Multi-scale audio super resolution via deep pyramid wavelet convolutional neural network

Binqiang Si, Dongqi Luo, and Jihong Zhu

Introduction: Audio super resolution (ASR), or audio bandwidth extension (ABE), is a basic problem in audio signal processing and has been studied in many works [1–3]. Given an audio with low sample rate, the goal of ASR is to recover the corresponding ground truth audio with high sample rate. ASR is of importance in many applications, e.g. multimedia communication and speech synthesis [4, 5]. For example, in communication applications, we can recover a clear version of the received low resolution audio, thus narrower communication bandwidth and less computational cost is required.

Convolutional neural networks (CNN) have exhibited superior performance to traditional methods in computer vision. Some recent works [1, 6] employed CNN to super resolve the audio. However, the audio signal is non-stationary, and the CNN based method always fails in capturing the detailed information. Hence, it is proposed to include the wavelet coefficients of the audio signal, and reconstruct the signal from these coefficients stage by stage rather.

The prediction errors of the wavelet coefficients are included to the loss function to force the model to capture the detail components. Experimental results show that the approach, training on the VCTK public dataset, achieves more appealing results than state-of-the-art methods.

Related work: ASR is a basic problem in audio signal processing. Various methods are proposed to solve the ASR problem, such like Gaussian mixture model based approaches [10–12], hidden Markov model based approaches [13, 14]. As the DNN shows powerful generalisation performance in many machine learning applications, such like computer vision and natural language processing, many DNN based approaches have been developed for ASR [1, 7, 15, 16].

CNN is a typical structure of DNN which shows powerful performance in many computer visio tasks, and was first employed in image super resolution (ISR). Chao et al. [17] used a fully convolutional neural network SRCNN to perform an ISR and achieve state-of-the-art performance, they also proved that SRCNN methods is equivalent to sparse-coding based methods. Kim et al. [18] employed a neural network to capture the high resolution audio, then the features are used to predict the wavelet coefficients of the high resolution audio.

Fig. 1 Architecture of PWCNN (×4). The resblocks extract features of the input low resolution audio, then the features are used to predict the wavelet coefficients of the high resolution audio.

Fig. 2 Illustration of the haar wavelet decomposition of an audio segment. The approximation maintains the global information of the signal, and the detail component depicts the local details.

Huang et al. [19] showed that image super resolution can benefit from wavelet information.

For ASR, the CNN based ASR approaches are also reported to outperform traditional methods. The CNN based ASR methods can be roughly divided into two classes: one is time domain based and another is frequency domain based. Similar to ISR, and use the raw data of the audio signal as input to predicts the high resolution signal. Kuleshov et al. [1] proposed a CNN based model with bottleneck architecture, they adopt the skip connection for residual-learning, and use the pixel shuffling layer for upsampling. Kim et al. [7] adopted a network similar to [1] as the generator and proposed a GAN based method, and use the unsupervised feature loss to improve the quality. The latter one focuses on the reconstruction of the spectrogram. Schmidt et al. [16] proposed long short term memory (LSTM) based method, which employed the LSTM to estimate the spectral envelope. Our model is time domain based, and use the wavelet coefficient to improve the estimation of detailed information.

Method:

Discrete wavelet transform: Given a discrete signal \( x \in \mathbb{R}^L \), the Mallat algorithm [20] gives a fast filtering implementation of wavelet decomposition, which is formulated as

\[
c[n] = (\downarrow 2)(x * \tilde{g})[n] = \sum_k x[k]*\tilde{g}[2n - k] \tag{1}
\]

\[
d[n] = (\downarrow 2)(x * \tilde{h})[n] = \sum_k x[k]*\tilde{h}[2n - k] \tag{2}
\]

where \( \tilde{g} \) and \( \tilde{h} \) are low pass and high pass filters of chosen wavelet transform, respectively, and \( c, d \) are the approximation and detail component respectively. \( \downarrow 2 \) denotes subsampling with stride 2. Figure 2 illustrates the haar wavelet decomposition of an audio segment.

Similarly, the Mallat algorithm inverse discrete wavelet transform is formulated as

\[
x[n] = g * (\uparrow 2)c[n] + h * (\uparrow 2)d[n] = \sum_k c[k]*g[k - 2n] + \sum_k d[k]*h[k - 2n] \tag{3}
\]
where \( g \) and \( h \) are low pass and high pass filters of chosen inverse wavelet transform, respectively. (\( f / 2 \)) is the zero-filling upsampling operator with stride 2.

**Network architecture:** To perform more precise super resolution of a low resolution signal \( x \), we propose to predict the wavelet coefficients of the desired high resolution signal stage by stage, Figure 1 shows the architecture of our method when the scaling factor is 4. We first use a resblock to extract features of the original low resolution signal \( x \), then pass the feature maps to two independent convolutional layers to predict the approximation and detail components of the next scale, finally we reconstruct the signal of the next scale, repeat the previous steps we can reconstruct the high resolution signal recursively.

The PWCNN is a fully convolutional neural network, it is composed by three component:

- **Embedding net.** The embedding net consists of resblocks, and the architecture of resblocks are shown in Figure 1. The embedding nets are employed to extract features of the input raw data and introduce the non-linearity to our model.
- **Convolutional layer.** Generally an audio signal has at most two channels, but we always need more feature maps to model the input signal, the convolutional layers are located between the input and the embedding nets to match the channels.
- **IDWT net.** As shown in Equation (3), to perform IDWT, we first up-samples the input signal with zeros filled, then convolve it with fixed filters. It can simply be implemented by transposed convolution.

We empirically set the size of convolutional kernels to three except the IDWT nets, and we take zero-padding to keep the size of feature maps unchanged.

**Loss function:** Given a dataset \( \mathcal{D} = \{ x_i, y_i \}_{i=1}^N \), where \( x_i \) are low resolution signals and \( y_i \) are corresponding ground truth high resolution signals, and we denote \( \hat{x} \) as the predicted high resolution signal. Let \( e^j_i \) and \( d^j_i \) be the \( j \)-level wavelet coefficients of \( x_i, e^j, d^j \) be the \( j \)-level wavelet coefficients of \( y_i \). The loss function of our method are combined of two parts

\[
\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{wave}
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \| x_i - \hat{x} \|_1 + \frac{1}{M} \sum_{j=1}^{J} \sum_{k=1}^{K} v_j \| d^j_i - \hat{d}^j_i \|_1
\]

(4)

where \( J = \log_2 S \in \mathbb{N}_0 \) is the level of wavelet decomposition, \( S \) is the scale of super resolution, \( v_j > 0 \) are the balance coefficients, \( \mathcal{L}_{rec} \) measures the error of reconstruction, \( \mathcal{L}_{wave} \) measures the error of the prediction of detail component and force the network to predict a correct wavelet coefficients.

**Experiments:**

**Dataset:** To evaluate our method, we perform several experiments on the VCTK dataset [21]. The VCTK dataset contains speech recordings of 109 native English speakers, all the recordings are collected by a 96 kHz microphone and downsamplored to 48 kHz via STPK. Note that the length of the original data \( L \) may not be divisible by the scale \( S \), thus we simply crop the first \( L - \text{mod}(L, S) \) samples of the original data, then use the cropped audio as the high resolution signal, and we obtain the low resolution data by low-pass filtering and downsampling.

| Scale | Metric | Spline | SSRNet | SSRRGAN | PWCNN |
|-------|--------|--------|---------|---------|-------|
| \( \times 2 \) | SNR    | 19.7   | 20.7    | 20.0    | 22.6  |
|        | LSD    | 4.4    | 3.1     | 2.1     | 1.0   |
| \( \times 4 \) | SNR    | 13.0   | 16.1    | 14.0    | 16.5  |
|        | LSD    | 8.0    | 3.5     | 2.7     | 1.4   |

Table 1. Performance comparison for multiple speaker task

Follow the similar procedure of [1], we construct two data set from the original VCTK dataset for single speaker and multi-speaker tasks. For single speaker task, we use the first 223 recordings of the first speaker for training and the last 8 recordings for testing. For multi-speaker task, we use the recordings of the first 99 speakers for training and test on the remaining recordings.

**Metrics:** In this letter, we adopt two frequently-used metrics to evaluate the quality of super resolution. The signal-to-noise-ratio (SNR) is defined as

\[
\text{SNR} = 20 \log_{10} \frac{|x|_2}{|x - \hat{x}|_2}
\]

(5)

SNR measures the quality of super resolution in time domain.

In frequency domain, we take the log-spectral distance (LSD) for metrics.

\[
\text{LSD} = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} (X(l, k) - \hat{X}(l, k))^2
\]

(6)

where \( X \) and \( \hat{X} \) denote the STFT coefficients of \( x \) and \( \hat{x} \) respectively. In this letter, we use the Hamming window and the window length is set to 2048, and the number of overlaps is set to 0.

**Results:** We conduct two experiments to compare our model and other baseline methods. We implement our model in PyTorch, we empirically set \( v_j = 1 \) for all \( j \). We use the Adam optimiser to train our model, the learning rate is set to \( 10^{-5} \). The data in the same mini-batch should be identical, hence we randomly crop segment with a length of 8000 for training. In the testing phase, we can use the full audio to evaluate the model.

Table 1 illustrates the performance comparison for multiple speaker task. The experimental results shows that our approach outperforms other methods. Table 2 shows the performance comparison for single speaker task, we observed that the SNR of method is lower than SSRRGAN when the scaling factor is 4, but the LSD of our method is more appealing. The reason that our model achieves better LSD is that our method focus on recovering the high frequency component of the ground truth audio signal, hence the recovered spectrogram is more precise. Figure 4 illustrates the spectrograms of the original audio and the SR audios on \( \times 4 \) SR tasks. We found that our method has a stronger estimation spectrogram capability, while SSRNet tends to produce an oversmooth spectrogram, which is why our model achieves a better LSD. Figure 5 compares the convergence curve of \( L_1 \) loss and \( L_2 \) loss with the same
settings, we observed that $L_1$ loss do achieves better result, while $L_2$ loss only achieves a SNR of 16.3 dB.

Conclusion and future work: In this letter, we propose a pyramid wavelet CNN (PWCNN) to solve the ASR problem. We propose to use the pyramid architecture to estimate the wavelet decomposition coefficients progressively, and use the estimation error of wavelet coefficients as a part of loss function to learn the estimation of detailed information. Experimental results show that our method achieves better performance than previous methods, for both ×2 and ×4 ASR tasks.

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 20 February 2021 Accepted: 25 March 2021
doi: 10.1049/el.12180

References

1. Kuleshov, V., Enam, S.Z., Ermon, S.: Audio super-resolution using neural nets. In: Workshop of International Conference on Learning Representation, pp. 1–5 (2017)

2. Lim, T.Y. et al.: Time-frequency networks for audio super-resolution. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 646–650. IEEE, Piscataway, NJ (2018)

3. Dong, J., Wang, W., Chambers, J.: Audio super-resolution using analysis dictionary learning. In: 2015 IEEE International Conference on Digital Signal Processing, pp. 604–608. IEEE, Piscataway, NJ (2015)

4. Eskimez, S.E., Koishida, K.: Speech super-resolution generative adversarial network. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3717–3721. IEEE, Piscataway, NJ (2019)

5. Chen, Z., et al.: An audio watermark-based speech bandwidth extension method. EURASIP Journal on Audio, Speech, and Music Processing 2013, 10 (2013)

6. Wang, H., Wang, D.: Time-frequency loss for cnn based speech super-resolution. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 861–865. IEEE, Piscataway, NJ (2020)

7. Kim, S., Sathe, V.: Adversarial audio super-resolution with unsupervised feature losses. (2019). https://openreview.net/forum?id=H1eH4n09KX. Accessed 19 April 2021.

8. Eskimez, S.E., Koishida, K., Duan, Z.: Adversarial training for speech super-resolution. IEEE J. Sel. Top. Signal Process. 13(2), 347–358 (2019)

9. Li, X. et al.: Speech audio super-resolution for speech recognition. In: Proc. Interspeech, pp. 3416–3420. International Speech Communication Association, Baixas (2019)

10. Pulakka, H. et al.: Speech bandwidth extension using gaussian mixture model-based estimation of the highband MEL spectrum. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5100–5103. IEEE, Piscataway, NJ (2011)

11. Pulakka, H., et al.: Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and a Gaussian mixture model. IEEE Trans. Audio Speech Lang. Process. 20(8), 2219–2231 (2012)

12. Wang, Y. et al.: Speech bandwidth extension based on GMM and clustering method. In: 2015 Fifth International Conference on Communication Systems and Network Technologies, pp. 437–441. IEEE, Piscataway, NJ (2015)

13. Jax, P., Vary, P.: Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden markov model. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1–L. IEEE, Piscataway, NJ (2003)

14. Song, G.-B., Martynovich, P.: A study of hmm-based bandwidth extension of speech signals. Signal Process. 89(10), 2036–2044 (2009)

15. Li, K. et al.: DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In: Sixteenth Annual Conference of the International Speech Communication Association, pp. 2578–2582. International Speech Communication Association, Baixas (2015)

16. Schmidt, K., Edler, B.: Blind bandwidth extension based on convolutional and recurrent deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5444–5448. IEEE, Piscataway, NJ (2018)

17. Dong, C., et al.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38(2), 295–307 (2015)

18. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654. IEEE, Piscataway, NJ (2016)

19. Huang, H. et al.: Wavelet-mlsrt: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1689–1697. IEEE, Piscataway, NJ (2017)

20. Mallat, S.G.: A theory for multisresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11(7), 674–693 (1989)

21. Veaux, C. et al.: Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research, Edinburgh (2017)