基于先验梅尔谱和神经声码器的语音丢包隐藏方法

黄晋维,鲍长春*,周 静
(北京工业大学信息学部语音与音频信息处理研究所,北京 100124)

摘 要: 对基于神经网络的丢包隐藏方法而言,输入特征是直接影响最终恢复效果的重要因素. 此外,如何通过丢包隐藏恢复高自然度的语音,也是亟待解决的难题. 为有效恢复丢包语音并提高自然度,本文提出了一种基于先验梅尔谱和神经声码器的语音丢包隐藏方法. 该方法采用一种非对称的编解码网络结构. 在编码端,用两个独立的编码网络分别从时域波形和梅尔谱中提取深层时频特征. 在解码端,将时频深层特征一同送入由时序自适应反归一化层构成的声码器中,以恢复丢失的语音信号并提高自然度. 仿真实验表明,该方法在语音感知质量和短时客观可懂度上均优于现有的两种丢包隐藏算法.

关键词: 丢包隐藏;先验梅尔谱;神经声码器;时序自适应反归一化层;时频特征
基金项目: 国家自然科学基金(No.61831019)
中图分类号: TN912 **文献标识码:** A **文章编号:** 0372-2112(2024)08-2581-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231056

A Speech Packet Loss Concealment Method Based on Priori Mel-Spectrum and Neural Vocoder

HUANG Jin-wei, BAO Chang-chun*, ZHOU Jing
(*Institute of Speech and Audio Signal Processing*, *Faculty of Information Technology*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: For the neural network-based speech Packet Loss Concealment (PLC), the input features are crucial factors that directly affect the final recovery performance. Additionally, the challenge of restoring high natural speech through PLC remains to be addressed. To effectively recover packet loss speech and improve its naturalness, this paper proposes a PLC method of speech signal based on the priori Mel-spectrum and neural vocoder. The proposed method adopts an asymmetric encoding and decoding network structure. At the encoding stage, this method utilizes two independent encoding networks to extract the latent time-frequency features from the waveform and Mel-spectrogram, respectively. At the decoding stage, the latent time-frequency features are jointly fed into a neural vocoder which is composed of several temporal adaptive denormalization layer to restore the lost speech signals and enhance the naturalness. Simulation experiments demonstrate that the proposed method outperforms two existing packet loss concealment algorithms in terms of perceptual evaluation of speech quality and short-time objective intelligibility.

Key words: packet loss concealment; Mel-spectrum; neural vocoder; temporal adaptive de-normalization layer; time-frequency features
Foundation Item(s): National Natural Science Foundation of China (No.61831019)

1 引言

随着计算机、互联网及芯片技术的快速发展,新兴语音通信技术为人们的日常生活提供了极大的便利. 目前,语音通信中的主流编码技术是先将语音信号按时序分割成包,再以语音包的形式进行编码传输. 为解决网络传输的延时与抖动所导致的丢包问题,语音通信会引入丢包隐藏(Packet Loss Concealment,PLC)技术,以恢复因丢包导致的语音损伤. 由于实时性的需求,语音通信中的PLC算法要满足因果性.

文献[1]首次提出了基于频域特征的因果PLC方

收稿日期:2023-11-14;修回日期:2024-03-05;责任编辑:赵克
*通讯作者:鲍长春

法的基本构架，即利用过去帧的频域特征来逐帧恢复丢包的语音信息．在此基础上，文献[2]提出了一种针对受损频谱特征的全带循环网络PLC方法，该方法利用神经网络增强受损的复数谱和幅度谱，然后通过两者联合训练，恢复丢失语音的复数谱．文献[3]利用对比学习方法，对提取的特征进行预训练，然后将预训练的特征输入至基于高保真生成对抗网络(Generative Adversarial Network for High Fidelity，HiFi-GAN)的声码器[4]中，以生成丢失语音的波形．此外，文献[5]和文献[6]同样将神经声码器应用于频域PLC方法中．这两种方法的结构类似，即均由预测模块与生成模块两部分构成，所不同的是文献[5]选用基于流的声码器[7]，而文献[6]则使用频域生成对抗式声码器[8]．

除了上述基于频域的因果PLC方法外，时域的端到端模型同样有着不错的恢复效果．基于时域的PLC模型省略了频域特征的分析-合成过程，直接通过时域语音样点来实现PLC任务．这种时序模型与循环神经网络高度适配，例如，文献[9]和文献[10]通过长短时记忆网络来恢复丢包语音，特别是文献[10]提出了基于隐藏单元的循环网络，对语音波形与丢包掩蔽一同建模，并根据丢包掩蔽调整循环网络神经元的状态值．另外，文献[11]提出了一种基于卷积循环网络的PLC模型，并应用了扩展和掩蔽训练来提升PLC的性能．

基于频域特征的模型与基于时域特征的模型均有各自的优点，故一些研究试图将两类特征结合起来实现PLC任务，例如，文献[12]提出了一种简易的时频结合的PLC方法，该方法将多尺度频域特征与上一帧时域波形级联送入网络中预测丢失的语音波形．文献[13]则提出了一种基于波形网络均衡器(Wave Network EQualizer，WaveNetEQ)的PLC方法，该方法利用生成的梅尔谱指导时域自回归模型[14]生成语音波形，此类方法有效地联合了时频特征，保障了语音恢复的有效性．此外，文献[15]提出一种基于时频结合判别网络的对抗式学习模型，可在低延迟条件下高质量地恢复丢失语音．

受上述方法启发，本文通过时频域特征和神经声码器来改善所恢复语音的质量与自然度．所提方法使用的声码器是风格Mel生成对抗网络(StyleMel Generative Adversarial Networks，StyleMelGAN)[16]，其风格迁移的思想源于文献[17]．文献[17]认为传统结构中的归一化层会倾向于“洗去”输入特征中隐含信息，故提出通过空间自适应的学习变换空间自适应反归一化层(SPatially Adaptive DEnormalization，SPADE)[17]，使用输入的语义信息来改善“洗去”问题．文献[18]提出通过两个独立的神经编码器，充分发挥了SPADE层双输入-双输出的结构优势．在SPA-DE基础之上，文献[16]提出时序自适应反归一化层(Temporal Adaptive DEnormalization，TADE)层，并将其应用于语音合成领域中．本文所提方法考虑结合文献[16]和文献[18]的研究思路，利用频域特征来指导时域特征的上采样合成过程，并通过基于TADE层的神经声码器来恢复高自然度的丢包语音．

2 基于先验梅尔谱的丢包隐藏方法

所提方法的PLC原理如图1所示．在训练阶段，所提方法首先将语音波形和对应的梅尔谱分别送入波形编码器和梅尔编码器中，其中，波形编码器通过一维卷积从语音波形中提取局部特征，而梅尔编码器则通过神经编解码网络预测并恢复缺失的梅尔谱．然后，将两者的编码输出一同送入基于TADE层的神经声码器中解码恢复丢包语音．在测试阶段，所提方法首先对输入帧进行判别，若未丢包，则直接输出；若发生丢包，则将历史缓存区内的语音信号送入网络隐藏模块中恢复丢失的语音．最后，所提方法将恢复的语音输出，并同时将其存储至历史缓存区中．

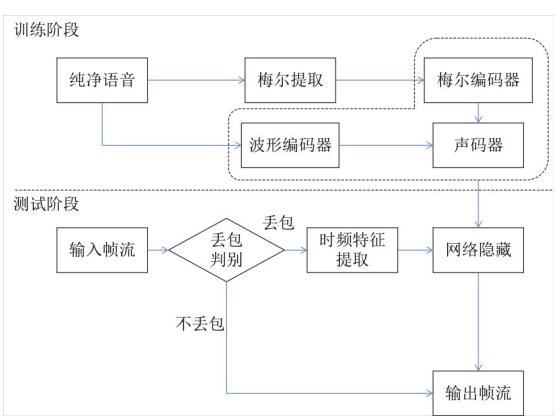

图1 所提丢包隐藏方法的原理框图

所提方法的网络结构如图2所示，整体结构可分成编码和生成两个模块．在虚线左侧的编码模块中，所提方法通过波形编码器和梅尔编码器，分别从历史缓存区的一段语音$\boldsymbol{s}_{\text{in}}$及其对应的梅尔谱$\boldsymbol{m}_{\text{in}}$中学习语音的局部时序波形特征$\boldsymbol{f}_{\text{s}}$及预测的梅尔特征$\boldsymbol{f}_{\text{m}}$；在虚线右侧的生成模块中，则将学习到的时序波形特征与先验的梅尔信息一同送入声码器中，恢复出丢失的语音波形$\boldsymbol{s}_{\text{pred}}$．图2中包含两个输出，其中梅尔特征$\boldsymbol{f}_{\text{m}}$仅在训练阶段中用于损失函数的计算，以辅助合成语音波形，而在测试阶段中被略去；预测波形$\boldsymbol{s}_{\text{pred}}$则为最终的输出帧流．

所提方法的编解码网络为非对称结构，主要体现

在两方面. 一方面,两个编码器是非对称的,由于所提取的两种深层特征的表征难度差异,梅尔编码器的网络结构远比波形编码器复杂;另一方面,编码模块和生成模块也是非对称的,这是因为生成高采样率的语音需要复杂的生成模块作为支撑,以提高丢包隐藏语音的质量和自然度.

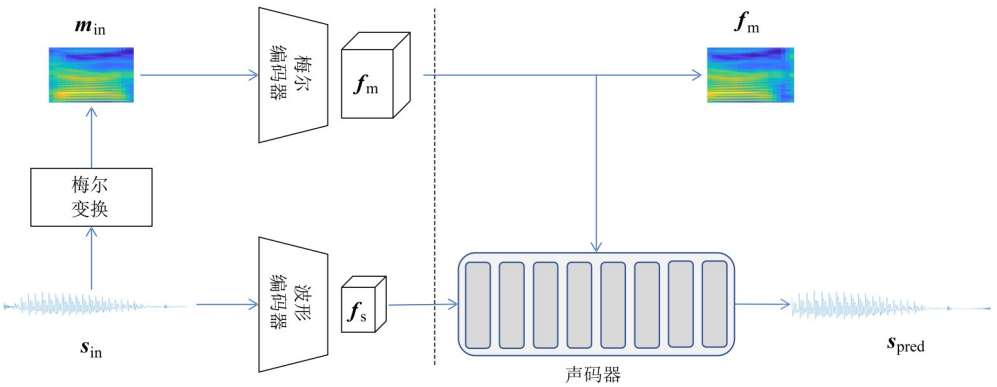

图2　所提方法的网络结构

本节后续内容将详细介绍所提方法各模块的基本原理及训练流程,即波形编码器、梅尔编码器、神经声码器、多尺度损失函数和整体网络的训练流程.

2.1　波形编码器

图2中的波形编码器由卷积网络和全连接网络构成,其沿着时间轴方向对语音信号进行一维卷积下采样,以此提取低维度的局部深层特征. 该特征对语音波形细节的恢复起着重要的作用.

设历史缓存区中语音段为 $\boldsymbol{s}_{\text{in}} \in \mathbb{R}^{1\times(NP)}$,其中 N 表示帧长,P 表示历史缓存区中所包含的帧数. 波形编码过程可以由式(1)表示:

$$\boldsymbol{f}_{\text{s}} = E_{\text{s}}\left(\boldsymbol{s}_{\text{in}}\right) \tag{1}$$

其中,$\boldsymbol{f}_{\text{s}} \in \mathbb{R}^{c\times\lambda}$ 为局部深层特征,c 表示卷积通道数(以下简称为通道数),λ 表示时间特征长度;E_{s} 表示波形编码器. 激活特征 $\boldsymbol{f}_{\text{s}}$ 后续将会被送至声码器中,来恢复丢包语音的波形细节.

在波形编码器 E_{s} 中,首先对输入语音 $\boldsymbol{s}_{\text{in}}$ 进行4倍下采样,然后进行卷积下采样,最后通过全连接层调整输出特征维度,得到激活特征 $\boldsymbol{f}_{\text{s}}$. 其中,每个卷积层中均包含着一个1维批量归一化层和一个含泄露的ReLU激活函数.

2.2　梅尔编码器

梅尔谱是一种高度结构性和高信息密度性的低维频域语音特征. 文献[13]表明,准确估计梅尔谱对PLC十分重要. 图2中所示的梅尔编码器采用时序卷积神经网络(Temporal Convolutional Network,TCN)作为核心组件,并设计为神经编解码结构.

在时序任务中,TCN仅利用卷积网络达到近似RNN的建模效果. 如图3所示,标准的TCN由输入层、隐藏层和输出层三部分组成. 网络的输入层特征为时间序列,从左至右表示从过去时刻至将来时刻. 红色节点表示当前时刻TCN的输出,其仅受由深蓝色节点所代表的历史输入特征的影响,而不受未来的和感受野外的浅蓝色节点影响. 深蓝色节点的范围越大,则影响输出的历史信息越多. 为扩大感受野范围,TCN在隐藏层中引入了可以产生空洞的膨胀卷积. 在图3中,深灰色节点会影响当前时刻网络输出,而浅灰色节点则会被“空洞”跳过,以此确保在不增加参数量的前提下,获取更多的历史信息. 注意,膨胀卷积的系数会随着层数的加深呈指数增长.

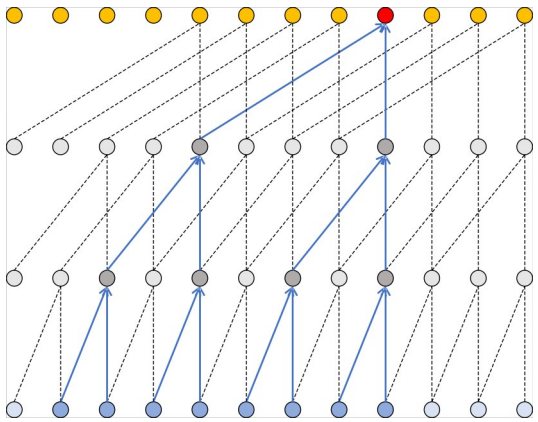

图3　TCN示意图

文献[19]提出使用残差网络结构来解决膨胀卷积可能会引起网络退化问题. 如图4所示,单个TCN残差块是由两类卷积构成的. 一类是1×1卷积层,通过调整卷积内核数量来影响输出的通道维度;另一类是膨胀卷积,位于不同层内的膨胀系数会呈指数增加. 此外,TCN残差块中添加了参数ReLU激活函数和1维批量归一化层,通过堆叠不同超参数的TCN则可以构成时序

卷积模块(Temporal Convolutional Module,TCM).

图2中的梅尔编码器的整体结构如图5所示,主要由三部分构成:神经编码器、神经解码器以及作为中间层的TCM. 其中,神经编码器与解码器均由二维卷积层构成,以便同时捕捉梅尔谱的帧间相关性和通道相关性. 除中间层TCM外,神经编解码器之间还有跳层连接. 这种跳层可以将神经解码层输出与对应的神经编码层的输出相级联,并送至下一个解码层中,以确保神经编解码网络的学习稳定性.

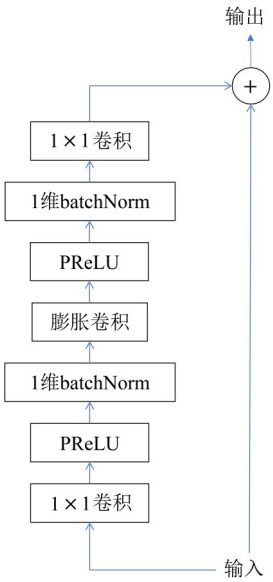

图4 TCN残差块的结构

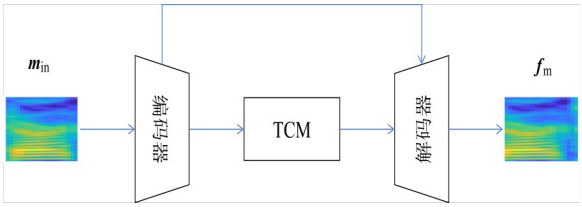

图5 梅尔编码器的整体结构

梅尔编码器的输入为与s_{in}相对应的原始梅尔谱$m_{in}\in\mathbb{R}^{c_{mel}\times\lambda}$,其中$c_{mel}$表示梅尔滤波器组的通道数,$\lambda$则表示梅尔谱的时间特征长度. 编码网络不断提高卷积通道数并降低梅尔通道数,但时间特征长度保持不变,以确保TCM可利用该维度上的特征相关性进行时序建模. 此外,编码器中的所有卷积层均与批量归一化层和参数ReLU激活函数相连.

编码特征在送入TCM前,会被整形为二维特征,以更好地契合TCM结构. 在所提方法中,TCM是由3个TCN堆叠组成,而TCN由指数膨胀的6个TCN残差块组成. 这些残差块中的膨胀系数分别为1、2、4、8、16和32.

梅尔编码的过程可由式(2)所表示,其中E_m表示梅尔编码器. 经过编码后,可以得到与输入语段相对应的预测梅尔特征$f_m\in\mathbb{R}^{c_{mel}\times\lambda}$.

$$f_m=E_m(m_{in}) \tag{2}$$

梅尔编码器的整体流程:首先,通过由7个二维卷积层构成的编码器提取出深层信息;然后,送入TCM中捕捉时序相关性;最后,通过由转置卷积构成的解码器得到预测的梅尔特征.

2.3 基于TADE残差块的神经声码器

神经声码器是所提方法的核心结构,由TADE层来构建. TADE层是一种基于线性反归一化映射的双输入-双输出网络结构,其反归一化特性主要体现在从指导特征中学习一组调制参数,并利用该参数对归一化后的激活输入进行线性调制. TADE层的双输入-双输出的特性体现在网络具有两个输入特征,其中指导特征将会对激活特征进行指导学习. 这一特性也是所提方法可进行时频特征结合学习的基础.

图6给出了TADE层的网络结构,其中绿线表示指导输入,蓝线表示激活输入. 首先,指导输入经过上采样后,送入一个一维卷积层(Conv1d)中;然后,将卷积结果重新送入两个独立的卷积层中,得到两个调制参数矩阵γ和β. 而激活输入则会送入实例归一化层(instanceNorm)中,得到归一化输出ω. 为了弥补归一化可能造成的信息丢失,TADE层对ω进行自适应的线性调制:$(\gamma+E)\odot\omega+\beta$,其中$E$表示单位矩阵,$\odot$表示矩阵的点乘运算.

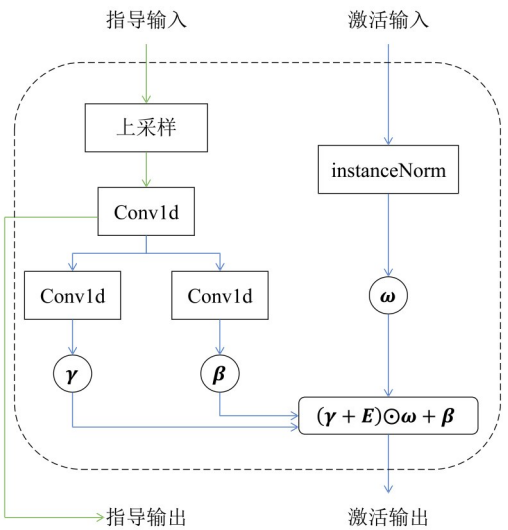

图6 TADE层的网络结构

图7中虚线框中给出了所提方法生成模块的基本单元,即TADE残差块. 与TADE层类似,TADE残差块同样为双输入-双输出残差结构. 在TADE残差块中,激

活输入和指导输入会经过两次TADE层学习映射. 首先,激活输入经过第一个TADE层后,将所输出的特征按维度均分为两部分,并分别进行独立卷积运算;然后,对应的卷积结果会被送入由Softmax和Tanh所构建的门控激活函数中,以获得输出;接着,该输出经过第二个TADE层后,被均分为两部分,分别经过一维扩张卷积(DiConv1d)和相同的门控激活函数;最后,该门控激活函数的输出将会与原来的激活输入相加,得到残差激活输出. 与之相对的,指导输入只会经过两次TADE层,以此尽可能地保留特征信息. 需要注意的是,只有第二个TADE层会对两种输入进行上采样操作.

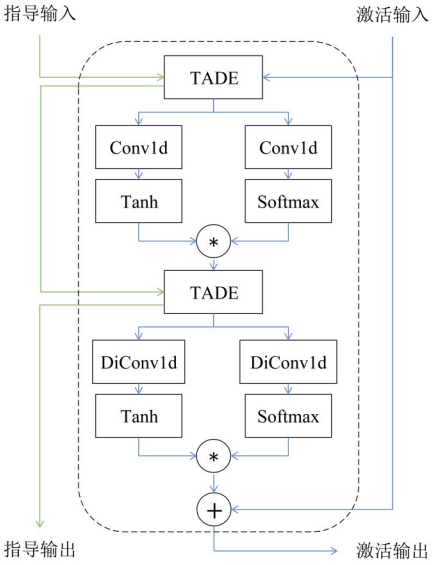

图7　TADE残差块的结构

TADE残差块的具体流程可表示为

$$\left[\boldsymbol{f}_{\mathrm{s}}', \boldsymbol{f}_{\mathrm{m}}'\right]=\mathrm{TADERes}\left(\boldsymbol{f}_{\mathrm{s}}, \boldsymbol{f}_{\mathrm{m}}\right) \tag{3}$$

其中,TADERes表示TADE残差块的运算;$\boldsymbol{f}_{\mathrm{s}}' \in \mathbb{R}^{c\times(\lambda d)}$为激活特征;$\boldsymbol{f}_{\mathrm{m}}' \in \mathbb{R}^{c\times(\lambda d)}$为指导特征,$d$表示当前TADE残差块的膨胀系数. 两种输入特征经过TADE残差块后,其时间维度λ会受到上采样的影响而增大.

图8给出了神经声码器的整体结构,其中包含了8个TADE残差块,并在末尾增添了一个最终激活模块. 最终激活模块由一个降低输出通道数的卷积层和一个激活函数构成. 该神经声码器的通道数为64,且最大膨胀系数为5. 最终激活输出$\boldsymbol{s}_{\mathrm{pred}} \in \mathbb{R}^{1\times(NP)}$的维度与输入波形同型,但$\boldsymbol{s}_{\mathrm{pred}}$包含当前丢失的一帧语音;而最终的指导输出将会被舍去,未用于损失函数的计算.

2.4　损失函数及训练流程

所提方法采用了分段式训练模型. 首先,通过预训练梅尔编码器E_{m}使其具有获取先验梅尔信息的能力;

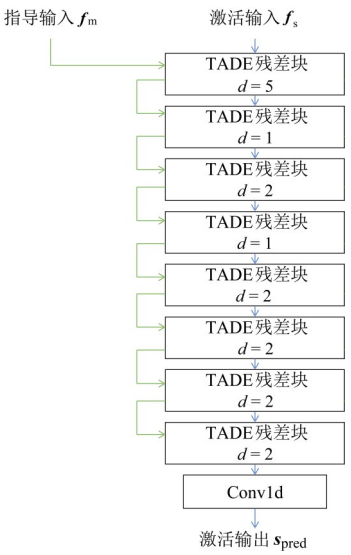

图8　所提方法的神经声码器结构

然后,对整体网络进行一致性训练. 在梅尔编码器的预训练中,E_{m}使用局部加权的L1范数作为损失函数,如式(4)所示. 该损失函数通过放大丢失位置的梅尔损失,来增强网络对于丢失信息的学习.

$$L_{\mathrm{mel}}=\left\|\boldsymbol{f}_{m}-\boldsymbol{m}_{\mathrm{real}}\right\|_{1}+\alpha_{\mathrm{mel}}\left\|\boldsymbol{f}_{\mathrm{m\_loss}}-\boldsymbol{m}_{\mathrm{real\_loss}}\right\|_{1} \tag{4}$$

其中,L_{mel}为指导梅尔损失;$\boldsymbol{f}_{\mathrm{m_loss}}$和$\boldsymbol{m}_{\mathrm{real_loss}}$分别为丢失语音所对应的预测梅尔特征和真实梅尔谱;$\alpha_{\mathrm{mel}}$为梅尔平衡系数.

对于网络整体的损失函数,本文使用多尺度损失函数,即包含指导梅尔损失、多分辨率频域损失、权重相位损失和波形损失.

多分辨率频域损失又可分为谱收敛损失和对数STFT幅度谱损失,分别如式(5)和式(6)所示:

$$L_{\mathrm{sc}}=\frac{\left\|\left|\boldsymbol{X}_{\mathrm{pred}}\right|-\left|\boldsymbol{X}_{\mathrm{real}}\right|\right\|_{F}}{\left\|\left|\boldsymbol{X}_{\mathrm{real}}\right|\right\|_{F}} \tag{5}$$

$$L_{\mathrm{stft}}=\mathrm{mean}\left(\left\|\log\left|\boldsymbol{X}_{\mathrm{pred}}\right|-\log\left|\boldsymbol{X}_{\mathrm{real}}\right|\right\|_{1}\right) \tag{6}$$

其中,$\|\cdot\|_{F}$表示弗罗贝尼乌斯范数,$|\boldsymbol{X}_{\mathrm{pred}}|$表示预测语音信号的幅度谱,$|\boldsymbol{X}_{\mathrm{real}}|$表示真实语音信号的幅度谱,mean(·)表示求取平均值. 多分辨率频域损失会在不同STFT参数下,将上述两种损失相加求和,得到

$$L_{\mathrm{mr}}=\frac{1}{\Gamma}\left[\sum_{\tau=1}^{\Gamma}\left(L_{\mathrm{sc}}^{\tau}+L_{\mathrm{stft}}^{\tau}\right)\right] \tag{7}$$

其中,L_{mr}表示多分辨率频域损失,Γ表示频域分辨率的个数,τ表示具体分辨率的参数. 多分辨率的具体参数设置与文献[20]一致.

当仅使用式(7)训练时,虽然网络可确保恢复语音的周期性,但生成语音与真实语音间存在时延. 这可能

是由于语音的相位信息不准确导致的. 因此,所提方法增添了权重相位损失,如式(8)所示.

$$L_{\mathrm{wp}}=\left\|\left|\boldsymbol{X}_{\mathrm{pred}}\right|\odot\left|\boldsymbol{X}_{\mathrm{real}}\right|-\mathrm{Re}\left\{\boldsymbol{X}_{\mathrm{pred}}\right\}\odot\mathrm{Re}\left\{\boldsymbol{X}_{\mathrm{real}}\right\}-\mathrm{Im}\left\{\boldsymbol{X}_{\mathrm{pred}}\right\}\odot\mathrm{Im}\left\{\boldsymbol{X}_{\mathrm{real}}\right\}\right\|_1 \tag{8}$$

其中,Re表示复数取实部,Im表示复数取虚部. 权重相位损失L_{wp}通过虚部与实部相减,计算恢复语音与真实语音间的相位损失.

此外,本文还采用了L1波形损失来增强恢复语音的波形匹配度,如式(9)所示.

$$L_{\mathrm{wav}}=\left\|\boldsymbol{s}_{\mathrm{pred}}-\boldsymbol{s}_{\mathrm{real}}\right\|_1+\alpha_{\mathrm{wav}}\left\|\boldsymbol{s}_{\mathrm{pred\_loss}}-\boldsymbol{s}_{\mathrm{real\_loss}}\right\|_1 \tag{9}$$

其中,$\boldsymbol{s}_{\mathrm{pred_loss}}$和$\boldsymbol{s}_{\mathrm{real_loss}}$分别表示丢失帧位置的恢复语音和真实语音,$\alpha_{\mathrm{wav}}$表示波形平衡系数.

将上述所有损失函数加权求和,最终获得所提方法的整体损失函数L_{total},如式(10)所示.

$$L_{\mathrm{total}}=L_{\mathrm{mr}}+\alpha'_{\mathrm{mel}}L_{\mathrm{mel}}+\alpha_{\mathrm{wp}}L_{\mathrm{wp}}+\alpha'_{\mathrm{wav}}L_{\mathrm{wav}} \tag{10}$$

其中,α'_{mel}、α_{wp}和α'_{wav}分别为各自损失函数的权重系数.

综上所述,所提PLC方法先分别利用波形编码器和梅尔编码器得到波形特征和预测的梅尔特征,然后将时频特征一同送入神经声码器中,预测丢失的语音. 整体训练过程可概括如下:

(1) 初始化网络参数;

(2) 利用式(4)训练先验梅尔编码器直至收敛;

(3) 将完成预训练的梅尔编码器嵌入整体网络中,并用式(10)进行训练,直至收敛.

3 实验

为验证所提方法的有效性,本节将其与4种参考方法进行对比:基于深度神经网络(Deep Neural Network,DNN)的PLC方法[1](记作DNN)、基于流声码器的PLC方法[5](记作FLOW)、基于波形相似叠加的PLC方法[21](记作WDNN)和基于时频特征结合的PLC方法[15](记作TFGAN). 此外,本文选择静音替代法作为基准方法(记作丢包). 在PLC性能测试中,将丢包率分别设置为5%、10%、15%、20%以及50%. 为评测连续丢包下的PLC性能,本文还设置了固定的丢包模型,即单帧丢包、双帧丢包以及三帧丢包,以评测各方法的连续丢帧下的PLC性能. 所有方法在相同的丢包数据集上进行测试,且均为因果的PLC实现方式,即均利用过去的语音帧信息来恢复当前的丢失语音帧.

3.1 实验设置

实验使用Librispeech语音数据集[22]构建网络的训练集和测试集. 其中,测试集语句数量为11 000句,测试集语句数量为500句. 语音采样率为16 kHz,帧间无叠接,窗长N为20 ms(320样点),历史缓存帧数P为9. 梅尔滤波器组通道数为80,最低频率为70 Hz,最高频率为8 kHz. 梅尔谱使用汉宁(Hanning)窗进行分帧,窗长为320样点,窗移为160样点. FFT长度为1 024样点. 卷积通道数c、时间特征长度λ和梅尔滤波器组的通道数c_{mel}分别设置为64、18和80.

为模拟真实丢包情况,本文通过Gilbert Elliott信道模型[23]生成模拟丢包测试集. 对于固定丢包数的模型,输入语音为纯净语音,丢失语音的长度是固定的,即丢失一帧、连续丢失两帧及连续丢失三帧. 所有方法均在相同的测试集上评测.

4种对比方法中,DNN方法由两个神经网络构成,分别用于恢复幅值和相位. 每个网络由4层全连接层组成,且每层的神经元个数为2 048并包含批量归一化层和Dropout层. 每一层的激活函数为ReLU,网络的优化方法是Adam,初始学习率设置为0.000 5,损失函数为均方误差函数,训练周期为40. FLOW方法主要由神经网络预测器、流神经声码器和平滑模块三部分构成. 在训练阶段和测试阶段中,流神经声码器的输入语音段的长度分别为16 000和1 920个样点. WDNN方法通过所设置的拉伸因子对信号波形进行拉伸,同时结合深度神经网络预测丢包语音的短时幅度谱. FLOW方法和WDNN方法的网络训练参数请参阅文献[5]和文献[21]. TFGAN方法是一种端到端的生成对抗性网络PLC方法. 其中,生成网络由膨胀残差卷积块构成,判别网络由时频判别模块构成. 在判别网络中,FFT长度为320. 网络优化的方法为Adam,初始学习率为0.000 1,网络稳定后降为0.000 05. TFGAN方法的损失函数同样包含多分辨率频域损失,具体参数设置与文献[15]一致. 网络训练周期为30. 注意,TFGAN方法的生成网络为基于时域波形的网络,而判别网络则为时频结合的网络.

所提方法的网络分成两阶段进行训练. 梅尔编码器的预训练中选用Adam优化方法,学习率为0.000 1,训练10个周期. 预训练结束后,将网络参数转储到完整的PLC网络中,并联合训练30周期. 网络输入帧数P为9. 优化方法仍为Adam,学习率为0.000 1. 损失函数权重系数α'_{mel}、α_{wp}、α'_{wav}、α_{mel}和α_{wav}分别为3、2、2、3和3.

3.2 实验结果与分析

实验的对比评测包括两方面:时域波形对比和客观质量评测对比. 其中,客观质量评测包括语音感知质量(Perceptual Evaluation of Speech Quality,PESQ),短时客观可懂度(Short-Time Objective Intelligibility,STOI),信噪比(Signal to Noise Ratio,SNR)和梅尔损失.

3.2.1　恢复的时域波形对比

图9给出了在单帧丢失情况下,不同PLC方法所恢复语音波形的对比示例. 其中,蓝色为真实语音波形,红色为恢复语音波形. 从图9可以看出,所有方法均能有效地填补缺失部分波形,且恢复波形与原始波形有着类似的周期结构. 如图9(*a*)所示,所提方法虽然前半段波形与真实波形贴合紧密,但后半段恢复效果有一定程度的下降. 其原因是梅尔特征预测准确性会随着时序而下降,进而影响神经声码器的恢复效果. 此外,还可看出所提方法恢复的波形较为平滑,这是由波形损失函数中的范数导致的. 由图9(*b*)可以看出,TFGAN方法的网络恢复侧重于基音周期和高频成分的恢复效果,而忽略了波形是否对齐. 这一问题是由损失函数中的波形损失与频域损失不平衡所导致的. 从图9(*c*)可以看出,WDNN方法所恢复的波形与真实波形贴合不紧密,但恢复了波形的周期性. 由图9(*d*)可以看出,虽然DNN方法所恢复的波形与真实波形看似紧密贴合,但易出现幅值突然增大或者缩小的问题. 这可能是以下两个原因导致的:L2损失函数在低频处预测失准,矩形窗导致了频域能量的泄露. 而在图9(*e*)中,FLOW方法恢复语音波形质量较差,严重失去了语音的波形特征.

图10给出了连续丢三帧时的波形恢复效果,所截取的语音段为一段能量逐渐下降的浊音语音,共三帧语音(960样点). 由图10(*a*)可以看出,与单帧恢复相似,所提方法的波形恢复效果会随时序而降低;在最后一帧中,虽然恢复波形已经与真实波形不再匹配,但仍然具有一定的周期性. 由图10(*b*)可以看出,TFGAN的恢复效果随时序下降得更为严重,这是因为该方法对波形的周期性恢复不佳. 虽然所提方法恢复语音的周期性要优于TFGAN,但此时两种方法所恢复的波形均比较糟糕,故两者的PLC性能应较为接近. 由图10(*c*)可看出,WDNN方法通过波形相似叠加法恢复了语音相位信息,在长时丢包情况下具有一定优势. 由图10(*d*)可以看出,DNN方法对波形的恢复相较于单帧丢失情况更为糟糕,这是因为在多帧丢失情况下,输入特征包含网络预测的波形成分,而DNN难以从不稳定的特征中来捕捉长时相关性. 由图10(*e*)可以看出,FLOW方法难以恢复出有效的语音波形,故此方法性能较差. FLOW方法丢包隐藏性能较差的主要原因是当无帧叠接时,由全连接层构成的神经网络预测器难以有效预测梅尔特征,从而使神经声码器难以预测丢失的语音波形.

此外,图11给出了丢包率为20%情况下各方法对语句波形恢复的整体对比. 其中,图11(*a*)为未丢包的真实语音,图11(*b*)为丢包后以零值填补的语音. 由图11(*b*)可以看出,语句中既存在十分明显的长时损失,又存在间断的短时损失. 由图10(*c*)可以看出,所提方法在浊音段中可有效恢复真实波形包络,且对比3.1 s左右的波形可以发现,所提方法恢复的幅值包络与真实语音最为接近. 然而,所提方法在清音段中的恢复效果并不理想,如语段末尾的清音段. 这是因为清音在短时内不具备周期性,使得网络难以准确学习到清音的映射关系,从而使得生成的语音类似于静音帧. 此外,由图11(*d*)~(*g*)可以看出,TFGAN恢复效果与所提方法相近,而DNN、WDNN和FLOW的方法要明显差于所提方法.

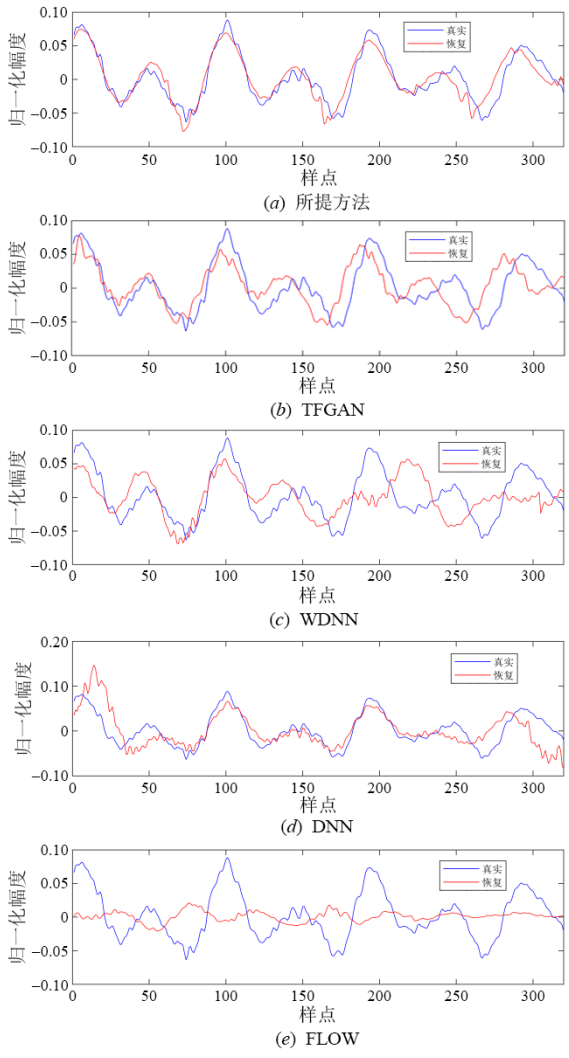

图9　单帧丢失下所恢复的语音波形对比

3.2.2　客观评测指标的对比

表1、表2和表3分别给出了不同丢包率情况下的PESQ、STOI和SNR评分,三者均为评分越高性能越优.

首先,从表1、表2和表3中可以看出,所提方法的PESQ、STOI和SNR评分均为最高(本文各表中的最优

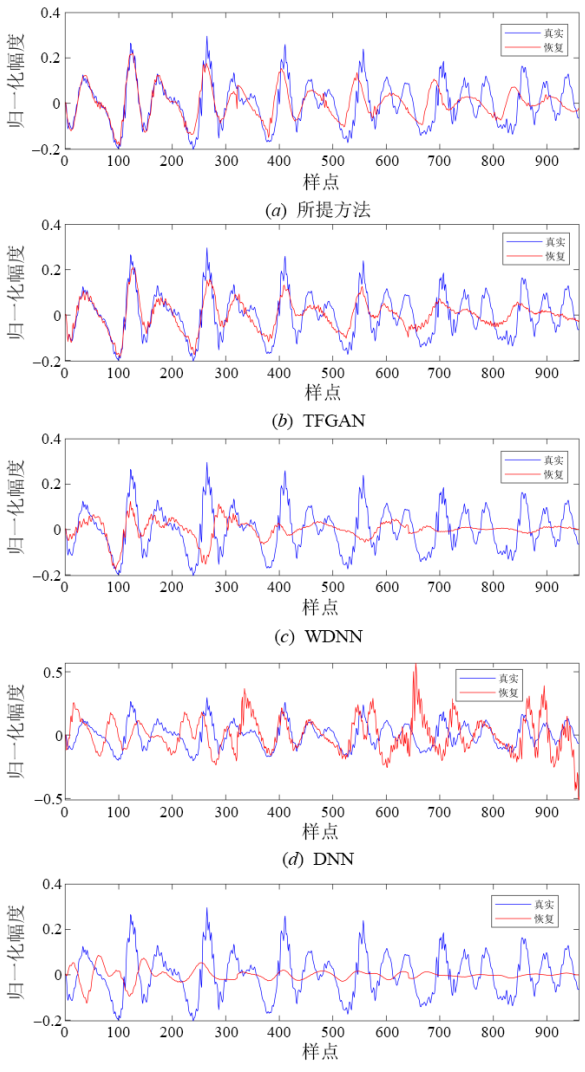

图10 多帧丢失下所恢复的语音波形对比

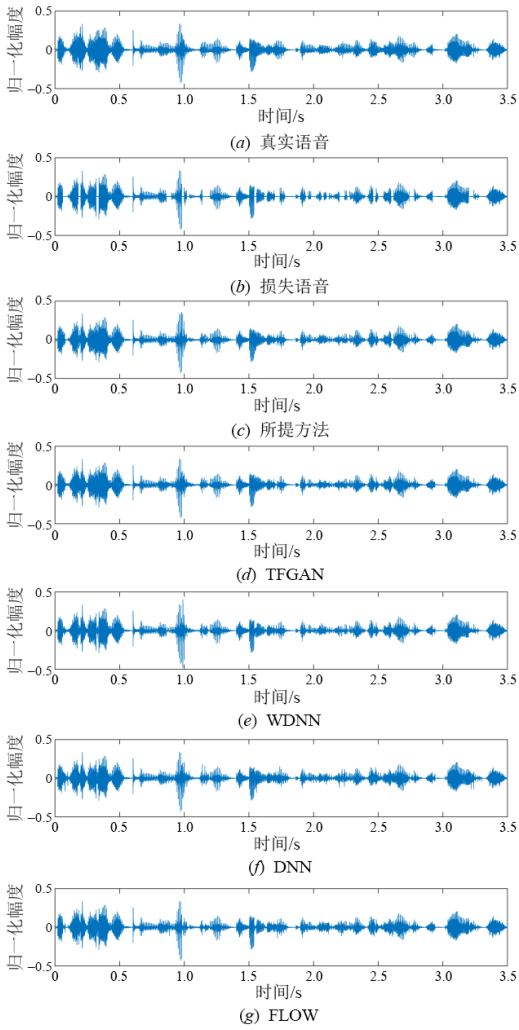

图11 20% 丢包率下所恢复语句的波形对比

结果由粗体数据表示),表明相较于参考方法而言,所提方法具有最优的客观评测结果. 其次,随着丢包率的上升,所提方法对PESQ和STOI的提升幅度明显增加. 即使是在20%丢包率的情况下,所提方法仍能将PESQ和STOI分别提升至2.6及0.93以上. 此外,由于恢复语音中包含着损伤噪声,因此在SNR评测指标下,FLOW、DNN和WDNN的性能甚至差于基准方法.

此外,表4给出了L1范数的梅尔损失,以评价所恢复语音的频域准确度. 梅尔损失越小,则恢复语音的频域准确度越高. 由表4中的结果可以看出,所提方法的梅尔损失最低,表明其频域准确度最高. 表明所提算法可以将时频特征高效结合,并使用TADE层将对齐的时频特征进行再耦合重构,从而恢复出具有周期性与局部细节(也就是高频成分)的语音波形. 而最具代表的TFGAN方法的生成网络以时域波形作为输入,以膨胀卷积层作为主要结构,故对丢包语音频域信息的恢复主要依赖于判别网络的性能,即当生成网络和判别网络间对特征域感知存在一定偏差时,就会导致丢包隐藏性能退化.

表1 不同丢包概率下的PESQ结果

概率	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
5%	2.89	3.00	3.06	3.17	3.40	**3.56**
10%	2.13	2.29	2.36	2.49	2.76	**2.97**
15%	1.73	1.93	2.00	2.11	2.40	**2.58**
20%	1.51	1.692	1.78	1.87	2.13	**2.33**
50%	1.13	1.27	1.31	1.36	1.50	**1.62**

表2 不同丢包率下的STOI结果

概率	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
5%	0.958 2	0.969 1	0.965 3	0.974 4	0.979 5	**0.981 3**
10%	0.916 8	0.941 1	0.938 4	0.949 5	0.963 3	**0.965 5**
15%	0.876 4	0.915 5	0.913 8	0.927 2	0.950 1	**0.951 5**
20%	0.835 9	0.886 8	0.891 6	0.904 8	0.936 0	**0.937 9**
50%	0.791 1	0.747 4	0.772 3	0.779 2	0.847 1	**0.855 4**

表3　不同丢包率下的SNR结果　　单位:dB

概率	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
5%	13.61	10.56	11.47	12.59	13.99	**14.53**
10%	9.91	8.03	7.99	9.53	10.40	**10.96**
15%	7.95	6.03	6.15	7.02	8.61	**9.02**
20%	6.31	4.62	4.71	5.73	7.08	**7.56**
50%	0.02	0.03	0.08	0.05	2.00	**2.52**

表4　不同丢包率下的L1范数梅尔损失结果

概率	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
5%	0.031	0.031	0.032	0.027	0.023	**0.022**
10%	0.064	0.053	0.062	0.055	0.046	**0.043**
15%	0.098	0.081	0.090	0.076	0.066	**0.062**
20%	0.138	0.113	0.118	0.112	0.086	**0.082**
50%	0.445	0.238	0.263	0.221	0.198	**0.190**

接下来,表5、表6和表7分别给出了不同丢包率情况下的PESQ、STOI和SNR评分结果. 可以看出,所提方法在丢单帧、丢双帧、丢三帧情况下的客观评测结果,所提方法优于参考方法. 从表5中可以看出,所提方法在单帧丢失情况下,性能远远强于其他方法,但是在连续丢失情况下,提升幅度则会相对较低. 表明恢复的丢失帧仍然会对所提方法的性能产生影响.

上述波形恢复对比和客观评分对比均表明,在不同的丢包情况下,所提方法较参考方法有着更优的丢包隐藏性能. 但是,为了评测所提方法运算效率,表8给出了在NVIDIA 2060super中进行推理的实时因子(越大表示运算效率越高). 由表8中可以看出,所提方法的运算效率差于DNN、WDNN、TFGAN这3种方法,但优于FLOW方法. 表明所提方法在提升PLC性能的同时,增大了计算复杂度.

表5　固定丢包模式下的PESQ评分

模式	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
单帧	2.24	2.05	2.21	2.24	2.53	**2.74**
双帧	1.32	1.57	1.69	1.77	1.88	**2.00**
三帧	1.22	1.38	1.50	1.58	1.59	**1.67**

表6　固定丢包模式下的STOI评分

模式	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
单帧	0.931 3	0.939 0	0.936 7	0.946 52	0.961 6	**0.967 0**
双帧	0.851 2	0.894 9	0.868 4	0.903 48	0.915 3	**0.929 0**
三帧	0.780 9	0.855 2	0.779 6	0.864 66	0.851 6	**0.894 4**

表7　固定丢包模式下的SNR评分　　单位:dB

模式	丢包	FLOW	DNN	WDNN	TFGAN	所提方法
单帧	9.92	9.13	9.11	9.01	10.52	**10.92**
双帧	6.92	6.29	6.27	6.28	7.19	**7.38**
三帧	5.11	4.60	4.65	4.67	5.39	**5.44**

表8　实时因子

FLOW	DNN	WDNN	TFGAN	所提方法
0.12	**19.98**	9.47	1.82	0.55

4　结论

本文提出了一种基于先验梅尔谱和神经声码器的语音PLC方法. 该方法联合时域波形编码器和频域梅尔编码器,可有效提取深层时频特征,并通过基于TADE残差块所构建的双输入-双输出神经声码器来预测丢包语音. 所提方法联合幅度与相位的损失函数,并通过多尺度信息来辅助丢包信息映射关系的学习. 实验结果验证了所提方法在不同丢包情况下的有效性. 此外,如何进一步提升运算效率,是需要进一步考虑的问题. 未来工作可考虑引入对抗性学习,从而降低网络参量,提升运算效率.

参考文献

[1] LEE B K, CHANG J H. Packet loss concealment based on deep neural networks for digital speech transmission[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(2): 378-387.

[2] NGUYEN V A, NGUYEN A H T, KHONG A W H. Improving performance of real-time full-band blind packet-loss concealment with predictive network[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.

[3] XUE H Y, PENG X L, LU Y. Contrast-PLC: Contrastive learning for packet loss concealment[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.

[4] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 17022-17033.

[5] ZHOU Y, BAO C C, HUANG J W, et al. A neural vocoder based packet loss concealment algorithm[C]//2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Piscataway: IEEE, 2022: 1-5.

[6] OU L L, CHEN Y P. Concealing audio packet loss using frequency-consistent generative adversarial networks[C]// 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). Piscataway: IEEE, 2022: 826-831.

[7] PRENGER R, VALLE R, CATANZARO B. Waveglow: A flow-based generative network for speech synthesis[C]// 2019 IEEE International Conference on Acoustics, Speech

and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 3617-3621.

[8] KIM J H, LEE S H, LEE J H, et al. Fre-GAN: Adversarial frequency-consistent audio synthesis[C]//Interspeech 2021. Brno: ISCA, 2021: 2197-2201.

[9] LOTFIDERESHGI R, GOURNAY P. Speech prediction using an adaptive recurrent neural network with application to packet loss concealment[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 5394-5398.

[10] MOHAMED M M, SCHULLER B W. ConcealNet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition[EB/OL]. (2005) [2023]. http://arxiv.org/abs/2005.07777.

[11] LIN J, WANG Y, KALGAONKAR K, et al. A time-domain convolutional recurrent network for packet loss concealment[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 7148-7152.

[12] VERMA P, MEZZA A I, CHAFE C, et al. A deep learning approach for low-latency packet loss concealment of audio signals in networked music performance applications[C]//2020 27th Conference of Open Innovations Association (FRUCT). Piscataway: IEEE, 2020: 268-275.

[13] STIMBERG F, NAREST A, BAZZICA A, et al. WaveNetEQ—Packet loss concealment with WaveRNN [C]//2020 54th Asilomar Conference on Signals, Systems, and Computers. Piscataway: IEEE, 2020: 672-676.

[14] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[EB/OL]. (2018) [2023]. http://arxiv.org/abs/1802.08435.

[15] WANG J, GUAN Y S, ZHENG C S, et al. A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission[J]. The Journal of the Acoustical Society of America, 2021, 150(4): 2577-2588.

[16] MUSTAFA A, PIA N, FUCHS G. StyleMelGAN: an efficient high-fidelity adversarial vocoder with temporal adaptive normalization[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 6034-6038.

[17] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 2332-2341.

[18] ZHANG W D, ZHU J W, TAI Y, et al. Context-aware image inpainting with learned semantic priors[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1-7.

[19] PANDEY A, WANG D L. TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 6875-6879.

[20] YANG G, YANG S, LIU K, et al. Multi-band melgan: Faster waveform generation for high-quality text-to-speech[C]//2021 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE, 2021: 492-498.

[21] JI Q, BAO C C, CUI Z H. Packet loss concealment based on phase correction and deep neural network[J]. Applied Sciences, 2022, 12(19): 9721.

[22] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2015: 5206-5210.

[23] MUSHKIN M, BAR-DAVID I. Capacity and coding for the gilbert-elliott channels[J]. IEEE Transactions on Information Theory, 1989, 35(6): 1277-1290.

作者简介

黄晋维 男,1998年出生,河北秦皇岛人. 北京工业大学硕士研究生. 研究方向为语音丢包隐藏.
E-mail: 1054061097@qq.com

鲍长春 男,1965年出生,内蒙古赤峰人. 中国电子学会会士. 现为北京工业大学教授、博士生导师. 主要研究方向为语音编码与语音增强. 中国电子学会会员编号:E190000343F.
E-mail: chchbao@bjut.edu.cn

周 静 男,1993年出生,四川广安人. 现为北京工业大学博士研究生. 主要研究方向为语音增强.
E-mail: zhoujing@emails.bjut.edu.cn