INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

QISTA-Net-Audio: Audio Super-resolution via Non-Convex ℓ_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

spybeiman@gmail.com, vacuityhu@iis.sinica.edu.tw, neilyenjulu@gmail.com,
yu.tsao@citi.sinica.edu.tw, lcs@iis.sinica.edu.tw

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. Comparison with the state-of-the-art method MfNet [1], in terms of measure metrics SNR, PESQ, and STOI, demonstrates the superior performance of our method.

Index Terms: audio super-resolution, speech bandwidth extension, non-convex optimization, deep learning, sparsity

1. Introduction

The audio super-resolution (ASR), similar to the image super-resolution problem where the audio samples can be analogous to the image pixels, is to reconstruct the high frequency (HF) signal from the low frequency (LF) one. Nevertheless, the ASR problem is challenging because the correlation between the HF and LF is low. Due to the limitation of transmission bandwidth and restriction of audio equipment, such as telephone and Bluetooth devices, the speech sampling rate is often limited at the user end. From the viewpoint of another user end, for example, the mobile phone, the sampling rate is usually higher than the telephone or the Bluetooth, leading to the requirement of ASR.

Formally, the audio super-resolution problem can be treated as a linear inverse problem (LIP). Let $x_0 \in \mathbb{R}^n$ be an high sampling rate audio signal, and let $A \in \mathbb{R}^{m\times n}$ be a downsampling matrix, where $m < n$. We obtain the corresponding low sampling rate audio signal y, where

$$y = Ax_0. \tag{1}$$

Traditionally, given y and A, solving x in (1) is equivalent to solving the following minimization problem:

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\mathcal{R}(\Psi(x)), \tag{2}$$

where Ψ denotes a kind of dictionary that can transform x to a sparse representation and $\mathcal{R}(z)$ is the regularization term corresponding to the structure of z. For example, since x is assumed to be an audio signal, we may adopt a discrete Fourier transform (DFT) as the dictionary, and the structure of $\Psi(x)$ could be expected to be sparse. In general, the ℓ_0-norm is used as the regularization term to measure the sparsity of a signal. However, due to the non-convexity and discontinuity of the ℓ_0-norm, the ℓ_1-norm is the most common relaxation option.

In this paper, to reduce the relaxation gap between the ℓ_0-norm and the ℓ_1-norm, we introduce the non-convex ℓ_q-norm regularization minimization problem as follows:

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|\Psi(x)\|_q^q, \tag{3}$$

where $\|\Psi(x)\|_q = \sum_i(|\Psi(x)_i|^q)^{\frac{1}{q}}$ and $0 < q < 1$, is called ℓ_q-quasi-norm, or ℓ_q-norm for short. Meanwhile, an efficient learning-based model for the ℓ_q-norm regularization minimization problem has been proposed by our earlier work [2, 3].

According to the DFT, we have two major information—magnitude and phase—in the spectrum domain. Magnitude and phase can be treated as signals in (2). More specifically, we define two ℓ_q-norm minimization problems corresponding to magnitude and phase, respectively, and propose a framework to reconstruct these two signals based on QISTA-Net-n [3], simultaneously. As illustrated in Fig. 1, we propose a novel learning-based method, dubbed QISTA-Net-Audio, for audio signal reconstruction.

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 ℓ_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-

http://dx.doi.org/10.21437/Interspeech.2021-670

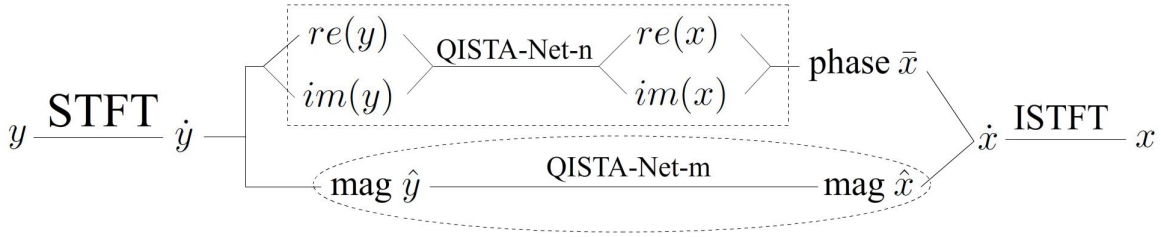

Figure 1: *Flowchart of proposed QISTA-Net-Audio. The dashed rectangle box is the reconstruction of the phase in Sec. 3.1. The dashed oval box is the reconstruction of the magnitude in Sec. 3.2.*

dicted high-resolution waveform with signal length n, and the raw waveform that we aim to find, respectively. Let $\dot{x}$ and $\dot{y}$ 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 $\hat{x}_0 \in \mathbb{R}^n$ be the magnitude of y, x, and x_0, respectively. Let $\bar{y} \in \mathbb{R}^m$ and $\bar{x} \in \mathbb{R}^n$ be the phase of y and x, respectively. Let $re(x)$ and $im(x)$ be the real part and the imaginary part of $\dot{x}$, respectively. Let I_n be the $n \times n$ identity matrix. Moreover, $\|\cdot\|_2$ means ℓ_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 [1, 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 $\dot{x}$, the complex-valued spectrum of x, in the frequency domain from $re(x)$ and $im(x)$ via (3) in Sec. 3.1. However, after predicting $\dot{x}$, we only reserve the phase information $\bar{x}$ from $\dot{x}$, 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 $\dot{x}$ and $\dot{y}$ be the complex-valued spectrum of x and y in the frequency domain, respectively, and let $\bar{x}$ and $\bar{y}$ be the phase of x and y, respectively.

To predict $\dot{x}$ from $\dot{y}$, where $\dot{y}$ consists of the $re(y)$ and $im(y)$, we can simply predict $re(x)$ and $im(x)$ by $re(y)$ and $im(y)$, respectively. Nevertheless, the correlation between real and imaginary will be ignored. With this consideration, we stack $re(y)$ and $im(y)$ together and treat it as an image $\tilde{Y} = [re(y)\,;\, im(y)]$ with the size $m \times 2$, which we call the RI image. Thus, we aim to predict the RI image $\tilde{X} = [re(x)\,;\, im(x)]$ of x from $\tilde{Y}$.

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 $\tilde{X}$. In [3], we proposed an efficient learning-based method called QISTA-Net-n by solving the ℓ_q-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 $\tilde{X}$.

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

$$\Psi^t(X) = \mathcal{C}_3^t\left(\mathrm{R}\left(\mathcal{C}_2^t\left(\mathrm{R}\left(\mathcal{C}_1^t\left(\mathcal{C}_0^t\left(X\right)\right)\right)\right)\right)\right),$$
$$\check{\Psi}^t(X) = \mathcal{C}_7^t\left(\mathcal{C}_6^t\left(\mathrm{R}\left(\mathcal{C}_5^t\left(\mathrm{R}\left(\mathcal{C}_4^t\left(X\right)\right)\right)\right)\right)\right),$$

where $\mathcal{C}_i^t$, $i = 0, 1, \cdots, 7$, are convolution operators, R is the Rectified Linear Unit (ReLU) function, and both $\mathcal{A}$ and $\mathcal{B}$ are

Algorithm 1 QISTA-Net-n for Phase Estimation

Require: $\tilde{Y} \in \mathbb{R}^{m\times 2}$
Ensure: $\bar{x} \in \mathbb{R}^n$
1: Initial $\tilde{X}^0 = \mathcal{B}\tilde{Y}$
2: **for** $t = 1$ to T **do**
3: $\quad \tilde{R}^t = \tilde{X}^{t-1} + \beta^t \mathcal{B}\left(\tilde{Y} - \mathcal{A}\tilde{X}^{t-1}\right)$;
4: $\quad \gamma_i^t = \frac{\lambda^t}{\left(|\Psi^t(\tilde{R}^t)_i| + \varepsilon_i\right)^{1-q}},\ \forall\, i \in [1:n]$;
5: $\quad \tilde{X}^t = \tilde{R}^t + \alpha^t \check{\Psi}^t\left(\eta(\Psi^t(\tilde{R}^t); \gamma^t) - \Psi^t(\tilde{R}^t)\right)$;
6: **end for**
7: $[re(x)\ ;\ im(x)] = \tilde{X}^T$
8: $\dot{x} = re(x) + i \cdot im(x) \in \mathbb{C}^n$
9: Calculate the phase $\bar{x}$ from $\dot{x}$

fully connected layers, and $\eta(\cdot;\cdot)$ is a component-wise soft-thresholding operator. The learning parameters of QISTA-Net-n are $\left\{\beta^t, \mathcal{B}, \mathcal{A}, \lambda^t, \alpha^t, \mathcal{C}_0^t, \mathcal{C}_1^t, \cdots, \mathcal{C}_7^t\right\}_{t=1}^T$. Since the size of $\tilde{Y}$ and $\tilde{X}$ are $m\times 2$ and $n\times 2$, respectively, all the kernels of the convolution operators are set to 2×2, and are operated with the stride of 1. The loss function here is MSE-loss plus an auxiliary loss

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

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

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

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

Nevertheless, as we have described previously, in order to further improve the quality of $\dot{x}$, we reserve the reconstructed phase only and improve the quality of magnitude by solving the non-convex ℓ_q-norm minimization problem in (3), with $\Psi = I_n$ 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 $\hat{y}$ and $\hat{x}$ be the magnitude of y and x, respectively. In this section, we predict $\hat{x}$ of the high-resolution signal by solving the non-convex ℓ_q-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 Ψ.

Specifically, we modify Algorithm 1 appropriately with $\Psi = I_n$ (fix both learning parameters Ψ^t and $\check{\Psi}^t$ as an identity matrix) and fix the learning parameter α^t as a constant 1. Moreover, the input and output are the magnitudes $\hat{y}$ and $\hat{x}$, 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 $\left\{\beta^t, \mathcal{B}, \lambda^t\right\}_{t=1}^T$ be the learning parameters. Note that, the matrix $\mathcal{B}$ is the learning parameter, whereas, as suggested in [3], A is a deterministic downsampling

Algorithm 2 QISTA-Net-m for Magnitude Reconstruction

Require: $\hat{y} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$
Ensure: $\hat{x} \in \mathbb{R}^n$
1: Initial $\hat{x}^{\text{init}} = \mathcal{B}\hat{y}$
2: **for** $t = 1$ to T **do**
3: $\quad \hat{r}^t = \hat{x}^{t-1} + \beta^t \mathcal{B}\left(\hat{y} - A\hat{x}^{t-1}\right)$;
4: $\quad \hat{x}_i^t = \eta\left(\hat{r}_i^t; \frac{\lambda^t}{\left(|\hat{r}_i^t| + \varepsilon_i\right)^{1-q}}\right),\ \forall\, i \in [1:n]$
5: **end for**

operator[1] to reduce the training time without loss of reconstruction performance.

The loss function here is MSE-loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\left\|\hat{x}_0 - \hat{x}^T\right\|_2^2.$$

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 $\dot{x}$. Since the magnitude $\hat{x}$ is always an approximately sparse signal, we only keep the phase of $\dot{x}$, 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 MfNet [1]. The reconstruction quality was measured in three metrics: the SNR (signal-to-noise ratio) (SNR $= 10\log_{10}\left(\frac{\|x_0\|_2^2}{\|x^* - x_0\|_2^2}\right)$), 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 $\varepsilon = 0.1 \cdot \mathbf{1}_{n\times 2}$, where $\mathbf{1}_{n\times 2} \in \mathbb{R}^{n\times 2}$ is a matrix with each component being equal to 1, and $q = 0.05$. The learning parameters of QISTA-Net-n were initialized as $\lambda^t = 10^{-5}$, $\beta^t = 10^{-1}$, $\alpha^t = 1$, and both $\mathcal{A}$ and $\mathcal{B}$, and all of $\mathcal{C}_i$, $i = 0, 1, \cdots, 7$ were initialized by xavier initializer [18], for all $1 \le t \le T$. 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 $\varepsilon = 0.1 \cdot \mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector with each component

[1] If the sampling rate of y is half of that of x, then $A_{i,2i-1}$ is 1 for all i, 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.*

	$\mathcal{C}_0$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$	$\mathcal{C}_7$
Input feature	1	32	32	32	32	32	32	32
Output feature	32	32	32	32	32	32	32	1

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.*

Method	SNR	PESQ	STOI
QISTA-Net-Audio	31.44	4.31	0.9958
MfNet	24.50	3.76	-
MfNet+P	24.55	3.88	-
MfNet+A	24.77	3.80	-
MfNet+C	24.70	3.82	-

being equal to 1, and $q = 0.05$. The learning parameters of QISTA-Net-m were initialized as $\lambda^t = 10^{-5}$, $\beta^t = 10^{-1}$, 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 $\dot{x}$, where the magnitude $\hat{x}$ of $\dot{x}$ achieves the reconstruction performance SNR=31.61 dB. On the other hand, the magnitude $\hat{x}$ obtained in Sec. 3.2, results in the performance SNR=33.98 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 predicated 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.

[3]The corpus is provided by Valentini *et al.*, which is publicly available. https://datashare.is.ed.ac.uk/handle/10283/2791

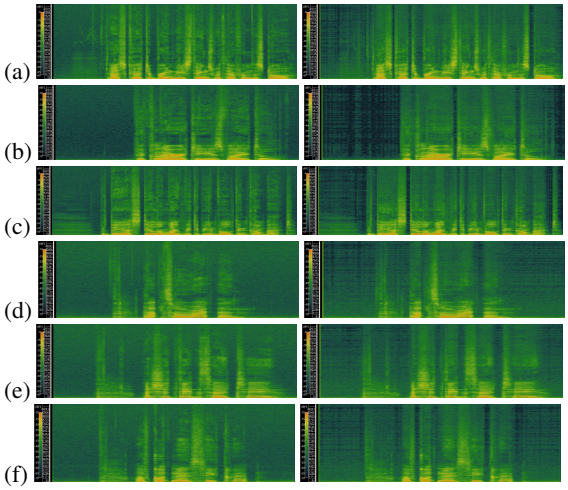

Figure 2: *Left: The spectrogram of the ground-truth. Right: The spectrogram of our prediction.*

6. References

[1] X. Hao, C. Xu, N. Hou, L. Xie, E. S. Chng, and H. Li, "Time-domain neural network approach for speech bandwidth extension," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2020.

[2] G. X. Lin and C. S. Lu, "QISTA-Net: DNN architecture to solve ℓ_q-norm minimization problem," *In Proc. IEEE Int. Workshop on Mach. Learn. Signal Process. (MLSP)*, Sep. 2020.

[3] G. X. Lin, S. W. Hu, and C. S. Lu, "QISTA-Net: DNN architecture to solve ℓ_q-norm minimization problem and image compressed sensing," *arXiv:2010.11363*, Oct. 2020.

[4] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Apr. 1979.

[5] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2005.

[6] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2001.

[7] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech," *In Proc. INTERSPEECH*, Aug. 2011.

[8] Y. Wang, S. Zhao, Y. Yu, and J. Kuang, "Speech bandwidth extension based on GMM and clustering method," *In Proc. IEEE Int. Conf. Commun. Syst. and Netw. Technol. (CSNT)*, Apr. 2015.

[9] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," *In Proc. IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2014.

[10] M. A. T. Turan and E. Erzin, "Synchronous overlap and add of spectra for enhancement of excitation in artificial bandwidth extension of speech," *In Proc. INTERSPEECH*, Sep. 2015.

[11] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015.

[12] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural nets," *In Proc. Int. Conf. Learn. Representations (ICLR)*, Apr. 2017.

[13] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2018.

[14] H. Wang and D. Wang, "Time-frequency loss for CNN based speech super-resolution," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2020.

[15] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE J. Sel. Topics Signal Process.*, May 2019.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ), a new method for speech quality assessment of telephone networks and codecs," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2001.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *In Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2010.

[18] X. Glorot and Y. Bengio, "Learning a deep convolutional network for image super-resolution," *In Proc. Mach. Learn. Res. (PMLR)*, vol. 9, pp. 249–256, 2010.

[19] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," *In Proc. ISCA Speech Synthesis Workshop (SSW)*, Sep. 2016.