TIME-FREQUENCY LOSS FOR CNN BASED SPEECH SUPER-RESOLUTION

Heming Wang[1] *and Deliang Wang*[1,2]
[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{wang.11401, wang.77}@osu.edu

ABSTRACT

Speech super-resolution (SR), also called speech bandwidth extension (BWE), aims to increase the sampling rate of a given lower resolution speech signal. Recent years have witnessed the successful application of deep neural networks in time or frequency domains, and deep learning has improved the performance considerably compared with conventional approaches. This paper proposes an autoencoder based fully convolutional neural network (CNN) that merges the information from both time and frequency domains. At the training time, we optimize the CNN using a new time-frequency loss (T-F loss), which combines a time domain loss and a frequency domain loss. The experimental results show that our model trained with the T-F loss achieves significantly better results than other state-of-the-art models, and yields balanced performance in terms of time and frequency metrics.

Index Terms— Super-resolution, bandwidth extension, deep learning, convolutional neural network, T-F loss.

1. INTRODUCTION

Audio super-resolution (SR) is the challenging task of recovering a high-resolution (HR) audio from the corresponding low-resolution (LR) audio input. From the spectral domain perspective, this task is also called *bandwidth extension*, i.e., extending from a narrowband to a wideband. This problem has been studied for decades. Due to the limitation of transmission bandwidth and restriction of audio equipment, such as telephone and bluetooth devices, speech resolution (and quality) is often limited (and low) at the user end. Speech bandwidth extension (BWE) is employed to recover the wideband signal. It has been demonstrated that this technique can also help with many other speech processing tasks, like speech coding and automatic speech recognition.

Early studies adopt signal processing techniques, such as the source-filter model [1]. To predict the upper band spectral envelopes, methods including codebook mapping [2] and linear mapping [3] have been proposed. Approaches from the statistical perspective consist of GMM [4, 5] and joint HMM/GMM [6, 7]. After the introduction of deep learning, we have seen advances in many topics in the speech field. In audio SR, recent studies have introduced deep learning techniques, which are shown to outperform conventional approaches. These have explored feedforward neural networks [8], recurrent neural networks with long short-term memory (LSTM) [9], CNN [10], waveform synthesizers like WaveNet [11] and SampleRNN [12], and generative adversarial networks (GAN) [13, 14]. A more detailed summary of related studies is given in Sect. 2.

Recent deep learning studies either work in the frequency domain or the time domain. Lim et al. [15] introduced a time-frequency network (TFNet) to jointly optimize the time and frequency domains of a signal. TFNet first trains two networks in each domain respectively, and then combines their respective results to reconstruct an HR signal. It outperforms methods that only use information from one domain. As TFNet essentially trains two deep networks at the same time, its structure is complicated and takes considerable computational resources to train. In addition, its spectral branch suffers from the issue of a limited receptive field.

In this paper, we propose a CNN to leverage cross-domain information. For this purpose, we introduce a new time-frequency loss (T-F loss) to facilitate training in both time and frequency domains. Our model operates in the time domain by taking an LR signal as the input, and outputs an reconstructed signal with a higher resolution. During training, the model minimizes the T-F loss. The experiments show that our new model performs well for both time and frequency domain metrics. Unlike TFNet, our CNN is relatively simple and efficient to train.

The rest of the paper is organized as follows. In Sect. 2, we provide a detailed description of the related prior studies. In Sect. 3 we present the network design, and the time-frequency loss function. In Sect. 4, the experimental setup, results, and comparisons are presented. Finally Sect. 5 concludes this paper.

2. RELATED WORK

Li et al. [8] appears to be the first work that introduced deep neural network (DNN) to address BWE. Their DNN is pre-trained as restricted Boltzmann machines and predicts the

This research was supported in part by an NIDCD (R01 DC012048) grant and the Ohio Supercomputer Center. We thank Ashutosh Pandey for discussions on AECNN.

wideband log power spectrum (LPS) from the narrowband LPS. The phase in the extended high-frequency range is produced by flipping and repeating the narrowband phase and adding a negative sign. The experimental results show that the DNN yields better results in terms of objective and subjective measures. Abel and Fingscheidt [16] proposed another frequency domain method, which utilized a DNN to estimate the lower-dimensional cepstral representation of the speech.

There are also studies that tackle the SR task from the time domain perspective. Inspired by the successful application of deep convolutional networks in image super-resolution, Kuleshov et al. [10] introduced AudioUnet, which is adapted from the image domain network [17, 18]. Their model is trained with signals in the time domain, and learns the mapping from pairs of LR signals and HR signals. AudioUnet is shown to outperform conventional methods and considerably improves speech quality. Another time domain research employed SampleRNN [12], which is a hierarchical recurrent neural network (RNN) used for audio waveform generation.

While the above studies have promising results, they only focus on information in one domain of signal representation. To combine the advantages of both time and frequency domain methods, Lim et al. [15] proposed a time-frequency network (TFNet). They adapted AudioUnet and built two networks, including one that is trained on pairs of LR and HR signals in the time domain and the other trained on pairs of short-time Fourier transform (STFT) magnitude in the frequency domain. The two networks are jointly optimized and a spectral fusion layer is utilized to combine the outputs of two branches. Experiments show that TFNet successfully merges information from both domains and outperforms methods that operate in one domain. Our research also focuses on incorporating cross-domain information. The major difference between our network and TFNet is that our model consists of a single network and operates only in the time domain, but it is optimized with a cross-domain T-F loss.

3. NETWORK DESIGN

Fig. 1 depicts the pipeline of our SR framework. Given an LR signal sampled by 8 KHz, we first upsample it to 16 kHz using the cubic spline interpolation [19], which corresponds to the baseline of image domain SR (bicubic upsampling). Then the upsampled signal and the HR signal are fed to our network as the input and target, respectively. We jointly optimize the network with our T-F loss, and reconstruct the HR signal after training is done.

3.1. AECNN

Our network structure is based on the autoencoder CNN (AECNN) by Pandey and Wang [20]. AECNN is a fully convolutional network composed of a series of encoder and decoder blocks. It includes skip connections to better reconstruct the final output from the encoder, as the output of encoder has a limited dimension. Rectified linear unit (ReLU) is used in each layer of the network as the nonlinearity, except for the last layer where we use the hyperbolic tangent (tanh) activation. A dropout ratio of 0.2 is employed for every one of three layers. Our CNN takes upsampled LR audio segments, each having 2048 samples as the input, and outputs HR audio segments of the same shape. One major change we introduce to AECNN is replacing the transposed convolution layers in decoder blocks with subpixel layers. A subpixel layer, first proposed by Shi et al. [17], is a upscaling layer implemented by convolution. It has been reported in [21] that transposed convolution layers tend to introduce artifacts to SR tasks in the image domain, and by applying subpixel layers these artifacts can be alleviated. We also find in our experiments that, by employing subpixel layers, we can accelerate the training process and improve the objective performance.

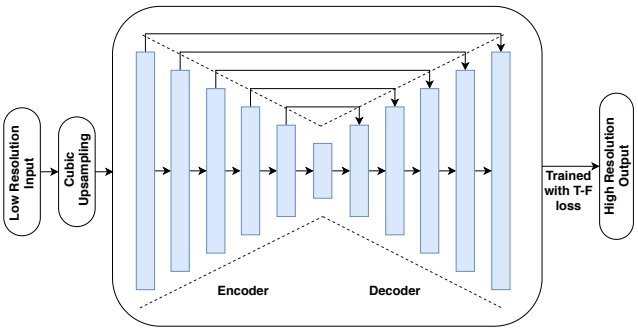


Fig. 1. Illustration of the super-resolution workflow and AECNN network structure.

3.2. Time-frequency loss

Our model is optimized with a T-F loss. The calculation of this loss function is illustrated in Fig. 2. The loss function consists of two parts: the time domain loss and the frequency domain loss. For the frequency domain loss, we adopt the loss function from [20]. Our network operates on a frame length of 2048 samples. This corresponds to a 128 ms long speech segment for a signal with a 16 kHz sampling rate. We utilize the overlap and add (OLA) method to combine the reconstructed output (denoted as SR) and calculate the loss in the frequency domain. The SR signal is first divided into frames of 512 samples, and frame shift of length 256. Then the obtained frames are multiplied by the Hamming window, so our analysis window size is 32 ms with a 50% overlap between neighboring frames. We calculate the STFT magnitude of the windowed frames and compare the results with the HR STFT magnitude. The frequency domain loss is obtained as the mean absolute error (MAE) between these two magnitudes, defined as

$$L_F(\hat{S}, S) = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} ||\hat{S}(m,k)| - |S(m,k)||, \quad (1)$$

where $\hat{S}$ and S are the STFT of SR and HR signals, respectively. We use m, k to index the frames and frequencies, respectively. For the time domain loss, we calculate the MAE of the SR and HR time series,

$$L_T(\hat{s}, s) = \frac{1}{N}\sum_{n=1}^{N}|\hat{s}(n) - s(n)|, \quad (2)$$

where $\hat{s}$ is the SR signal, and s is the ground truth HR signal. We use n to index time samples. The T-F loss is the linear combination of the time domain loss L_T and the frequency domain loss L_F, which is calculated as:

$$L = \alpha L_T + (1-\alpha)L_F. \quad (3)$$

As shown in Eq. 3, we combine L_T and L_F with a coefficient α. The value of α is set to 0.85, obtained by a grid search.

Essentially, our model is a time domain model but optimized with a T-F loss. We calculate the loss function using the OLA method to combine consecutive frames, as we find in experiments that the OLA method has better performance than simple concatenation. This result is expected, since 2048 samples are too large a frame size to satisfy the stationary assumption for short-time signal processing. The reason for calculating the STFT magnitudes is two-fold. First, by visualizing the magnitude and phase after STFT, we can observe T-F structure in the magnitude spectrogram, but not in the phase spectrogram. Second, the experiments in [20] show that a frequency domain loss function using both real and imaginary parts of the STFT does not perform as well as the one that employs only STFT magnitudes.

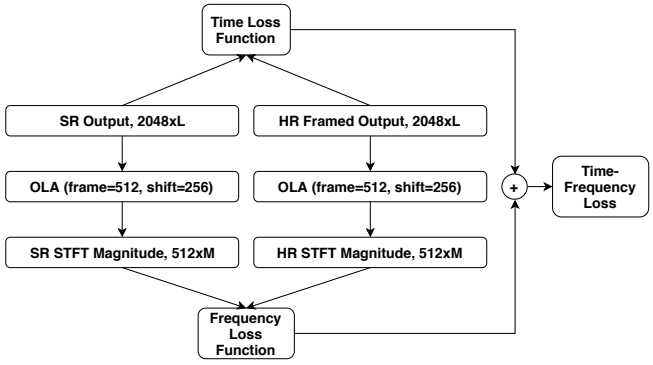


Fig. 2. Schematic diagram showing the process of calculating the T-F loss. L denotes the number of 2048-sample frames, and M is the number of 512-sample frames.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

We train our CNN using the ADAM optimizer with a learning rate of 0.0003. The network is trained with a batch size of 16 for 100 epochs. We add an early stopping criterion such that the training process will stop if the validation loss has not improved for 3 consecutive epochs. To ensure the input and target signal for our CNN have the same length, we preprocess the source input with the cubic spline interpolation.

We evaluate the SR performance with two objective metrics, the signal to noise ratio (SNR) and the log-spectral distance (LSD) [22]. They reflect the performance from the time domain and the frequency domain, respectively. SNR is defined as,

$$\mathrm{SNR}(\hat{s}, s) = 10\log_{10}\frac{\sum_{n=1}^{N} s(n)^2}{\sum_{n=1}^{N}[\hat{s}(n) - s(n)]^2}. \quad (4)$$

LSD measures the distance between two signals in the frequency domain, which is defined as follows:

$$\mathrm{LSD}(\hat{S}, S) = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{K}\sum_{k=1}^{K}[\log_{10}\frac{\hat{S}(\omega,k)^2}{S(\omega,k)^2}]^2}. \quad (5)$$

4.2. Results and Comparisons

4.2.1. TIMIT dataset

We first evaluate and analyze the performance of our model on the TIMIT corpus [23]. TIMIT is a standard corpus which contains English utterances from 630 speakers with a 16 kHz sampling rate. We choose 4620 utterances as the training dataset, and 1153 utterances to construct the validation dataset. The TIMIT core test subset consisting of 192 utterances is used as our test dataset. The test dataset consists of 24 speakers that are not included in the training and validation datasets, so we can assess the ability to generalize to new speakers. To create the LR, HR pairs for our CNN, we downsample the signal to 8 kHz for each file in the dataset as the LR signal.

We compare the quantitative performance of our model with four other deep SR models. These are DNN-BWE by Li et al. [8], AudioUnet by Kuleshov [10], TFNet by Lim et al. [15] and SampleRNN by Ling et al. [12]. See Sect. 2 for more description of each comparison method. We have successfully implemented DNN-BWE and TFNet, and adopt the code provided by the authors to implement AudioUnet. As we have difficulty in replicating the results of SampleRNN, we copy their reported results on the TIMIT core test dataset. We train all the models on the TIMIT dataset and follow the training setup described in Sect. 4.1.

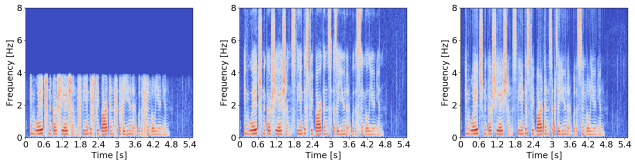
Table 1 shows the super-resolution results of the proposed CNN, as well as the four comparison methods, on the TIMIT dataset. The upscaled signal obtained by the cubic spline interpolation (available in SciPy) is used as the baseline without deep SR. Our model has improved over the spline baseline by 4.64 dB in terms of SNR, and cut LSD by 65.1%. The results show that our network consistently improve over other deep learning methods for SNR and LSD metrics. Compared

Table 1. Evaluation and comparison results of SR methods on the TIMIT dataset.

	SNR	LSD
Spline	15.48	2.27
DNN-BWE	17.37	1.56
AudioUnet	18.59	0.89
TFNet	18.91	0.87
SampleRNN	19.00	0.83
Proposed	**20.12**	**0.79**

with the state-of-the-art model (SampleRNN) on the TIMIT dataset, we have improved SNR by 1.12 dB and decreased LSD by 0.04.

Fig. 3 illustrates the output of our super-resolution model on a TIMIT utterance ("In wage negotiations, the industry bargains as a unit with a single union"). Comparing the spectrograms we can observe that the missing high-frequency components in the LR spectrogram are recovered very well by our model.



LR Spectrogram HR Spectrogram SR Spectrogram

Fig. 3. SR results visualized using spectrograms. From left to right are the spectrograms that correspond to the LR input, the ground truth HR signal and the reconstructed SR signal.

4.2.2. VCTK dataset

To further evaluate our model and compare it with prior studies, we run experiments on the VCTK dataset [24], which contains 44 hours of speech data from 108 different speakers with a 16 kHz sampling rate. Our experiments follow the setup of two tasks in previous studies [10, 15]. One is the *single-speaker* task, which trains the model on one specific speaker from the VCTK dataset. We select the first 223 recordings of that speaker as the training dataset, and test on the last 8 recordings of the same speaker. The other task is the *multi-speaker* task. We train our model on the first 99 VCTK speakers and run tests on the 8 remaining speakers. We compare the performance of our model with four baselines: the cubic spline and three deep learning methods (DNN-BWE, AudioUnet and TFNet) using their reported results. SampleRNN is not included in this comparison because it was not evaluated on this corpus by its authors. The results are summarized in Table 2.

The ratio in Table 2 indicates the upscaling ratio. $ratio = 2$ means we upscale speech signals from 8 kHz to 16 kHz, and $ratio = 4$ indicates the task is upscaling from 4 kHz to 16 kHz. As shown in the table, our model considerably outperforms other baselines at the upscale ratio of 2. Compared with the cubic spline baseline, Our model improves SNR by approximately 5 dB, and cut LSD to below 0.9 for both tasks. Moreover, our CNN significantly outperforms AudioUnet and DNN-BWE for both metrics. Additionally, Our model performs better at the upscale ratio of 4 over the baselines, and we see a substantial improvement in terms of LSD. However, the SNR gap is not as big as for the ratio 2 case.

Table 2. Experimental results comparison on the VCTK dataset at upscaling ratio 2 and 4.

		VCTK_S		VCTK_M	
Model	Ratio	SNR	LSD	SNR	LSD
Spline	2	20.3	1.95	19.7	1.91
DNN-BWE	2	20.1	1.61	19.9	1.56
AudioUnet	2	21.1	1.39	20.7	1.35
TFNet	2	N/A	N/A	N/A	N/A
Proposed	2	**25.4**	**0.81**	**24.0**	**0.89**
Spline	4	14.8	3.56	13.0	3.51
DNN-BWE	4	15.9	2.13	14.9	2.52
AudioUnet	4	17.1	1.56	16.1	1.52
TFNet	4	18.5	1.3	17.5	1.27
Proposed	4	**19.3**	**0.93**	**18.1**	**0.97**

4.2.3. Comparison of loss functions

To examine the superiority of our T-F loss, we conduct a study of different loss functions on the TIMIT corpus following the setup described in 4.1. As shown in Table 3, only using a frequency domain loss achieves the best LSD performance, but SNR performance is poor. Only using a time domain loss has a similar phenomenon. It has a high SNR value, but the performance is mediocre in terms of LSD. Our T-F loss combines the strengths of loss functions in time and frequency domains, and has a balanced performance for both metrics.

Table 3. Comparison of loss functions. From left to right are the results of the time domain mean squared error (MSE) loss, the time domain MAE loss, the frequency domain loss, and the T-F loss.

	T loss (MSE)	T loss (MAE)	F loss	T-F loss
SNR	19.69	20.11	12.75	**20.12**
LSD	0.88	0.95	**0.78**	0.79

5. CONCLUSION

In this paper, we have proposed a novel CNN model for speech super-resolution that combines the strengths of both time and frequency domain methods. The proposed CNN is fed with time domain signals and optimized using a T-F loss. The experimental results demonstrate that our model significantly outperforms the existing approaches and has a balanced performance in SNR and LSD metrics. Furthermore, our approach is computationally efficient and avoids complex network design.

6. REFERENCES

[1] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proceedings of ICASSP*, 1979, vol. 4, pp. 428–431.

[2] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proceedings of ICASSP*, 2005, vol. 1, pp. I–805.

[3] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," 2001, vol. 1, pp. 665–668.

[4] A.H. Nour-Eldin and P. Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech," in *Proceedings of INTERSPEECH*, 2011, p. r1188.

[5] Y. Wang, S. Zhao, Y. Yu, and J. Kuang, "Speech bandwidth extension based on GMM and clustering method," in *Proceedings of CSNT*, 2015, pp. 437–441.

[6] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proceedings of IWAENC*, 2014, pp. 1–5.

[7] M.T. Turan and E. Erzin, "Synchronous overlap and add of spectra for enhancement of excitation in artificial bandwidth extension of speech.," in *Proceedings of INTERSPEECH*, 2015, pp. 2588–2592.

[8] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *Proceedings of ICASSP*, 2015, pp. 4395–4399.

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

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

[11] M. Wang, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, D. Yu, and H. Meng, "Speech super-resolution using parallel WaveNet," in *Proceedings of ISCSLP*, 2018, pp. 260–264.

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

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

[14] D. Haws and X. Cui, "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *Proceedings of ICASSP*, 2019, pp. 6780–6784.

[15] T.Y. Lim, R.A. Yeh, Y. Xu, M.N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *Proceedings of ICASSP*, 2018, pp. 646–650.

[16] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 71–83, 2017.

[17] W. Shi, J. Caballero, F. Huszár, J. Totz, A.p. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of CVPR*, 2016, pp. 1874–1883.

[18] C. Dong, C.C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[19] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998.

[20] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[21] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.

[22] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[23] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[24] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.