基于离散小波包变换与胶囊生成对抗网络的语音超分辨率算法

陈习坤,杨俊美
(华南理工大学电子与信息学院,广东广州 510640)

摘 要: 目前主流的语音超分辨率(Speech Super-Resolution,SSR)算法是使用卷积神经网络(Convolutional Neural Networks,CNN)把低分辨率(Low-Resolution,LR)语音信号转换为高分辨率(High-Resolution,HR)的语音信号.但只使用普通的CNN所带来的效果通常比较平滑且缺少细节信息.生成对抗网络(Generative Adversarial Networks,GAN)的引入可以很好地解决这一问题.此外,胶囊网络(Capsule Networks,CapsNet)可以将空间信息编码为特征,这样与GAN结合可以更好地判断数据的真假.离散小波变换(Discrete Wavelet Transform,DWT)是一种正交多分辨分析的工具,它在信号处理方面有很出色的表现.小波变换的一个扩展是离散小波包变换(Discrete Wavelet Packet Transform,DWPT),它在某些应用中提供了更有效的信号分析.本文提出一种基于DWPT和胶囊生成对抗网络(CapsGAN)的SSR网络架构Wavelet-SRGAN.对比实验结果表明,本文所提的算法能以最少的参数实现与现有先进算法相当的性能.在算法上有几个核心步骤:(1)在生成器网络中加入DWPT层;(2)在鉴别器上加入胶囊网络;(3)训练时加入小波损失.

关键词: 语音超分辨率;生成对抗网络;离散小波变换;离散小波包变换;小波损失
基金项目: 国家自然科学基金(No.61871188,No.61801133)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)04-1039-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220395

Speech Super-Resolution Algorithm Based on Discrete Wavelet Packet Transform and Capsule Generative Adversarial Network

CHEN Xi-kun, YANG Jun-mei
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

Abstract: The currently popular algorithms of speech super-resolution (SSR) use convolutional neural networks (CNN) to transform the low-resolution (LR) speech signal into high-resolution (HR) speech signal. However, the HR signal reconstructed from the ordinary CNN network is usually smooth and lack of details. Generative adversarial networks (GAN) can effectively solve this problem and generate high-quality speech signal. In addition, capsule networks (CapsNet) can encode the spatial information into features, and the combination with GAN will effectively improve the ability of discriminator. Moreover, discrete wavelet transform (DWT) is a tool for orthogonal multi-resolution analysis, which has excellent performance in signal processing. An extension of DWT is discrete wavelet packet transform (DWPT), which provides more efficient signal analysis in many applications. Based on the above mentioned DWPT and capsule generative adversarial networks (CapsGAN), we propose an SSR network architecture in this paper, named as Wavelet-SRGAN. Comparative experiment results show that the proposed Wavelet-SRGAN can achieve comparable performance against current state-of-the-art methods with the least amount of parameters. The key steps and main contributions of our algorithm are as follows: (1) adding a DWPT layer to the generator networks; (2) imbedding a capsule network in the discriminator; (3) additional wavelet loss is considered in the training process.

Key words: speech super-resolution; generative adversarial networks; discrete wavelet transform; discrete wavelet packet transform; wavelet loss
Foundation Item(s): National Natural Science Foundation of China (No.61871188, No.61801133)

收稿日期:2022-04-12;修回日期:2022-06-25;责任编辑:朱梅玉

1 引言

一般语音信号采样是以 8 kHz 为采样频率，按照国际标准 ITU-T Rec.G.711[1]，分别采用 α 律和 μ 律 PCM 编码律对采样进行量化，得到的比特率为 64 kb/s. 如果在高压缩比下使用感知音频编码器，传输信道中会出现宽度限制. 例如，MP3 已经成为非常流行的语音存储和传输格式. 感知音频编码器以较低的分辨率存储信号信息，从而实现了较高的编码效率，但是这仅对人类听觉系统而言较好. MP3 方案在实施时，信号中引入大量失真，而这失真的频谱信息是听不见的. 现在对于非常高的压缩比或者非常低的比特率，编码算法无法将所有失真保持在全带宽信号的屏蔽阈值以下. 通常在高频部分减少带宽，可以为带宽受限的信号实现指定的比特率. 这样严格限制了采样频率除去高频成分得到的语音信号，与自然语音相比质量明显下降. 这样高频信息丢失的语音信号降低了语音的透明度和清晰度，从而给听众带来低沉的听觉感受，造成较差的体验. SSR 是将语音从低分辨率转换为高分辨率的任务. 在数字信号处理文献中，它也被称为人工带宽扩展（Artificial Bandwidth Extension，ABE）、采样率转换（Sample-Rate Conversion，SRC）或重采样（Resampling）. SSR 是改变离散信号的采样率以获得相同底层连续信号的新的离散表示的过程. 更高的采样率包含更丰富的信息，因为更大的频率范围可以在数据中表示. 大多数如音乐和电影的多媒体应用程序，都使用最高分辨率的语音，因为它们捕获了各种乐器的更高层次的细节和纹理.

目前而言，大多数 SSR 和 ABE 算法都是围绕语音信号将 8 kHz 采样信号转化为 16 kHz 的信号，使用的是传统的码本映射[2,3]、高斯混合模型[4]和隐马尔可夫模型[5]. 随着深度神经网络（Deep Neural Networks，DNN）大面积应用于图像超分辨率等计算机视觉任务，其强大的拟合能力和快速运行速度给 SSR 算法带来了另一种思路. 与传统的 SSR 和 ABE 算法方法相比，端到端深度神经网络不需要特征工程可以直接从 LR 信号 $\boldsymbol{X}_{\mathrm{LR}}$ 生成 HR 信号 $\boldsymbol{X}_{\mathrm{HR}}$. 首先著名的 U-net[6]在医学图像分割的成功应用启发了 Kuleshov 等人[7]，他们提出了经典的 SSR 模型 AudioUnet. Kuleshov 等人[8]通过引入 TFiLM 层，利用循环层学习到的信息对卷积的特征映射块进行调制，解决了卷积在 U-net 中远程依赖建模方面的局限性. 在文献[9]中，Lim 等人使用 AudioUnet 的结构结合 $\boldsymbol{X}_{\mathrm{LR}}$ 的时域和频域一起生成 $\boldsymbol{X}_{\mathrm{HR}}$ 的波形信号. Wang 等人[10]则使用 $\boldsymbol{X}_{\mathrm{LR}}$ 的短时傅里叶变换（Short-Time Fourier Transform，STFT）频域 L1 损失函数和时域的 L1 损失函数结合的 TF-loss 训练 AudioUnet 网络的 TF-Unet. 文献[11]提出了一种新的类似于快速傅里叶变换的结构 FFTnet 并且使用了 Mel 频率损失函数去训练网络. 生成对抗网络（Generative Adversarial Networks，GAN）[12]的提出给超分辨率算法网络的训练提供了一种新的思路. 文献[13]提出的 SRGAN 在图像超分辨率上细节信息处理有了很大的提升. 文献[14,15]使用 GAN 在语音增强和语音生成方面有了不错的效果，这说明了 GAN 对语音领域的提升效果. 之后的文献[16,17]分别基于 AudioUnet 和 FFTnet 结合 GAN 来训练网络. 之后的图像超分辨率研究[18,19]把 GAN 与 Hinton[20]提出的胶囊网络（Capsules Networks，CapsNet）结合，使图像超分辨率有了不错的提升效果，这指出了本文的一个研究方向. 在文献[21]中，Huang 等人提出基于 DWPT 的 CNN 网络并结构结合小波损失来训练网络，可以实现 8 倍以上的超分辨率结果，这给了本文算法一个思路. 文献[22,23]中 Scaled Dot-Product 注意力机制对序列处理提供了新的想法. 这种 CNN 结构不仅可以起到类似于循环神经网络（Recurrent Neural Networks，RNN）的效果，并且其基于 CNN 的基本结构可以有很快的运行速度. 这也给了本文算法很大的启发.

本文所提出的算法直接对时域语音波形进行处理，基于小波损失和生成对抗网络的网络架构，可以实现较好的 2 倍和 6 倍超分辨率结果. 相比于之前的基于 AudioUnet 和 FFTnet 的优化提升结构，本文提出一种新的基于注意力机制的架构，并且提出了基于 DWPT 的小波损失函数来替换原来的频域损失函数，带来性能上的提升. 基于 CapsNet 的 GAN 具有较强的鲁棒性，这提出了一种新的训练语音 GAN. 本文算法整合了注意力机制和 DWPT 等新的处理语音信号的方法，相比于之前的算法，取得了更好的重构性能.

2 Wavelet-SRGAN 网络架构

与目前的大多数 SSR 网络不一样，Wavelet-SRGAN 网络架构运用了注意力机制和小波包分解. 图 1 描述了整个网络使用流程. 首先，使用 Bicubic 插值将 $\boldsymbol{X}_{\mathrm{HR}}$（8 kHz）插值到采样率为 16 kHz 或者 48 kHz 的信号. 其次输入用 GAN 来训练带有 DWPT 分解且参数为 θ_G 的生成器 G_{θ_G} 来进行超分辨率预测. 相比频域运算的网络，本文提出的网络架构是端到端型运算，具有较快的速度.

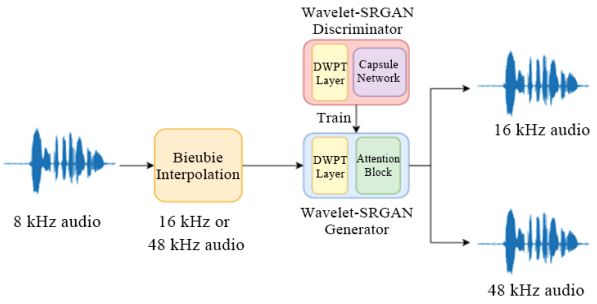


图 1 Wavelet-SRGAN

2.1 离散小波包变换

DWT本质上是一组带通滤波器. 其通过反复使用一组低通和高通滤波器来实现滤波. 在实现DWT时使用了几种滤波器,它们在系数的数量、期望的频率响应、信号表示中的误差等方面各有优缺点. DWT滤波器是线性时不变系统,其输出可以使用卷积关系计算,卷积关系涉及系统的输入、输出和脉冲响应. 在DWT的实现中,滤波器的输出需要下采样. 与DFT不同,对于给定的滤波器和输入,可以使用多组DWT系数. 在对滤波器输出进行下采样时,我们通常采用偶数索引值. 奇数索引值也是有效的DWT系数. 此外,滤波器系数可以按给定顺序或相反顺序使用. 输入数据的边界扩展可以以不同的方式执行. 在任何级别的DWT分解中,只有近似部分(频谱的低频部分)被进一步分解. DWPT使用的也是同样的带通滤波器,只不过在运算上有区别. 在DWPT中,细节部分(频谱的高频部分)也被进一步分解. 与独特的DWT表示不同,这种分解会导致信号的许多表示. 其优点是,可以使用冗余来选择与某些标准有关的信号的最佳表示. 这种方法为某些应用提供了比DWT更好的解决方案. 在文献[24]中,Mallat提出了DWT的快速算法,DWPT也可以使用这一快速算法. 在一次DWPT分解中,输入信号分别于分解高通 $\boldsymbol{G}(n)$ 和分解低通滤波器系数 $\boldsymbol{H}(n)$ 进行卷积运算,之后进行二抽取运算. 其输入与输出关系可以表示为

$$\widehat{\boldsymbol{X}}_{\mathrm{a}}=\boldsymbol{X}(n)*\boldsymbol{H}(n)=\sum_{k}\boldsymbol{H}(n-k)\boldsymbol{X}(k) \tag{1}$$

$$\widehat{\boldsymbol{X}}_{\mathrm{d}}=\boldsymbol{X}(n)*\boldsymbol{G}(n)=\sum_{k}\boldsymbol{G}(n-k)\boldsymbol{X}(k) \tag{2}$$

$$\boldsymbol{X}_{\mathrm{a}}=\widehat{\boldsymbol{X}}_{\mathrm{a}}(2n) \tag{3}$$

$$\boldsymbol{X}_{\mathrm{d}}=\widehat{\boldsymbol{X}}_{\mathrm{d}}(2n) \tag{4}$$

其中,$\boldsymbol{X}_{\mathrm{a}}$ 表示分解后的平均项,$\boldsymbol{X}_{\mathrm{d}}$ 表示分解后的细节项. 之后 $\boldsymbol{X}_{\mathrm{a}}$ 和 $\boldsymbol{X}_{\mathrm{d}}$ 可以接着进行这一分解过程 $\boldsymbol{X}_{\mathrm{aa}}$,$\boldsymbol{X}_{\mathrm{ad}}$,$\boldsymbol{X}_{\mathrm{da}}$ 和 $\boldsymbol{X}_{\mathrm{dd}}$ 项. 图2示例了DWPT的运算过程. 在DWPT层中,本文选择的是Daubechies小波家族的Db1(Haar)和Db4小波并且经过4次分解,所以输入长度为 l 的序列后输出长度变为 $\frac{l}{2^4}$. 其中,Db1和Db4的分解高通和分解低通滤波器系数可参考文献[25,26].

2.2 生成器架构

Wavelet-SRGAN的生成器 G_{θ_G} 的任务是生成输入Bicubic插值后的 $\boldsymbol{X}_{\mathrm{LR}}$ 的高频部分. 如图3所示,本文设计的网络架构具有两个支路. 对于卷积层而言,k 为卷积核数,s 为卷积运算的步长,n 为通道数. 本文的DWPT层使用的是进行四次DWPT分解,所以信号的频谱会产生16个相等的子频带项. 这16个子带分别进入

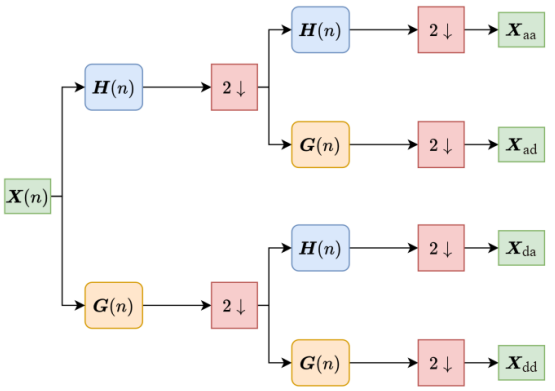


图2　DWPT图

32个通道、步长为1、卷积核长度为3的卷积层,之后按通道连接成为一个张量. 之后按图3结构,进行张量运算.

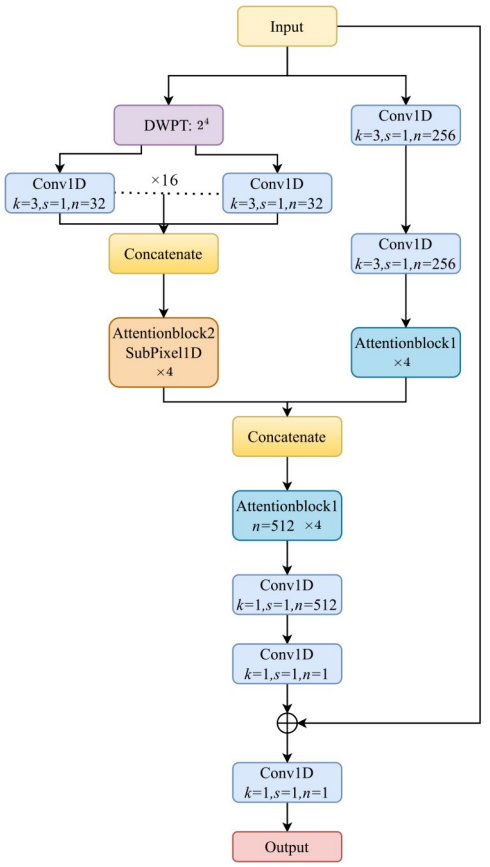


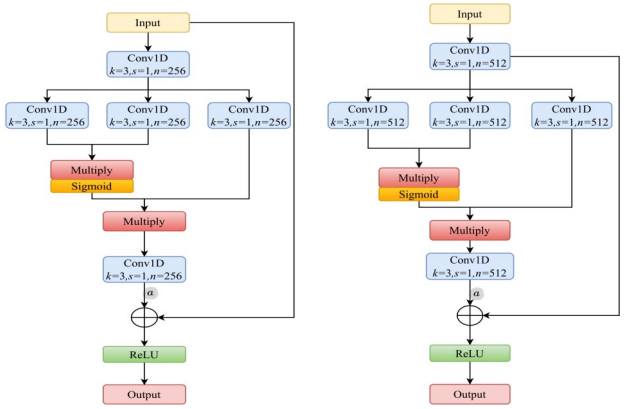
图3　生成器 G_{θ_G} 结构图

图4示例了Attention1和Attention2网络块的架构. 本文提出的注意力机制是不同于传统的Scaled Dot-Product注意力机制. Scaled Dot-Product注意力机制的矩阵乘积形式对于一维较长的语音序列而言,训练时造成的张量维度太大. 所以Attention1和Attention2使用的是Multiply注意力机制并且取消了Softmax使用Sigmoid归一化函数. 对于SubPixel层而言,其在图像超

分辨是一种更优的Upsampling层. 在文献[27,28]中,SubPixel层将通道和张量长度重组进行Upsampling,减少了重建时候的伪影. 总的来说,本文的重建网络是将尺寸为L的插值语音信号$\boldsymbol{X}$,先分解为(N_w,l)的小波平均和细节项$\hat{\boldsymbol{C}}=\left(\hat{\boldsymbol{C}}_1,\hat{\boldsymbol{C}}_2,\cdots,\hat{\boldsymbol{C}}_{N_w}\right)$(其中,$N_w$表示分解后的个数,$l$对应分解和的长度并且对应不同的数字频率带信息),之后转化为尺寸为L的原始语音信号$\hat{\boldsymbol{Y}}$. 总的来说,本文的网络可以定义为

$$\hat{\boldsymbol{Y}}=\phi\left(\boldsymbol{X},\hat{\boldsymbol{C}}\right)=\phi\left\{\boldsymbol{X},\left(\hat{\boldsymbol{C}}_1,\hat{\boldsymbol{C}}_2,\cdots,\hat{\boldsymbol{C}}_{N_w}\right)\right\}=\phi\left(\boldsymbol{X},\varphi_i(\boldsymbol{X})\right) \tag{5}$$

其中,φ_i: $\mathbf{R}^{N_W\times l}\rightarrow\mathbf{R}^L$, $i=1,2,\cdots,N_W$为信号到小波系数的映射;ϕ为插值信号到超分辨信号的映射.



(a) Attention1 (b) Attention2

图4 Attention1和Attention2

2.3 鉴别器架构

为了区分真实的$\boldsymbol{X}_{\mathrm{HR}}$和生成的$\boldsymbol{X}_{\mathrm{SR}}$样本,本文设计了一个参数为$\theta_D$的鉴别器网络$D_{\theta_D}$. CapsGAN用于图像超分辨率研究[18,19],它利用胶囊网络来输出最后的评分,所以本文设计的鉴别器D_{θ_D}也有相同的原理. 鉴别器D_{θ_D}的架构如图5所示,其中也包含了DWPT层. 基于波形和小波包分解的鉴别器D_{θ_D}有助于降低噪声和梳状滤波伪影. 本文使用2个跨步卷积来减少可训练的参数数量和一个CapsNet层. 胶囊网络(CapsNet)由胶囊的计算单元组成. 每个胶囊是一组嵌套在一起的神经元,用向量表示特定特征的实体参数. 这个向量表示对象的位姿参数,向量的长度对应于该特定特征存在的概率. 每个向量乘以一个权重矩阵来预测每个高级特征对应的向量. 然后,动态路由评估这些预测的一致性,并计算第二层每个胶囊的最终向量. 与CNN不同的是,CapsNet能够学习特征的层次结构和对象之间的几何关系. 这个独特的特性使CapsNet成为一个位姿不变的网络. 因此,在一些分类任务中,它的准确率和学习速度[20]都超过了CNN. 本文采用的结构是Primary

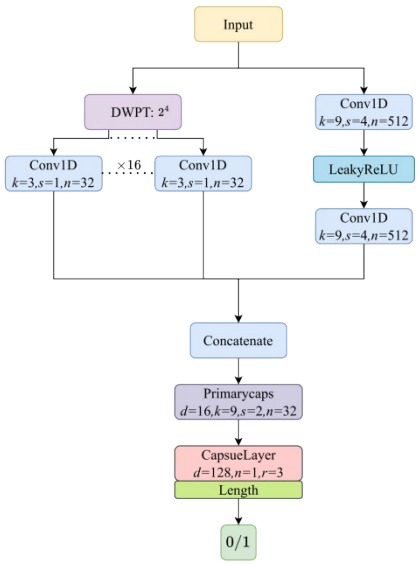


图5 鉴别器D_{θ_D}结构图

胶囊层有16个胶囊,每个胶囊有32个神经元. 最后一层包含1个胶囊和128个神经元. 动态路由在三个迭代中执行. 由于CapsNet的输出是一向量并且激活函数使用的是squash,其定义为

$$\boldsymbol{v}_j=\frac{\left|\boldsymbol{s}_j\right|^2}{1+\left|\boldsymbol{s}_j\right|^2}\cdot\frac{\boldsymbol{s}_j}{\left|\boldsymbol{s}_j\right|} \tag{6}$$

其中,$\boldsymbol{v}_j$是胶囊j的向量输出,$\boldsymbol{s}_j$是它的总输入,所以其输出向量是一个长度为0到1的向量. 为了解决对抗的最小最大值问题,Goodfellow等人[12]进一步定义了一个鉴别器网络D_{θ_D}与生成器G_{θ_G}交替优化的损失函数:

$$\min_{\theta_G}\max_{\theta_D}E\left[\log\left(D_{\theta_D}\left(\boldsymbol{X}_{\mathrm{HR}}\right)\right)\right]+E\left[\log\left(1-D_{\theta_D}\left(G_{\theta_G}\left(\boldsymbol{X}_{\mathrm{LR}}\right)\right)\right)\right] \tag{7}$$

所以对于D_{θ_D}输出而言,可以去掉最后的全连接层和Sigmoid函数,直接由向量的长度来代表D_{θ_D}的判断.

2.4 优化损失函数

本文定义的损失函数由三部分组成:(1)样本的L1(Mean Absolute Error, MAE)损失l_{content},也就是内容损失;(2)样本小波包分解后的L1损失l_{wavelet};(3)鉴别器D_{θ_D}的对抗损失$l_{\mathrm{adversarial}}$. 总的损失函数为这些函数的线性组合,其定义如下:

$$l_{\mathrm{loss}}=l_{\mathrm{content}}+l_{\mathrm{wavelet}}+10^{-2}l_{\mathrm{adversarial}} \tag{8}$$

2.4.1 内容损失

本文使用的内容损失函数是信号的MAE损失. 使用n作为序列时间样本索引,N代表截取的样本序列的总长度,内容损失函数定义为

$$l_{\mathrm{content}}\left(\boldsymbol{X}_{\mathrm{SR}},\boldsymbol{X}_{\mathrm{HR}}\right)=\frac{1}{N}\sum_{n=1}^{N}\left|\boldsymbol{X}_{\mathrm{SR}}(n)-\boldsymbol{X}_{\mathrm{HR}}(n)\right| \tag{9}$$

仅仅最小化MAE损失很难捕捉高频纹理细节,从而产生较差的感知结果. 由于高频小波系数可以描述纹理细节,本文将超分辨率问题从原始语音空间转化为小波域,并引入基于小波的损失来帮助细节重建.

2.4.2 小波损失

考虑n级小波包分解,其中n决定了超分辨率的尺度因子r和小波系数个数N_w,则$r=2^n$,$N_w=2^n$. 设$\boldsymbol{C}=\left(\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_{N_w}\right)$和$\hat{\boldsymbol{C}}=\left(\hat{\boldsymbol{C}}_1,\hat{\boldsymbol{C}}_2,\cdots,\hat{\boldsymbol{C}}_{N_w}\right)$分别代表真实和预测的小波系数,那么小波损失函数可以定义为

$$l_{\text{wavelet}}\left(\hat{\boldsymbol{C}},\boldsymbol{C}\right)=\frac{1}{M}\sum_{m=1}^{M}\left\{\lambda_1\left|\widehat{\boldsymbol{C}}_1(m)-\boldsymbol{C}_1(m)\right|+\sum_{i=2}^{N_w}\lambda_i\left|\widehat{\boldsymbol{C}}_i(m)-\boldsymbol{C}_i(m)\right|\right\} \tag{10}$$

其中,设置不同的小波损失系数λ的目的是平衡不同频带小波系数的重要性. 对于高频系数的权重较大的局部纹理,可以给予更多的关注. 对其中近似信息项约束$\frac{1}{M}\sum_{m=1}^{M}\left\{\lambda_1\left|\widehat{\boldsymbol{C}}_1(m)-\boldsymbol{C}_1(m)\right|\right\}$是捕获全局的拓扑信息,保证了原始输入信息不丢失且有利于训练的稳定性.

2.4.3 对抗损失

除了目前的内容损失和小波损失外,本文还加入了对抗损失. 对抗损失的目的是让生成器G_{θ_G}通过欺骗鉴别器D_{θ_D},使生成的语音更多地驻留在自然语音上. 鉴别器D_{θ_D}对训练样本的给出的概率为$D_{\theta_D}(G_{\theta_G}(\boldsymbol{X}_{\text{LR}}))$. 根据文献[13]提及的更好的梯度特性,本文训练时使用BinaryCrossentropy作为对抗训练的损失函数. 因此,对抗损失定义为

$$l_{\text{adversarial}}=-\log\left(D_{\theta_D}\left(G_{\theta_G}\left(\boldsymbol{X}_{\text{LR}}\right)\right)\right) \tag{11}$$

3 实验结果与分析

为了验证本文算法的有效性,本文使用采样率为48 kHz的VCTK数据集[29]来进行模型的训练和测试. 其中,VCTK数据集包括110位不同口音的英语使用者发出的语音数据. 每位演讲者读出大约400个句子,这些句子选自报纸、彩虹段落和用于演讲口音档案的启发段落. 每个演讲者都有一组不同的根据增加上下文和语音覆盖率的贪心算法选择的报纸文本. 对于对比试验,我们训练了文献[7]中的AudioUnet,使用文献[10]中的时频损失函数来训练AudioUnet得到的TF-Unet,训练了文献[8]中提及使用的TFiLM层的AudioUnet即TFiLM-Unet以及文献[11]中使用Mel频率损失的FFTnet.

3.1 实验设置与客观评价指标

本文的网络结构以及复现的网络结构都是使用Tensorflow[30]这一深度学习架构来实现的. 复现模型与本文提出模型的预训练都使用Adam[31]优化器,初始学习率α_1为0.000 1. 本文设置batchsize为2,对模型迭代训练500×10^3次,学习率按0.9的指数每5×10^3次衰减一次. 然后开始50×10^3次的对抗训练. 对抗训练时,鉴别器D_{θ_D}选择SGD[32]优化器,学习率α_2设置为0.001. 而生成器G_{θ_G}则使用Adam优化器,学习率α_3初始设为0.000 01并且按0.5的指数每10×10^3次衰减一次. 本文选择客观评价指标有如下几个.

(1)信噪比(Signal-to-Noise Ratio,SNR)为信号处理文献中使用的一个基本评价标准,其定义如下:

$$\text{SNR}\left(\boldsymbol{X}_{\text{SR}},\boldsymbol{X}_{\text{HR}}\right)=10\log\left(\frac{\left\|\boldsymbol{X}_{\text{HR}}\right\|^2}{\left\|\boldsymbol{X}_{\text{SR}}-\boldsymbol{X}_{\text{HR}}\right\|^2}\right) \tag{12}$$

(2)对数功率谱距离[33](Log-Spectral Distance,LSD),计算公式如下:

$$\text{LSD}\left(\boldsymbol{X}_{\text{SR}},\boldsymbol{X}_{\text{HR}}\right)=\frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{X}_{\text{SR}}}(l,k)-\widehat{\boldsymbol{X}_{\text{HR}}}(l,k)\right)^2} \tag{13}$$

其中,l和k为帧和频率的索引,$\widehat{\boldsymbol{X}_{\text{SR}}}$和$\widehat{\boldsymbol{X}_{\text{HR}}}$分别为$\boldsymbol{X}_{\text{SR}}$和$\boldsymbol{X}_{\text{HR}}$的对数功率谱. 对数功率谱幅值的计算公式为$\boldsymbol{S}=\log|\boldsymbol{X}|^2$,其中,$\boldsymbol{X}$为分帧后信号的STFT谱. 本文引入这一频域的评价指标来测试频域性能.

(3)高频对数功率谱距离(High Frequency Log-Spectral Distance,HFLSD),计算公式如下:

$$\text{HFLSD}\left(\boldsymbol{X}_{\text{SR}},\boldsymbol{X}_{\text{HR}}\right)=\frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\boldsymbol{X}_{\text{SR}}}(l,j)-\widehat{\boldsymbol{X}_{\text{HR}}}(l,j)\right)^2} \tag{14}$$

其中,j表示频率的索引.

(4)尺度不变信号失真比(Scale-Invariant Signal-to-Distortion Ratio,SI-SDR)[34,35],它测量的是一个均匀的比例因子的采样保真度. 提出SI-SDR的目的是作为误差函数,原因是SNR这种传统的误差度量会根据信号自身值和估计值进行缩放,进而增大SNR. 所以SI-SDR确保残差确实与目标正交,可以缩放目标或缩放估计. 对目标进行缩放,使残差与它正交,所以SI-SDR可以定义为

$$\text{SI}-\text{SDR}\left(\boldsymbol{X}_{\text{SR}},\boldsymbol{X}_{\text{HR}}\right)=10\log\left(\frac{\left\|\frac{\boldsymbol{X}_{\text{SR}}^{\text{T}}\boldsymbol{X}_{\text{HR}}}{\left\|\boldsymbol{X}_{\text{HR}}\right\|^2}\boldsymbol{X}_{\text{HR}}\right\|^2}{\left\|\frac{\boldsymbol{X}_{\text{SR}}^{\text{T}}\boldsymbol{X}_{\text{HR}}}{\left\|\boldsymbol{X}_{\text{HR}}\right\|^2}\boldsymbol{X}_{\text{HR}}-\boldsymbol{X}_{\text{SR}}\right\|^2}\right) \tag{15}$$

(5)VGG距离(VGG-dis),指由预先训练的VG-

Gish[35,36]网络计算出的高分辨率语音$\boldsymbol{X}_{\mathrm{HR}}$与超分辨率语音$\boldsymbol{X}_{\mathrm{SR}}$嵌入层之间的L2距离. 由于VGGish网络运行于Mel频谱图上,并接受了广泛的语音分类任务的训练,因此预期VGG距离能反映感知到的语音质量.

(6)语音质量感知评估[37](Perceptual Evaluation of Speech Quality,PESQ),这个指标的设计考虑了人类的听觉感知,并显示出与主观听力测试的相关性高于单纯的L1或L2距离. 由于PESQ最高只支持16 kHz采样率信号,所以6倍超分辨率不作为评价指标.

3.2 实验结果与分析

为了公平对比,本文选择一样的测试语音数据集. 表1和表2分别是2倍和6倍超分辨率的指标结果. 本文从以下的6个指标和参数量评价Wavelet-SRGAN模型. 本实验分别使用了两种Daubechies小波家族离散小波Db1和Db4. 表1所示是2倍超分辨率的结果,本文算法模型在SNR和SISDR指标中取得较好的结果. 其中,基于Db1离散小波变换的SRGAN在这个两个指标最好,分别是21.04 dB和21.16 dB. 同时本文算法也取得了较好的PESQ分数. 本文由于使用的是基于时间域的结构和损失函数,所以在LSD和HFLSD这个两个指标上弱于使用Mel频率损失的FFTnet结构. 表2所示为6倍超分辨率的结果,可以看出,同样使用Db1小波的SRGAN取得了最好的SNR和SISDR,分别为18.50 dB和18.62 dB. 在6倍超分辨率的结果中,本文使用的VGG-dis也取得了最小的距离,可以看出使用GAN来优化的语音,其特征更加接近真实的语音信号.

表1 2倍超分辨率算法指标对比

模型	指标						
	Parameters	SNR	LSD	HLSD	SI-SDR	PESQ	VGG-dis
AudioUnet[1]	70 977 154	19.09	7.12	7.85	19.13	3.82	0.151
TFiLM-Unet[8]	68 221 186	20.06	6.38	6.99	20.14	3.78	**0.141**
TF-Unet[10]	70 977 154	18.73	6.01	6.62	18.76	3.73	0.147
FFTnet[11]	**4 014 081**	16.07	**5.67**	**6.24**	16.01	3.57	0.186
Db1Waveletnet	34 501 647	**21.04**	5.84	6.42	21.15	**3.94**	0.161
Db4Waveletnet	34 501 647	20.86	6.30	6.91	20.97	3.82	0.170
Db1-SRGAN	34 501 647	**21.04**	5.83	6.41	**21.16**	**3.94**	0.160
Db4-SRGAN	34 501 647	20.80	6.31	6.92	20.91	3.82	0.171

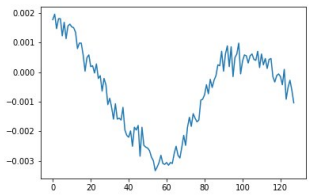
表2 6倍超分辨率算法指标对比

模型	指标					
	Parameters	SNR	LSD	HLSD	SI-SDR	VGG-dis
AudioUnet[1]	70 977 154	18.40	7.68	8.11	18.42	0.185
TFiLM-Unet[8]	70 977 154	17.94	7.68	8.02	17.94	0.232
TF-Unet[10]	68 221 186	16.08	7.07	7.30	16.02	0.189
FFTnet[11]	**4 014 081**	14.28	**6.76**	**6.82**	14.35	0.217
Db1Waveletnet	3 450 647	18.48	7.37	7.82	18.50	0.184
Db4Waveletnet	34 501 647	18.32	7.74	8.26	18.33	0.190
Db1-SRGAN	34 501 647	**18.50**	7.39	7.85	**18.52**	**0.183**
Db4-SRGAN	34 501 647	18.37	7.75	8.28	18.38	0.188

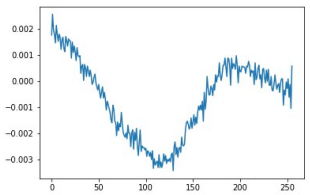
图6为各个模型生成的2倍超分辨率语音时域结果图,可以看出,AudioUnet和TFiLM-Unet生成的波形结果有些平滑. 使用频率损失函数的FFTnet和TF-Unet在时域的表现比AudioUnet和TFiLM-Unet较好. 但是FFTnet生成结果较差,高频分量明显较多. 本文使用Wavelet-SRGAN生成的时域结果较好.

图7为6倍语音超分辨率的时域结果. AudioUnet和TFiLM-Unet生成的波形结果过于平滑,缺少较多高频分量. 而FFTnet引入过多的高频分量,有些失真. TF-Unet表现较好但是在一些位置有些失真. 本文使用Wavelet-SRGAN生成的时域结果较好,体现了原波形的形状,也恢复了较多高频分量.

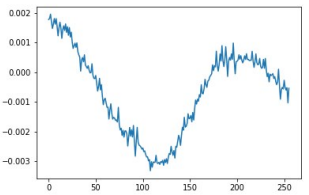
图8为2倍超分辨率语音信号的频谱图. 本文取语音信号使用汉宁窗取256个点为一帧,重叠数为128个点,使用512个点的FFT绘制出结果. 可以看出,降采样后损失较多频率成分. 使用AudioUnet和TFiLM-Unet模



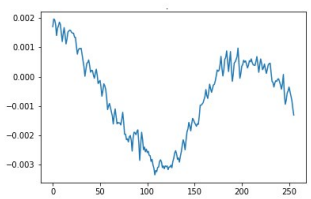
(*a*) 8 kHz低分辨率



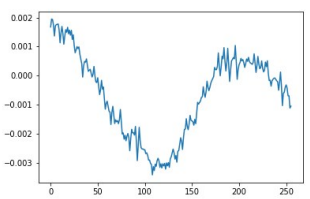
(*b*) 16 kHz高分辨率



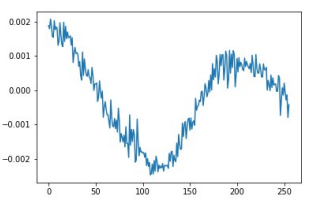
(*c*) AudioUnet



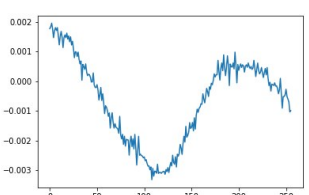
(*d*) TFiLM-Unet



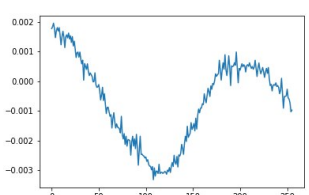
(*e*) TF-Unet



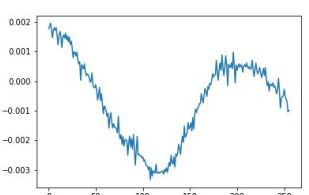
(*f*) FFTnet



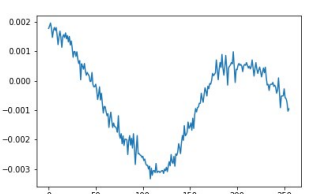
(*g*) Db1Waveletnet



(*h*) Db4Waveletnet

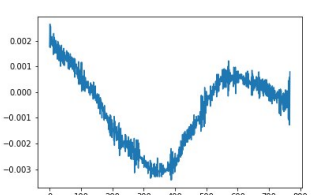


(*i*) Db1-SRGAN

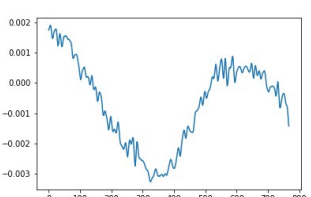


(*j*) Db4-SRGAN

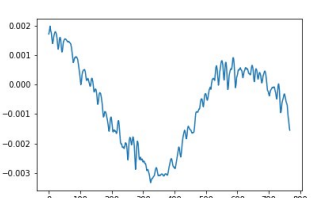
图6　2倍超分辨率的波形图



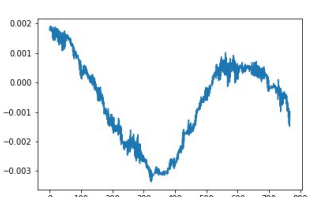
(*a*) 48 kHz高分辨率



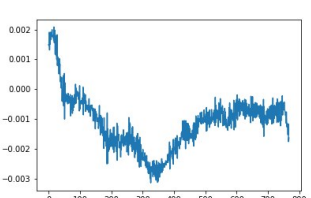
(*b*) AudioUnet



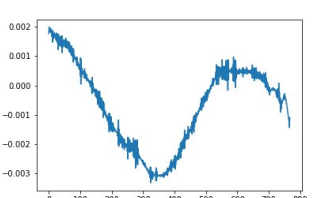
(*c*) TFiLM-Unet



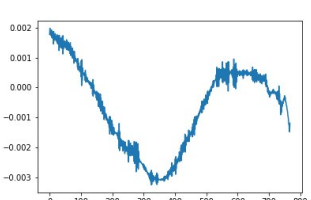
(*d*) TF-Unet



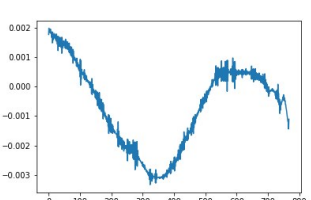
(*e*) FFTnet



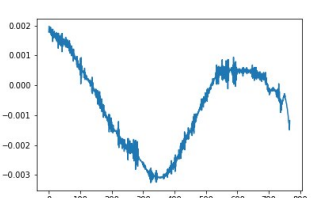
(*f*) Db1Waveletnet



(*g*) Db4Waveletnet



(*h*) Db1-SRGAN



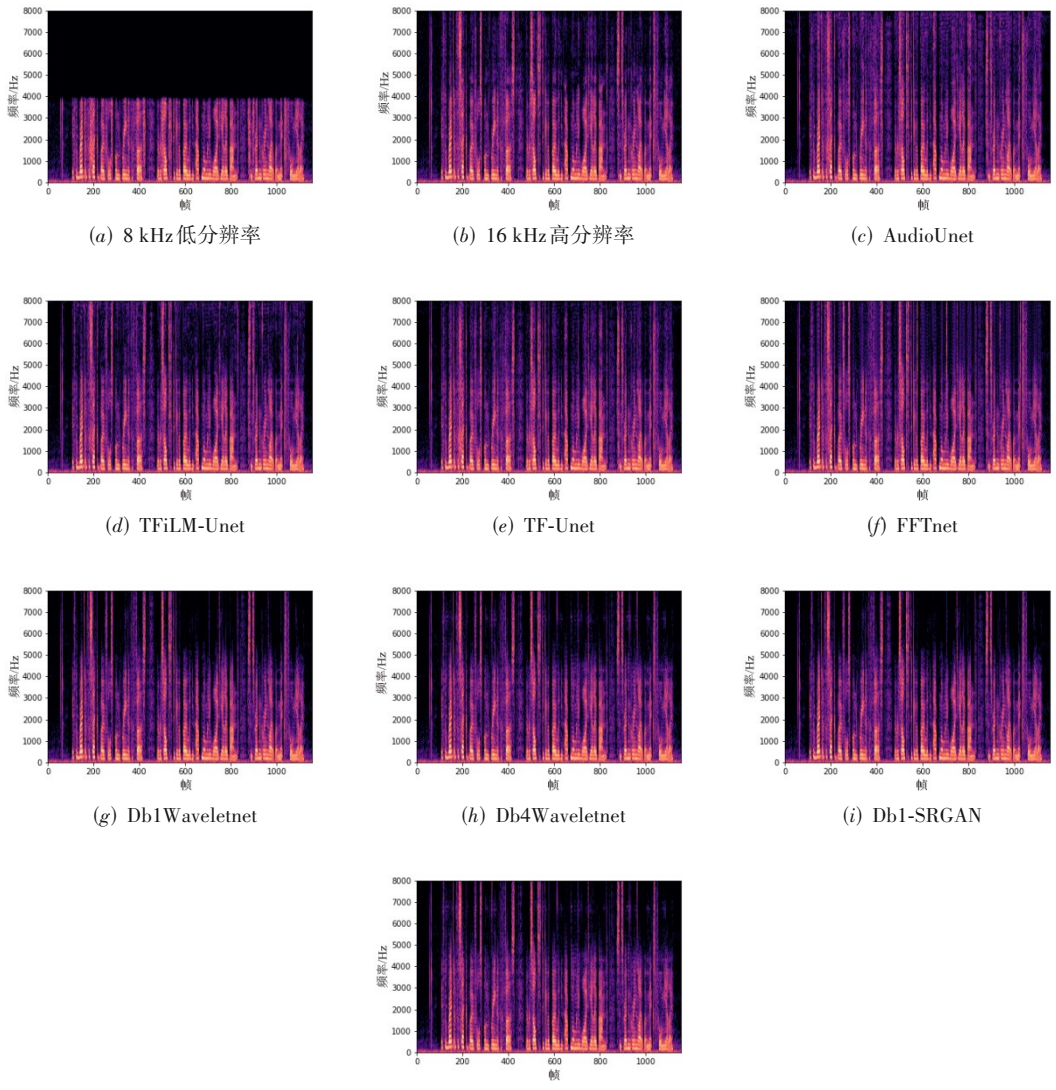
(*i*) Db4-SRGAN

图7　6倍超分辨率的波形图

型恢复的频谱结果较好,而FFTnet和TF-Unet恢复的频谱有些过度补充.虽然本文的模型有些地方频谱恢复得较空,但是由表1可以看出,这LSD和HFLSD差距不大.这更加说明了使用Wavelet-SRGAN没有过度恢复频率成分.

图9为6倍超分辨率的频谱结果图.本文是从8 kHz采样率恢复到48 kHz采样率,从图8的窄带频谱和图9中的48 kHz宽带频率谱的对比中可以看出,损失的频率成分十分多.在使用频率损失的TF-Unet和FFTnet时频率恢复较好,在基于时域的AudioUnet,TFiLM-Unet和本文提出的Wavelet-SRGAN中,频率恢复较差.这是因为本文算法模型是完全基于时间域的,不用在训练时使用傅里叶变换,完全基于常数滤波器卷积层.而AudioUnet和TFiLM-Unet损失函数使用的是MSE,这使时域波形图过于平滑.而且TFiLM-Unet的TFiLM模块引入了LSTM层,这种循环网络结构带来额外的训练时间.而本文模型完全基于卷积神经网络,使用注意力机制来代替LSTM层,这减少了训练时间.

在国际标准中,统一使用平均主观意见分(Mean Opinion Score,MOS)来评价语音质量.MOS用一个数字表示,从1分到5分,其中1分为最差,5分为最好.本次算法MOS得分测试通过微信问卷打分方式,收集到32人的测试结果.其中志愿者年龄集中在20~25岁,男女比例均衡.具体的得分情况如表3所示.可以看出,在2倍超分辨率情况下,使用离散小波Db1的SRGAN算法得到的语音信号获得了最高的MOS得分,最接近于真实的高分辨率算法.在6倍超分辨率情况下,本文使用较小参数量的网络可以接近最好的AudioUnet的MOS得分.



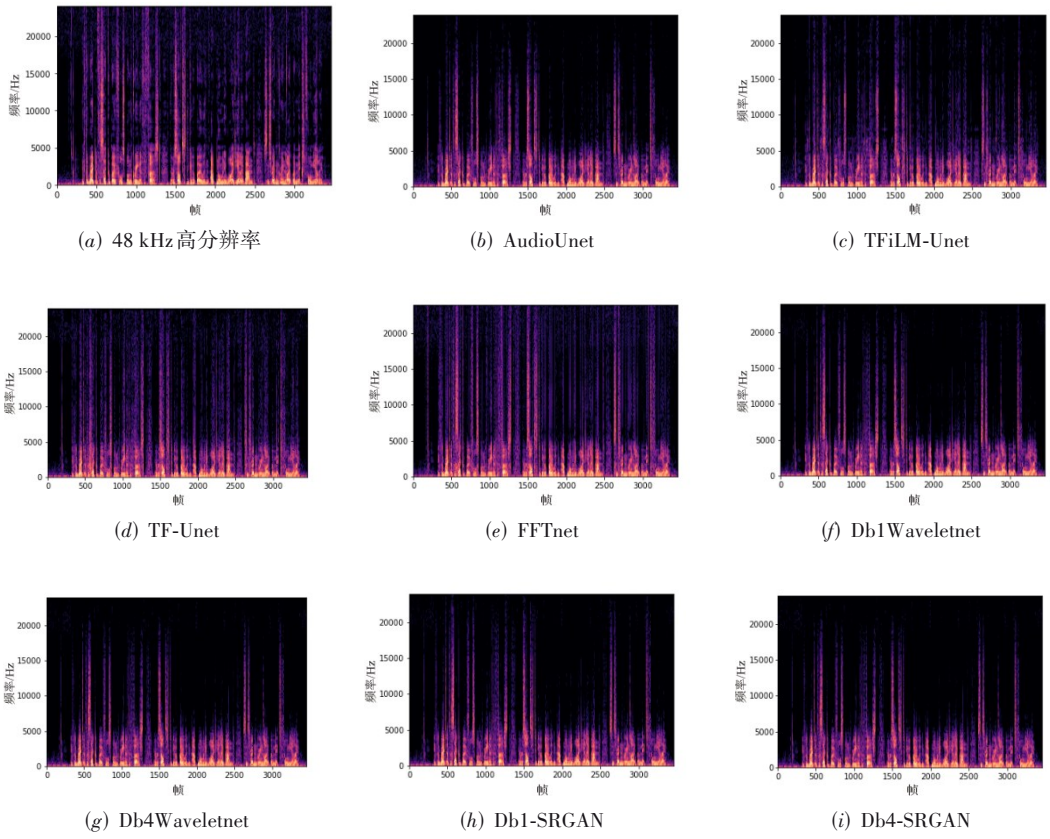
(a) 8 kHz低分辨率 (b) 16 kHz高分辨率 (c) AudioUnet

(d) TFiLM-Unet (e) TF-Unet (f) FFTnet

(g) Db1Waveletnet (h) Db4Waveletnet (i) Db1-SRGAN

(j) Db4-SRGAN

图8 2倍超分辨率的频谱图



(*a*) 48 kHz高分辨率　(*b*) AudioUnet　(*c*) TFiLM-Unet

(*d*) TF-Unet　(*e*) FFTnet　(*f*) Db1Waveletnet

(*g*) Db4Waveletnet　(*h*) Db1-SRGAN　(*i*) Db4-SRGAN

图9　6倍超分辨率的频谱图

表3　超分辨率算法MOS得分

模型	倍数	
	2倍	6倍
High Resolution	4.25	4.69
AudioUnet[1]	3.16	**3.88**
TFiLM-Unet[8]	3.56	3.69
TF-Unet[10]	3.25	2.78
FFTnet[11]	2.88	2.81
Db1Waveletnet	3.88	3.41
Db4Waveletnet	3.88	3.53
Db1-SRGAN	4.19	3.59
Db4-SRGAN	4.03	3.66

4　结论

本文算法改进了目前使用的LSTM层的TFiLM模块,提出一种新的适用于语音超分辨率的注意力机制;提出了一种新的损失函数,在基于传统的MAE损失函数中加入小波损失;使用小波损失去替换传统的频率损失,来捕捉恢复语音信号部分的细节部分变化,这带来一种新的思考方式. 而且本文提出了一种结合时域和小波系数的CapsGAN网络. 胶囊网络输出的并非一个常量而是向量,这可以很好地使用在鉴别器中,以克服传统的GAN的不稳定性. 而且本文的训练集是多人语言数据集,这也体现了所提模型的泛化性能. 下一步则是使用例如散射变换其他的小波变换去优化模型.

参考文献

[1] International Elecommunication Union. Pulse code modulation (PCM) of voice frequencies[EB/OL]. (1988-11-25) [2022-04]. https://www.itu.int/rec/T-REC-G.711-198811-I/en.

[2] ISER NTERNATIONAB, SCHMIDT G. Neural networks versus codebooks in an application for bandwidth extension of speech signals[C]//8th European Conference on Speech Communication and Technology (Eurospeech 2003). Geneva: ISCA, 2003: 565-568.

[3] QIAN Y, KABAL P. Wideband speech recovery from narrowband speech using classified codebook mapping[C]// Proceedings of the 9th Australian International Conference on Speech Science & Technology. Melbourne: Australian Speech Science & Technology Association Inc, 2002: 106-111.

[4] LIU X, BAO C C, JIA M S, et al. A harmonic bandwidth

extension based on Gaussian mixture model[C]//IEEE 10th International Conference on Signal Processing Proceedings. Beijing: IEEE, 2010: 474-477.

[5] JAX P, VARY P. Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model[C]//2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). Hong Kong: IEEE, 2003: I.

[6] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234-241.

[7] KULESHOV V, ENAM S Z, ERMON S. Audio super resolution using neural networks[EB/OL]. (2017-08-02)[2022-04]. https://arxiv.org/abs/1708.00853.

[8] BIRNBAUM S, KULESHOV V, ENAM Z, et al. Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations[EB/OL]. (2019-09-14)[2022-04]. https://arxiv.org/abs/1909.06628.

[9] LIM T Y, YEH R A, XU Y J, et al. Time-frequency networks for audio super-resolution[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018: 646-650.

[10] WANG H M, WANG D L. Time-frequency loss for CNN based speech super-resolution[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 861-865.

[11] FENG B, JIN Z Y, SU J Q, et al. Learning bandwidth expansion using perceptually-motivated loss[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 606-610.

[12] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[13] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 105-114.

[14] FU S W, LIAO C F, TSAO Y, et al. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement[EB/OL]. (2019-05-13)[2022-04]. https://arxiv.org/abs/1905.04874.

[15] SU J Q, JIN Z Y, FINKELSTEIN A. HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks[EB/OL]. (2020-06-10)[2022-04]. https://arxiv.org/abs/2006.05694.

[16] CHEN X K, YANG J M. Speech bandwidth extension based on Wasserstein generative adversarial network[C]//2021 IEEE 21st International Conference on Communication Technology (ICCT). Tianjin: IEEE, 2021: 1356-1362.

[17] 徐峰, 李平. 基于FFTNet-GAN的音频超分辨率方法研究[J]. 信号处理, 2021, 37(1): 59-65.
XU F, LI P. Research on audio super-resolution method based on FFTNet-GAN[J]. Journal of Signal Processing, 2021, 37(1): 59-65. (in Chinese)

[18] MAJDABADI M M, KO S B. MSG-CapsGAN: Multi-scale gradient capsule GAN for face super resolution[C]//2020 International Conference on Electronics, Information, and Communication (ICEIC). Barcelona: IEEE, 2020: 1-3.

[19] MAJDABADI M M, CHOI Y, DEIVALAKSHMI S, et al. Capsule GAN for prostate MRI super-resolution[J]. Multimedia Tools and Applications, 2022, 81(3): 4119-4141.

[20] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[EB/OL]. (2017-10-26) [2022-04]. https://arxiv.org/abs/1710.09829.

[21] HUANG H B, HE R, SUN Z N, et al. Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 1698-1706.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000-6010.

[23] KUMAR R, KUMAR K, ANAND V, et al. NU-GAN: High resolution neural upsampling with GAN[EB/OL]. (2020-10-22)[2022-04]. https://arxiv.org/abs/2010.11362.

[24] MALLAT S G. A theory for multiresolution signal decomposition: The wavelet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7): 674-693.

[25] RUCH D K, VAN FLEET P J. Wavelet Theory: An Elementary Approach with Application[M]. Hoboken: John Wiley & Sons, Inc., 2009.

[26] DAUBECHIES I. Ten Lectures on Wavelets[M]. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

[27] SHI W Z, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las vegas: IEEE, 2016: 1874-1883.

[28] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts[J/OL]. Distill, (2016) [2022-04]. http://distill.pub/2016/deconv-checkerboard.

[29] YAMAGISHI J, VEAUX C, MacDonald K. CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92) [EB/OL]. (2019-11-13) [2022-04]. https://datashare.ed.ac.uk/handle/10283/3443.

[30] ABADI M, BARHAM P, CHEN J M, et al. TensorFlow: A system for large-scale machine learning[C]//Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Savannah: USENIX Association, 2016: 265-283.

[31] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-10-22) [2022-04]. https://arxiv.org/abs/1412.6980.

[32] BOTTOU L. Large-Scale Machine Learning with Stochastic Gradient Descent[M]//19th International Conference on Computational Statistics. Paris: Physica Heidelberg, 2010: 177-186.

[33] GRAY A, MARKEL J. Distance measures for speech processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24(5): 380-391.

[34] LE ROUX J, WISDOM S, ERDOGAN H, et al. SDR-half-baked or well done?[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 626-630.

[35] LI Y P, TAGLIASACCHI M, RYBAKOV O, et al. Real-time speech frequency bandwidth extension[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 691-695.

[36] HERSHEY S, CHAUDHURI S, ELLIS D P W, et al. CNN architectures for large-scale audio classification[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017: 131-135.

[37] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City: IEEE, 2002: 749-752.

作者简介

陈习坤　男,1998年生于江西赣州. 华南理工大学电子与信息学院研究生. 研究方向为语音超分辨率、语音带宽扩展.
E-mail: 2363862709@qq.com

杨俊美(通讯作者)　女,1979年生于山东济南. 华南理工大学电子与信息学院硕士生导师. 研究方向为智能信号处理、自适应滤波、图像超分辨率重建、语音去混响等.
E-mail: yjunmei@scut.edu.cn