TIME-DOMAIN NEURAL NETWORK APPROACH FOR SPEECH BANDWIDTH EXTENSION

Xiang Hao\(^1\), Chenglin Xu\(^2\), Nana Hou\(^2\), Lei Xie\(^1\), Eng Siong Chng\(^2\), Haizhou Li\(^3\),\(^4\)

\(^1\) School of Computer Science, Northwestern Polytechnical University, Xi’an, China
\(^2\) School of Computer Science and Engineering, Nanyang Technological University, Singapore
\(^3\) Department of Electrical and Computer Engineering, National University of Singapore, Singapore
\(^4\) Machine Listening Lab, University of Bremen, Germany

ABSTRACT

In this paper, we study the time-domain neural network approach for speech bandwidth extension. We propose a network architecture, named multi-scale fusion neural network (MfNet), that gradually restores the low-frequency signal and predicts the high-frequency signal through the exchange of information across different scale representations. We propose a training scheme to optimize the network with a combination of perceptual loss and time-domain adversarial loss. Experiments show the proposed multi-scale fusion network consistently outperforms the competing methods in terms of perceptual evaluation of speech quality (PESQ), signal to distortion rate (SDR), signal to noise ratio (SNR), log-spectral distance (LSD) and word error rate (WER). More promisingly, the multi-scale fusion network requires only 10% of the parameters of the time-domain reference baseline.

Index Terms— speech bandwidth extension, multi-scale fusion, neural networks, deep learning

1. INTRODUCTION

Speech bandwidth extension, which expands narrowband signal to wideband signal, has been widely studied for many years [1]. This technique plays an important role in many practical scenarios, e.g., narrowband signal enhancement [2] and automatic speech recognition (ASR) [3]. Other applications include speech audio compression and text-to-speech synthesis [4].

There have been a large variety of methods for speech bandwidth extension. In general, these approaches can be classified into two categories, namely rule-based and statistical approaches. The rule-based methods generate the high frequency spectrum based on the acoustic knowledge of the speech signal [5], while the statistical approaches assume that there exists a non-linear relationship between the spectral features of low frequency and high frequency components. The statistical approaches try to model such relationship by learning a mapping function between the narrowband signal and the wideband signal.

The statistical methods can be implemented either in frequency domain or in time domain. One of the frequency domain methods is to predict the spectral envelope of the high frequency part. Linear predictive coding (LPC) [6], Gaussian mixture model [7, 8], hidden Markov model [9, 10], and neural networks [11, 12, 13, 14, 15, 16] have been used to estimate the spectral envelope. These methods, however, face a common problem that the excitation, which is usually unknown, to reconstruct the signal. One of the most recent advances in speech signal processing is the ability to directly model raw signal in the time domain using neural networks [23, 24, 25], that avoids the phase estimation problem. The idea of time-domain neural network approach has opened up a new direction for speech bandwidth extension [26, 27, 4, 28, 29]. Time-frequency network (TFNet) [30] represents one of the successful implementations.

In this paper, we investigate a multi-scale fusion network (MfNet) to improve the performance of speech bandwidth extension in time domain. MfNet is inspired by the idea of time-domain speech bandwidth extension [26] as well as the work of image super-resolution [31]. As we know, multi-scale learning achieves better performance by capturing multi-resolution information. For example, in source separation and audio classification, with different convolutional filter size [32] or a cascade of wavelet filter banks [33] that learn multi-scale representations, one observes performance improvement. In this work, we study how MfNet performs speech bandwidth extension by aggregating the speech information across different scale representations, in analogy to multi-resolution image information in supervision.

Note that the training objective plays a key role in neural network performance. In the field of speech bandwidth extension, mean squared error (MSE) is used usually. In addition, perceptual loss is adopted in [28] and adversarial loss is used in [15, 16, 21]. The input data and adversarial loss calculation in the prior work of speech bandwidth extension is in frequency domain. In this paper, we would like to study the networks that take time-domain speech as input, and measure the training objective also in time-domain. We will study the effect of different loss functions that include perceptual loss and adversarial loss.

2. MULTI-SCALE FUSION NEURAL NETWORK

2.1. Architecture of MfNet

Our aim is to reconstruct a 16kHz wideband signal from a 8kHz narrowband signal. Suppose we have a 16kHz wideband signal \( \hat{x} = [\hat{x}_1, \ldots, \hat{x}_t, \ldots, \hat{x}_T] \). Same as in [26], by applying Chebyshev low-pass filter, we can get a 8kHz narrowband signal \( \bar{x} = [\bar{x}_1, \ldots, \bar{x}_t, \ldots, \bar{x}_T/2] \). Then, we use bicubic interpolation to generate a “fake” 16kHz wideband signal \( \hat{x} = [\hat{x}_1, \ldots, \hat{x}_t, \ldots, \hat{x}_T] \). MfNet takes the “fake” 16kHz signal as input, and is trained to produce 16kHz signal using the original 16kHz wideband signal as the advent of deep learning (DL), many approaches have studied how to estimate the high-frequency spectrum directly [17, 18, 19, 20, 21, 22]. Such techniques require the phases of the high frequency component, which is usually unknown, to reconstruct the signal. One of the most recent advances in speech signal processing is the ability to directly model raw signal in the time domain using neural networks [23, 24, 25], that avoids the phase estimation problem. The idea of time-domain neural network approach has opened up a new direction for speech bandwidth extension [26, 27, 4, 28, 29]. Time-frequency network (TFNet) [30] represents one of the successful implementations.

*Lei Xie is the corresponding author, lxie@nwpu.edu.cn.
target in a supervised training.

The architecture of MiNet is illustrated in Fig 1. We use $C_l^r$ to
represent the feature maps, where $l$ indicates this feature map is the
output of the $l$-th layer, $r$ indicates the time resolution or scale. In our
approach, the convolution doesn’t change the feature size, while the
downscaling can be achieved by a 1-dimension convolution of stride
2 that halves the time resolution. In this way, the neural network
consists of feature maps of different scales. The multi-scale fusion
block, which is used to aggregate information among the different
scale representations, is composed of convolution, downscaling, and
upsampling operation. For upsampling operation, we use convolution
to smooth input feature, and then performs bilinear interpolation in
the time direction by a factor of two. Suppose the feature maps of
the $l$-th layer are $\{C_1^l, \ldots, C_p^l, \ldots, C_t^l\}$. Through multi-scale
fusion block, the feature maps of the $(l+1)$-th layer are
$\{C_1^{l+1}, \ldots, C_{r}^{l+1}, \ldots, C_t^{l+1}\}$, where $\mathbf{C}_t^{l+1} = \frac{1}{8} \sum_{i=1}^{8} S(C_i^l)$. The $S(\cdot)$ is one
of convolution, downscaling, upsampling, and is used to resize feature
maps.

To better explain the multi-scale fusion block, we show an ex-
ample in Fig 2(a), which uses the feature maps $\{C_1^1, C_2^2, C_3^3\}$
to calculate the feature map $C_3^2$. The concatenation block with detailed
structure is shown in Fig 2(b), which is another way to aggregate the
information among the different scale representations. We first use
upsampling operation to resize different scale representations, then we
further concatenate these high-level feature maps along the channel
dimension. At last, we aggregate the information through a convolu-
tional layer.

2.2. Training Objective
As Section 2.1 definition, the input signal and original wideband
signal is $\mathbf{x}$ and $y$. Furthermore, we denote the predicted wideband
signal as $\hat{y}$. Thus, we have $\mathbf{y} = G(\mathbf{x})$, where $G(\cdot)$ is the map-
ing function represented by neural network. We explore the use of
several loss functions, namely are time-domain loss, perceptually-
motivated loss, adversarial loss (GAN loss) and a composite loss
which combines perceptually-motivated loss and GAN loss.

2.2.1. time-domain loss
The time-domain loss between predicted wideband signal and origi-
nal wideband signal is formulated as:

$$L_t(x, y) = \frac{1}{M} \sum_{m=1}^{M} \|x - y\|_2,$$  \hspace{1cm} (1)

where $M$ is the batch size.

2.2.2. perceptually-motivated loss
The intuition of using perceptually-motivated loss is that the mel
scale spectrograms approximate the perceived auditory information
by humans in psychoacoustic experiments. The perceptual loss is
defined as the L1 loss of the mel-spectrogram between the predicted
wideband signal and original wideband signal. This loss function is
described as:

$$L_p(x, y) = \|\text{Mel}(x) - \text{Mel}(y)\|_1,$$  \hspace{1cm} (2)

where $\text{Mel}(\cdot)$ is the mel-spectrum transformation. The spectrogram,
derived from STFT, is transformed into mel-spectrogram based on
triangular filters ranging from 3.8kHz to 8kHz. We calculate the loss
function over this frequency range to emphasize our objective, that is
to recover the high frequency part of the signal. In order to make sure
that the lower frequencies are intact, the final perceptually-motivated
loss is defined as:

$$L_p(x, y) = L_t(x, y) + \lambda_f L_f(x, y),$$  \hspace{1cm} (3)

where $\lambda_f$ is the weighting parameter for the perceptual loss and it’s
value is 0.001.

2.2.3. adversarial loss
In a Generative Adversarial Network (GAN) [34], the adversarial
model plays the two-player minimax game between a generator and
a discriminator. The generator, $G$, captures the data distribution and
maximize the similarity of real data and generated data, while the
discriminator, $D$, estimates the probability that a sample is from nat-
ural speech as opposed to synthetic speech generated by $G$. Suppose
that we adopt our proposed MiNet as the generator $G$, and the archi-
tecture of $D$ is same as in [35]. The value function can be defined as
follows:

$$V(D, G) = \mathbb{E}_{x \sim \rho_x} [\log D(x)] + \mathbb{E}_{z \sim \rho_z} [\log(1 - D(G(z)))],$$  \hspace{1cm} (4)

where $G$ is trained to minimize this value function, $D$ is trained to
maximize it. Equation (4) is to minimizing the Jensen-Shannon di-
vergence between the real data distribution and the distribution of

![Fig. 2. (a) an example of the process in which multi-scale fusion block aggregates multi-scale information; (b) The detailed structure of the concatenation block.](image-url)
generated data. However, this neural network model is notoriously difficult to train. [36] proposes to minimize the Wasserstein-1 distance. Thus, the value function is described as:

$$V_{WGAN}(D, G) = \mathbb{E}_{x \sim P_x}[D(x)] - \mathbb{E}_{\tilde{x} \sim P_{\tilde{x}}}[D(G(\tilde{x}))].$$  \tag{5}

In Equation (5), $D$ must be 1-Lipschitz. Weight clipping [36] and gradient penalty [37] are two methods to enforce this constraint. In our experiments, gradient penalty is used. Finally, the GAN loss is given as follows:

$$L_G(x, y) = L_l(x, y) - \lambda_a D(y),$$  \tag{6}

$$L_D(x, y) = D(y) - D(x) + \lambda (\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2,$$  \tag{7}

where $\alpha$ is a random number which sampled from a uniform distribution $U[0, 1]$, $\lambda$ and $\lambda_a$ is set to 10 and 0.001 respectively.

2.2.4. composite loss

The composite loss is a combination of perceptual loss and adversarial loss. The discriminator loss $L_D$ is same as used in our adversarial loss, and the generator loss is formalized as:

$$L_G(x, y) = L_l(x, y) + \lambda_f L_f(x, y) - \lambda_a D(y),$$  \tag{8}

where $\lambda_f$ and $\lambda_a$ is set to 0.001.

3. EXPERIMENTS

3.1. Datasets

To evaluate the performance of the proposed method, the corpus of Valentini-Botinhao [38] of 86 different speakers, is adopted in our experiments. A total of 84 different speakers are included in the training set and 2 other speakers in the test set. In our experiments, we split the original wave files into several segments with 50% overlap and every segment consists of 16,384 samples. We randomly select 133,096 and 14,789 segments from this training set for training and examining the convergence situation, respectively. The test set, consisting of 824 sentences, is used to evaluate the performances of different approaches.

3.2. Comparative Study

We implemented several systems both in the frequency domain and time domain as the benchmarking references for the proposed MfNet methods. They are described next in detail.

- **Spline**: simple bicubic interpolation.
- **LSM**: frequency domain method [17]. The input of this method is the log-spectrum of the narrowband signal and the output is the high-frequency log-spectrum of the wideband signal. This model has 3 hidden layers and every hidden layer has 2,048 hidden nodes.
- **DRCNN**: time domain method [26]. This approach is an encoder-decoder architecture, which consists of a series of upsampling and downsampling blocks. Notably, each convolutional layer has filters with the same size. We can consider the re-sampling as extracting features with filters at different time resolution, which is different from re-scaling, which keeps the time resolution the same, but shifts the filters at different stride. The method of upsampling is Subpixel shuffling. In the downsampling stage, the number of channels is [128, 256, 512, 512] and corresponding filter size is [65, 33, 17, 9]. In the upsampling stage, the number of channels is [1024, 1024, 512, 256] and corresponding filter size is [9, 17, 33, 65].
- **MfNet**: proposed MfNet architecture trained with the basic time-domain loss.
- **MfNet+P**: proposed MfNet architecture trained with the perceptually-motivate loss.
- **MfNet+A**: proposed MfNet architecture trained with the adversarial loss.
- **MfNet+C**: proposed MfNet architecture trained with the composite loss which combines perceptual loss and GAN loss.

3.3. Evaluation Metrics

We implement two evaluation metrics. One is the group of signal-based evaluation indicators, that includes perceptual evaluation of speech quality (PESQ), signal to distortion ratio (SDR) and log-spectral distance (LSD), another is the word error rate (WER) of automatic speech recognition (ASR). This ASR system is an end-to-end model and trained by the data of LibriSpeech [39]. For a reference speech utterance $\bar{x} = [x_1, \ldots, x_T]$ and the corresponding predicted speech $\bar{y} = [y_1, \ldots, y_T]$, the SNR is calculated according to the following formula:

$$\text{SNR} = 10 \log \frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T} (x_t - y_t)^2}.$$  \tag{9}

The LSD is defined as follows:

$$\text{LSD} = \frac{1}{T} \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} (X(l, k) - Y(l, k))^2.$$  \tag{10}

where $X(l, k)$ and $Y(l, k)$ are the log-spectral power magnitudes of $x$ and $y$, respectively. The $k$ and $l$ indexes frequency and frame. The $K$ represents the numbers of frequency in a frame and the $L$ is the total number of frames of a speech utterance.

3.4. Experiment Results

3.4.1. signal-based evaluation results

The evaluation results of different models are summarized in Table 1. In general, the proposed MfNet methods consistently outperform the baseline systems. More specifically, we first focus on MfNet whose training objective is the time-domain loss as defined in Equation (1). We observe that MfNet consistently outperforms Spline, LSM, and DRCNN baselines on all signal-based evaluation metrics. We further observe that these MfNet methods outperform the time-domain DRCNN baseline method in both low-frequency and high-frequency part. The LSD on both high and low frequency part are also summarized in Table 1. The results suggest that MfNet can restore low-frequency part well and learn better representations to predict the high-frequency part. The results also validate our claim that multi-scale information, through multi-scale fusion unit to aggregate different bandwidth representations, can help to achieve better bandwidth extension performance.
We further investigate the improvement of the proposed MfNet with different training objective. We firstly add the perceptual based L1 loss of high-frequency bands of the mel-spectrogram between the estimated signal and the original wideband signal into the time-domain loss with a weight, as defined in Equation (3). We name MfNet with the perceptual-motivated loss as “MfNet+P”. MfNet+P method achieves a 12.8% and 3.2% relative improvement over MfNet in terms of the LSD error on high-frequency bands and PESQ. This verifies our motivation that adding perceptual information into the training objective loss improves the perceptual quality of the estimated signal.

We further study whether using a GAN scheme would improve the perceptual quality and intelligibility of the estimated signal, where the weighted GAN loss in time-domain is defined as Equation (6) and (7). We name this system as “MfNet+A”. We observe that the MfNet+A outperforms the MfNet in terms of SNR, SDR, PESQ and LSD LF. However, the performance of the MfNet+A is worse than the MfNet+P in terms of PESQ and LSD. The results suggest that the perceptual loss is very good at improving the of perceptual quality of the estimated signal.

Finally, we interpolate the perceptual loss and the adversarial loss, as defined in Equation (8), and we name it as “MfNet+C”. We observe that the MfNet+C balances the importance of the perceptual loss and adversarial loss, and achieves balanced performance between MfNet+P and MfNet+A in terms of SNR, SDR, PESQ and LSD.

To facilitate the comparison, we also visualize the spectrogram of different approaches as shown in Figure 3. In the area A, it’s obvious that our proposed methods do better in restoring high-frequency information. Also, we find that none of the methods performs well in area B, probably due to the fact that this part represents a consonant with low energy, in particular, whose low frequency energy is not informative. We will explore some phonotactics information to solve this problem in the future.

3.4.2. speech recognition results

We now conduct experiments using the ASR system on bandwidth extended speech signals. Also, we use this ASR system to decode the original 16kHz wideband signal x. The word error rates are reported in Table 2. We observe that MfNet has a clear advantage over Spline, LSM and DRCNN methods. We also observe that training MfNet with different loss functions reduces the WER, in particular MfNet+P achieves the best result. We believe that re-training the ASR with bandwidth-extended speech will lead to further performance improvement.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>SNR (dB)</th>
<th>SDR</th>
<th>PESQ</th>
<th>LSD Full</th>
<th>LSD LF</th>
<th>LSD HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spline</td>
<td>–</td>
<td>21.88</td>
<td>26.09</td>
<td>3.75</td>
<td>2.14</td>
<td>1.34</td>
<td>2.67</td>
</tr>
<tr>
<td>LSM [17]</td>
<td>13.38M</td>
<td>21.26</td>
<td>25.57</td>
<td>2.76</td>
<td>1.80</td>
<td>1.10</td>
<td>2.26</td>
</tr>
<tr>
<td>DRCNN [26]</td>
<td>56.41M</td>
<td>23.95</td>
<td>27.76</td>
<td>3.55</td>
<td>1.84</td>
<td>0.47</td>
<td>2.60</td>
</tr>
<tr>
<td>MfNet</td>
<td>5.96M</td>
<td>24.50</td>
<td>28.08</td>
<td>3.76</td>
<td>1.61</td>
<td>0.19</td>
<td>2.26</td>
</tr>
<tr>
<td>MfNet+P</td>
<td>5.96M</td>
<td>24.55</td>
<td>28.20</td>
<td>3.88</td>
<td>1.40</td>
<td>0.21</td>
<td>1.97</td>
</tr>
<tr>
<td>MfNet+A</td>
<td>5.96M</td>
<td>24.77</td>
<td>28.48</td>
<td>3.80</td>
<td>1.72</td>
<td>0.18</td>
<td>2.42</td>
</tr>
<tr>
<td>MfNet+C</td>
<td>5.96M</td>
<td>24.70</td>
<td>28.47</td>
<td>3.82</td>
<td>1.46</td>
<td>0.21</td>
<td>2.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>SNR (dB)</th>
<th>SDR</th>
<th>PESQ</th>
<th>LSD Full</th>
<th>LSD LF</th>
<th>LSD HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spline</td>
<td>–</td>
<td>21.88</td>
<td>26.09</td>
<td>3.75</td>
<td>2.14</td>
<td>1.34</td>
<td>2.67</td>
</tr>
<tr>
<td>LSM [17]</td>
<td>13.38M</td>
<td>21.26</td>
<td>25.57</td>
<td>2.76</td>
<td>1.80</td>
<td>1.10</td>
<td>2.26</td>
</tr>
<tr>
<td>DRCNN [26]</td>
<td>56.41M</td>
<td>23.95</td>
<td>27.76</td>
<td>3.55</td>
<td>1.84</td>
<td>0.47</td>
<td>2.60</td>
</tr>
<tr>
<td>MfNet</td>
<td>5.96M</td>
<td>24.50</td>
<td>28.08</td>
<td>3.76</td>
<td>1.61</td>
<td>0.19</td>
<td>2.26</td>
</tr>
<tr>
<td>MfNet+P</td>
<td>5.96M</td>
<td>24.55</td>
<td>28.20</td>
<td>3.88</td>
<td>1.40</td>
<td>0.21</td>
<td>1.97</td>
</tr>
<tr>
<td>MfNet+A</td>
<td>5.96M</td>
<td>24.77</td>
<td>28.48</td>
<td>3.80</td>
<td>1.72</td>
<td>0.18</td>
<td>2.42</td>
</tr>
<tr>
<td>MfNet+C</td>
<td>5.96M</td>
<td>24.70</td>
<td>28.47</td>
<td>3.82</td>
<td>1.46</td>
<td>0.21</td>
<td>2.05</td>
</tr>
</tbody>
</table>

Table 1. The SNR, SDR, PESQ and LSD of different methods evaluated on test set. LSD Full, LSD LF and LSD HF shows the LSD value calculated for the whole spectrogram, low-frequency range and high-frequency range respectively.

3.4.3. Comparison of network complexity

In general, a network with a larger number of parameter offers better performance. We would like to compare the number of parameters of different networks. The statistics are summarized in Table 1. Notably, the architecture of our proposed MfNet only has requires about 10% of the parameters of the time domain DRCNN baseline. The compact structure of the proposed MfNet is an obvious advantage, especially when it also provides better performance. These results further prove that the proposed MfNet, which learns representations from different scales, has an advantage in estimating wideband signal.

4. CONCLUSION

In this work, we first show the promising ability of multi-scale fusion neural network for speech bandwidth extension. Based on this neural network structure, we explore the effect of different loss functions and propose a composite loss. Compared with a simple interpolation method, a frequency domain method and a time domain method, the proposed approaches can consistently achieve better performance in terms of SNR, SDR, PESQ, LSD and WER. In addition, an obvious advantage is the MfNet needs fewer parameters, compared with the time domain baseline, to achieve better performance. In the future, we will explore other techniques which have been proved useful in audio generation task, such as u-law compress, dilated convolution and temporal convolutional module and further combines multi-scale neural network with these techniques.

5. REFERENCES


