计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(11): 2725−2736, 2025

基于梅尔谱与压缩激励加权量化的语音神经编解码方法

周俊佐[1] 易江燕[1] 陶建华[2,3] 任 勇[1] 汪 涛[1]
[1](中国科学院自动化研究所 北京 100190)
[2](清华大学自动化系 北京 100084)
[3](北京信息科学与技术国家研究中心(清华大学) 北京 100084)
(zhoujunzuo2023@ia.ac.cn)

Neural Speech Codec Method Based on Mel Spectrogram and Squeeze-Excitation-Weighted Quantization

Zhou Junzuo[1], Yi Jiangyan[1], Tao Jianhua[2,3], Ren Yong[1], and Wang Tao[1]
[1] (*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2] (*Department of Automation*, *Tsinghua University*, *Beijing* 100084)
[3] (*Beijing National Research Center for Information Science and Technology* (*Tsinghua University*), *Beijing* 100084)

Abstract At present, end-to-end speech neural codecs, represented by SoundStream, have demonstrated outstanding performance in reconstructed speech quality. However, these methods require extensive convolutional computations, leading to lengthy encoding times. To address this issue, we introduce a neural speech codec method based on Mel spectrogram and squeeze-excitation-weighted quantization. This method aims to maintain high speech perceptual quality while reducing computational costs and increasing operational speed, thereby minimizing latency. Specifically, we utilize Mel spectrogram features as input, capitalize on the temporal compression properties during Mel spectrogram extraction, and combine a lower-layer convolutional encoder to simplify the computation process. Additionally, inspired by squeezed excitation network concepts, we extract excitation weights for each dimension of the output features from the encoder's final layer. These weights are used as the weighting coefficients for each dimension of the compressed features when calculating codebook distances in the quantizer, thus enabling the learning of correlations between features and enhancing the performance of quantization. Experimental results on the LibriTTS and VCTK datasets indicate that this method significantly enhances the computational speed of the encoder and improves the reconstructed speech quality at lower bit rates (≤3 Kbps). For instance, at a bitrate of 1.5 Kbps, the real-time factor (RTF) of encoding computations can increase by up to 4.6 times. Regarding perceptual quality, at a bitrate of 0.75 Kbps, objective metrics such as short-time objective intelligibility (STOI) and virtual speech quality objective listener (VISQOL) show an average improvement of 8.72% compared with the baseline. Additionally, ablation studies not only demonstrate that the optimization effect of compressed excitation weight methods is inversely correlated with bitrate, but also reveal that, compared with the periodic activation function Snake, the Relu activation function can significantly speed up processing while maintaining comparable speech perceptual quality.

Key words speech codec; Mel-spectrogram; squeeze-and-excitation networks; residual vector quantization; generative adversarial network

收稿日期: 2024-05-21; **修回日期**: 2025-03-27

基金项目: 中国科学院战略性先导科技专项(XDB0500103); 国家自然科学基金项目(62322120, U21B2010, 62306316, 62206278)
This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0500103) and the National Natural Science Foundation of China (62322120, U21B2010, 62306316, 62206278).

通信作者: 易江燕(jiangyan.yi@nlpr.ia.ac.cn)

摘 要 目前,以 SoundStream 等为代表的端到端语音神经编解码器在重建语音感知质量方面展现了优异性能. 然而,这些方法需要大量的卷积计算,从而导致较长的编码时间. 为缓解上述问题,提出基于梅尔谱和压缩激励加权量化的神经语音编解码方法. 该方法旨在保持较高语音感知质量的同时降低计算代价,加快运行速度,从而减少时延. 具体而言,采用梅尔谱特征作为输入,借助梅尔谱提取过程中时域压缩的性质,并结合低层卷积编码器以简化运算过程. 此外,借鉴压缩激励网络思想,提取了编码器最后一层输出特征各维度的激励权重,将其作为量化器中计算码本距离时压缩特征各维度的权重系数,由此学习特征间的相关性并优化量化性能. 在 LibriTTS 和 VCTK 数据集上的实验结果表明,该方法显著提升了编码器计算速度,且能在较低比特率时(≤3 Kbps)提升重建语音质量. 以比特率 1.5 Kbps 时为例,编码计算实时率(real-time factor, RTF)最多可提升 4.6 倍. 对于感知质量,以 0.75 Kbps 为例,短时客观可懂度(short-time objective intelligibility, STOI)、虚拟语音质量客观评估(virtual speech quality objective listener, VISQOL)等客观指标相较基线平均可提升 8.72%. 此外,消融实验不仅表明压缩激励权重方法的优化效果与比特率呈反相关,而且发现 Relu 激活函数相较周期性质激活函数 Snake 而言,在语音感知质量相当的情况下,能大量加快运行速度.

关键词 语音编解码;梅尔谱图;压缩激励网络;残差矢量量化;生成对抗网络

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440329 **CSTR:** 32373.14.issn1000-1239.202440329

语音编解码是移动通信、互联网通信等众多领域的重要技术之一[1-5]. 语音信号在传输过程中经历了发送端的信号处理和特征提取,随后通过数据压缩传输至接收端,最终接收端通过解码器将恢复的特征解码重建成语音波形,这构成了典型的语音编解码系统. 其包括编码器、量化器和解码器 3 个模块. 传统的语音编解码器大多基于数字信号处理方法,针对不同适用条件结合一些专家知识精心设计和选择构建,例如利用心理声学和语音合成等领域的知识来提高编码效率等[6-10]. 然而,这些方法不仅适用性受到限制,其生成的语音质量也有限[11]. 以神经网络为代表的机器学习方法最初仅被应用于语音降噪等编解码的后处理阶段[12-13]. 随着“数据驱动”模式下深度学习技术的进步,这些方法从辅助优化逐渐转变为编解码器本身的核心组件之一,不仅便于设计,而且展现了出色的性能,在不同网络带宽条件下均能解码得到较高质量的重建语音[14-21].

由于解码过程与语音合成领域的声码器同属于波形的生成过程,因此,早期的工作尝试直接用神经声码器模型实现语音解码[22-23]. 2023 年,Petermann 等人[17] 使用 WaveNet[24] 作为语音生成模型,利用其自回归生成能力,显著提高了解码语音的质量. 2018 年,Kankanahalli[25] 实现了神经语音编解码系统的端到端优化,该系统无需手动特征工程,全面优化宽带语音编码管道中的各个步骤(包括压缩、量化和解压缩),大幅提升了系统的适应性. 2019 年,Gârbacea 等人[22] 同时采用基于矢量量化变分自动编码 VQ-VAE[26] 和 WaveNet 解码器的神经网络架构进行语音编解码. VQ-VAE 通过将编码特征离散化以完成数据压缩过程,提升了量化器的性能. 2021 年,Kleijn 等人[23] 推出了 Lyra,对经过 KLT 变换(Karhunen–Loève transform, KLT)的梅尔谱进行矢量量化,并采用 WaveGRU 作为解码器,实现了高效的波形恢复. Zeghidour 等人[11] 开发的 SoundStream 利用生成对抗网络(generative adversarial network, GAN)的模式进行对抗性训练,并引入残差矢量量化的技巧,使单个模型能够处理不同的比特率,从而适用于各种网络带宽. 此外,这还是一种对波形采用全卷积编解码器的端到端编解码系统. 2022 年,Défossez 等人[27] 开发的 Encodec 在 SoundStream 基础上更进一步,引入了轻量级的语言模型、熵编码等技术进行优化,且可以通过分别处理左右声道来压缩立体声音频. 而 Ratnarajah 等人[28] 则专为高效压缩多声道语音提出 M3-AUDIODEC,展示了神经编解码器在多声道音频编码上的显著提升. 2023 年,Yang 等人[29] 在 SoundStream 的基础上提出 Hifi-codec,采用分组残差矢量量化方法,进一步提升了重建语音质量. 上述语音神经编解码方法普遍采用卷积编码器直接从波形中学习特征,获得了很好的重建语音质量,但卷积编码器通过调节卷积单元中下采样倍数完成时域帧的逐步压缩,在提取优秀潜在特征的同时,以一定的卷积计算量为代价.

梅尔谱特征作为声学领域的经典手工特征,符

合人耳听觉的感知特性[30-31]. 尽管过去的一些编解码器方法[23]曾利用梅尔谱图作为编码器输入，但其缺乏量化方法优化[32]，相应的编解码设计使得相较SoundStream等波形方法重建音频质量有所不足.

为了在保持重建语音质量的同时降低语音编码计算开销，本文提出了基于梅尔谱与压缩激励加权量化的语音神经编解码方法. 由于梅尔谱的提取过程伴随着时域帧的逐步压缩，因此，本文对梅尔谱运用卷积编码器进行特征提取，采用的卷积层数更少；降低计算规模，从而减少时延，以平衡更加多样的用户需求和更为稳定的用户体验. 对于量化器，由于卷积编码器各个输出通道信息量具有差异[33]，该不均匀性将影响矢量量化过程中各个通道维度的重要程度. 因此，本文借鉴了压缩激励网络（squeeze-and-excitation networks, SENet）思想[26]，提取编码器最后一层各维度的激励权重，使其作为量化器中计算码本距离时各维度的权重系数. 即通过编码器的自适应学习捕捉特征之间的相互依赖性，减少冗余信息[26]，从而确定对量化器更重要的通道，提升量化性能.

本文在LibriTTS[34]和VCTK[35]数据集上进行实验，结果表明使用梅尔谱图作为输入特征，结合低层卷积编码器，且采用压缩激励的方法利用梅尔谱图经过低层编码器输出特征信息量的不均匀性，不仅普遍可以降低时延，并且能在较低比特率环境中（$\leqslant$3 Kbps）提升感知质量. 以感知质量较好的基线Hifi-codec[29]为基准，比特率为1.5 Kbps时，本文方法编码计算的实时率（real time factor, RTF）最多可提升至4.6倍. 此外，较低比特率的情况下，本文方法的感知质量超越所有基准模型. 特别是在0.75 Kbps时，与最佳的Encodec[27]相比，短时客观可懂度（short-time objective intelligibility, STOI）[36]和虚拟语音质量客观评估（virtual speech quality objective listener, VISQOL）[37]的平均提升率为8.72%.

本文探究了不同比特率下各压缩率与码本数目组合的消融实验. 发现比特率提高时，增加压缩率不利于模型的训练效果和性能，应当同时兼顾码本数目进行参数选择. 此外，通过对不同比特率和不同输入特征进行压缩激励权重的消融实验，本文探索了压缩激励权重方法的性质. 其优化效果与比特率呈反相关，且相较波形特征，更适用于梅尔谱编解码器. 最后，本文还对神经解码器网络中的激活函数进行了消融实验，通过比较Relu激活函数和具有周期特性的Snake激活函数[38-39]，研究结果表明：在保持语音质量相当的情况下，Relu激活函数能显著提高运行速度. 特别是在低比特率环境下（$\leqslant$3 Kbps），Relu函数在保持语音质量几乎不变的同时，实时率（RTF）均能提升2倍以上，因此更适合实际需求.

1 相关工作

本节从梅尔谱图、压缩激励网络、残差矢量量化3个方面介绍相关工作.

1.1 梅尔谱图

梅尔谱图作为一种语音处理中常用的前端特征，其原理根植于人耳对频率感知的非线性特性，特别是对低频信号更为敏感. 为了模拟这种听觉特性，人们引入了梅尔标度，这种非线性对数变换针对频率标度进行转换，将语音语谱图的频率维度应用梅尔标度即为梅尔谱图[30,40]. 该转换过程涉及语谱图与多个梅尔滤波器相乘，而语音的语谱图则源自对语音序列进行短时傅里叶变换并提取幅度谱. 在短时傅里叶变换中，帧移操作对时域信号进行压缩，因此梅尔谱图的时域帧数远远少于原始波形数据点，为后续的编解码过程提供了便利.

1.2 压缩激励网络

众所周知，卷积网络能够通过融合各层局部感受野中的空间和信道信息来构造信息特征. 压缩激励网络（SENet）[26]的本质是注意力机制在卷积网络领域中的应用. 作为一个简化的结构，它可以插入卷积网络中，并对通道关系之间的相互依赖性加以关注，此前在视觉等众多领域其适用性都得到了验证[41-42]，这种子结构也适用于语音场景[43]. 具体而言，通过对卷积的每个输出通道预测一个常数权重，并对该通道加权. 这种注意力机制以轻量级的计算代价让模型更偏向信号的最具信息量的部分.

1.3 残差矢量量化

在语音编解码器的量化环节中，编码器提取的特征需要被压缩，矢量量化是常用的方法之一[22]. 通过建立一张码本，将连续的特征空间转化为离散的token. 采用的码本数目越多，解码器恢复的语音质量就越高，但更多的码本也会消耗传输时的网络带宽. 因此，对于神经编解码器，不同的网络带宽环境需要专门训练各自数目的码本，这大大增加了使用和训练的成本. 对此，人们设计了残差矢量量化（residual vector quantization, RVQ）[11]结构，如图1所示. 该结构通过将上一量化层$q_i(\cdot)$离散化后与其输入标准值y_i的残差作为下一量化层$q_{i+1}(\cdot)$的输入，最终累加各量化层离散化后的token作为输入连续特征x的重建值x'，即

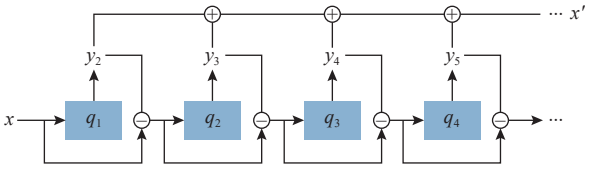


Fig. 1 The structure of residual vector quantization

图 1 残差矢量量化结构

$$y_1 = x, \tag{1}$$

$$y_{i+1} = q_{i+1}(y_i - q_i(y_i)), \tag{2}$$

$$x' = \sum_{i=1}^{N} y_i, \tag{3}$$

其中量化层$q_i(\cdot)$是在码本空间中选择距离最近的条目, 由于传输时仅记录条目的序号, 从而实现信息压缩的效果. 上述过程, 残差堆积层数N越多, 离散化精度也越高. 此外, 残差结构具有相当的灵活性. 可以 1 次训练多层码本, 而推理时根据带宽限制只选用前几层($n \leqslant N$)进行累积重建, 大大扩宽了应用范围.

为了加强码本空间的表示能力, 可以采用对同一量化层码本拆分成多组的形式, 即分组矢量残差量化(group residual vector quantization, GRVQ)方法[29], 进一步深化了量化能力.

本文工作中将使用压缩激励机制对量化部分进行优化.

2 方 法

2.1 整体框架

本文采取的系统整体框架如图 2 所示, 生成部分分为编码器、量化器和解码器 3 个组件. 此外, 还包括鉴别器部分. 语音波形提取梅尔谱后, 输入卷积编码器. 卷积编码器遵循与 Hifi-codec 相似的结构[29], 但采用较少的卷积模块数目, 即由首尾的 1 维卷积层和中间的卷积单元组成. 每个卷积单元由 3 个残差单元[44] 和 1 个下采样层组成[11]. 解码器则由首尾的 1 维卷积层和中间 4 个卷积单元组成[11].

量化器接受编码器的输出特征和压缩激励权重. 在编码器压缩率给定的情况下, 灵活调节码本数目以实现不同比特率下的调节. 在比特率相对较高时(需要 4 个及以上的码本数目时), 采取 Hifi-codec 的策略, 即 GRVQ, 为了训练的稳定, 组数设置为 2. 而码本数目低于 4 个时, 则取消分组, 退化为普通残差量化.

对于鉴别器, 本实验参考声码器 Hifi-GAN 的策略, 采用对语音波形直接检验的多尺度鉴别器(multi-scale discriminator, MSD)与多周期鉴别器(multi-period discriminator, MPD)[45]. 前者是通过核大小为 4 的步幅平均池化(strided average pooling)层对波形序列进行 2 倍和 4 倍的下采样之后的语音序列进行操作. 后者是将 1 维的原始语音序列T按照固定周期p抽样为长为p、宽为T/p的 2 维数据, 然后对重塑后的数据应用 2 维卷积. 鉴别器最终输出判定情况与中间层检测时的特征图, 以用于后续生成对抗损失和特征损失计算.

2.2 梅尔编解码与压缩激励加权量化

编解码系统中压缩率是从波形采样点到进入量化器前时间帧的下采样倍数[27], 由于本文采用梅尔谱特征, 因此帧移(*hop_size*)本身具有压缩的作用. 编码器卷积单元中下采样层卷积核的采样步幅(stride)则承担了剩余的压缩效果. 解码器卷积单元中所有的上采样倍数均来自卷积核的采样步幅. 因此需要

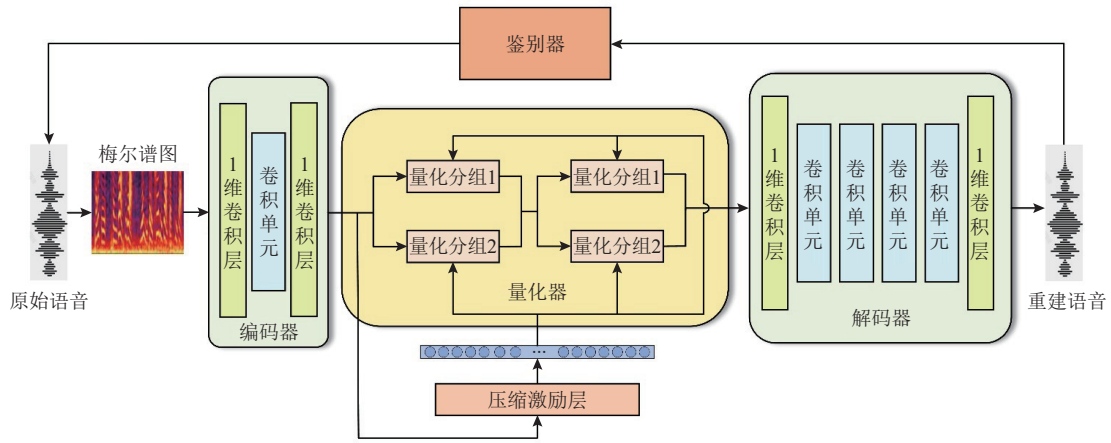


Fig. 2 Structure diagram of neural speech codec method based on Mel spectrogram and squeeze-excitation-weighted quantization

图 2 基于梅尔谱与压缩激励加权量化的语音神经编解码方法结构图

满足：

$$r = hop\_size \times \prod_i S_{\mathrm{enc},i} = \prod_i S_{\mathrm{dec},i,} \tag{4}$$

其中r是压缩率，$S_{\mathrm{enc},i}$和 $S_{\mathrm{dec},i}$分别指编码器和解码器第i个卷积单元中下采样层卷积核的采样步幅.

式(4)的成立使得编码器端所需的卷积单元数量相较于直接从波形中进行下采样时减少. 尽管计算梅尔谱图的过程包含短时傅里叶以及梅尔滤波器相乘的运算，这也会带来一定的时间消耗，但其计算量低于具有相同压缩程度的卷积单元所进行的卷积运算，后续的实验证明了这一点.

对于量化过程中的压缩激励权重. 本文希望可以通过动态地分配通道维度的权重来提高网络表达能力，具体而言，我们对卷积编码器的最后一层进行全局平均池化，全局空间信息被聚合到一个信道描述符中，即压缩过程. 继而通过先后 2 个全连接及中间的 Relu 层，接连 Sigmoid 函数，该步骤能学习被挤压通道之间的非线性交互，即激励过程[26]. 最终得到与编码器输出通道数 C_{enc} 相同的 1 维向量作为各通道评价分数($Score_1, Score_2, \cdots, Score_{C_{\mathrm{enc}}}$). 如图 2 所示，作为通道注意力权重，该分数并非与编码器最后 1 层相乘，而是保留进入量化器模块. 尽管 GRVQ 中分组的操作使得量化时实际的码本嵌入宽度与待查询的编码器输出通道数C_{enc}并不相等. 但是每轮残差后，各个分组码本会合并，合并后的总码本嵌入宽度 $W = C_{\mathrm{enc}}$. 某帧($x_1, x_2, \cdots, x_{C_{\mathrm{enc}}}$)查询码本时，需要计算该帧与码本中各条目($c_1, c_2, \cdots, c_W$)的距离并选取最近的条目，我们定义该距离为

$$d = \sum_{i=1}^{W} Score_i \times (x_i - c_i)^2. \tag{5}$$

2.3 损失函数

本文实验整体框架采用 GAN 结构，分为鉴别器损失和生成器损失 2 部分[46].

对于包含多个子鉴别器的鉴别器组，K是鉴别器数目，D_i代表多周期鉴别器(MPD)或多尺度鉴别器(MSD)的第i个子鉴别器，x 为原始波形样本，$\hat{x}$ 即生成器输出的波形，定义鉴别器的对抗损失：

$$L_{\mathrm{adv_D}} = \frac{1}{K}\sum_{i=1}^{K}[(D_i(x)-1)^2 + D(\hat{x})^2]. \tag{6}$$

同时，生成器的对抗损失：

$$L_{\mathrm{adv_G}} = \frac{1}{K}\sum_{i=1}^{K}(D_i(\hat{x})-1)^2. \tag{7}$$

定义生成部分的重建波形的梅尔谱图$\hat{m}$与原始样本梅尔谱图m之间的 L1 距离为重建损失(reconstruction loss) L_{rec}：

$$L_{\mathrm{rec}} = \|m - \hat{m}\|. \tag{8}$$

此外，对于鉴别器D_i的第 l 个中间层，w为其输入的中间特征，$\hat{w}$为该层输出，我们定义总体的 E 特征损失(feature loss) L_{feat}[27]：

$$L_{\mathrm{feat}} = \frac{1}{KL}\sum_{i=1}^{K}\sum_{l=1}^{L}\frac{\|D_i^l(w) - D_i^l(\hat{w})\|}{mean\left(\|D_K^l(w)\|\right)}. \tag{9}$$

量化过程中，第n组第i个量化器 $q_{n,i}$，对于其待查帧 $z_{n,i}$，定义承诺损失(commitment loss) L_{c}[27]：

$$L_{\mathrm{c}} = \sum_{n,i}\|z_{n,i} - q_{n,i}(z_{n,i})\|_2^2. \tag{10}$$

综上所述，整体损失定义为

$$L_{\mathrm{total}} = \lambda_{\mathrm{feat}}L_{\mathrm{feat}} + \lambda_{\mathrm{c}}L_{\mathrm{c}} + \lambda_{\mathrm{adv_G}}L_{\mathrm{adv_G}} + \lambda_{\mathrm{adv_D}}L_{\mathrm{adv_D}} + \lambda_{\mathrm{rec}}L_{\mathrm{rec}}, \tag{11}$$

其中λ_{feat}, λ_{c}, $\lambda_{\mathrm{adv_G}}$, $\lambda_{\mathrm{adv_D}}$, λ_{rec}是超参数.

3 实验设置

3.1 训练数据集与测试数据集

本实验采用的训练数据集是 LibriTTS 英文多说话人数据集[34]，该数据集由 24 kHz 的 2 456 位说话人组成，总计持续时长为 585 h. 我们的评估数据集源自 VCTK[22]，该数据集与 LibriTTS 一样同属于英文多说话人数据集. 它包含 110 位英语多说话人录制的 48 kHz 语音. 我们从中随机抽取 100 条语音，并将其降采样为 24 kHz.

3.2 实验细节与基线

由于神经编解码器的压缩率、码本数量和其实现的比特率之间相互制约，确定其中 2 个参数后，剩余的参数也随之确定. 码本数量和压缩率均将影响模型的训练过程. 为了调节不同码本数目进行各比特率下的性能比较，主实验中将压缩率r统一设置为 320，梅尔谱图频域维度为 80，*hop_size* 设为 160，对于压缩率的影响，后续亦将设计消融实验探究.

此外，编码器卷积单元下采样倍数设置为 2. 解码器卷积单元上采样倍数逐个设置为 8, 5, 4, 2. 以编码器唯一的下采样卷积单元为例，其网络参数如表 1 所示. Conv 1 进行下采样，此外，Conv 2~3, Conv 4~5, …, Conv 18~19 等相邻 2 层卷积分别组成残差连接结构. 训练时，批大小取 16，学习率为 0.000 2. 超参数 λ_{feat}, λ_{c}, $\lambda_{\mathrm{adv_G}}$, $\lambda_{\mathrm{adv_D}}$, λ_{rec} 设置为 1, 10, 1, 1, 45. 编码器输出特征 512 维. 单个码本的索引范围为 [0, 1 023].

Table 1 Network Architecture of Convolutional Unit

表 1 卷积单元网络架构

卷积单元	卷积核数/步长/空洞数
Conv 1	4/2/1
Conv 2	11/1/1
Conv 3	11/1/1
Conv 4	11/1/3
Conv 5	11/1/1
Conv 6	11/1/5
Conv 7	11/1/1
Group Norm	
Conv 8	7/1/1
Conv 9	7/1/1
Conv 10	7/1/3
Conv 11	7/1/1
Conv 12	7/1/5
Conv 13	7/1/1
Group Norm	
Conv 14	3/1/1
Conv 15	3/1/1
Conv 16	3/1/3
Conv 17	3/1/1
Conv 18	3/1/5
Conv 19	3/1/1
Group Norm	

因此我们的模型和基线进行了比特率分别为 0.75 Kbps, 1.5 Kbps, 3 Kbps, 4.5 Kbps, 6 Kbps 的实验, 分别对应 1, 2, 4, 6, 8 个数目的码本. 所有实验均在单卡 2080Ti 上训练了 25 个 epoch, 测试时统一采用 CPU 运行(Intel® Xeon® CPU E5-2640 v4 @ 2.40 GHz). 本实验采取 Hifi-codec[20], Encodec[19], SoundStream[18] 模型作为基线, 其中 Encodec 和 SoundStream 采用 Yang 等人[29] 在 Hifi-codec 工作中重新实现的代码①. 除了上述神经编解码器, 本实验也与当前广泛应用的 Opus 进行了比较.

3.3 评价指标

本实验客观指标采用 RTF 衡量时延, 其定义为输入语音的时长与编码所需时间的比率[11]. 因此, 该指标与编码器的运行速度正相关. 评测中, 无论是本文方法还是基线, 编码时间均包含了从波形输入到生成编码索引的整个过程, 这也包括了波形到梅尔谱图的转换时间.

本实验采用 STOI[36]、VISQOL[37] 以及 WARP-Q[47] 评价语音的感知质量. 其中 WARP-Q 专门被提出应用于神经语音编解码器的评测指标, 本文采用其原始评分值进行比较. 本实验主观指标采用 MOS 评分, 该数据由 5 位志愿者对 100 条随机语音从 1~5 打分完成评测.

4 实验结果与讨论

4.1 实验结果

本文的所有测试实验和消融实验都在 VCTK 数据集上进行. 采用训练之外的同类别数据集进行测试的实验模式不仅符合实际的应用场景, 也对模型的泛化能力要求更高. 最终, 各神经编解码器在 VCTK 上的结果如表 2 所示, 可见本文方法在时间指标, 即编码器端的 RTF 相对所有同类神经编解码器在所测的任意比特率下都具有优势. 以基线里平均感知质

Table 2 Metrics Evaluated on the VCTK Dataset for Each Neural Codec

表 2 各神经编解码器在 VCTK 数据集上的评价指标

比特率/Kbps	码本数	模型	RTF_{enc} ↑	STOI ↑	WARP-Q ↓	VISQOL ↑	MOS ↑
0.75	1	SoundStream	18.817	0.703	2.988	2.134	1.80
		Encodec	19.188	0.736	2.917	2.226	2.28
		Hifi-codec	5.506	0.634	3.024	2.028	1.88
		本文模型	**36.656**	**0.747**	**2.604**	**2.536**	**2.40**
1.5	2	SoundStream	15.438	0.736	2.860	2.315	2.52
		Encodec	18.923	0.777	2.694	2.492	3.20
		Hifi-codec	6.507	0.751	2.709	2.520	2.84
		本文模型	**36.447**	**0.801**	**2.371**	**2.958**	**3.62**
3	4	SoundStream	15.760	0.757	2.798	2.407	2.36
		Encodec	18.550	0.79	2.625	2.584	3.16
		Hifi-codec	6.549	0.839	2.125	3.311	3.52
		本文模型	**35.632**	**0.846**	**1.997**	**3.334**	**3.88**
4.5	6	SoundStream	20.036	0.790	2.711	2.651	3.16
		Encodec	22.119	0.809	2.535	2.758	3.40
		Hifi-codec	6.562	**0.862**	1.912	**3.562**	3.62
		本文模型	**22.338**	0.859	**1.798**	3.492	**3.94**
6	8	SoundStream	20.145	0.799	2.502	2.694	3.50
		Encodec	21.423	0.818	1.852	2.799	3.44
		Hifi-codec	5.673	**0.878**	1.852	**3.665**	3.92
		本文模型	**23.493**	0.865	**1.776**	3.552	**3.98**

注: 黑体数值表示最优结果, "↑"表示数值越大越好, "↓"表示数值越小越好.

① https://github.com/yangdongchao/AcademiCodec

量较好的 Hifi-codec 作为基准，比特率为 1.5 Kbps 时编码实时率提升幅度最大，最多可提升 4.6 倍. 而比特率为 3 Kbps 时提升幅度较小，依然达到了 3.15 倍. 对于语音感知质量指标 STOI, WARP-Q, VISQOL，在比特率较低(0.75 Kbps, 1.5 Kbps, 3.0 Kbps)时表现良好，能超过所有基线模型，但比特率较高时感知质量却有所不足. 以极低比特率 0.75 Kbps 时为例，相较最好的 Encodec 客观感知质量平均提升 8.72%. 但比特率为 6 Kbps 时，相较 Hifi-codec, STOI, VISQOL 指标已经低于基线，3 种客观指标平均降低了 0.46%. 主观 MOS 的测定结果也进一步支持了该结论. 从残差矢量量化的原理分析，所用的码本越多，残差层数就越深，但是除去第 1 层接收的是编码器输出特征以外，后续接收的是前层量化的残差. 因此，由编码器计算的压缩激励权重所反映的通道之间的相关性和重要性对于后续残差层的效用有所削弱，甚至不再利于量化，后续的消融实验也进一步证实了这点，即对于重建语音感知质量，本文方法更有利于较低比特率，比特率较高时失效.

此外，表 3 展示了本文方法与目前实际场景中广泛使用的通用编解码方法 Opus 的对比结果. 由于 Opus 的最低支持带宽为 6 Kbps，并且其帧大小为 2.5 ms, 5 ms, 10 ms 等值，因此该比较实验存在一定的局限. 所有编解码器均在 6 Kbps 的带宽条件下进行测试，神经编解码器采用 8 码本，压缩率取 320 倍. 结果显示，Opus 在 STOI 指标上具有一些优势.

Table 3 Evaluation Comparison of Our Method and the General Codec Opus

表 3 本文方法与通用编解码器 Opus 的评测对比

模型	STOI ↑	WARP-Q ↓	VISQOL ↑	MOS ↑
Opus-10ms	0.703	2.610	2.476	3.32
Opus-5.0ms	0.858	2.383	3.037	3.58
Opus-2.5ms	**0.956**	1.976	3.475	3.86
本文模型	0.865	**1.776**	**3.552**	**3.98**

注：黑体数值表示最优结果，“↑”表示数值越大越好，“↓”表示数值越小越好.

本实验中原始语音和重建语音的梅尔谱图对比如图 3 所示. 以 6 Kbps 和 0.75 Kbps 时为例，分别代表了高比特率与低比特率场景. 将图 3(b)(c)、图 3(d)(e)与图 3(a)相比较，本文方法的重建语音梅尔谱图均与原始语音更为贴近，这意味着更相似的听感体验.

4.2 关于压缩率的消融探究

为了研究压缩率对所提出编解码器音质的影响，我们在比特率为 3 Kbps 和 6 Kbps 的条件下进行了消融实验. 实验中，较高的压缩率需要增加码本的数目. 具体的实验配置如表 4 所示，实验结果见表 5. 在 3 Kbps 条件下，随着压缩率和码本数的增加，语音质量指标均有所提升. 然而，在 6 Kbps 条件下，尽管压缩率和码本数增加，音质和语音可懂度却出现下降. 这表明在更高的比特率中，扩大压缩率导致的码本数目增大并不总是对模型的训练和最终性能产生积极的影响. 这一结果强调了在不同比特率下，需要谨

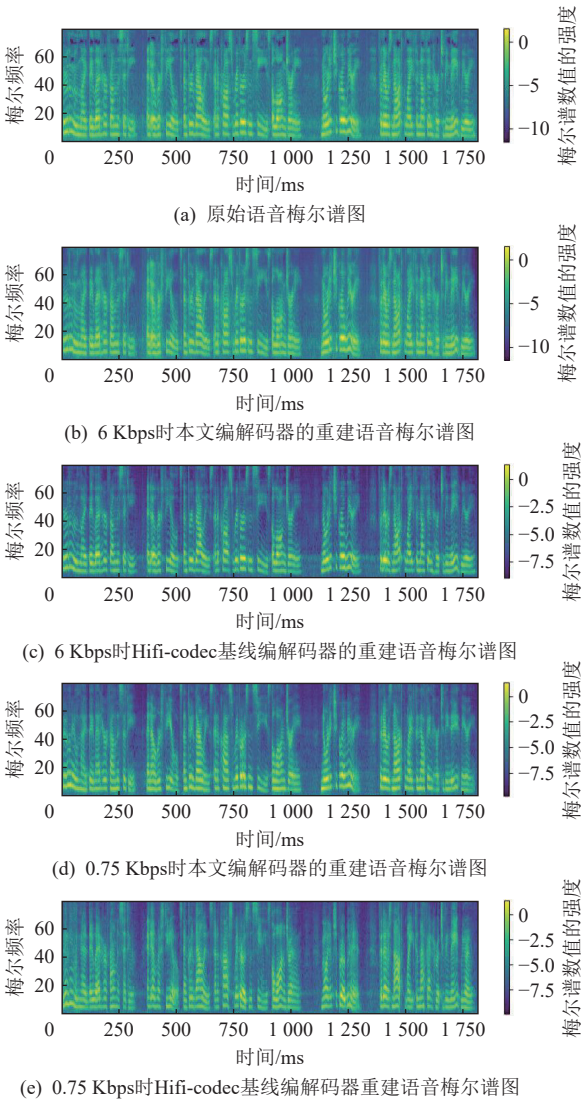


Fig. 3 Comparison of Mel spectrogram

图 3 梅尔谱图对比

Table 4 Experimental Configuration of Different Compression Rates

表 4 不同压缩率的实验配置

压缩率	编码器下采样倍数	解码器上采样倍数	解码器各层卷积核
180	2	5, 4, 4, 2	10, 8, 8, 4
320	2	8, 5, 3, 2	16, 11, 7, 4
640	2	10, 8, 4, 2	20, 16, 8, 4

Table 5 Ablation Experiments on Different Compression Ratios and Codebook Count

表 5 对于不同压缩率和码本数量的消融实验

比特率/Kbps	压缩率	码本数	STOI ↑	WARP-Q ↓	VISQOL↑
3	180	2	0.843	2.028	3.314
	320	4	0.846	1.997	3.334
	640	8	**0.854**	**1.852**	**3.427**
6	180	4	**0.898**	**1.608**	**3.786**
	320	8	0.865	1.776	3.552
	640	16	0.835	1.848	3.387

注：黑体数值表示最优结果，“↑”表示数值越大越好，“↓”表示数值越小越好.

慎选择压缩率和码本数.

4.3 关于压缩激励机制的消融探究

为了探究压缩激励方法的有效性和适用边界，我们将比特率设置为 1.5 Kbps, 3 Kbps, 6 Kbps 三个档次进行了消融实验. 为了进一步对比，我们将输入特征分别设为梅尔谱图和波形进行消融探究，其中波形实验仿照 Hifi-codec 模型并恢复了多层卷积编码器结构. 最终结果分别如表 6 和表 7 所示，图 4 和图 5 分别对表 6 和表 7 中音频质量的各项指标进行了可视化，其中 WARP-Q 越低音频质量越高.

上述图表显示梅尔谱特征下，压缩激励权重(SE-weight)的方法在 1.5 Kbps 和 3 Kbps 的比特率下表现出一定的有效性，但在 6 Kbps 时效果开始减弱；而在波形特征下，该方法仅在 1.5 Kbps 的极低比特率下具有一定的效果，而在 3 Kbps 时其效果已显著降低. 同时，表 4 和表 5 均显示压缩激励方法会引发轻微的时延损耗，因此实际使用中应结合带宽环境进行取舍. 此外，我们可以认为，相较波形特征而言，梅尔谱特征作为输入(Mel-Input)与压缩激励权重方法更为适配，大多数情形下，这样组合在速度和质量上更具优势. 该现象也说明梅尔卷积编码器各个输出通道信息量之间的差异性会大于波形卷积编码器. 这是符合认知的，因为梅尔谱频率维度之间本身就具有显著差异，人语音的能量更多地集中在低频区域，且人耳对不同频带的感知不同，比如对低频感知更加明显，对高频信息感知较为模糊，这样的差异性将更大地开发低比特率时压缩激励权重方法的潜能.

Table 6 Ablation Experiments on Squeeze-Excitation-Weighted Mechanism (Input Characteristic is Mel Spectrogram)

表 6 压缩激励加权机制的消融实验（输入特征为梅尔谱图）

比特率/Kbps	码本数	指标	Mel-Input+SE-weight（本文）	Mel-Input
1.5	2	RTF_{enc} ↑	36.447	**42.515**
		STOI ↑	**0.801**	0.795
		WARP-Q ↓	2.371	**2.318**
		VISQOL ↑	**2.958**	2.926
3.0	4	RTF_{enc} ↑	35.632	**44.498**
		STOI ↑	**0.846**	0.837
		WARP-Q ↓	**1.997**	2.035
		VISQOL ↑	**3.334**	3.310
6.0	8	RTF_{enc} ↑	23.493	**32.345**
		STOI ↑	0.865	**0.866**
		WARP-Q ↓	1.776	**1.733**
		VISQOL ↑	**3.552**	3.499

注：黑体数值表示最优结果，“↑”表示数值越大越好，“↓”表示数值越小越好.

Table 7 Ablation Experiments on Squeeze-Excitation-Weighted Mechanism (Input Characteristic is Waveform)

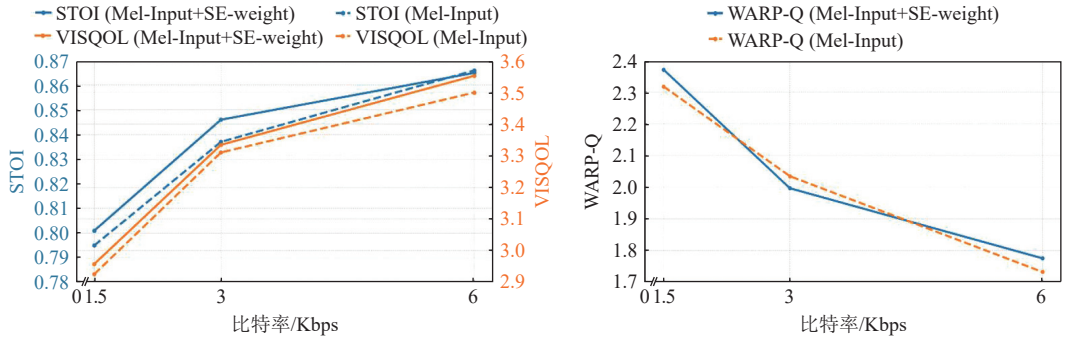
表 7 压缩激励加权机制的消融实验（输入特征为波形）

比特率/Kbps	码本数	指标	Wave-Input+SE-weight	Wave-Input
1.5	2	RTF_{enc} ↑	5.527	**6.507**
		STOI ↑	**0.753**	0.751
		WARP-Q ↓	**2.526**	2.709
		VISQOL ↑	**2.630**	2.520
3.0	4	RTF_{enc} ↑	5.702	**6.549**
		STOI ↑	0.837	**0.839**
		WARP-Q ↓	2.147	**2.125**
		VISQOL ↑	**3.329**	3.311
6.0	8	RTF_{enc} ↑	5.369	**5.673**
		STOI ↑	0.871	**0.878**
		WARP-Q ↓	2.020	**1.852**
		VISQOL ↑	3.477	**3.665**

注：黑体数值表示最优结果，“↑”表示数值越大越好，“↓”表示数值越小越好.

4.4 关于激活函数的消融探究

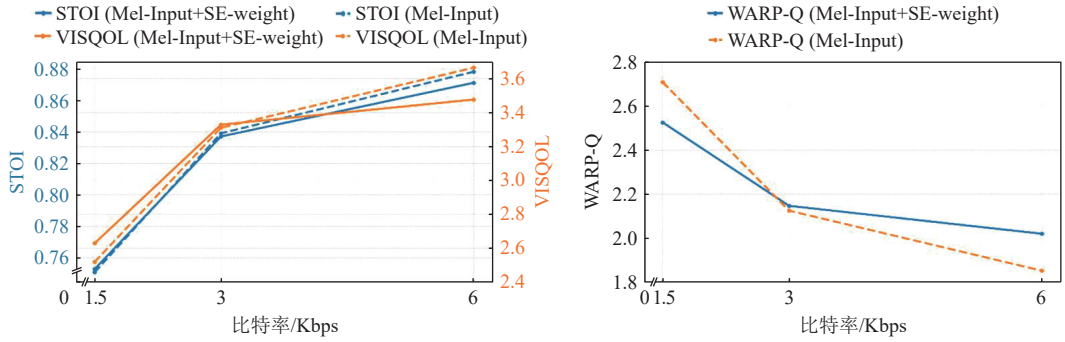
本节旨在讨论激活函数对梅尔谱编解码器语音感知质量与时延的影响. 实验中将在梅尔谱编解码器中进行 Relu 激活函数与 Snake 激活函数的对比. Snake 函数作为一种具有周期性特征的激活函数，能够有效地适应语音波形高周期性的性质. 在声码器和波形编解码器任务中，Snake 激活函数已被证实能显著提升语音质量[38-39]. 消融实验结果如表 8 所示，对于梅尔谱编解码器，Relu 激活函数能在语音质量与 Snake 激活函数相当的情况下，运行速度均具有明显提升. 尤其是比特率较低的情况下，以 0.75 Kbps 和 1.5 Kbps 为例，Relu 激活函数和 Snake 激活函数在



(a) STOI和VISQOL指标 (b) WARP-Q指标

Fig. 4 Ablation experiments on squeeze-excitation-weighted mechanism (Input characteristic is Mel spectrogram)

图 4 关于压缩激励加权机制的消融实验（输入特征为梅尔谱图）



(a) STOI和VISQOL指标 (b) WARP-Q指标

Fig. 5 Ablation experiments on squeeze-excitation-weighted mechanism (Input characteristic is waveform)

图 5 关于压缩激励加权机制的消融实验（输入特征为波形）

Table 8 Ablation Experiments on Activation Function

表 8 关于激活函数的消融实验

比特率/Kbps	码本数	指标	Relu（本文）	Snake
0.75	1	RTF_{enc}↑	**36.656**	11.590
		STOI ↑	**0.747**	0.746
		WARP-Q ↓	2.604	**2.568**
		VISQOL ↑	2.536	**2.538**
1.50	2	RTF_{enc}↑	**36.447**	11.784
		STOI ↑	0.801	**0.810**
		WARP-Q ↓	2.371	**2.224**
		VISQOL ↑	2.958	**3.073**
3.00	4	RTF_{enc}↑	**35.632**	10.821
		STOI ↑	0.846	**0.851**
		WARP-Q ↓	1.997	**1.989**
		VISQOL ↑	**3.334**	3.234
4.50	6	RTF_{enc}↑	**22.338**	5.915
		STOI ↑	0.859	**0.892**
		WARP-Q ↓	1.798	**1.563**
		VISQOL ↑	3.492	**3.638**
6.00	8	RTF_{enc}↑	**23.493**	4.967
		STOI ↑	0.865	**0.900**
		WARP-Q ↓	1.776	**1.448**
		VISQOL ↑	3.552	**3.658**

注：黑体数值表示最优结果，“↑”表示数值越大越好，“↓”表示数值越小越好.

语音质量上几乎没有差距，RTF 分别能提高 2.16 倍和 2.09 倍，因此 Relu 激活函数更适合应用场景实际需求.

5 结 论

本文采用了使用梅尔谱做为编码器输入特征，并采用低层卷积编码器，采用压缩激励的方法利用了梅尔谱图经过低层编码器输出特征各通道信息量的不均匀性. 本文在 LibriTTS 和 VCTK 数据集上进行实验，结果表明，该方法在提升编码器端运行速度上具有优势，减少了时延. 此外，在较低比特率的场景中，重建的语音相比波形编解码器基线具有更好的感知质量. 通过消融实验，本文探究了压缩激励权重方法在不同带宽和不同输入特征中的适用情况，进一步确定了压缩激励权重更适合低比特率条件的结论. 此外，本文还对编解码器的激活函数进行了消融探究，采用的 Relu 激活函数相比周期性 Snake 激活函数在运行速度上更具突出优势. 在未来，可以根据对梅尔谱图的先验知识，设计更高效的、结合其特点的编解码系统.

作者贡献声明: 周俊佐提出了算法思路、完成实验并撰写论文;易江燕提出指导意见并修改论文;陶建华参与实验;任勇实现了方法和实验设计;汪涛负责方法和实验设计,并修改论文.

参 考 文 献

[1] De Andrade J F, De Campos M L R, Apolinario J A. Speech privacy for modern mobile communication systems[C]//Proc of the 33rd IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2008: 1777−1780

[2] Haneche H, Ouahabi A, Boudraa B. Compressed sensing-speech coding scheme for mobile communications[J]. Circuits, Systems, and Signal Processing, 2021, 40(10): 5106−5126

[3] Budagavi M, Gibson J D. Speech coding in mobile radio communications[J]. Proceedings of the IEEE, 1998, 86(7): 1402−1412

[4] Bessette B, Salami R, Lefebvre R, et al. The adaptive multirate wideband speech codec (AMR-WB)[J]. IEEE Transactions on Speech and Audio Processing, 2002, 10(8): 620−636

[5] Cox R V, Kroon P. Low bit-rate speech coders for multimedia communication[J]. IEEE Communications Magazine, 1996, 34(12): 34−41

[6] Huang Yongfeng, Liu Chenghao, Tang Shanyu, et al. Steganography integration into a low-bit rate speech codec[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(6): 1865−1875

[7] Valin J M, Terriberry T B, Montgomery C, et al. A high-quality speech and audio codec with less than 10-ms delay[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 18(1): 58−67

[8] Hicsonmez S, Sencar H T, Avcibas I. Audio codec identification from coded and transcoded audios[J]. Digital Signal Processing, 2013, 23(5): 1720−1730

[9] Dietz M, Multrus M, Eksler V, et al. Overview of the EVS codec architecture[C]//Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 5698−5702.

[10] Valin J M, Vos K, Terriberry T. Definition of the opus audio codec[EB/OL]. 2012[2024-12-26]. https://datatracker.ietf.org/doc/html/rfc6716

[11] Zeghidour N, Luebs A, Omran A, et al. SoundStream: An end-to-end neural audio codec[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2021, 30: 495−507

[12] Biswas A, Jia D. Audio codec enhancement with generative adversarial networks[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 356−360

[13] Stimberg F, Narest A, Bazzica A, et al. WaveNetEQ—Packet loss concealment with waveRNN[C]//Proc of the 54th Asilomar Conf on Signals, Systems, and Computers. Piscataway, NJ: IEEE, 2020: 672−676

[14] Xiao Wei, Liu Wenzhe, Wang Meng, et al. Multi-mode neural speech coding based on deep generative networks[C]//Proc of the 24th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2023: 819−823

[15] Wu Y-C, Gebru I D, Marković D, et al. Audiodec: An open-source streaming high-fidelity neural audio codec[C]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023: 1−5

[16] Jiang Xue, Peng Xiulian, Zhang Yuan, et al. Disentangled feature learning for real-time neural speech coding[C/OL]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023[2025-02-05]. https://ieeexplore.ieee.org/document/10094723

[17] Petermann D, Jang I, Kim M. Native multi-band audio coding within hyper-autoencoded reconstruction propagation networks[C/OL]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023[2025-02-05]. https://ieeexplore.ieee.org/document/10094593

[18] Lim H, Lee J, Kim B H, et al. End-to-end neural audio coding in the MDCT domain[C/OL]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023[2025-02-05]. https://ieeexplore.ieee.org/document/10096243

[19] Kleijn W B, Lim F S C, Luebs A, et al. Wavenet based low rate speech coding[C]//Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2018: 676−680

[20] Jang Inseon, Yang Haici, Lim W, et al. Personalized neural speech codec[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 991−995

[21] Du Zhihao, Zhang Shiliang, Hu Kai, et al. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 591−595

[22] Gârbacea C, Oord A, Li Yazhe, et al. Low bit-rate speech coding with VQ-VAE and a WaveNet decoder[C]//Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 735−739

[23] Kleijn W B, Storus A, Chinen M, et al. Generative speech coding with predictive variance regularization[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 6478−6482

[24] Oord A, Dieleman S, Zen Heiga, et al. WaveNet: A generative model for raw audio[J]. arXiv preprint, arXiv: 1609.03499, 2016

[25] Kankanahalli S. End-to-end optimized speech coding with deep neural networks[C]//Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2018: 2521−2525

[26] Oord A, Vinyals O. Neural discrete representation learning[C/OL]//Proc of the 31st Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2017[2025-02-05]. https://dl.acm.org/doi/10.5555/3295222.3295378

[27] Défossez A, Copet J, Synnaeve G, et al. High fidelity neural audio compression[J]. arXiv preprint, arXiv: 2210.13438, 2022

[28] Ratnarajah A, Zhang Shixiong, Yu Dong. M3-AUDIODEC: Multi-channel multi-speaker multi-spatial audio codec[J]. arXiv preprint, arXiv: 2309.07416, 2023

[29] Yang Dongchao, Liu Songxiang, Huang Rongjie, et al. Hifi-codec: Group-residual vector quantization for high fidelity audio codec[J]. arXiv preprint, arXiv: 2305.02765, 2023

[30] O'shaughnessy D. Speech Communications: Human and Machine[M]. Piscataway, NJ: IEEE, 1999

[31] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal processing, 1980, 28(4): 357−366

[32] Hasanabadi M R. MFCC-GAN codec: A new AI-based audio coding[J]. arXiv preprint, arXiv: 2310.14300, 2023

[33] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]// Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 7132−7141

[34] Zen Heiga, Dang V, Clark R, et al. LibriTTS: A corpus derived from librispeech for text-to-speech[J]. arXiv preprint, arXiv: 1904.02882, 2019

[35] Liu Zhaoyu, Mak B. Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers[J]. arXiv preprint, arXiv: 1911.11601, 2019

[36] Taal C H, Hendriks R C, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]// Proc of the 35th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2010: 4214−4217

[37] Hines A, Skoglund J, Kokaram A, et al. VISQOL v3: An open source production ready objective speech and audio mestric[J]. arXiv preprint, arXiv: 2004.09584, 2020

[38] Kumar R, Seetharaman P, Luebs A, et al. High-fidelity audio compression with improved rvqgan[C/OL]//Proc of the 38th Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2024[2025-02-05]. https://openreview.net/forum?id=qjnl1QUnFA

[39] Lee S, Ping W, Ginsburg B, et al. BigVGAN: A universal neural vocoder with large-scale training[J]. arXiv preprint. arXiv: 2206.04658, 2022

[40] Dietz M, Multrus M, Eksler V, et al. Overview of the EVS codec architecture[C]//Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 5698–5702

[41] Gao Weiwei, Shan Mingtao, Song Nan, et al. Detection of microaneurysms in fundus images based on improved YOLOv4 with SENet embedded[J]. Journal of Biomedical Engineering, 2022, 39(4): 713−720 (in Chinese)
(高玮玮, 单明陶, 宋楠, 等. 嵌入 SENet 的改进 YOLOv4 眼底图像微动脉瘤自动检测算法[J]. 生物医学工程学杂志, 2022, 39(4): 713−720)

[42] Chen Qiang, Liu Li, Han Rui, et al. Image identification method on high speed railway contact network based on YOLO v3 and SENet[C]//Proc of the 38th Chinese Control Conf (CCC). Piscataway, NJ: IEEE, 2019: 8772−8777

[43] Wang Chenglong, Yi Jiangyan, Tao Jianhua, et al. Global and temporal-frequency attention based network in audio deepfake detection[J]. Journal of Computer Research and Development, 2021, 58(7): 1466−1475 (in Chinese)
(王成龙, 易江燕, 陶建华, 等. 基于全局-时频注意力网络的语音伪造检测[J]. 计算机研究与发展, 2021, 58(7): 1466−1475)

[44] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770−778

[45] Kong J, Kim J, Bae J. Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[C]//Proc of the 34th Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2020: 17022−17033

[46] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C/OL]//Proc of the 28th Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2014[2025-02-05].https://www.researchgate.net/publication/263012109_Generative_Adversarial_Networks

[47] Jassim W A, Skoglund J, Chinen M, et al. WARP-Q: Quality prediction for generative neural speech codecs[C]//Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 401−405

Zhou Junzuo, born in 2000. Master candidate. His main research interest includes speech synthesis.
周俊佐, 2000 年生. 硕士研究生. 主要研究方向为语音合成.

Yi Jiangyan, born in 1984. PhD, master supervisor. Her main research interests include speech information processing, speech generation and identification, and continuous learning.
易江燕, 1984 年生. 博士, 硕士生导师. 主要研究方向为语音信息处理、语音生成与鉴别、持续学习.

Tao Jianhua, born in 1972. PhD, PhD supervisor. His main research interests include intelligent information fusion and processing, speech processing, affective computing, and big data analysis.
陶建华, 1972 年生. 博士, 博士生导师. 主要研究方向为智能信息融合与处理、语音处理、情感计算、大数据分析.

Ren Yong, born in 1998. PhD candidate. His main research interest includes speech synthesis.

任 勇, 1998 年生. 博士研究生. 主要研究方向为语音合成.

Wang Tao, born in 1996. PhD. His main research interest includes speech synthesis.

汪 涛, 1996 年生. 博士. 主要研究方向为语音合成.

《计算机研究与发展》投稿指南

征稿范围

计算机体系结构; 计算机网络与通信; 网络与信息安全; 人工智能; 软件技术; 并行与分布式计算; 图形图像、信息检索等应用技术; 其他计算机相关领域.

征稿类型

1)学术论文: 有创新学术见解的研究成果的完整论述, 对该学术领域的发展有积极意义, 或者支持我国学术生态和产业生态的重要成果.

2)综述: 对新兴的、活跃的学术研究领域或技术开发领域的评述.

3)快报短文: 对目前国内外计算机领域新思想、新观点、新技术的解读文章.

征稿要求

1)本刊只接收中文稿, 不受理英文稿.学术论文建议不超过 15 页, 综述不超过 20 页, 短文不超过 4 页.

2)要求来稿未在正式出版物上发表过, 不存在一稿多投问题.

3)作者在投稿时, 需同时向编辑部提交所有作者手写签字确认的“投稿声明”扫描版, 如不提供不予受理.对于违反投稿声明相关条款的来稿, 本刊将视情节做严肃处理, 后果自负.

4)本刊双盲评审, 作者投稿时上传系统的电子版中需要隐去作者、单位、基金、作者简介等信息, 初审通过后给编辑部邮寄的纸质版本需要包含前述署名信息.改后复审的稿件在提交修改版本以及修改说明时, 同样需符合双盲评审要求.

5)稿件评审通过后, 作者需提交单位审核盖章的“不涉密证明”和所有作者签字确认的“著作权转让声明”.

6)投稿时间在 6 个月以上的, 若没有收到编辑部发出的审理结果通知, 在书面通知编辑部并收到编辑部确认回复之后, 作者方可自行处理此稿件.

7)无论何种原因撤回投稿, 需以书面形式通知编辑部并确保所有作者知情.