arXiv:2301.03801v1 [cs.SD] 10 Jan 2023

UnifySpeech：用于零样本文本转语音和语音转换的统一框架

Haogeng Liu[1], **Tao Wang**[1], **Ruibo Fu**[2], **Jiangyan Yi**[2],
Zhengqi Wen[2]中国科学院 北京，中国

Jianhua Tao[1]
系　自动化系，清华大学人工智能学院，中国科学院大学，中国 北京，中国

摘要

文本转语音（TTS）和语音转换（VC）是两种不同的任务，二者都旨在根据不同的输入模态生成高质量的语音。鉴于二者的相似性，本文提出了UnifySpeech，首次将文本转语音和语音转换纳入统一框架。该模型基于这样一个假设：语音可以解耦为三个相互独立的组成部分：内容信息、说话人信息和韵律信息。文本转语音和语音转换都可以视为从输入中提取这三类信息，并据此重建语音。对于文本转语音，语音内容信息源自文本；而对于语音转换，它源自源语音。因此，除了语音内容提取模块之外，两种任务的其余模块均可共享。我们采用向量量化和域约束来弥合文本转语音与语音转换的内容域之间的差异。客观评估和主观评估结果表明，通过结合这两种任务，文本转语音获得了更强的说话人建模能力，而语音转换则具备了出色的语音内容解耦能力。

索引词： 解耦、零样本学习、文本转语音、语音转换、向量量化

1 引言

克隆目标说话人的声音是一项颇具吸引力的技术，可应用于多种场景（Sisman 等，2020），例如娱乐创作、个性化移动助手、安全领域等。最理想的语音克隆方式是，仅提供一段相对较短的未见目标说话人的语音作为参考，便能合成该目标说话人的任意语音，这称为零样本语音克隆。在语音研究领域，文本转语音（TTS）和语音转换（VC）是实现语音克隆的两种主流方式（Sisman 等，2020）。因此，近年来已有多种用于零样本文本转语音和语音转换的技术被提出（Gorodetskii 和 Ozhiganov，2022；Tang 等，2021）。

然而，尽管文本转语音和语音转换是语音克隆中两种重要且输出形式相同的方法，但两者的研究多少有些独立，彼此之间的互动并不多。不过，它们都属于语音合成任务。从语音生成的角度来看，我们将目标说话人语音中的信息分为三类：（1）语音内容，即音素的特征，或语音转换中的语音后验图（PPG），表示语音所承载的内容。（2）说话人信息，表示说话人的特征，与说话人的发音器官有关。（3）韵律信息，涵盖语调、重音和语音节奏。根据 FastSpeech2（Ren 等人，2020），基频、能量和时长信息可以很好地反映这些特征。由于文本转语音直接从音素中提取语音内容，因此比语音转换更容易获取与说话人无关的内容信息。随着语音转换接触到更多说话人，它有望获得更强的说话人建模能力。因此，通过将文本转语音和

语音转换整合到一个统一框架中，并结合两者的训练数据，可以帮助模型更好地学习这三类信息。

遗憾的是，在同一统一框架中研究文本转语音和语音转换具有挑战性，因为语音内容是从不同模态中提取的。具体而言，文本转语音中的语音内容通过音素信息获得，而文本转语音中的音素与语音并非等长序列，需要借助注意力机制对二者进行对齐。然而，注意力机制常常会受到说话人信息的影响，因此无法学习到与说话人完全无关的语音内容表征。而在语音转换中，源语音和目标语音在语音内容上是对齐的，因此可以直接从源语音中提取语音内容，这与文本转语音不同。与语音内容信息相比，说话人信息和韵律信息可以在文本转语音和语音转换中使用同一个网络进行建模。因此，这里的难点在于保持文本转语音和语音转换中的语音内容一致。随着语音合成的发展，近期提出的 Adaspeech2（Yan 等人，2021）能够将文本信息与语音信息结合起来，从而有效地将语音内容与输入解耦。统一框架因此成为可能。

总体而言，本文的主要贡献如下：

- 我们提出了 UnifySpeech，一个用于文本转语音和语音转换的统一框架。语音转换使无标签数据能够参与训练过程，从而让文本转语音接触到更多说话人。文本转语音增强了语音转换中的语音内容解耦能力。因此，两个流程都能从另一个流程中受益。
- 我们采用向量量化和域约束来弥合文本转语音与语音转换的内容域之间的差距。消融实验表明了该方法的有效性。
- 我们进行了广泛的实验，包括零样本文本转语音和零样本语音转换。结果表明，联合训练的文本转语音性能优于 StyleSpeech（Min 等人，2021），联合训练的语音转换则具备更强的语音解耦能力。

本文的演示可见：https://liuhaogeng.github.io/UnifySpeech/。

2 背景

本节将简要回顾本研究的背景，包括神经文本转语音模型和语音转换模型，以及用于文本转语音和语音转换任务的零样本学习。

2.1 文本转语音任务

文本转语音任务旨在对文本与语音之间的映射关系建模，这是一个涉及两个长度不等序列的建模问题。根据模型中文本与语音之间的对齐机制（Battenberg 等人，2020），端到端文本转语音模型可分为两类：1）使用神经网络学习文本与语音之间的对齐信息，例如 Tacotron（Wang 等人，2017b）中的局部敏感注意力。研究者已提出多种对注意力机制的改进方法。此外，受自动语音识别中的连接时序分类（Kim 等人，2017）启发，glow-TTS（Kim 等人，2020）和 VITS（Kim 等人，2021）能够自动学习对齐信息。2）将音素的时长信息（Ren 等人，2019）作为先验知识引入，通过对文本进行上采样来实现对齐（McAuliffe 等人，2017）。由于基于文本上采样得到的信息独立于说话人，因此能够将语音内容与语音很好地分离。因此，本文引入时长信息来构建文本转语音模型。

2.2 语音转换任务

语音转换可以看作两个步骤（Sisman 等，2020）。首先，从源语音中提取与说话人无关的语音内容信息，然后将目标说话人信息嵌入语音内容中，以重建目标说话人的语音。根据提取语音内容的方式，语音转换模型可分为两类：（1）基于文本的方法。（2）信息瓶颈方法。第一种方法需要额外的预训练自动语音识别模型。由于自动语音识别采用有监督的方式进行训练，因此需要大量配对文本和语音。此外，流水线建模容易累积误差，从而影响

系统的性能。因此，许多研究工作都聚焦于后者。通过对信息瓶颈施加一些限制，可以将不同类型的信息解耦。然而，如果信息瓶颈无法得到良好解耦，模型性能将显著下降。

2.3 面向文本转语音和语音转换的零样本学习

基于文本转语音和语音转换的零样本学习研究，重点关注如何提取有效的说话人信息，并将其嵌入文本转语音或语音转换模型中，以进行联合训练或分段训练。典型的说话人特征包括 i-vector（Wang 等，2017a）、d-vector（Variani 等，2014）、x-vector（Snyder 等，2018）等。此外，通过专门设计的网络结构提取说话人风格信息的模块，例如 GST（Wang 等，2018）和 VAE（Van Den Oord 等人，2017），也能取得良好效果。

3 UNIFYSPEECH

本节将介绍 UnifySpeech 用于零样本文本转语音和语音转换任务的细节。我们首先介绍 UnifySpeech 的核心思想：语音因子分解，然后说明 UnifySpeech 的形式化定义。最后，我们将描述 UnifySpeech 的模型结构。

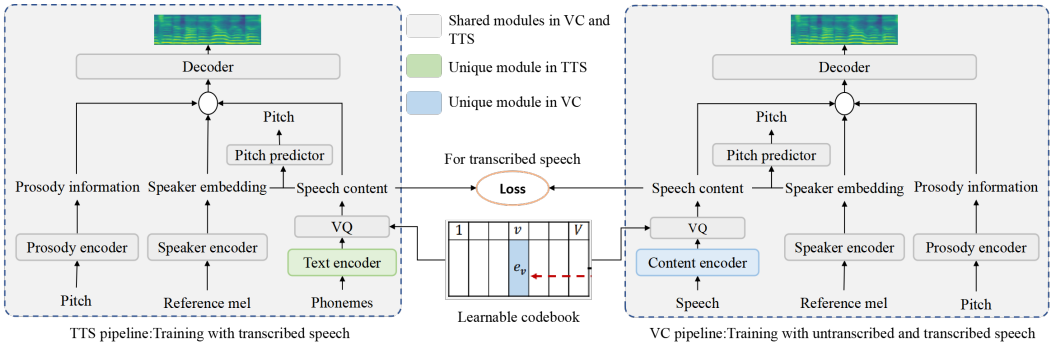

图 1：UnifySpeech 的结构。

3.1 语音表示解耦

可控、可迁移的语音生成任务的核心，是先将生成语音的各个组成部分解耦，再对每个部分进行控制和迁移。虽然一些端到端模型可以直接对文本与语音之间的关系（文本转语音）或语音到语音任务（语音转换）进行建模，但由于端到端模型的各个组成部分相互耦合，模型的迁移学习面临很大困难。因此，我们首先将语音生成任务解耦为三个独立的组成部分，然后将它们输入解码器以合成语音。这一思路也是 UnifySpeech 的主要架构。下文将详细介绍这三个组成部分及其来源。

语音内容：为了生成清晰易懂的语音信号，对准确的语音内容信息进行建模至关重要。语音内容属于语言信息，与说话人无关。由于文本转语音和语音转换的输入信号类型不同，提取语音内容的方式也有所不同。对于文本转语音，语音内容来源于文本。首先，为了学习文本的上下文信息，我们使用文本编码器对文本进行编码，以获得上下文表示。然后，根据音素的时长信息对上下文表示进行上采样，从而获得语音内容信息。对于语音转换，由于源语音与目标语音已对齐，我们直接使用内容编码器从源语音中提取语音内容信息。

说话人：说话人信息包括说话人的特征，例如音色、音量等。我们可以从目标说话人提供的语音中提取说话人信息。这

在文本转语音和语音转换任务中很常见，因此说话人提取网络可以在这两项任务中共享。由于说话人提取网络无需文本即可直接从语音中提取信息，因此可以使用语音转换任务中大量没有文本标注的数据进行训练，这有助于提升模型的迁移学习能力。

韵律：韵律信息表示说话人如何表达内容信息。它独立于说话人信息，与表达方式有关。由于基频信息可以反映语音节奏，因此我们使用基频信息来提取韵律信息。与说话人信息类似，韵律信息可以同时用于文本转语音和语音转换。此外，在训练过程中，我们可以从真实语音中获取基频信息，但在推理阶段没有真实值。因此，我们在训练阶段提出了一个基频预测模块（Ren 等人，2020），以语音内容信息和说话人信息为输入，预测基频信息。

3.2 使用向量量化的语音内容

由于在文本转语音和语音转换中获取语音内容的方式不同，因此两个域之间很容易产生偏差。如果发生这种偏差，将对某些下游共享模块（如基频预测器和解码器）造成严重损害。因此，为了尽可能确保两个语音内容域的一致性，我们从两个方面对其进行重构：

- 首先，我们使用共享码本对连续特征空间进行量化。
- 其次，我们利用有标签数据缩小两个离散特征空间之间的差异。

具体过程如下。假设文本转语音流程中的文本编码器得到的向量为 $C_p = (C_p^1, C_p^2, \cdots, C_p^T)$，长度为 T；语音转换流程中的内容编码器得到的向量为 $C_s = (C_s^1, C_s^2, \cdots, C_s^{T'})$，长度为 T'。需要注意的是，我们在文本编码器中加入了长度调节模块，以解决文本序列与语音序列长度不匹配的问题；该模块由 FastSpeech（Ren 等人，2019）提出。因此，如果文本和语音是配对的，$T' = T$。向量 C_p 和 C_s 是公式中的连续向量序列。由于连续特征的表示范围很大，C_p 和 C_s 难以匹配。我们借鉴了向量量化变分自编码器（Van Den Oord 等人，2017）中针对潜变量的离散化方法。具体而言，对于每个时间步 t，可以通过在字典中查找最近的预定义离散化嵌入向量，将 C_p 中的连续潜在表示 C_p^t 映射为 $\overline{C}_p^t$：

$$\overline{C}_p^t = e_k, \quad k = argmin_j \left\| C_p^t - e_j \right\|_2 \tag{1}$$

其中，e_j 是码本字典中的第 j 个嵌入向量，而 $j \in 1, 2, \cdots, V$。由于选择距离最小的条目会导致操作不可微，因此可以使用直通梯度估计器来近似梯度，表示如下：

$$\bar{h}_t = h_t + e_v - \mathrm{sg}\left(h_t\right), \quad v = \arg\min_k \left\| h_t - e_k \right\|_2 \tag{2}$$

其中，sg(·) 是停止梯度操作（Van Den Oord 等人，2017），该操作会在反向传播期间将其输入视为常量。

经过向量量化后，可以得到量化序列 $\overline{C}_p = (\overline{C}_p^1, \overline{C}_p^2, \cdots, \overline{C}_p^T)$ 和 $\overline{C}_s = (\overline{C}_s^1, \overline{C}_s^2, \cdots, \overline{C}_s^T)$。需要注意的是，当 C_p 和 C_s 被量化为 $\overline{C}_p$ 和 $\overline{C}_s$ 时，它们共享同一个码本 e。这样做的优点在于，由于文本转语音流程中的语音内容特征 $\overline{C}_p$ 与说话人无关，共享同一个码本有助于在语音转换流程中学习与说话人无关的语音内容特征 $\overline{C}_s$，这对语音转换任务至关重要。

尽管两个流程使用相同的码本进行编码，但无法保证编码后两个领域之间不存在偏差。因此，为了进一步消除两个域之间的偏差，我们使用文本转语音流程中的有标签语音来监督两个域的训练。具体而言，对于配对的文本和语音数据，我们约束量化序列 $\overline{C}_p$ 和 $\overline{C}_s$ 之间的特征距离：

$$\mathcal{L}_{\text{pair}} = \left\| \overline{C_p} - \overline{C_s} \right\|_2^2 \tag{3}$$

通过这种方式，我们可以利用有限的有标签 d 高效地缩小领域差异 ata.

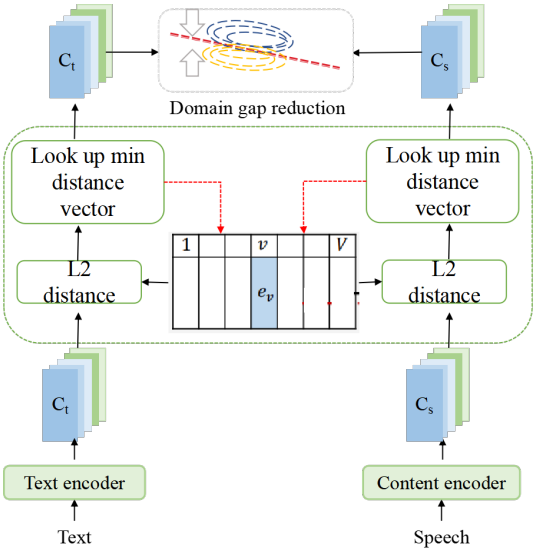

图 2：矢量量化操作的结构。

3.3 UNIFYSPEECH 流水线

图 2 展示了我们提出的 UnifySpeech 架构概览。它由一个序列到序列文本转语音模型和一个序列到序列语音转换模型组成。其核心思想是共享大部分模块参数（说话人编码器、韵律编码器、解码器和音高预测器），并将文本转语音和语音转换中的语音内容映射到同一空间。如上所述，UnifySpeech 使我们能够在有标签数据和无标签数据的拼接数据上训练模型。对于使用有标签数据的监督训练，可以分别训练两个模型，使其预测语音与真实值之间的损失最小化。对于使用无标签数据的无监督训练，可以训练语音转换流程，并与文本转语音共享参数。

为进一步阐明训练过程，我们将该框架展开如下。

3.3.1 文本转语音流程

将文本和语音序列对记为 $(x,\ y,\ F0) \in D$，其中 D 是配对文本和语音语料库。文本序列 x 中的每个元素表示一个音素或字符，而语音序列 y 中的每个元素表示一个语音帧。 $F0$ 是 y 的基频信息。经过三个编码器后得到的表示分别是语音内容 C、说话人 S 和韵律 P。

然后，将这三个部分相加并输入解码器，以获得预测语音 y'。此外，为了在干扰阶段获得基频信息，我们使用内容信息和说话人信息来预测基频。上述过程可表示为：

$$y' = \text{decoder}(C + S + P) \tag{4}$$

$$F0' = \text{pitch_predictor}(C + S) \tag{5}$$

因此，文本转语音过程中的重构损失包括两部分：

$$\mathcal{L}_{rec}^{VC} = MSE(y, y') + MSE(F0, F0') \tag{6}$$

此外，我们使用语音转换流水线中的内容编码器提取文本转语音训练语音的内容表示 C_s，并通过计算 C_s 和 C_p 的距离损失来缩小两个域之间的距离，具体说明见第3.2节。

$$\mathcal{L}_{\text{pair}} = \left\| \overline{C_p} - \overline{C_s} \right\|_2^2 \tag{7}$$

文本转语音流程的损失可表示为：

$$\mathcal{L}^{TTS} = \mathcal{L}_{rec}^{TTS} + \mathcal{L}_{\text{pair}} \tag{8}$$

其中， MSE 表示均方误差。

3.3.2 语音转换流程

Den注意所有无标签或有标签的语音（y， $F0$） $\in Y$。我们首先从中提取三种信息 $(y, F0)$ ，即语音内容 C_s、说话人 S_s 和韵律 P_s。

然后，与文本转语音流程类似，我们使用共享解码器和基频预测模块来预测语音信号 y' 和 $F0'$，这与公式1和公式2类似。

语音转换流程的损失仅包括重构损失，可表示为：

$$\mathcal{L}^{VC} = MSE(Y, Y') + MSE(F0, F0') \tag{9}$$

其中， MSE 表示均方误差。

3.3.3 训练过程

借助这种统一框架，文本转语音和语音转换可以通过联合训练相互学习。算法的详细内容如下。

过程 1 UnifySpeech 训练算法

1: **输入**：语音与文本配对数据集 (x, y)，仅包含语音的数据集 y'2: **重复**3: # A. 使用语音-文本数据对的文本转语音流程 1. 抽取语音与文本配对数据 (x, y) 2. 从 x 中提取语音内容信息，用于域损失 3. 根据文本生成预测语音 y、基频 F0 和语音内容 4. 计算文本转语音的损失 $\mathcal{L}_{rec}^{TTS}$4: # B. 使用仅语音数据的语音转换流程 1. 抽取语音与文本配对数据 (x, y) 2. 从 y 中提取语音内容信息，用于域损失 3. 计算两个流程的域损失 $\mathcal{L}_{pair}$4. 从仅包含语音的数据集

5 0 y' 中抽取

6 $\mathcal{L}^{VC}$

语音 y' 。根据语音生成预测语音 y、基频 F 和语音内容。计算语音转换的损失 5: # C. 损失组合： 1. 将所有损失 ($\mathcal{L}_{rec}^{TTS}$, $\mathcal{L}_{pair}$, $\mathcal{L}^{VC}$) 合并为一个损失变量 2. 计算文本转语音和语音转换参数的梯度 3. 使用梯度下降优化更新文本转语音和语音转换参数 6: **直到**收敛

4 实验与结果分析

本节中，我们通过实验评估所提出的方法。实验从两个方面展开：零样本文本转语音、零样本语音转换。

4.1 数据集

我们使用两个数据集分别模拟有标签数据和无标签数据。VCTK 数据集是一个英语数据集，包含 44 小时的语音和 109 位说话人，用作有标签数据。每位说话人约有 400 个句子。LibriTTS（Zen et al., 2019）用作无标签数据，其中包含来自 2484 位说话人的 585 小时语音数据。我们仅使用 LibriTTS 中的语音数据，并舍弃文本，以进行无监督训练。这样可以模拟我们能够获取大量无标签语音的场景。所有实验均采用 16 位、22050 Hz 的采样率。我们通过汉宁窗处理、12.5 毫秒帧移、50 毫秒帧长和 1024 点傅里叶变换提取 80 维梅尔频谱图。在本实验中，我们使用 hifigan（Kong et al., 2020）作为声码器。

4.2 模型细节

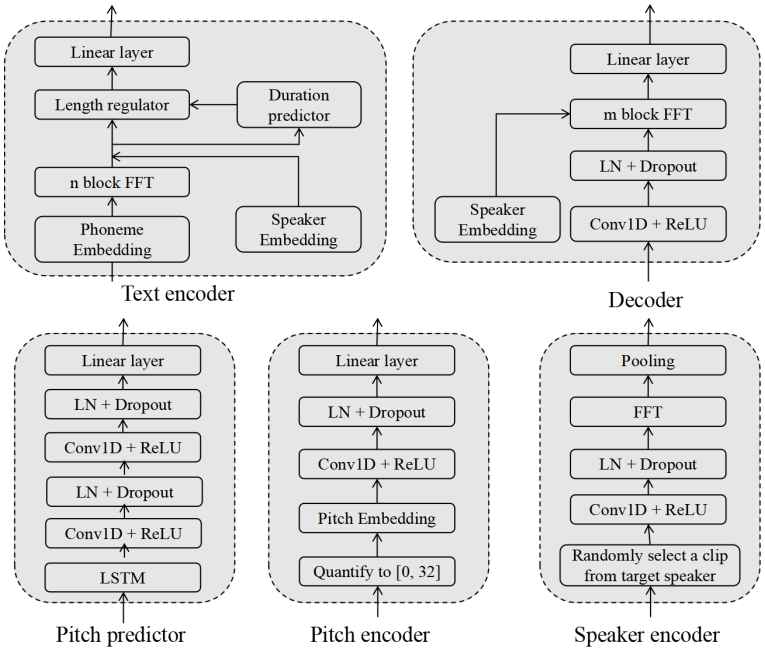

图 3：UnifySpeech 中各模块的结构。FFT（Ren 等人，2019）指前馈变换器。

我们提出的方法中各模块的细节如图 3 所示。具体而言，为了使文本编码器和内容编码器提取的语音内容序列长度相同，我们在文本编码器中加入了长度调节器，其设计受 FastSpeech（Ren 等人，2019）的启发。时长预测器的结构与 FastSpeech 中的相同。解码器和内容编码器的结构相似，但输入和输出的维度相反。对于韵律编码器，我们首先将每一帧的基频量化为 32 个可能值，并根据其取值将其编码为可学习的嵌入向量。此外，我们将音高预测器的输出改为一个分布。这样，音高预测就变成了分类任务，从而降低了帧级音高预测的难度。我们采用 StyleSpeech 中的说话人编码器，并使用风格自适应层归一化（SALN）（Min 等人，2021）方法融合特征，以确保比较公平。

文本编码器中前馈变换器（Vaswani 等，2017）（FFT）模块的数量为 4，解码器模块中则为 6。每个 FFT 模块中，隐藏状态的维度为 256。所有一维卷积的卷积核大小均设为 3。Dropout 率设为 0.5。解码器最后一个线性层的维度为 256。各编码器（文本编码器、基频编码器、内容编码器）最后一个线性层的维度为 256。使用 Adam优化器（Kingma 和 Ba，2014）更新参数。初始学习率为 0.001，随后学习率按指数规律衰减。

4.3 零样本文本转语音

我们首先进行了零样本文本转语音任务。我们从 VCTK 中选取了四位在训练过程中未使用过的说话人作为目标说话人。对于每位说话人，我们随机选取约 20 个句子作为目标。随后，我们计算合成语音与真实语音之间的基频均方根误差（基频 RMSE）、梅尔倒谱失真（Kubichek，1993）（MCD）、浊音/清音标志错误率（V/UV）以及基频相关系数（基频 CORR），作为客观指标。

表1：零样本文本转语音的客观评估结果。

模型	基频均方根误差（Hz）	梅尔倒谱失真（dB）	V/UV	基频相关系数
UnifySpeech	**17.84**	**2.51**	**16.9**%	**0.93**
StyleSpeech	19.02	2.63	18。06%	0.92

我们还进行了主观评估，以比较语音的质量和相似度。我们采用平均意见分（MOS）评估自然度，并采用相似度平均意见分（SMOS）评估相似度。两项指标均采用1至5分制，并报告95%置信区间（CI）。

表2：各模型的平均意见分（MOS）。带语音转换（VC）表示该模型经过联合训练。

模型	MOS	SMOS
GT	4.32 ± 0.15	–
真实梅尔频谱 + 声码器	4.09 ± 0.15	–
StyleSpeech	3.52 ± 0.13	3.82 ± 0.13
UnifySpeech-TTS（带语音转换）	$\mathbf{3.76 \pm 0.12}$	$\mathbf{3.95 \pm 0.13}$

为了更好地展示我们方法的有效性，我们展示了 VCTK 和 LibriTTS 中说话人嵌入向量的 t-SNE 投影（Van der Maaten 和 Hinton，2008）。图 4 展示了说话人可视化结果。对于已见说话人（x）和未见说话人（o），相应的说话人嵌入向量各自形成一个簇，并与其他说话人明显区分开来。不同说话人之间的边界清晰。这表明说话人编码器表现良好。

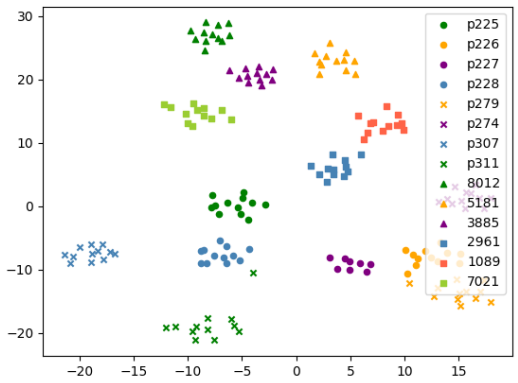

图 4：生成语音的说话人可视化，其中圆形和三角形表示未见说话人，x 和方形表示已见说话人。

4.4 零样本语音转换

我们开展了零样本语音转换任务，使用未见说话人的语音作为参考语音。由于缺少平行语料库，我们仅进行了主观评估。但语音转换和文本转语音共享相同的

说话人编码器，图 4 中的零样本文本转语音也可以作为参考。
与零样本文本转语音任务相同，我们采用平均意见分（MOS）评估自然度，采用相似度平均意见分（SMOS）评估相似度。两项指标均按 1 至 5 分制评分，并报告 95% 置信区间（CI）。

表 3：语音转换模型的平均意见分（MOS）。标注“With TTS”表示该模型经过联合ly训练。

模型	MOS	SMOS
GT	4.32 ± 0.15	–
真实梅尔频谱 + 声码器	4.09 ± 0.15	–
UnifySpeech-VC（不含文本转语音）	$\mathbf{3.63 \pm 0.13}$	1.31 ± 0.06
UnifySpeech-VC（含文本转语音）	3.58 ± 0.12	$\mathbf{3.31 \pm 0.13}$

可以发现，单独训练语音转换流程时，其性能较差。换言之，它不具备区分能力。但与文本转语音联合训练能够提升其语音解耦能力，进而间接增强说话人建模能力，后文将对此进行分析。

4.5 消融实验

为探究联合训练是否有效，我们分别训练文本转语音和语音转换部分，并对它们进行客观评估。我们还移除了向量量化部分并测试模型。对于文本转语音，我们计算合成语音与真实语音之间的基频均方根误差（基频的均方根误差）、梅尔倒谱失真（Kubichek，1993）（梅尔倒谱失真）、浊音/清音（浊音/清音标志的错误率）以及基频相关系数（基频的相关系数），将其作为客观指标。

表 4：零样本文本转语音的客观评估结果。

模型	基频均方根误差（Hz）	梅尔倒谱失真（db）	V/UV	基频相关系数
UnifySpeech-TTS（不含语音转换）	19.31	2.55	16.9%	0.91
UnifySpeech-TTS-novq（含语音转换）	19.41	2.58	17。2%	0.92
UnifySpeech-TTS（含语音转换）	**17.84**	**2.51**	**16.9**%	**0.93**

对于语音转换，我们分别计算来自相同说话人（同说话人平均余弦相似度）和不同说话人（不同说话人平均余弦相似度）的嵌入向量的平均余弦相似度（Lei 等人，2022）。随后将两者的比值用作评估指标。

表 5：来自相同说话人和不同说话人的平均余弦相似度之比。Unseen 或 seen 表示说话人是否来自测试集。With 或 without TTS 表示语音转换流程是否使用文本转语音流程进行训练。

模型	$\frac{S-ACS}{D-ACS}$（未见）	$\frac{S-ACS}{D-ACS}$（已见）
UnifySpeech-VC（使用文本转语音）	**2.8**	**6.0**
UnifySpeech-VC（不使用文本转语音）	1.0	1.0

为验证 VQ 是否能弥合文本转语音（TTS）与语音转换（VC）中语音内容部分之间的差异，我们计算了来自 TTS 和 VC 的音素表示之间的 L2 距离。由于在 VC 中，许多帧表示同一个音素，因此我们选取这些帧的聚类中心作为相应的音素表示。我们从验证集中随机选取 4 个句子进行实验。

表 6：文本转语音和语音转换中的音素表示之间的 L2 距离。 S_1 表示 $sentence_1$。

模型	S1	S2	S3	S4	平均值
UnifySpeech-VC（不使用向量量化）	0.464	0.476	0.474	0.553	0.492
UnifySpeech-VC（使用向量量化）	**0.152**	**0.205**	**0.208**	**0.205**	**0.193**

4.6 结果分析

上述结果表明，联合训练实际上提升了文本转语音（TTS）的说话人建模能力和语音转换（VC）的语音解耦能力。此外，VQ 增强了音素与语音中相同内容表示的一致性，确保模型能够正确运行。

对于文本转语音，与语音转换共享模块使无标签数据能够参与其训练过程。随着更丰富的说话人风格模式，说话人风格建模能力也得到了增强。

对于语音转换而言，单独训练时，它不具备说话人建模能力。由于自监督训练过程旨在降低源音频的重构损失，因此目标说话人的参考音频在这一过程中并不关键。该过程只需内容编码器和解码器即可完成，这可能就是来自不同说话人的说话人嵌入都十分相似的原因。联合训练时，文本损失发挥了正则化因子的作用，使内容编码器仅从源语音中提取内容信息（保留说话人信息，丢弃其他信息）。这使得包含丰富说话人信息的参考语音成为重构过程中不可或缺的部分。因此，模型的说话人建模能力得到了提升。

5 结论

本文提出了 UnifySpeech，一种用于文本转语音和语音转换的统一框架。两项任务都能从另一项任务中受益。由于使用大量无标签数据进行训练，它们的少样本建模能力以及合成语音的质量均有所提升。未来，我们将致力于进一步提高合成语音的质量，并使生成语音的风格更加接近目标说话人。

参考文献

Eric Battenberg、RJ Skerry-Ryan、Soroosh Mariooryad、Daisy Stanton、David Kao、Matt Shannon 和 TomBagby。用于稳健长语音合成的位置相对注意力机制。载于*ICASSP 2020-2020 IEEE 国际声学、语音与信号处理会议 (ICASSP)*，第 6194–6198 页。IEEE，2020 年。

Artem Gorodetskii 和 Ivan Ozhiganov。通过动态卷积注意力实现零样本长语音克隆。*arXiv* 预印本 *arXiv:2201.10375*，2022 年。

Jaehyeon Kim、Sungwon Kim、Jungil Kong 和 Sungroh Yoon。Glow-TTS：一种通过单调对齐搜索实现文本转语音的生成流。神经信息处理系统进展，33：8067–8077，2020 年。

Jaehyeon Kim、Jungil Kong 和 Juhee Son。结合对抗学习的条件变分自编码器，用于端到端文本到语音。载于国际机器学习会议，第 5530–5540 页。PMLR，2021 年。

Suyoun Kim、TakaakiHori 和 Shinji Watanabe。基于多任务学习的联合 CTC-注意力端到端语音识别。载于*2017年IEEE声学、语音与信号处理国际会议 (ICASSP)*，第 4835–4839 页。IEEE，2017 年。

Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014 年。

Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。Hifi-gan：用于高效、高保真语音合成的生成对抗网络。神经信息处理系统进展，33:17022–17033，2020 年。

Robert Kubichek。用于客观语音质量评估的梅尔倒谱距离测量。载于*IEEE 太平洋地区通信、计算机与信号处理会议论文集*，第 1 卷，第 125–128 页。IEEE，1993 年。

Yi Lei、Shan Yang、Jian Cong、Lei Xie 和 Dan Su。Glow-wavegan 2：高质量零样本文本转语音合成与任意到任意语音转换。*arXiv* 预印本 *arXiv:2207.01832*，2022。

Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本-语音对齐。发表于 *Interspeech*，第 2017 卷，第 498–502 页，2017。Dongchan Min、Dong Bok Lee、Eunho Yang 和 Sung Ju Hwang。Meta-StyleSpeech：多说话人自适应文本转语音生成。发表于 国际机器学习会议，第 7748–7759 页。PMLR，2021。Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。FastSpeech：快速、稳健且可控的文本转语音。神经信息处理系统进展，32，2019。Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。FastSpeech 2：快速、高质量的端到端文本转语音。*arXiv* 预印本 *arXiv:2006.04558*，2020。Berrak Sisman、Junichi Yamagishi、Simon King 和 Haizhou Li。语音转换及其挑战概述：从统计建模到深度学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29:132–157，2020。David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur。X 向量：用于说话人识别的稳健深度神经网络嵌入。发表于*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 5329–5333 页。IEEE，2018。Chuanxin Tang、Chong Luo、Zhiyuan Zhao、Dacheng Yin、Yucheng Zhao 和 Wenjun Zeng。用于音频旁白中基于文本插入的零样本文本转语音。*arXiv* 预印本 *arXiv:2109.05426*，2021。Aaron Van Den Oord、Oriol Vinyals 等。神经离散表示学习。神经信息处理系统进展，30，2017。Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。机器学习研究杂志，9(11)，2008。Ehsan Variani、Xin Lei、Erik McDermott、Ignacio Lopez Moreno 和 Javier Gonzalez-Dominguez。用于小型文本相关说话人验证的深度神经网络。发表于*2014 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 4052–4056 页。IEEE，2014。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制就是你所需要的一切。神经信息处理系统进展，30，2017。Shuai Wang、Yanmin Qian 和 Kai Yu。说话人嵌入编码了什么？发表于 *Interspeech*，第 1497–1501 页，2017a。Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio 等。Tacotron：迈向端到端语音合成。*arXiv* 预印本 *arXiv:1703.10135*，2017b。Yuxuan Wang、Daisy Stanton、Yu Zhang、RJ-Skerry Ryan、Eric Battenberg、Joel Shor、Ying Xiao、Ye Jia、Fei Ren 和 Rif A Saurous。风格令牌：端到端语音合成中的无监督风格建模、控制与迁移。发表于 国际机器学习会议，第 5180–5189 页。PMLR，2018。Yuzi Yan、Xu Tan、Bohan Li、Tao Qin、Sheng Zhao、Yuan Shen 和 Tie-Yan Liu。AdaSpeech 2：利用未转录数据进行自适应文本转语音。发表于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6613–6617 页。IEEE，2021。Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J Weiss、Ye Jia、Zhifeng Chen 和 Yonghui Wu。LibriTTS：一个源自 LibriSpeech 的文本转语音语料库。*arXiv* 预印本 *arXiv:1904.02882*，2019。