QISTA-Net-Audio: Audio Super-resolution via Non-Convex $\ell_q$-norm Minimization

Gang-Xuan Lin$^1$, Shih-Wei Hu$^2$, Yen-Ju Lu$^1$, Yu Tsao$^1$, and Chun-Shien Lu$^{12}$

$^1$Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
$^2$Institute of Information Science, Academia Sinica, Taipei, Taiwan

Abstract

Audio super-resolution (ASR) aims to reconstruct the high-resolution signal from its corresponding low-resolution one, which is hard while the correlation between them is low.

In this paper, we propose a learning model, QISTA-Net-Audio, to solve ASR in a paradigm of linear inverse problem. QISTA-Net-Audio is composed of two components. First, an audio waveform can be presented as a complex-valued spectrum, which is composed of a real and an imaginary part, in the frequency domain. We treat the real and imaginary parts as an image, and predict a high-resolution spectrum but only keep the phase information from the viewpoint of image reconstruction. Second, we predict the magnitude information by solving the sparse signal reconstruction problem. By combining the predicted magnitude and the phase together, we can recover the high-resolution waveform.

The contributions of this paper are summarized as follows.

1. We propose a novel framework to reconstruct the high sampling rate audio signal by the information from the magnitude and the phase, both of which are obtained via learning-based methods.

2. We predict the spectrum from the available real and imaginary parts. However, instead of reconstructing the real part and the imaginary part separately, our novel idea aims to combine both of them together and treat it as an image (upper part of Fig. 1) in order not to lose the correlation between them. We then adopt the QISTA-Net-n [3] to reconstruct the image.

3. In order to improve the reconstruction performance, we predict the magnitude (bottom part of Fig. 1) by solving a non-convex $\ell_q$-norm minimization problem, which leads to better performance than the one obtained from the real and imaginary parts. We modify QISTA-Net-n [3] appropriately, called QISTA-Net-m, to reconstruct the magnitude.

1.1. Notations

Let $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, and $x_0 \in \mathbb{R}^n$ be the low-resolution waveform in the time domain with signal length $m$, the pre-
dicted high-resolution waveform with signal length $n$, and the raw waveform that we aim to find, respectively. Let $\hat{x}$ and $\hat{y}$ be the complex-valued spectrum of $x$ and $y$ in the frequency domain, respectively. Let $\hat{y} \in \mathbb{R}^m$, $\hat{x} \in \mathbb{R}^n$, and $x_0 \in \mathbb{R}^n$ be the magnitude of $y$, $x$, and $x_0$, respectively. Let $\hat{y} \in \mathbb{R}^m$ and $\hat{x} \in \mathbb{R}^n$ be the phase of $y$ and $x$, respectively. Let $\text{re}(x)$ and $\text{im}(x)$ be the real part and the imaginary part of $x$, respectively. Let $I_n$ be the $n \times n$ identity matrix. Moreover, $\|\cdot\|_2$ means $\ell_2$-norm, and the $\|\cdot\|_F$ means Frobenius norm.

2. Related Work

Audio super-resolution has been studied for decades. Early studies include signal processing techniques [4], upper band spectral envelopes prediction [5, 6] and statistical techniques [7, 8, 9, 10]. For the past few years, deep learning has received considerable attention due to its outstanding performance in diverse applications, including audio super-resolution [11, 12, 13, 14].

To the best of our knowledge, [11] is the first one in introducing deep learning to address speech bandwidth extension. But, the extended high-frequency phase is reconstructed naively by flipping and repeating the narrowband phase and adding a negative sign. Besides, [15] proposed a generative adversarial network for audio super-resolution, using the same strategy as [11] to reconstruct the extended high-frequency phase.

Inspired by the convolutional neural network (CNN) in image super-resolution, [12] introduced AudioUnet to treat audio signals as image signals. Based on AudioUnet, [13] proposed a time-frequency network (TFNet), which is the first one considering both spectrum domain and time domain information. TFNet adopted AudioUnet to upsample the low-resolution signal for obtaining its high-resolution counterpart in the time domain, and in the meantime, predicted the magnitude information in the frequency domain. The final result is then obtained by combining both these signals. Similar to TFNet, we adopt our own QISTA-Net-n [3] and build two networks to reconstruct the magnitude and the phase, respectively. Moreover, the authors [13] used both information in the time and frequency domains to reconstruct the high-resolution signal. On the contrary, our study uses both the magnitude and phase in the frequency domain for high-frequency signal recovery.

Another convolution-based method is the multi-scale fusion neural network (MfNet), which was proposed by [1]. The authors heuristically stack and connect multiple convolution operators to fuse different scale convolution outputs, and it consistently outperforms the competing methods in terms of perceptual evaluation of speech quality (PESQ) and signal to noise ratio (SNR). Therefore, in this paper, we mainly compare with MfNet to verify our method.

3. Proposed Method

The flowchart of proposed method is illustrated in Fig. 1. Since an audio waveform can be represented as a complex-valued spectrum in the frequency domain through the transformation methods such as STFT (short-time Fourier transform). Therefore, it can be represented in terms of the magnitude and phase components, or represented in terms of the real and imaginary parts.

We propose to predict $\hat{x}$, the complex-valued spectrum of $x$, in the frequency domain from $\text{re}(x)$ and $\text{im}(x)$ via (3) in Sec. 3.1. However, after predicting $\hat{x}$, we only reserve the phase information $\hat{x}$ and ignore the magnitude. This is because the magnitude part can be better estimated by considering the magnitude as an approximately sparse signal. To improve the quality of the magnitude $\hat{x}$, we treat recovery of the magnitude part as a sparse signal reconstruction problem in Sec. 3.2. We will see later in Sec. 4.4 that magnitude recovery in Sec. 3.2 outperforms that in Sec. 3.1.

Combining both the predicted magnitude from Sec. 3.2 and predicted phase from Sec. 3.1, we finally obtain the high sampling rate waveform $x$ in Sec. 3.3.

3.1. Reconstructing the Phase

Let $\hat{x}$ and $\hat{y}$ be the complex-valued spectrum of $x$ and $y$ in the frequency domain, respectively, and let $\hat{x}$ and $\hat{y}$ be the phase of $x$ and $y$, respectively. To predict $\hat{x}$ from $\hat{y}$, where $\hat{y}$ consists of $\text{re}(y)$ and $\text{im}(y)$, we can simply predict $\text{re}(x)$ and $\text{im}(x)$ by $\text{re}(y)$ and $\text{im}(y)$, respectively. Nevertheless, the correlation between real and imaginary will be ignored. With this consideration, we stack $\text{re}(y)$ and $\text{im}(y)$ together and treat it as an image $\hat{Y} = [\text{re}(y) : \text{im}(y)]$ with the size $m \times 2$, which we call the RI image. Thus, we aim to predict the RI image $\hat{X} = [\text{re}(x) : \text{im}(x)]$ of $x$ from $\hat{Y}$.

Since we treat a real part and an imaginary part of an audio as an image, then we adopt an image reconstruction method to construct $\hat{X}$. In [3], we proposed an efficient learning-based method called QISTA-Net-n by solving the $\ell_p$-norm minimization problem (3) for image reconstruction. In this subsection, we adopt QISTA-Net-n, which is described in Algorithm 1 and illustrated in the dashed rectangle box in Fig. 1, to construct the RI image $\hat{X}$.

In Algorithm 1, $\Psi^t$ and $\tilde{\Psi}^t$ are denoted, respectively, as

$$
\Psi^t(X) = C^t_0 \left( R \left( C^t_1 \left( R \left( C^t_2 \left( X \right) \right) \right) \right) \right)
$$

$$
\tilde{\Psi}^t(X) = C^t_0 \left( C^t_1 \left( R \left( C^t_2 \left( X \right) \right) \right) \right)
$$

where $C^t_i, i = 0, 1, \cdots, 7$, are convolution operators, $R$ is the Rectified Linear Unit (ReLU) function, and both $A$ and $B$ are
where $\bar{\tilde{y}}$ is the identity operator, $\bar{\Psi} = \bar{\Psi}_0$. The loss function here is MSE-loss plus an auxiliary loss

$$\mathcal{L} = \mathcal{L}_{MSE} + \delta \mathcal{L}_{aux} = \frac{1}{n} \| \tilde{X}_0 - \tilde{X}^T \|^2_F + \delta \sum_{t=1}^T \| \bar{\Psi}^t \circ \bar{\Psi} - I \|^2_F,$$

where $\tilde{X}^T$ is the output of the $T$-th layer of the network, $\tilde{X}_0$ is the RI image of $x_0$ (notice that $\tilde{X}_0$ is the initialization in Algorithm 1), $I$ is the identity operator, $\| \cdot \|_F$ is the Frobenius norm, and $\delta = 0.01$. The auxiliary loss is implemented by

$$\mathcal{L}_{aux} = \sum_{t=1}^T \| \bar{\Psi}^t(\bar{\Psi}^t(\tilde{R}^t)) - \tilde{R}^t \|^2_F.$$

Via Algorithm 1, we predict the spectrum $\tilde{x}$ from $\tilde{y}$, and then obtain the phase $\hat{x}$ from $\tilde{x}$.

Nevertheless, as we have described previously, in order to further improve the quality of $\hat{x}$, we reserve the reconstructed phase only and improve the quality of magnitude by solving the non-convex $\ell_p$-norm minimization problem in (3), with $\bar{\Psi} = \bar{\Psi}_n$ being an identity matrix, as described in the next section.

### 3.2. Reconstructing the Magnitude

Given an audio signal, we have two major information, magnitude and phase, in the Fourier domain. Let $\tilde{y}$ and $\tilde{x}$ be the magnitude of $y$ and $x$, respectively. In this section, we predict $\tilde{x}$ of the high-resolution signal by solving the non-convex $\ell_p$-norm minimization problem (3) [3]. Since the magnitude component is said to be (nearly) sparse due to sparse representation of STFT, an identity matrix is simply chosen as the dictionary $\bar{\Psi}$.

Specifically, we modify Algorithm 1 appropriately with $\bar{\Psi} = \bar{\Psi}_n$ (fix both learning parameters $\bar{\Psi}^t$ and $\bar{\Psi}$ as an identity matrix) and fix the learning parameter $\alpha^t$ as a constant 1. Moreover, the input and output are the magnitudes $\tilde{y}$ and $\tilde{x}$, respectively. The resultant algorithm is called QISTA-Net-m (QISTA-Net-n for magnitude estimation), as shown in Algorithm 2.

In QISTA-Net-m, we let $\bar{\beta}^t = \bar{\beta} \bar{B}$ be the learning parameters. Note that, the matrix $\bar{B}$ is the learning parameter, whereas, as suggested in [3], $\bar{A}$ is a deterministic downsampling operator\(^1\) to reduce the training time without loss of reconstruction performance.

The loss function here is MSE-loss

$$\mathcal{L}_{MSE} = \frac{1}{n} \| \hat{x}_0 - \hat{x}^T \|^2_F.$$

The magnitude reconstruction is shown in the dashed oval box in Fig. 1.

### 3.3. Reconstructing the High Sampling Rate Waveform: QISTA-Net-Audio

In Sec. 3.1, we predict the complex-valued spectrum $\tilde{x}$. Since the magnitude $\hat{x}$ is always an approximately sparse signal, we only keep the phase of $\hat{x}$, and improve the quality of $\hat{x}$ as described in Sec. 3.2. That is, we adopt the estimated phase and magnitude from Sec. 3.1 and Sec. 3.2, respectively. We finally obtain the predicted high-resolution waveform $x$ by applying the ISTFT (inverse STFT). Such a resultant $x$ is the final output of our framework, which we call QISTA-Net-Audio.

### 4. Experimental Results

In this section, we show the performance of the proposed method QISTA-Net-Audio\(^2\) in reconstructing the high sampling rate waveforms. We compare QISTA-Net-Audio with MiNet [1]. The reconstruction quality was measured in three metrics: the SNR (signal-to-noise ratio) (SNR = $10 \log_{10} \left( \frac{\sum^{|y|} |y|^2}{\sum^{|\tilde{y} - y|} |\tilde{y} - y|^2} \right)$), the PESQ (perceptual evaluation of speech quality) [16], and the STOI (short-time objective intelligibility measure) [17].

#### 4.1. Parameter Setting

In this paper, the frequency spectrum of each time domain waveform are represented in the Short-Time Fourier Transform (STFT) domain. The window size of STFT is 256 with 50% overlap.

The constant parameters in QISTA-Net-n were $\varepsilon = 0.1 \cdot \mathbf{I}_{8 \times 2}$, where $\mathbf{I}_{n \times 2} \in \mathbb{R}^{n \times 2}$ is a matrix with each component being equal to 1, and $q = 0.05$. The learning parameters of QISTA-Net-n were initialized as $\lambda^t = 10^{-0.5}$, $\beta^t = 10^{-1}$, $\alpha^t = 1$, and both $\bar{A}$ and $\bar{B}$, and all of $\bar{C}_i$, $i = 0, 1, \ldots, 7$ were initialized by xavier initializer [18], for all $1 \leq t \leq T$. The numbers of input features and output features of all the convolution layers are listed in Table 1.

In addition, the constant parameters in QISTA-Net-m were $\varepsilon = 0.1 \cdot \mathbf{I}_n$, where $\mathbf{I}_n \in \mathbb{R}^n$ is a vector with each component\(^3\)  

\(^1\)The sampling rate of $y$ is half of that of $x$, then $A_{t,2t-1} = 1$ for all $n$, and 0, otherwise.

\(^2\)All of our implementation codes can be downloaded from https://github.com/spybeiman/QISTA-Net-Audio/
Table 1: The list of input feature and output feature.

<table>
<thead>
<tr>
<th>Input feature</th>
<th>(\ell_0)</th>
<th>(\ell_1)</th>
<th>(\ell_2)</th>
<th>(\ell_3)</th>
<th>(\ell_4)</th>
<th>(\ell_5)</th>
<th>(\ell_6)</th>
<th>(\ell_7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output feature</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 2: Comparison between QISTA-Net-Audio and MfNet [1] with 2X upsampling in terms of average SNR (dB), PESQ, and STOI. The results of [1] were directly excerpted from the paper.

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>QISTA-Net-Audio</td>
<td>31.44</td>
<td>4.31</td>
<td>0.9958</td>
</tr>
<tr>
<td>MfNet</td>
<td>24.50</td>
<td>3.76</td>
<td>-</td>
</tr>
<tr>
<td>MfNet+P</td>
<td>24.55</td>
<td>3.88</td>
<td>-</td>
</tr>
<tr>
<td>MfNet+A</td>
<td>24.77</td>
<td>3.80</td>
<td>-</td>
</tr>
<tr>
<td>MfNet+C</td>
<td>24.70</td>
<td>3.82</td>
<td>-</td>
</tr>
</tbody>
</table>

being equal to 1, and \(q = 0.05\). The learning parameters of QISTA-Net-m were initialized as \(\lambda^t = 10^{-5}\), \(\beta^t = 10^{-3}\), and \(\mathcal{B}\) was initialized by xavier initializer [18].

4.2. Datasets for Training and Testing

For a fair comparison, we choose the same voicebank corpus\(^3\) [19] as in MfNet [1]. For each audio waveform in the training dataset (in a total of 84 speakers) and test dataset from [19], we apply STFT to obtain our training dataset and test dataset, respectively. We randomly select 133096 and 14789 samples from our training dataset for training and validation, respectively [1].

4.3. Performance Comparison

The comparison results are shown in Table 2. The “-” in Table 2 means the authors did not provide the results. It is obvious to observe that our method obtains significantly better quality than [1]. We show the spectrogram of the predicted waveforms for visual inspection in Fig. 2. As we can see, the high-frequency information is well predicted. The testing dataset consists of 824 sentences in a total of 2 different speakers, where Fig. 2 (a)-(c) and (d)-(f) were randomly selected from the two speakers, respectively.

4.4. Ablation Study

On the one hand, in Sec. 3.1, we predict the complex-valued spectrum \(\hat{x}\), where the magnitude \(\hat{x}\) of \(\hat{x}\) achieves the reconstruction performance \(\text{SNR}=31.61\, \text{dB}\). On the other hand, the magnitude \(\hat{x}\) obtained in Sec. 3.2, results in the performance \(\text{SNR}=33.98\, \text{dB}\). Therefore, we adopt the phase from Sec. 3.1 and the magnitude from Sec. 3.2 to predict the high-resolution waveform \(x\).

5. Conclusions

The challenging issue of audio super-resolution (ASR) in reconstructing the high resolution signal from its low-resolution counterpart originates from the low correlation between them. In this paper, we propose a learning model, QISTA-Net-Audio, to solve ASR in a paradigm of linear inverse problem. The characteristics of our method include 1) the phase part is predicted by treating the available real and imaginary information as an image for image recovery and 2) the magnitude part is directly treated as a sparse signal recovery problem. Comparison with state-of-the-art demonstrates the effectiveness of our method.

6. References


---

\(^3\)The corpus is provided by Valentini et al., which is publicly available. https://datashare.is.ed.ac.uk/handle/10283/2791


