high sample rate audio that results in better ASR performance than standard BWE methods, including previous deep learning based approaches. While we have focused on ASR as a potential application our proposed system can potentially be extended to many other fields, such as noise cancellation or speech enhancement, and applications such as audio style conversion and text-to-audio generation.

7. References

[1] V. Atti, V. Krishnan, D. Dewasurendra, V. Chebiyyam, S. Subasingha, D. J. Sinder, V. Rajendran, I. Varga, J. Gibbs, L. Miao *et al.*, "Super-wideband bandwidth extension for speech in the 3GPP EVS codec," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5927–5931.

[2] P. S. Nidadavolu, C.-I. Lai, J. Villalba, and N. Dehak, "Investigation on bandwidth extension for speaker recognition," *Proc. Interspeech 2018*, pp. 1111–1115, 2018.

[3] S. 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*, vol. 1. IEEE, 2001, pp. 665–668.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[5] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003. [Online]. Available: http://dx.doi.org/10.1016/S0165-1684(03)00082-3

[6] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *in Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA02)*. Citeseer, 2002.

[7] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5029–5033.

[8] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Speech Coding Proceedings, 1999 IEEE Workshop on*. IEEE, 1999, pp. 171–173.

[9] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5469–5473.

[10] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrow-band speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[12] C. Donahue, J. McAuley, and M. Puckette, "Synthesizing audio with generative adversarial networks," *arXiv preprint arXiv:1802.04208*, 2018.

[13] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural nets," in *ICLR (Workshop Track)*, 2017.

[14] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[15] Z. Yi, H. R. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation." in *ICCV*, 2017, pp. 2868–2876.

[16] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[18] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2721–2725.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *arXiv preprint*, 2017.