Towards Robust Speech Super-Resolution

Heming Wang, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Speech super-resolution (SR) aims to increase the sampling rate of a given speech signal by generating high-frequency components. This paper proposes a convolutional neural network (CNN) based SR model that takes advantage of information from both time and frequency domains. Specifically, the proposed CNN is a time-domain model that takes the raw waveform of low-resolution speech as the input, and outputs an estimate of the corresponding high-resolution waveform. During the training stage, we employ a cross-domain loss to optimize the network. We compare our model with several deep neural network (DNN) based SR models, and experiments show that our model outperforms existing models. Furthermore, the robustness of DNN-based models is investigated, in particular regarding microphone channels and downsampling schemes, which have a major impact on the performance of DNN-based SR models. By training with proper datasets and preprocessing, we improve the generalization capability for untrained microphone channels and unknown downsampling schemes.**

***Index Terms*—Speech super-resolution, bandwidth extension, convolutional neural network, robust speech super-resolution.**

I. Introduction

FOR bandwidth-limited signal transmission and equipment such as bluetooth and telephony, only low-frequency components of speech signals are preserved. Narrow bandwidth or low resolution degrades speech quality, or even intelligibility. Speech super-resolution (SR) aims to increase the waveform resolution of such speech by generating high-frequency components. It is also referred to as speech bandwidth extension (BWE) viewed from the spectral perspective. SR or BWE is beneficial to many speech processing tasks, such as text-to-speech synthesis [32], automatic speech recognition [3], [23] and speech enhancement [8].

Early studies in this field use signal processing methods. A source-filter model is introduced to extend the bandwidth by finding the high-frequency residual signal and the spectral envelope individually [8], [29]. To predict upper band spectral envelops from narrowband speech, codebook methods are employed to map narrowband speech representations and the corresponding upperband envelopes [7], [42], [47]. Gaussian mixture models (GMMs), and joint hidden Markov model and GMM have been exploited for estimation in codebook mapping [4], [33], [46]. These statistical methods yield better results compared with deterministic mapping, but tend to yield overly smoothed spectra [27]. The introduction of deep learning advances many areas of signal processing. For speech SR, deep neural networks (DNNs) have demonstrated superior performance. DNN studies include multiplayer perceptrons [6], [21], [41] to predict vocal tract filter parameters or the original log-power spectrum [23], [28], recurrent neural networks with long-short-term memory [15], convolutional neural networks (CNNs) [14], speech waveform synthesizers like WaveNet [16], [35], [51] and SampleRNN [26], [31] for conditional speech generation, and generative adversarial networks [10], [17], [24]. A more detailed summary of the related work is given in Section II. In general, deep learning based approaches show better performance compared with statistical approaches.

Current DNN-based SR models operate in matched settings, where recording conditions are fixed, and high-resolution (HR)/low-resolution (LR) pairs are obtained in the same way during training and testing. We observe that these models fail to generalize to different experimental settings. Specifically, the performance of such models degrades on speech databases with different recordings, or on LR signals generated by a different downsampling scheme. Therefore, it is important to investigate the robustness of SR models to such factors, and achieve robust SR. In this paper, we address speech SR in the time domain by employing a CNN model to reconstruct speech with higher sampling rates. We propose a cross-domain loss, which not only produces excellent performance in terms of signal-to-noise ratio (SNR) and log-spectral distance (LSD) [13], but also removes unwanted artifacts in generated speech. Additionally, the proposed CNN can operate in real-time.

Another contribution of this paper is an examination of the sensitivity of DNN-based SR models to different recording channels and downsampling schemes. By employing different microphone impulse responses and performing cross-corpus experiments, we demonstrate that microphone channel is a major factor that affects SR performance. We also note that models trained with different downsampling schemes exhibit different levels of performance, and do not generalize to another way of downsampling. We show how to improve robustness to these variations with a proper training strategy. As a result, our model generalizes to untrained speech corpora and data acquisition schemes. A preliminary version of this study was published in ICASSP 2020 [50], but the present paper goes far beyond the earlier version. The preliminary version adopts a time-frequency loss, and this version proposes a novel cross-domain loss that

Manuscript received September 17, 2020; revised December 13, 2020 and January 17, 2021; accepted January 18, 2021. Date of publication January 25, 2021; date of current version June 16, 2021. This work was supported in part by an NIDCD (R01 DC012048) Grant and the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Roland Badeau. *(Corresponding author: Heming Wang.)*

Heming Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wang.11401@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering, and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2021.3054302

further improves SR performance. Robustness is not addressed in [50], but is a major topic in the current investigation. In addition, more evaluations and comparisons are provided in this paper.

The rest of the paper is organized as follows. In Section II, we review related prior studies, which also serve as baselines in our comparisons. In Section III we present the network design and loss functions. Section IV provides experimental results and comparisons. In Section V, the robustness of the model is examined in terms of microphone channels and downsampling schemes. Section VI concludes the paper.

II. Related Work

Li and Lee utilize a DNN to address speech BWE [23]. They employ log-power spectrum (LPS) as features, and predict the wideband LPS from the narrowband LPS. A DNN is pretrained as a restricted Boltzmann machine and optimized by a mean-squared error (MSE) loss between the predicted features and target features. The phase of the upperband is produced by flipping the narrowband phase and adding a negative sign. In another BWE study, Abel and Fingscheidt employ a DNN to estimate the lower-dimensional cepstral representation of spectral envelopes [2]. The upperband phase is obtained by copying the phase of narrowband spectrum. Experiments show that these DNN-based BWE approaches yield better results than traditional approaches.

Inspired by the successful application of CNNs in image SR, Kuleshov *et al.* introduce AudioUNet [22], which is adapted from an image domain network [9], [44]. This is an end-to-end autoencoder model that takes the raw waveform as input and outputs the predicted SR waveform. This method operates in the time domain and thus does not need to estimate the phase separately. It outperforms conventional approaches and considerably improves the quality of reconstructed speech.

While the above studies show promising results, they only focus on one representation domain. Lim *et al.* propose a time-frequency network (TFNet) that incorporates information from both time and frequency domains [25]. TFNet is built from two AudioUNets, where one is trained with LR and HR waveforms and the other is trained with the magnitudes of short-time Fourier transform (STFT). These two networks are jointly optimized and a spectral fusion layer is utilized to combine the output of two branches. The STFT magnitude is obtained by combining estimates from the two branches, and an estimate of the STFT phase is obtained through the time branch. Experiments show that TFNet successfully leverages the cross-domain information and outperforms AudioUNet.

III. Model Description

Suppose we are given an LR speech segment s_{lr} at a sampling rate fs_{lr}. The goal of speech SR is to reconstruct a speech signal s_{hr} at a sampling rate fs_{hr}, such that $fs_{hr} > fs_{lr}$, i.e. restoring high-frequency components. The ratio fs_{hr}/fs_{lr} is referred to as the *downsampling factor*, which is typically an integer 2 or 4. For instance, LR signals may be standard telephone speech signals sampled at 8 kHz, and HR signals are 16 kHz. To reconstruct the HR signal, we learn a DNN model f that takes the LR signal s_{lr} as the input. With parameters of the model denoted as θ, the model produces the corresponding reconstruction s_{sr}:

$$s_{sr} = f(\theta, s_{lr}) \tag{1}$$

Fig. 1 depicts the overall pipeline of the proposed SR model. We first upsample LR signals to the desired sampling rate using cubic spline interpolation [11]. Then the upsampled signal and HR signal are fed into our model as the input and desired output, respectively. The proposed neural network is based on the autoencoder CNN (AECNN) by Pandey and Wang [37]. AECNN is a fully convolutional network composed of a series of encoder and decoder blocks, with skip connections to better reconstruct encoder outputs. Parametric rectified linear units (PReLUs) are used to each layer, except for the last layer which is linear. Dropout is employed every three layers, as illustrated in Fig. 1. Our CNN takes as the input upsampled segments, each having 2048 samples and with a 50% overlap between consecutive frames, and outputs the corresponding segment HR estimates. The network is trained with raw waveforms and minimizes a loss derived from STFT. One change we introduce to AECNN is to decoding blocks, where we replace transposed convolution layers with subpixel layers. A subpixel layer, proposed by Shi *et al.* [44], is an upscaling layer implemented by convolution operations. It has been reported in [34] that using transposed convolution for upsampling layers can lead to artifacts in the outputs of image SR, often referred to as checkerboard artifacts. By applying subpixel layers these artifacts are alleviated for image SR. We observe that employing subpixel layers accelerates the training progress and slightly improves speech SR results. Note that subpixel layers are also used in AudioUNet in upsampling blocks [22].

Our model is optimized with a cross-domain loss. For a real valued signal s of length N in the time domain, the discrete Fourier transform (DFT) amounts to multiplying by a complex-valued matrix D,

$$S = Ds \tag{2}$$

where D is a $N \times N$ matrix, and S is the DFT of s. We express the complex formula in (2) in real and imaginary parts. Extracting the real and imaginary part from matrix D, we obtain D_r and D_i, respectively. Equation (2) can be expressed as:

$$S_r = D_r s \tag{3}$$

$$S_i = D_i s \tag{4}$$

where S_r and S_i represent the real and imaginary part of the DFT in real values. Then $S = (D_r + iD_i)s = S_r + iS_i$, and i denotes the imaginary unit. The DFT magnitude can be expressed as $S_{mag} = \sqrt{S_r^2 + S_i^2}$.

The first loss explored is a frequency-domain loss defined as the mean absolute error (MAE) between two STFT magnitudes,

$$L_F(\hat{S}, S) = \frac{1}{M}\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}|\hat{S}_{mag}(m,k) - S_{mag}(m,k)| \tag{5}$$

Here $\hat{S}, S$ denote the STFT of SR and HR segments, respectively, and m, k index time and frequency, respectively. This loss is denoted as L_F.

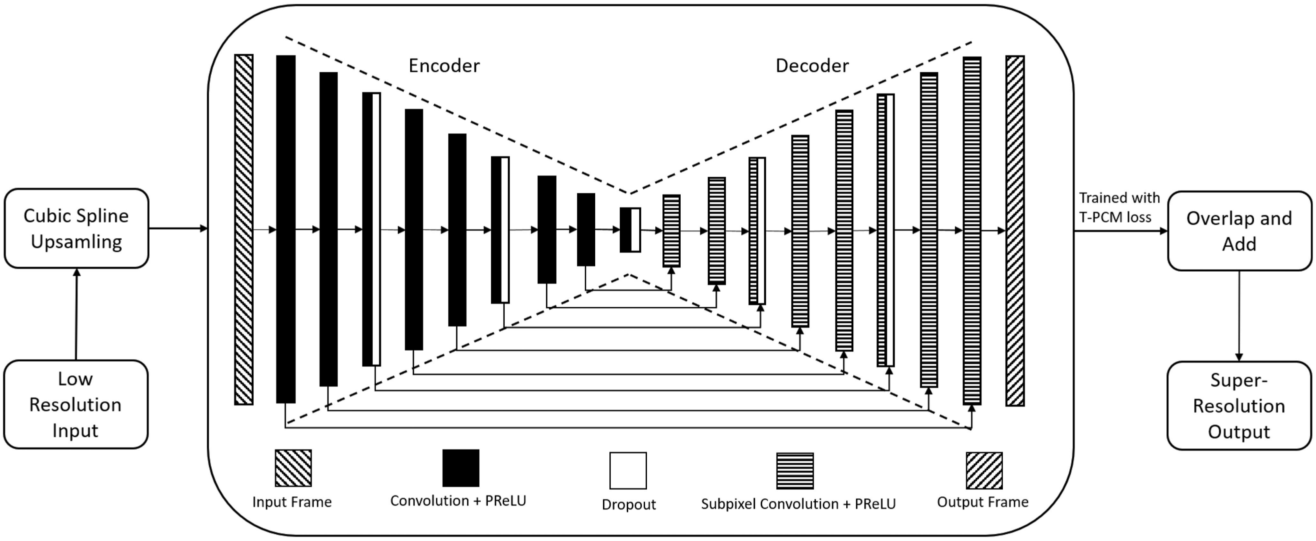


Fig. 1. Illustration of the super-resolution pipeline and our AECNN network structure.

L_F only optimizes STFT magnitudes and the phase estimation of SR signal is ignored. L_F does not perform well on time-domain metrics such as SNR and generates an unwanted artifact in reconstructed speech. To address the phase estimation, we investigate three other loss functions. The first one adds a term that measures real and imaginary parts of STFT, denoted as $L_{RI\text{-}MAG}$:

$$L_{RI\text{-}MAG}(\hat{S}, S) = L_F + L_{RI} \tag{6}$$

$$L_{RI}(\hat{S}, S) = \frac{1}{M}\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}(|\hat{S}_r(m,k) - S_r(m,k)| + |\hat{S}_i(m,k) - S_i(m,k)|) \tag{7}$$

The second one incorporates a phase constraint to combine a time-domain loss with a frequency-domain loss, and uses time-domain estimation to compensate for phase,

$$L_{TF} = \alpha L_T + (1-\alpha)L_F \tag{8}$$

$$L_T(\hat{s}, s) = \frac{1}{N}\sum_{n=1}^{N}|\hat{s}(n) - s(n)| \tag{9}$$

where $\hat{s}$, s are the time-domain SR and HR signals of length N, respectively. We use a coefficient α to combine two loss terms L_T and L_F. This loss is called time-frequency loss L_{TF}, and the value of α is set to 0.85 to balance the magnitude difference between frequency and time losses [50].

L_{TF} and $L_{RI\text{-}MAG}$ improve SNR scores and alleviate artifacts phenomenon; however, the artifacts are not completely removed by either of them. The third loss function we investigate is inspired by a recent study of similar artifacts in speech enhancement [38], for which we take the STFT magnitudes of both SR signals and residual signals into account. The residual signal, denoted as s_{re}, is obtained by subtracting an upsampled signal (denoted as s_{up}) from its corresponding HR signal. In the frequency domain, we have the following relationship,

$$S_{hr} = S_{up} + S_{re} \tag{10}$$

where S_{up} is a spectral vector in the complex domain. If we only optimize the magnitude of S_{hr}, there can be infinite candidates of S_{hr} that satisfy Equation (10). If we optimize the magnitudes of S_{re} and S_{hr} simultaneously, the infinite number of candidates for S_{hr} is reduced to two due to a triangular magnitude relation (see [38]). Therefore, optimizing both residual and SR signals imposes a phase constraint on CNN optimization.

$$L_{PCM}(\hat{S}, S, S_{up}) = L_F(\hat{S}, S) + L_F(\hat{S} - S_{up}, S - S_{up}) \tag{11}$$

This loss function is denoted as PCM (phase constrained magnitude) [38]. L_{PCM} takes the STFT of upsampled signal S_{up} as an additional input, and is composed of two frequency loss functions for speech and its corresponding residual. L_{PCM} can effectively remove unwanted artifacts. Worth noting are previous studies showing that it is simpler for DNN optimization with residual terms added [18], [45].

The SNR performance of L_{PCM} is improved compared with L_F, but still not optimal compared with L_{TF}. To improve time-domain performance, we introduce L_T,

$$L_{T\text{-}PCM}(\hat{s}, s, \hat{S}, S, S_{up}) = \beta L_T(\hat{s}, s) + (1-\beta)L_{PCM} \tag{12}$$

This loss is denoted as T-PCM, and we use a coefficient β, set to 0.6 in this study, to combine loss terms from the two domains.

Fig. 2 illustrates the calculation of $L_{T\text{-}PCM}$ on a segment of 2048 samples, which corresponds to a 128 ms long segment for 16 kHz sampling frequency. We apply the overlap and add (OLA) method when calculating the loss within each segment, because segments of 128 ms are too long to satisfy the stationarity assumption for short-time signal processing. For frequency-domain loss calculation, we take into consideration both signal and residual segments. Framed segments are first divided into frames of 512 samples with a frame shift of 256 samples, corresponding to an analysis window of 32 ms with

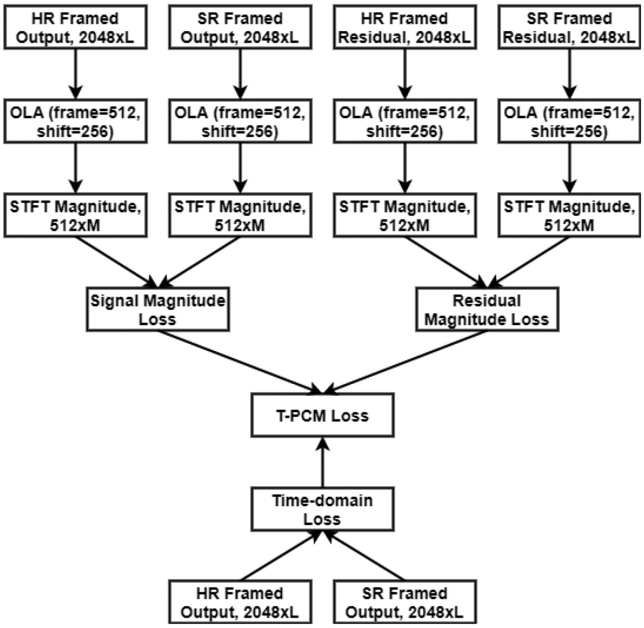


Fig. 2. Schematic diagram showing the process of calculating the T-PCM loss. L denotes the number of 2048-sample frames, and M represents the number of 512-sample frames.

a 50% frame overlap. Then we multiply these frames with a Hamming window. The STFT magnitudes for both signals and residuals are calculated on windowed frames to define the magnitude loss. For the time-domain loss calculation, we calculate the MAE for framed SR segments and HR segments. Loss terms from both domains are added to define the T-PCM loss.

IV. Evaluation and Comparison

A. Experimental Setup

We evaluate our model on two datasets, TIMIT [12] and VCTK [48]. TIMIT is a standard corpus containing speech recordings from 630 speakers with a 16 kHz sampling rate. From the training part of the corpus, we choose 4620 utterances as the training dataset, and 1153 utterances as the validation set. We select a subset of the TIMIT core test set for test purposes, which consists of 192 utterances from 24 speakers that are not included in the training and validation datasets, thus enabling us to assess the generalization ability to untrained speakers. The VCTK corpus contains 44 hours of speech recordings from 108 speakers with a 48 kHz sampling rate. For a fair comparison, we follow the task design of [22] and [25]. The first task uses speech data of one specific speaker (speaker p225). The other task is multi-speaker, for which we train and test using the whole corpus. Following the description in [25], we split the data to 88%, 6%, and 6% for training, validation and testing purposes. We also make sure that there is no speaker overlap between training, validation and testing for the multi-speaker task. LR signals are obtained by first applying a Chebynov type I low-pass filter and then subsampling. A silence filter that discards samples below an energy threshold of 0.05 is performed to stabilize training and ensure faster convergence.

For preprocessing, all the utterances are first resampled to a 16 kHz sampling rate if their original sampling rate is higher than 16 kHz. Then each utterance is normalized to zero mean and unit variance. Note this is different from our preliminary version [50], where we apply a uniform normalization (rescaling each utterance to the range -1.0 to 1.0). We observe that mean and variance normalization (MVN) improves cross-corpus generalization. We divide each utterance into frames of 2048 samples (128 ms), and with an overlap of 1024 samples between consecutive frames. LR inputs are upsampled using cubic spline interpolation so that input and desired output signals for our CNN have the same length. During the reconstruction stage, we combine consecutive frames using the OLA method.

We evaluate the SR performance with three objective metrics: SNR, LSD, and PESQ for wideband speech [5]. SNR is a time-domain metric, defined as,

$$\mathrm{SNR}(\hat{s}, s) = 10 \log_{10} \frac{\sum_{n=1}^{N} s(n)^2}{\sum_{n=1}^{N} [\hat{s}(n) - s(n)]^2} \tag{13}$$

LSD, a frequency-domain metric, measures the logarithmic distance between two magnitude spectra in dB.

$$\mathrm{LSD}(\hat{S}, S) = \frac{1}{M} \sum_{m=1}^{M} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left[\log_{10} \frac{\hat{S}_{mag}(m,k)^2}{S_{mag}(m,k)^2} \right]^2} \tag{14}$$

When the two spectra are the same, LSD will be 0 dB, the smallest possible distance. Given a reference and a degraded audio signal, PESQ for wideband speech is a standard metric of perceptual speech quality with a value range from 1.04 to 4.64. Higher PESQ indicates better listening quality.

B. Comparison Models

We compare with four other deep SR models described in Section II. The first baseline is the spectral domain model by Li and Lee [23], referred to as DNN-BWE. This model takes 9 frames (4 preceding and 4 succeeding) as the input and predicts the current STFT frame. We follow the implementation details in the original work, which uses 256-sample frames with a frame shift of 128 samples, and a 4-layer DNN with 2048 hidden units for training. The second baseline is by Abel and Fingscheidt (denoted as DNN-Cepstral) [1], [2], and we follow the original description by using a 256-sample frame length and a 128-sample frame shift for narrowband speech. Their DNN has 3 hidden layers with 256 units in each. The third comparison model is the waveform-based model proposed by Kuleshov *et al.* [22], denoted as AudioUNet. Following their default setting in the publicly available code,[1] we set the number of filters in the encoding layers to 128, 256, 512, and 512, and the filter size to 65, 33, 17, and 9. The setting for decoder layers is similar except for the reverse order and doubled filter size. AudioUNet operates on 2048-sample segments, with a 50% overlap between consecutive segments. The last baseline is TFNet by Lim *et al.* [25], which consists of two AudioUNets. But for both branches, the number of filters is halved to reduce

[1] https://github.com/kuleshov/audio-super-res

TABLE I
EXPERIMENTAL RESULTS FOR SR MODELS EVALUATED ON TIMIT

	SNR	LSD	PESQ
Spline	15.48	2.27	2.56
DNN-BWE	17.05	1.05	2.78
DNN-Cepstral	16.27	0.97	2.79
AudioUNet	18.59	0.89	2.94
TFNet	18.91	0.87	3.12
AECNN	19.63	0.72	3.59
Proposed	**20.18**	**0.72**	**3.65**

TABLE II
EXPERIMENTAL RESULTS FOR SR MODELS EVALUATED ON VCTK WITH DOWNSAMPLING FACTOR OF 2 AND 4

		VCTK_S			VCTK_M		
Model	R	SNR	LSD	PESQ	SNR	LSD	PESQ
Spline	2	19.07	1.99	3.84	18.89	2.08	3.53
DNN-BWE	2	19.04	1.40	3.85	18.80	1.38	3.56
DNN-Cepstral	2	19.89	1.25	3.85	19.09	1.34	3.59
AudioUNet	2	20.82	1.36	3.90	19.94	1.32	3.68
TFNet	2	21.11	1.24	3.91	19.84	0.99	3.72
Proposed	2	**22.44**	**0.94**	**4.17**	**22.08**	**0.88**	**3.91**
Spline	4	15.33	3.13	3.07	13.42	2.99	3.13
DNN-BWE	4	15.30	1.47	3.27	13.53	1.38	3.24
DNN-Cepstral	4	15.47	1.44	3.28	13.87	1.36	3.25
AudioUNet	4	17.29	1.41	3.40	16.65	1.40	3.39
TFNet	4	18.35	1.33	3.49	17.32	1.22	3.48
Proposed	4	**18.86**	**0.94**	**3.51**	**18.13**	**0.95**	**3.64**

TABLE III
COMPARISON OF VARIOUS LOSS FUNCTIONS ON THE TIMIT DATASET

Loss	SNR	LSD	PESQ
MAE	20.05	0.94	3.22
MSE	19.98	0.89	3.23
F	10.88	0.72	3.57
RI	5.48	0.72	3.54
TF	**20.27**	0.76	3.54
RI-MAG	18.34	0.72	3.51
PCM	15.88	**0.71**	3.63
T-PCM	20.18	0.72	**3.65**

parameters. To be consistent with the settings in their paper, a frame size of 8192 samples with a 75% overlap is employed for the network input.

For our proposed network, the kernel size is set to 11 for all convolutional layers. The number of channels for each encoding layer is set to 64, 64, 64, 128, 128, 128, 256, 256, and 256, and the decoding layers have the same numbers except in the reverse order. Our network is trained with a mini-batch size of 32 for 100 epochs, and optimized with the T-PCM loss. Dropout ratio is set to 0.2 for dropout layers. The Adam optimizer [20] with a learning rate of 0.0003 is used for stochastic gradient descent based optimization. The learning rate is halved if the loss has not improved for 3 consecutive epochs on the validation set. We add an early stopping criterion such that the training process stops if the validation loss has not improved for 6 successive epochs. For other deep SR baselines, the training setup follows their original descriptions.

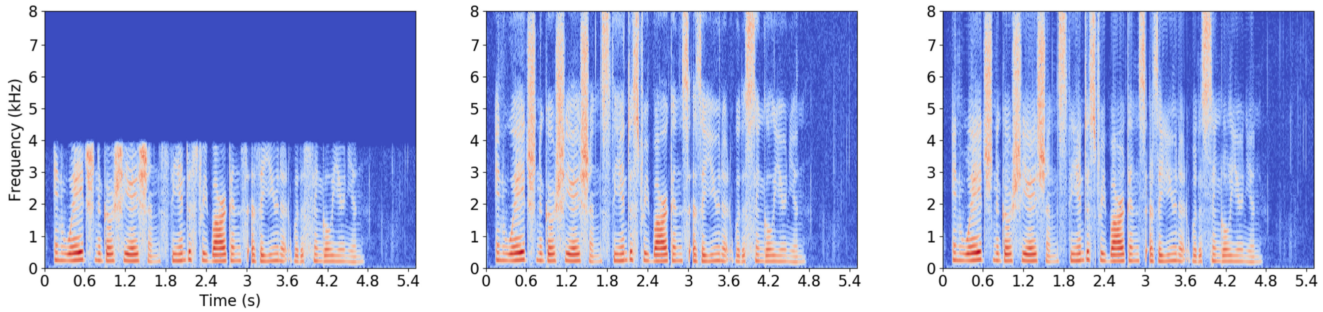
C. Results and Comparisons

Table I presents the results of our proposed CNN, as well as the other baselines, on the TIMIT dataset which is downsampled to 8 kHz to create LR signals. The first row corresponds to the objective scores by applying cubic spline interpolation, which is a conventional signal processing baseline and outputs limited but stable improvement for SR. Our model improves over the cubic spline method by 4.7 dB in terms of SNR, and cuts LSD by 68.3%. PESQ is improved by nearly 1.1, which is a large improvement for speech quality. Compared with the other four deep learning baselines, we see consistent improvement over all three metrics. Specifically, compared with the best-performing baseline of TFNet, our model improves SNR by around 1.3 dB, LSD by 17.2%, and PESQ by 0.5. Also our model slightly improves over the AECNN baseline in terms of SNR and PESQ.

The VCTK results are reported in Table II. The ratio R is the downsampling factor, where $R = 2$ implies upsampling from 8 kHz to 16 kHz, and $R = 4$ represents upsampling from 4 kHz to 16 kHz. VCTK_S represents the single-speaker task, and VCTK_M represents the multi-speaker task. As shown in Table II, our model shows superior performance for both downsampling factors. For $R = 2$, compared with the spline baseline, our model improves SNR by over 3.0 dB, and cuts LSD to below 1.0 for both tasks. We also see a consistent improvement in PESQ. For $R = 4$, we observe similar improvements for all three metrics over the spline method. For instance, for the multi-speaker task, we see SNR boosted by around 4.7 dB, LSD cut to below 1.0, and PESQ increased by 0.5. Our model also consistently improves over other deep SR baselines. Compared with the strongest baseline of TFNet, our proposed model yields better SNR, LSD and PESQ scores for both tasks under the two downsampling factors. For the multi-speaker task under $R = 2$, for example, SNR is improved by around 2.2 dB, LSD by 11.1%, and PESQ by around 0.2.

To examine the advantage of our proposed T-PCM loss function, we compare different loss functions on the TIMIT corpus. As shown in Table III, the time-domain losses L_{MSE} and L_{MAE} give good SNR values, but not LSD performance. By introducing spectral loss terms (L_F, L_{RI}, $L_{RI\text{-}MAG}$, L_{TF}, L_{PCM} and $L_{T\text{-}PCM}$), the PESQ values increase, indicating that spectral loss leads to better speech quality. Although L_F and L_{RI} have good LSD scores, very low SNR scores are observed. Although L_{RI} manifests both phase and magnitude, the real and imaginary parts are related as the cosine and sine, respectively, of the phase multiplied by the magnitude [52]. By including a magnitude term, $L_{RI\text{-}MAG}$ is found to improve both phase and magnitude estimation over L_{RI}. Additionally, due to poor phase estimation, for L_F and L_{RI} we observe unwanted artifacts in reconstructed speech for certain signals obtained by decimating downsampling (see Section V-B). The L_{TF} and $L_{RI\text{-}MAG}$ losses show balanced performance in terms of SNR, LSD, and PESQ, and artifacts are alleviated but still exist. For L_{PCM} and $L_{T\text{-}PCM}$, we do not hear the artifact in reconstructed speech.

Table IV compares the numbers of trainable parameters of our proposed model and the other baselines, which show that our model achieves strong performance with a relatively small number of parameters. Fig. 3 illustrates the output of our SR model on a sample TIMIT utterance ("In wage negotiations, the industry bargains as a unit with a single union"). Comparing



(a) LR Spectrogram (b) HR Spectrogram (c) SR Spectrogram

Fig. 3. Spectrograms of SR results: (a). LR input, (b). Ground truth HR signal and (c). Reconstructed SR signal.

TABLE IV
NUMBER OF TRAINABLE PARAMETERS FOR DIFFERENT SR MODELS, WHERE M INDICATES MILLION

	Number of Parameters
DNN-BWE	11.2 M
DNN-Cepstral	0.25 M
AudioUNet	70.9 M
TFNet	58.8 M
Proposed	10.2 M

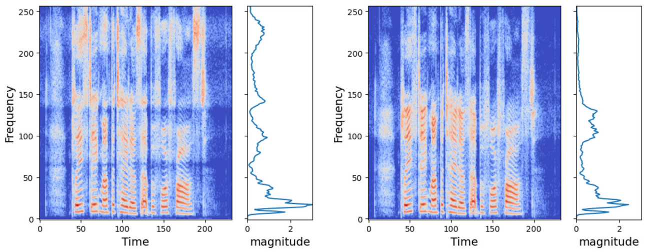
the spectrograms we can observe that missing high-frequency components in the LR spectrogram are recovered well by our model.

V. ON ROBUSTNESS

Although recent SR studies show promising performance under matched experimental settings, whether models trained on one corpus can be generalized to other corpora and whether different downsampling schemes affect the robustness of the model are yet to be investigated. Real-world applications often require SR models to be insensitive to such factors. This section examines the important issue of robustness.

A. Corpus Channels

A speech corpus typically contains speech signals recorded in a fixed environment. Taking TIMIT for example, all recordings are collected in the same anechoic room with a single microphone. Although this setting guarantees a uniform quality, it likely introduces signal characteristics unique to the specific experimental setting, impeding the generalizability of the trained models on one corpus. The recording characteristic of a corpus is referred to as *corpus channel* [39]. To validate this analysis, we randomly choose two microphone impulse response (MIR) functions from Vintage Mics,[2] and convolve them with TIMIT utterances. As shown in Fig. 4, the energy distributions over frequency are distinct for the same utterance when convolved with different MIRs. Table V further illustrates channel effects, where our model trained with original TIMIT utterances shows degraded performance when tested on utterances convolved with the two MIRs. The last row is the average result of testing separately on utterances convolved with 20 randomly picked MIRs.



(a) MIR 1 Spectrogram (b) MIR 2 Spectrogram

Fig. 4. Spectrograms of an utterance convolved with two different MIRs, together with energy distributions along frequency.

TABLE V
MODEL TRAINED ON ORIGINAL TIMIT UTTERANCES TESTED ON DATA CONVOLVED WITH DIFFERENT MIRS

Model	SNR	LSD	PESQ
Spline	15.48	2.27	2.56
Original	20.18	0.72	3.65
Test on MIR1	12.53	1.38	2.41
Test on MIR2	13.75	0.87	2.99
Average of 20 MIRs	14.76	1.01	2.82

To systematically investigate channel effects, we conduct cross-corpus experiments on four different databases using the deep SR models evaluated in the previous section. Experimental settings and DNN architectures are as described in Section IV-A and IV-B. The four datasets are TIMIT [12], Wall Street Journal (WSJ) [40], LIBRIspeech [36], and IEEE [19]. For the WSJ corpus, speakers read Wall Street Journal articles plus spontaneous dictations. Two sets of microphones are utilized for the recordings: a close-talking Sennheiser HMD414 and a secondary microphone which may vary. WSJ recordings are sampled at a 16 kHz sampling rate, and contain a small amount of background noise. For our experiments, we use 12 736 utterances from 100 speakers to train, 1206 utterances from 10 speakers to validate, and 651 utterances from 8 speakers to test. LibriSpeech consists of 1000 hours of 16 kHz English speech recordings, which are derived from reading audiobooks in the LibriVox project. The recordings are collected from volunteers

[2] https://www.audiothing.net/impulses/vintage-mics /

TABLE VI
EXPERIMENTAL RESULTS FOR CROSS-CORPUS SR USING THE FOUR BASELINES AND PROPOSED MODEL

	TIMIT			WSJ			LIBRI			IEEE		
Model / Training Dataset	SNR	LSD	PESQ	SNR	LSD	PESQ	SNR	LSD	PESQ	SNR	LSD	PESQ
Spline	15.48	2.27	2.56	15.89	2.23	2.43	12.43	1.87	2.53	20.75	1.98	2.90
DNN-BWE / TIMIT	**17.05**	**1.05**	**2.78**	14.19	1.32	2.03	11.42	1.63	1.29	18.54	1.39	1.73
DNN-BWE / WSJ	15.51	1.18	2.63	**16.71**	**1.19**	**2.73**	12.62	1.17	2.49	19.22	1.23	2.42
DNN-BWE / LIBRI	15.51	1.18	2.65	16.62	1.28	2.66	**13.21**	**1.17**	**2.55**	19.24	1.29	2.44
DNN-BWE / IEEE	15.09	1.12	1.99	12.06	1.83	1.45	9.78	1.68	1.05	**19.94**	**0.94**	**3.19**
DNN-Cepstral / TIMIT	**16.27**	**0.97**	**2.79**	16.06	0.97	2.41	13.15	1.53	2.07	18.97	1.40	2.08
DNN-Cepstral / WSJ	16.61	0.96	2.72	**17.03**	**0.92**	**2.98**	13.38	1.16	2.44	19.65	1.18	2.79
DNN-Cepstral / LIBRI	16.95	0.97	2.77	16.77	0.95	2.70	**13.55**	**1.13**	**2.77**	20.06	1.16	2.93
DNN-Cepstral / IEEE	15.80	0.97	2.72	15.49	1.03	2.64	12.42	1.22	2.36	**20.10**	**0.98**	**3.22**
AudioUNet / TIMIT	**18.59**	**0.89**	**2.94**	7.38	1.91	1.63	9.98	1.33	1.63	12.85	1.57	1.29
AudioUNet / WSJ	17.86	1.01	2.78	**18.00**	**1.00**	**3.08**	17.70	1.06	2.67	23.39	1.05	3.33
AudioUNet / LIBRI	18.11	0.97	2.76	16.90	1.02	2.72	**17.86**	**1.04**	**2.83**	22.57	1.11	3.35
AudioUNet / IEEE	15.95	0.95	2.21	14.77	1.02	2.21	14.35	1.27	2.02	**22.55**	**1.02**	**3.49**
TFNet / TIMIT	**18.91**	**0.87**	**3.12**	11.81	1.19	1.84	9.27	1.25	1.44	11.61	1.55	1.32
TFNet / WSJ	17.96	1.03	2.82	**18.57**	**0.99**	**3.22**	17.93	1.03	2.87	20.74	1.14	3.18
TFNet / LIBRI	18.60	0.90	2.78	16.96	0.99	2.76	**18.13**	**1.00**	**2.91**	21.86	1.04	3.32
TFNet / IEEE	15.81	0.94	2.27	14.80	1.02	2.27	15.07	1.07	2.09	**23.63**	**0.86**	**3.79**
Proposed / TIMIT	**20.18**	**0.72**	**3.65**	13.85	0.96	2.15	13.72	0.97	2.13	23.37	1.02	2.99
Proposed / WSJ	19.35	0.92	3.08	**21.31**	**0.74**	**3.64**	18.62	0.91	3.06	25.83	0.90	3.59
Proposed / LIBRI	19.58	0.92	3.23	17.85	0.93	2.98	**18.94**	**0.91**	**3.28**	25.86	0.99	3.60
Proposed / IEEE	17.40	0.88	2.44	16.04	0.91	2.41	15.67	0.98	2.27	**26.47**	**0.65**	**4.14**
Proposed / Mixed	19.61	0.76	3.55	20.31	0.76	3.52	18.45	0.96	3.00	26.19	0.85	3.85

across the world, so LibriSpeech has various recording environments and thus contains diverse channels. In this paper, the LibriClean subset (denoted as LIBRI) of the LibriSpeech corpus is chosen for our experiments. From the LIBRI corpus, we select 28 539 utterances for training, 2703 utterances for validation and 2620 utterances for testing, with no speaker overlap. IEEE contains 720 phonetically balanced English sentences uttered by a male speaker with a sampling frequency of 25 kHz. We randomly select 576 utterances for training, 72 utterances for validation and the remaining 72 utterances are reserved for testing purposes. For all datasets, utterances are first resampled to 16 kHz, and LR signals are generated at 8 kHz by applying the subsampling scheme.

Table VI summarizes the results of cross-corpus SR experiments. Each row represents one model trained on a particular dataset and tested on all four datasets. Each column shows the results on a specific dataset. As shown in the table, our model outperforms all the other baselines for all four datasets. As expected, the best performance is observed when training and testing are done on the same corpus. We observe that the generalization ability of each model differs with training dataset, and the models trained with datasets that contain diverse channels are more robust when testing on untrained corpora. Specifically, training with WSJ or LIBRI shows comparable objective scores when tested on untrained corpora. Training with IEEE or TIMIT, however, shows poor performance on untrained corpus channels. This observation is more obvious for AudioUNet, TFNet and the proposed network. For the spectral-domain models of DNN-BWE and DNN-Cepstral, the generalization advantage of WSJ and LIBRI mainly manifests in LSD and PESQ. These models perform relatively poorly for the time-domain metrics of SNR, likely because they focus on magnitude optimization. In addition, we expect that training on multiple corpora should enhance robustness to the trained corpora. To verify this, we train our model by randomly selecting 10 000 utterances from the training sets of the four datasets (TIMIT, WSJ, LIBRI and IEEE), and then test on their test sets. The results are given in the last row of Table VI, and show that this strategy yields good performance for all four datasets.

We remark that the size of the corpus does not seem to be a key factor for generalization. Although IEEE has only about a sixth of the utterances of the TIMIT dataset, the models trained on TIMIT do not display better generalization. Especially time-domain models (AudioUNet, TFNet, Proposed) trained with TIMIT perform even worse than trained with IEEE on untrained corpora. Another remark is that the models trained with WSJ and LIBRI have comparable robustness even though LIBRI contains more microphone channels. This may result from the fact that WSJ contains more background noise, which is another factor that affects cross-corpus generalization.

B. *Downsampling Schemes*

Most of the existing speech SR networks are trained with simulated datasets, where LR/HR pairs are generated by applying a specific downsampling scheme. In the real world, however, the pre-assumed downsampling scheme may not match the LR/HR relationship. Our experiments indicate DNN-based models are sensitive to different downsampling schemes, and this affects the generalization capability of supervised SR models.

We divide downsampling schemes into three categories. The first one is referred to as *subsampling*, which is the default setting of the MATLAB [30] downsample function. Subsampling decreases the sampling rate by discarding samples at fixed intervals. The second category is *decimating*, where one first applies a low-pass filter and then subsamples to acquire the desired LR signal. This is the default setting for the MATLAB decimate, resample functions and the SciPy [49] decimate function. The first two methods operate in the time domain, and the third category named *FFT* operates in the frequency domain. The FFT

TABLE VII
EXPERIMENTAL RESULTS OF DIFFERENT DOWNSAMPLING SCHEMES EVALUATED ON TIMIT

	Subsampling			Decimating			FFT		
Model	SNR	LSD	PESQ	SNR	LSD	PESQ	SNR	LSD	PESQ
Trained with Subsampling	**20.18**	**0.72**	**3.65**	15.32	1.28	3.15	16.02	0.88	3.52
Trained with Decimating	-10.85	2.51	1.03	**17.94**	**0.75**	**3.95**	-12.33	2.48	1.03
Trained with FFT	14.55	0.84	2.59	15.62	1.35	3.32	**17.28**	**0.72**	**3.92**
Trained with Random	19.57	0.75	3.51	17.07	0.77	3.81	16.74	0.77	3.88

scheme transforms an HR signal to the Fourier domain, leaves out high-frequency parts above the cutoff frequency, and then transforms back to obtain the corresponding LR signal. This is the default setting of the resample function in SciPy.

Table VII provides a comparison of the three categories of downsampling schemes on TIMIT using the proposed model. In the first three rows, we use the model trained with one specific downsampling scheme to test on data obtained by all three downsampling schemes. We observe that the decimating scheme performs the worst when tested on untrained schemes, with a drastic drop in objective scores (even negative SNR values). Although we see a degradation for subsampling and FFT, the drop is not nearly as severe as decimating.

Subsampling is simple and efficient among the three schemes. However, according to the sampling theorem of Shannon [43], this method introduces an unwanted artifact (referred to as aliasing) during the downsampling process as there are components with frequencies higher than the Nyquist frequency. Decimating solves this problem by first applying a low-pass filter. By default decimating uses Chebyshev Type I infinite impulse response filter of order 8 as the anti-aliasing filter in both MATLAB and SciPy package. We investigate two other low-pass filters (Butterworth and Bessel) for decimating. Experiments show that SR performance is highly affected by the type of low-pass filters. This indicates the models learned using decimating schemes carry unwanted characteristics of specific filters, which limit their application to realistic signals. The FFT method also avoids the aliasing phenomenon, and is a better choice for generalization purposes since it does not involve any filter. However, the computational expense is higher than the other two schemes.

It is important to develop a model that is robust against downsampling schemes. To achieve this, we introduce a random downsampling strategy: for each HR signal, the corresponding LR signal is generated by randomly picking one downsampling scheme from the three categories. By doing so, we make sure that models learn the essential features for SR, not the acoustic properties of specific downsampling techniques. The last row in Table VII provides the results when training with random downsampling. The results demonstrate that the model trained in this way is capable of producing satisfactory SR performance regardless of how LR signals are generated.

VI. CONCLUDING REMARKS

In this paper, we propose a novel CNN model for speech super-resolution that combines the strengths of both time and frequency domain approaches. The proposed CNN operates on time-domain signals, but is optimized using a cross-domain loss. Different loss functions have been investigated, and evaluation results show that the proposed T-PCM loss leads to better performance and avoids annoying artifacts in reconstructed speech. Experimental results on various datasets have demonstrated that our model significantly outperforms other DNN methods. Furthermore, our model is computationally efficient with a relatively small number of parameters. Also, as the proposed CNN model operates frame by frame, using no future (or past) information, it is a causal system.

We have also examined the robustness for deep learning based SR models. Specifically, we have investigated the effects of corpus channels and downsampling schemes. We have demonstrated that training with datasets that contain diverse channels and a random downsampling strategy improves model robustness. For future work, we plan to study how to improve the robustness of SR models to other factors such as background noise and room reverberation.

ACKNOWLEDGMENT

The authors thank Ashutosh Pandey for the discussion on AECNN and loss functions.

REFERENCES

[1] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 71–83, Jan. 2018.

[2] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5469–5473.

[3] A. Albahri, C. S. Rodriguez, and M. Lech, "Artificial bandwidth extension to improve automatic emotion recognition from narrow-band coded speech," in *Proc. 10th Int. Conf. Signal Process. Commun. Syst.*, 2016, pp. 1–7.

[4] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 1–5.

[5] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II: Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778, 2002.

[6] C. V. Botinhao, B. Carlos, L. P. Caloba, and M. R. Petraglia, "Frequency extension of telephone narrowband speech signal using neural networks," in *Proc. Comput. Eng. Syst. Appl.*, 2006, pp. 1576–1579.

[7] H. Carl, "Bandwidth enhancement of narrowband speech signals," in *Proc. EUSIPCO*, 1994, pp. 1178–1181.

[8] 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, Signal Process.*, vol. 1, 2001, pp. 665–668.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, pp. 295–307, Feb. 2016.

[10] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 347–358, May 2019.

[11] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM J. Numer. Anal.*, vol. 17, pp. 238–246, 1980.

[12] 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 STIN*, vol. 93, 1993, Art no. 27403.

[13] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

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

[15] Y. Gu, Z. H. Ling, and L. R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Proc. INTERSPEECH*, 2016, pp. 297–301.

[16] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, "Speech bandwidth extension with WaveNet," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 205–208.

[17] D. Haws and X. Cui, "CycleGAN bandwidth extension acoustic modeling for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6780–6784.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] *IEEE*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[21] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

[22] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural nets," in *Proc. Workshop Int. Conf. Learn. Representations*, 2017, pp. 1270–1279.

[23] K. Li and C. H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4395–4399.

[24] S. Li, S. Villette, P. Ramadas, and D. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5029–5033.

[25] 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, Signal Process.*, 2018, pp. 646–650.

[26] 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/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 883–894, May 2018.

[27] Z. H. Ling *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[28] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Proc. INTERSPEECH*, 2015, pp. 2598–2602.

[29] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, pp. 428–431.

[30] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB*. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2004.

[31] S. Mehri *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," 2016, *arXiv:1612.07837*.

[32] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech," in *Proc. INTERSPEECH*, 2014, pp. 2494–2498.

[33] 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*, 2011, Art no. r 1188.

[34] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," Distill, 2016. [Online]. Available: http://doi.org/10.23915/distill.00003

[35] A. V. Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[37] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[38] A. Pandey and D. L. Wang, "Dense CNN with self-attention for time-domain speech enhancement," 2020, *arXiv:2009.01941*.

[39] A. Pandey and D. L. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2489–2499, 2020.

[40] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992.

[41] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

[42] J. Sadasivan, S. Mukherjee, and C. Seelamantula, "Joint dictionary training for bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5925–5929.

[43] C. E. Shannon, "Communication in the presence of noise," in *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[44] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no 1, 2017.

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

[47] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. I– 805.

[48] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *Univ. Edinburgh. The Centre for Speech Technol. Res.,* 2017.

[49] P. Virtanen *et al.* J, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[50] H. Wang and D. L. Wang, "Time-frequency loss for CNN based speech super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2020*, pp. 861–865.

[51] M. Wang *et al.*, "Speech super-resolution using parallel WaveNet," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 260–264.

[52] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

Heming Wang received his Bachelor degree in physics in 2016, and M.S. degree in applied mathematics in 2018 from the University of Waterloo, Ontario, Canada. He is currently working toward the Ph.D. degree at the Ohio State University. His research interests lie in speech super-resolution and deep learning.

DeLiang Wang, biography not available at the time of publication.