

# 基于非对称结构的神经网络语音编解码器研究

李晋鑫<sup>1</sup>, 别红霞<sup>2</sup>

<sup>1</sup> 北京邮电大学人工智能学院, 北京 100876

<sup>2</sup> 北京邮电大学人工智能学院, 北京 100876

**摘要:** 随着深度学习技术的飞速发展, 基于神经网络的语音处理算法层出不穷。神经网络语音编解码器作为一项重要的语音通信技术, 受到学术界和工业界的广泛关注。然而, 目前主流的神经网络语音编解码器采用对称的编码器-解码器结构, 这种对称结构的语音编解码器存在模型结构冗余, 计算效率较低。本文将一个基于对称结构的神经网络语音编解码器作为基线模型并作出改进, 提出了一种基于非对称结构的神经网络语音编解码器。通过使用一个单分支的编解码网络进行语音编码, 与强大的双分支解码器组成非对称结构, 有效地减少了编码器的模型参数量, 加快了模型推理速度。在 LJSpeech 数据集上的语音生成质量实验表明, 相较于基线模型, 本文模型在保证语音生成质量的条件下, 减少了约 1/3 的编码器模型参数量。

**关键词:** 人工智能; 神经网络语音编解码器; 非对称编码器-解码器结构

**中图分类号:** TP18; TN912.33

## Research on Asymmetric Architecture-based Neural Speech Codec

LI Jin-Xin<sup>1</sup>, BIE Hong-Xia<sup>2</sup>

<sup>1</sup> Department of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876

<sup>2</sup> Department of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876

**Abstract:** With the rapid advancement of deep learning technologies, neural network-based speech processing algorithms have proliferated. Neural speech codecs, as a key speech communication technology, have attracted widespread attention from both academia and industry. However, mainstream neural speech codecs typically adopt a symmetric encoder-decoder architecture, which suffers from structural redundancy and low computational efficiency. In this paper, we build upon a baseline model with a symmetric architecture and propose an improved asymmetric architecture-based neural speech codec. By employing a single-branch encoder for speech encoding and combining it with a powerful

**基金项目:** 无

**作者简介:** 李晋鑫 (2000-), 男, 硕士研究生, 主要研究方向: 神经网络语音编解码器、神经声码器。通信作者: 别红霞 (1971-), 女, 教授, 主要研究方向: 多媒体信息智能与传输、工业大数据智能、智能边缘计算。

dual-branch decoder, the proposed asymmetric architecture effectively reduces encoder parameters and accelerates model inference speed. Experiments on the LJSpeech dataset demonstrate that, compared to the baseline model, the proposed model reduces encoder parameters by approximately one-third while maintaining comparable speech generation quality.

**Key words:** Artificial Intelligence; Neural Speech Codec; Asymmetric Encoder-Decoder Architecture

## 0 引言

主流的神经网络语音编解码器由编码器、量化器以及解码器组成。编码器对语音信号进行特征编码与压缩,编码后的语音特征经过量化器进行向量量化,得到离散的语音编码。解码器从离散语音编码生成高质量的语音信号。SoundStream[1]是一个经典的神经网络语音编解码器,它在编码器和解码器中分别使用4个卷积模块对语音信号进行编解码,实现了比Opus[2]和EVS[3]等非神经网络的语音编解码器的语音生成质量提升。除此之外,Encodec[4]、HiFi-Codec[5]等较为先进的神经网络语音编解码器也都使用类似的网络结构。此外,还有一些神经网络语音编解码器对语音的时频谱图进行编解码,再利用逆短时傅里叶变换(Inverse Short-Time Fourier Transform, ISTFT)生成语音信号。其中,APCodec[6]利用幅度谱和相位谱实现语音编解码,在编码器和解码器中分别使用两个神经网络分支对幅度谱和相位谱进行编解码。然而,这些主流的神经网络语音编解码器使用对称的编码器-解码器结构,即编码器和解码器的模型结构镜像对称,模型参数量相近。最近的一些研究<sup>[7, 8, 9]</sup>表明,编码器和解码器在语音编解码的过程中扮演的角色不同,与编码器相比,解码器通常需要更强大。这表明基于对称结构的神经网络语音编解码器存在潜在的改进空间,可以通过合理的设计减少编码器的模型参数量。

本文对APCodec语音编解码器的模型结构作出改进,提出了一种基于非对称结构的神经网络语音编解码器。具体而言,编码器使用一个单分支的编码网络对一组幅度谱和相位谱进行联合特征编码,解码器使用两分支的解码网络分别生成幅度谱和相位谱,最后利用ISTFT重建语音信号。单分支的编码器和双分支的解码器组成了非对称结构,有效地减少了编码器的模型参数量,提高了模型推理速度。

## 1 相关工作

### 1.1 短时傅里叶变换

短时傅里叶变换(Short-Time Fourier Transform, STFT)是一种分析信号频谱变化的数字信号处理方法,广泛应用于语音信号处理领域<sup>[10]</sup>。在对语音信号的频谱分析过程中,首先对语音信号进行分帧、加窗,将其分成帧级的离散序列,再对每个离散序列进行STFT。STFT的三个主要计算参数包括帧长 $W$ 、帧移 $H$ 和傅里叶变换长度 $L$ 。对于时长为 $t$ ,采样率为 $f_s$ 的一

段语音信号  $x \in \mathbb{R}^T$ , 其中  $T = f_s \cdot t$ 。由 STFT 得到的频谱为  $S \in \mathbb{C}^{F \times N}$ , 其中  $F$  对应频谱特征的频域维度,  $N$  对应频谱特征的时域维度。 $F$  和  $N$  与 STFT 计算参数的关系如下:

$$F = \frac{L}{2} + 1, N = \left\lceil \frac{T}{H} \right\rceil \quad (1)$$

STFT 谱是复数谱, 将其按照复数的幅度和相角可以分解为一对幅度谱  $A \in \mathbb{R}^{F \times N}$  和相位谱  $P \in \mathbb{R}^{F \times N}$ 。通过设置不同的 STFT 计算参数, 可以得到不同分辨率的语音频谱特征。不同分辨率的频谱特征具有不同的时域分辨率和频域分辨率, 呈现了语音中不同的时频特征信息<sup>[11]</sup>。此外, STFT 是可逆的, 使用相同计算参数的 ISTFT 可以将幅度谱和相位谱恢复为原始语音信号。

## 1.2 对称结构的神经网络语音编解码器: APCodec

艾杨等人提出的 APCodec[6] 是一个基于幅度谱和相位谱的对称结构的神经网络语音编解码器, 其模型结构如图 1(a) 所示。APCodec 使用两个相互独立的编码分支分别对一组幅度谱和相位谱进行特征编码与压缩, 每个编码分支包含一个网络结构相同的子编码器 (sub-encoder)。每个子编码器中包含 1 个输入一维卷积、2 个层归一化 (Layer Normalization, LN) [12] 层、1 个的 modified ConvNeXt v2[13] 网络、1 个线性层以及 1 个下采样卷积层。子编码器输出的幅度谱和相位谱特征在频域维度方向拼接, 以实现幅度特征和相位特征的融合, 并使用一个后处理一维卷积进一步降低融合特征的频域维度。编码器的输出特征紧接着输入到基于残差向量量化 (Residual Vector Quantization, RVQ) [1] 的量化器中, 得到离散语音编码。APCodec 使用与编码器镜像对称的解码器结构, 两个独立的解码分支分别生成幅度谱和相位谱。与编码器的主要区别在于, 解码器的子解码器 (sub-decoder) 使用基于一维反卷积的上采样层对特征进行时域上采样。

APCodec 使用端到端的方式训练, 训练使用的损失包括生成对抗网络 (Generative Adversarial Network, GAN) 相关损失、量化损失和频谱损失。对于 GAN 相关损失, APCodec 使用多周期判别器 (multi-period discriminator, MPD) [14] 和多分辨率判别器 (multi-resolution discriminator, MRD) [15] 作为判别器。对抗损失  $L_{adv-G}$  和  $L_{adv-D}$  分别用于生成器和判别器, 特征匹配损失  $L_{FM}$  用于计算判别器输出的特征图之间的损失。量化损失  $L_Q$  用于减小模型的量化误差。参考艾杨等人先前的研究工作<sup>[16, 17]</sup>, 频谱损失  $L_{spec}$  由幅度谱损失  $L_A$ 、相位谱损失  $L_P$ 、STFT 谱损失  $L_S$  以及梅尔谱损失  $L_M$  构成。完整的生成器损失  $L_G$  定义如下:

$$L_G = \lambda_{spec} L_{spec} + \lambda_Q L_Q + L_{adv-G} + L_{FM} \quad (2)$$

其中,  $\lambda_{spec}$  和  $\lambda_Q$  为超参数。

对于 48kHz 采样率的语音信号, APCodec 能够在 6kbps 的语音传输速率下获得比 SoundStream[1]、Encodec[4]、HiFi-Codec[5] 和 AudioDec[7] 更好的语音生成质量。并且 APCodec 能够对 16kHz 和 24kHz 的语音进行编解码, 语音传输速率更低。尽管 APCodec 实现了高质量的语音生成, 但其模型结构仍存在改进空间。

## 2 本文方法

本文对 APCodec[6] 的编码器结构进行了改进, 提出了一种基于非对称结构的神经网络语音编解码器。本文模型结构见图 1(b), 图中用蓝色模块表示与 APCodec 不同的改进模块。

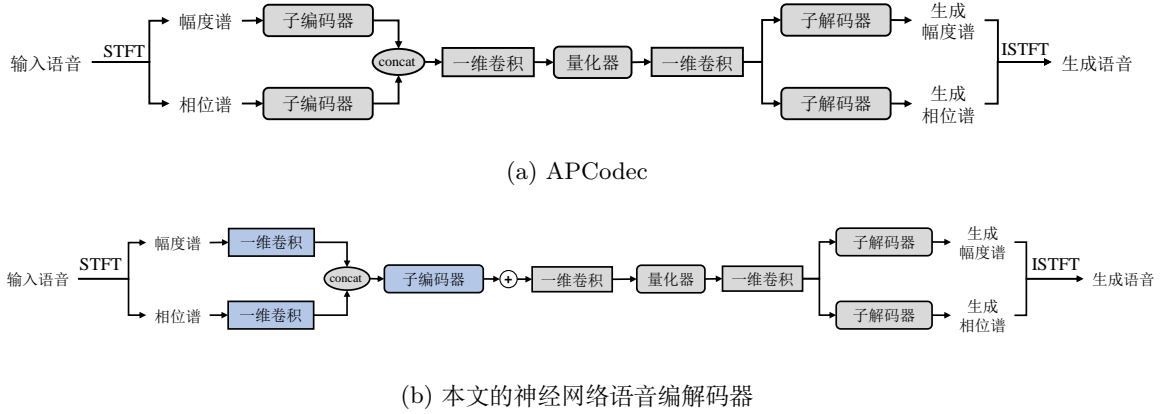


图 1: APCodec 与本文的神经网络语音编解码器模型结构对比

### 2.1 非对称框架

利用 STFT 得到的一对幅度谱和相位谱分别表示语音信号的幅度信息和相位信息。从图 2(a) 和图 2(b) 中可以看到, 语音的幅度谱具有结构性较强, 是基于深度学习的语音处理算法中常用的语音时频谱图。与之相比, 相位谱的结构性较差。基于幅度谱和相位谱的这种结构差异, APCodec[6] 在编码器中使用两个独立的编码分支分别对幅度谱和相位谱进行特征编码。然而, 利用 STFT 提取的幅度谱和相位谱并非两个互不相关的时频谱图。现有研究<sup>[18, 19]</sup>表明, 基于 STFT 的幅度谱导数与基于 STFT 的相位谱导数之间存在变换关系。如图 2(c) 和图 2(d) 中绿色框标注的部分所示, 幅度谱的导数和相位谱的导数存在结构相似性。因此, APCodec 在编码器中使用两个完全独立的编码分支存在模型结构冗余。基于这一理论, 本文的神经网络语音编解码器在编码器中使用一个公共的编码分支对输入的幅度谱和相位谱进行联合编码与压缩, 有效地减少了编码器的模型参数量。

为了实现高质量的语音生成, 本文的解码器使用与 APCodec 相同的双分支结构, 即使用两个独立的解码分支分别生成幅度谱和相位谱。由于子编码器和子解码器分别是编码器和解码器中参数量最大的模块, 只使用一个子编码器的编码器和使用两个子解码器的解码器组成了非对称结构, 编码器更加轻量化, 模型推理速度更快。

### 2.2 模型结构

本文的神经网络语音编解码器使用帧长为  $W$ 、帧移为  $H$ 、傅里叶变换长度为  $L$  的 STFT 提取一对幅度谱  $A \in \mathbb{R}^{F \times N}$  和相位谱  $P \in \mathbb{R}^{F \times N}$  作为编码器的输入特征。编码器利用两个卷积

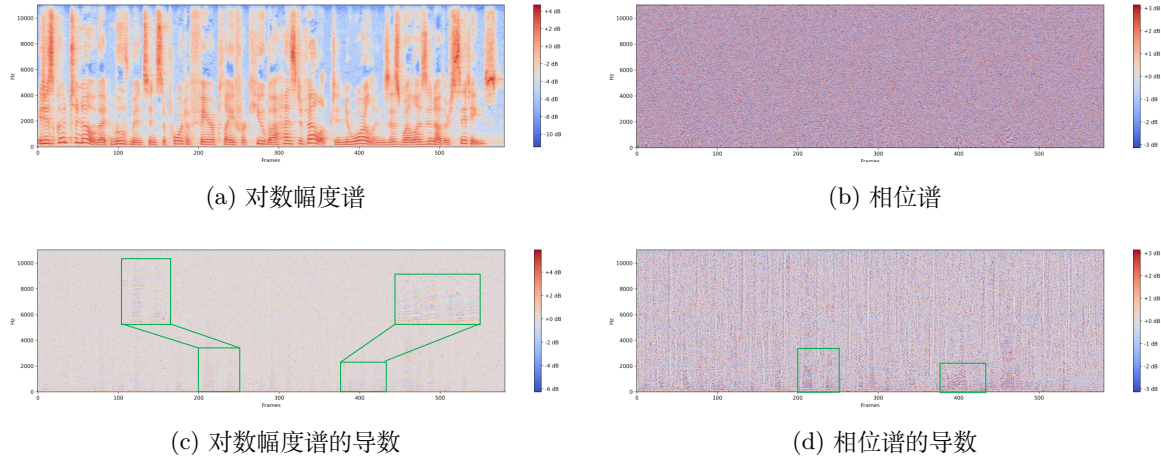


图 2: 对数幅度谱、相位谱、对数幅度谱导数及相位谱导数对比

核大小为 7、输入通道数为  $F$ 、输出通道数为 256 的预处理一维卷积分别对  $A$  和  $P$  进行频域特征降维, 并使用特征拼接的方式将幅度谱和相位谱特征在频域维度进行特征融合, 得到融合后的特征。融合后的幅度谱和相位特征通过一个与 APCodec[6] 相似的子编码器进行深层编码与压缩。子编码器中包含一个输入通道数为 512, 输出通道数为 256 的线性层, 用于对融合特征进行频域维度变换。维度变换后的特征经过 8 个输入特征维度为 256, 中间特征维度为 512 的 ConvNeXt v2[13] 模块对幅度谱和相位谱特征进行深层特征编码。ConvNeXt v2 模块前后各使用了一个 LN 层, 经过 ConvNeXt v2 模块后的特征维度保持不变。子编码器最后使用一个卷积核大小为 7 的下采样卷积层进行时域维度下采样, 得到子编码器的输出特征  $M \in \mathbb{R}_m^F \times \frac{N}{n}$ 。其中,  $m$  和  $n$  分别表示幅度谱和相位谱在频域维度和时域维度的下采样倍率。子编码器的输出特征  $M$  再经过一个卷积核大小为 7, 输出通道数为 32 的后处理一维卷积进一步减小输出特征的频域维度。

此外, 本文的解码器、量化器和判别器的结构与 APCodec 保持一致。

## 3 实验

### 3.1 实验设置

**数据集:** LJSpeech 数据集 [20] 是一个公开的英文数据集, 包含 13100 个音频片段。这些片段提取自 7 本非小说类书籍, 书籍的出版时间介于 1884 年至 1964 年之间, 属于公共领域。该数据集的语音内容由 LibriVox 项目在 2016 年至 2017 年录制, 所有音频均由一位女性朗读者完成, 并附有文本转录。每个语音片段的长度从 1 秒到 10 秒不等, 总时长约 24 小时, 音频采样率为 22050Hz。LJSpeech 数据集非常适用于语音生成任务。本文随机选择了 90% 条语音片段作为训练集和验证集, 其余 10% 作为测试集。



**模型参数和训练细节:** 对于输入的幅度谱和相位谱, STFT 的帧长为 320 ( $W = 320$ )、帧移为 40 ( $H = 40$ )、傅里叶变换长度为 1024 ( $L = 1024$ ), 根据公式 1, 幅度谱和相位谱的频域维度大小为 513 ( $F = 513$ )。子编码器的频域下采样倍率为 2 ( $m = 2$ ), 时域下采样倍率为 8 ( $n = 8$ ), 输入语音经过编码器后, 总时域下采样倍率为 320 ( $H_h \cdot n$ )。本文使用与 APCodec[6] 相同的模型训练方式。训练使用单张 NVIDIA RTX 4090 GPU, 训练步数为 600k, batchsize 设置为 16。每一批次训练使用的语音片段长度为 7960, 使用 AdamW 优化器 [21] 对模型进行训练优化, 初始学习率为 0.0002, 学习率在每个 epoch 按 0.999 的衰减因子进行衰减。

本文将 3 个主流的神经网络语音编解码器作为对比模型, 分别是 AudioDec[7]、HiFi-Codec[5] 和 APCodec[6]。这些语音编解码器的描述如下:

1. AudioDec 通过引入两阶段训练范式和模块化结构, 加快了模型训练, 并且能够适应不同的语音处理场景。借助 AudioDec 的官方开源代码<sup>1</sup>, 本文复现了 AudioDec v1 模型。第一个训练阶段的训练步数为 300k, batchsize 大小设置为 16; 第二个训练阶段的训练步数为 300k, batchsize 大小设置为 16。
2. HiFi-Codec 的编码器和解码器结构借鉴了 Encodec[4] 和 SoundStream[1] 的模型结构, 量化器使用了分组残差向量量化 (Group-residual Vector Quantization, GRVQ), 有效地提高了语音生成质量。借助 HiFi-Codec 的官方开源代码<sup>2</sup>, 本文复现了该语音编解码器并作为对比模型。HiFi-Codec 以端到端的方式训练, 模型训练步数为 600k, batchsize 大小设置为 16。
3. 1.2 节介绍了 APCodec。本文借助 APCodec 的官方开源代码<sup>3</sup>, 复现了该语音编解码器并作为对比模型。APCodec 的训练步数为 600k, batchsize 大小设置为 16。

**评价指标:** 本文使用五个客观评价指标评估生成语音的质量, 包括 Fréchet 语音距离 (Fréchet Audio Distance, FAD)、有声/无声部分的 F1 分数 (Voiced/Unvoiced F1 Score, V/UV F1)、多尺度短时傅里叶谱失真 (Multi-Resolution STFT Distance, MSTFT)、尺度不变信噪比 (Scale Invariant Signal to Noise Ratio, SI-SNR) 以及 PESQ。这些指标的计算方式参考 Amphion<sup>4</sup> [22]。此外, 本文还使用 UTMOS[23] 作为平均意见得分 (Mean Opinion Score, MOS) 的替代, 它是在积累了一定的 MOS 数据后, 在生成语音上训练得到的模型。

本文使用模型参数量评估语音编解码器的模型大小, 单位为百万 (Million, M)。模型参数量越小, 表明模型结构越简单。实时率 (Real Time Factor, RTF) 是模型在 GPU 或 CPU 上的语音处理时间与语音时长的比值, 在语音生成任务中常用于评估语音编解码器模型的语音生成速度。RTF 越小, 表明模型的语音处理速度越快。当 RTF 小于或等于 1 时, 可以认为语音处理是实时的<sup>[24]</sup>。本文评估模型性能使用单张 NVIDIA GeForce RTX 4090 GPU, CPU 使用单核的 Intel(R) Xeon(R) Platinum 8336C。

<sup>1</sup><https://github.com/facebookresearch/AudioDec>

<sup>2</sup><https://github.com/yangdongchao/AcademiCodec>

<sup>3</sup><https://github.com/YangAi520/APCodec>

<sup>4</sup><https://github.com/open-mmlab/Amphion>

### 3.2 语音生成质量对比实验

语音生成质量对比实验将本文提出的基于非对称结构的神经网络语音编解码器与 3 个对比模型进行对比。通过调整这些模型的时域下采样倍率和量化器参数使得它们都处在同一个语音传输速率下进行比较。

表 1: 各个语音编解码器在 LJSpeech 测试集上的语音生成质量评价得分

模型	FAD ↓	V/UV F1 ↑	MSTFT (Hz)↓	SI-SNR (dB)↑	PESQ ↑	UTMOS ↑
AudioDec	0.62	0.931	1.329	-29.8	2.77	3.58
HiFi-Codec	0.51	0.961	1.021	1.4	3.29	4.30
APCodec	0.52	0.962	<b>0.957</b>	2.4	<b>3.39</b>	<b>4.37</b>
本文模型 w 单分支解码器	0.63	0.951	1.032	-0.3	3.03	4.10
本文模型	<b>0.46</b>	<b>0.964</b>	0.973	<b>2.6</b>	3.36	4.34

各个模型在 LJSpeech 测试集上的语音质量评价指标得分见表 1。可以看到, 本文模型在多个语音质量评价指标上优于对比模型, 例如反映语音波形重建质量的 FAD 指标和反映语音相位重建质量的 SI-SNR 指标。表明本文模型能够达到与主流神经网络语音编解码器相当的语音生成质量。

### 3.3 语音生成速度对比实验

语音生成速度对比实验比较了本文模型与 APCodec 的编码器模型参数量以及编码器在 CPU 和 GPU 上的推理速度, 实验结果如表 2 所示。由于本文在编码器中仅使用了一个子编码器, 因此编码器的模型参数量比 APCodec 的编码器少了约 1/3, 并且在 CPU 和 GPU 上的 RTF 更小, 表明本文的编码器的推理速度更快。结合表 1 中与 APCodec 的语音生成质量对比实验结果, 可以得到, 本文模型能够在减少约 1/3 编码器模型参数数量的条件下, 达到与 APCodec 相当的语音生成质量, 表明了本文提出的非对称结构的神经网络编解码器的优越性。

表 2: 编码器模型参数量与推理速度对比结果

模型	编码器模型参数量 (M)↓	RTF(CPU) ↓	RTF(GPU) ↓
APCodec	6.75	0.110	0.00122
本文模型	<b>4.43</b>	<b>0.061</b>	<b>0.00063</b>

为了验证本文在非对称结构中使用双分支解码器的必要性。在本文模型的基础上, 将解码

器改为仅使用一个公共的解码分支。使用单分支解码器的语音编解码器在 LJSpeech 测试集上的语音生成质量评价得分见表 1 第 4 行。可以看到, 使用单分支解码器的语音编解码器在所有语音质量评价得分上的表现都显著下降, 这表明大参数量的双分支解码器有利于实现高质量的语音生成。

## 4 结论

为了解决主流的神经网络语音编解码器中存在的模型结构冗余问题, 本文提出了一种基于非对称结构的神经网络语音编解码器。编码器只使用一个公共的编码分支对幅度谱和相位谱进行联合特征编码与压缩, 解码器使用两个独立的解码分支分别重建幅度谱和相位谱, 用于实现高质量的语音生成。单分支的编码器与双分支的解码器组成了非对称结构, 有效地减少了编码器的模型参数量, 加快了模型推理速度。在 LJSpeech 数据集上的对比实验证明了本文模型相较于主流神经网络语音编解码器模型的优越性。

## 参考文献 (References)

- [1] Zeghidour N, Luebs A, Omran A, et al. Soundstream: An end-to-end neural audio codec[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 495-507.
- [2] J.-M. Valin, K. Vos, and T. B. Terriberry. "Definition of the Opus Audio Codec." IETF RFC 6716, 2012, <https://tools.ietf.org/html/rfc6716>.
- [3] Dietz M, Multrus M, Eksler V, et al. Overview of the EVS codec architecture[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 5698-5702.
- [4] Défossez A, Copet J, Synnaeve G, et al. High fidelity neural audio compression[J]. arXiv preprint arXiv:2210.13438, 2022.
- [5] Yang D, Liu S, Huang R, et al. Hifi-codec: Group-residual vector quantization for high fidelity audio codec[J]. arXiv preprint arXiv:2305.02765, 2023.
- [6] Ai Y, Jiang X H, Lu Y X, et al. APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
- [7] Wu Y C, Gebru I D, Marković D, et al. Audiodec: An open-source streaming high-fidelity neural audio codec[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.



- [8] Ju Z, Wang Y, Shen K, et al. NaturalSpeech 3: zero-shot speech synthesis with factorized codec and diffusion models[C]//Proceedings of the 41st International Conference on Machine Learning. 2024: 22605-22623.
- [9] Ji S, Jiang Z, Wang W, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling[J]. arXiv preprint arXiv:2408.16532, 2024.
- [10] 赵晓雷. 基于短时傅里叶变换的语音信号处理研究 [J]. 舰船电子工程, 2018, 38(4): 19-22.
- [11] 白燕燕, 胡晓霞. 基于 MATLAB 的语音短时谱的分析 [J]. 电子测试, 2019(18): 44-45.
- [12] Lei Ba J, Kiros J R, Hinton G E. Layer normalization[J]. ArXiv e-prints, 2016: arXiv: 1607.06450.
- [13] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders." in Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16 133 – 16 142.
- [14] Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in neural information processing systems, 2020, 33: 17022-17033.
- [15] Jang W, Lim D, Yoon J, et al. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation[C]//Proc. Interspeech 2021. 2021: 2207-2211.
- [16] Ai Y, Ling Z H. APNet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2145-2157.
- [17] Du H P, Lu Y X, Ai Y, et al. APNet2: high-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra[C]//National Conference on Man-Machine Speech Communication. Singapore: Springer Nature Singapore, 2023: 66-80.
- [18] Shimauchi S, Kudo S, Koizumi Y, et al. On relationships between amplitude and phase of short-time Fourier transform[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 676-680.
- [19] Zheng N, Zhang X L. Phase-aware speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 27(1): 63-76.
- [20] Ito K, Johnson L. The LJ Speech Dataset[EB/OL]. 2017[2025-02-25]. Available: <https://keithito.com/LJ-Speech-Dataset/>.

- [21] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.
- [22] Zhang X, Xue L, Gu Y, et al. Amphion: an Open-Source Audio, Music, and Speech Generation Toolkit[C]//2024 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2024: 879-884.
- [23] Saeki T, Xin D, Nakata W, et al. Utmos: Utokyo-sarulab system for voicemos challenge 2022[J]. arXiv preprint arXiv:2204.02152, 2022.
- [24] 王晶, 徐亮, 陈晓娇, 等. 基于神经网络的低码率语音编码技术研究综述 [J]. 信号处理, 2024, 40(12): 2261-2280.