arXiv:2305.05401v1 [cs.SD] 9 May 2023

听声学唱：通过对语音录制进行无监督学习，打造可控虚拟歌手

薛伟[1], **王怡文**[2],**刘启峰**[3,1],**郭毅可**[1] 香港科技大学， [2] 香港珠海学院，[3] 中国科学院香港创新研究院

摘要

数字人正在虚拟世界中崛起，其形象已与真人难以区分。赋予数字人音频相关能力至关重要，因为声音承载着丰富的个人特征。我们的目标是打造一种可控的音频形态虚拟歌手；然而，要通过监督建模来控制歌声的所有不同因素，例如音色、速度、音高和歌词，极其困难，因为准确标注所有这些信息需要投入大量人力。本文提出一种框架，只需以完全无监督方式“聆听”任意内容的干净语音录制，就能将一个人的声音数字化，并且仅使用说话录音也能预测歌声。我们开发了一种基于变分自编码器（VAE）的框架，利用一组预训练模型将音频编码为代表歌声不同因素的各类隐藏嵌入，再将这些嵌入解码为原始音频。通过操纵代表不同因素的隐藏嵌入，可以控制生成的歌声，还可以通过在不同音色之间插值，进一步生成新的虚拟歌手。各类实验的评估结果证明了所提方法的有效性。该方法是制作人工智能合唱团的关键技术，并于2022年7月助力香港的人机共生交响乐团演出。

1 引言

我们正在迈入一个新时代，现实世界与虚拟世界的界限日益模糊，人们之间的地理隔阂以及人类与 AI 之间的鸿沟也逐渐消失。这进一步促进了人类与 AI 之间的相互启发与共同创造，推动科学与艺术不断拓展边界。数字人可以被打造得与真人难以区分；生成自然且个性化的声音至关重要，因为声音传递的不仅是交流所需的内容信息，也包括音色、口音和语调等个性化信息。本文旨在打造能够歌唱的音频形式数字人，即虚拟歌手，其应用范围广泛，涵盖娱乐、虚拟助手、文化传承和数字永生。

机器生成自然语音，传统上是通过文本转语音（TTS）这一问题来实现的：根据指定内容的文本合成语音波形。早期的 TTS 方法试图估计过度简化、用于模拟发声器官物理结构的线性滤波器 [1; 2; 3], , 而当前主流方法则以监督方式训练深度神经网络（DNN），在保留音色的同时，对不同内容和动态变化下的语音信号进行建模。通常，这些方法首先利用声学建模网络将文本转换为时频梅尔频谱图 [4; 5; 6; 7; 8; 9], , 然后采用声码器 [10; 11; 12; 13]

将梅尔频谱图转换为时域波形。常用的声学建模网络包括 Tacotron [4],DurIAN [6], 和 FastSpeech [8; 9], , 声码器则包括 WaveNet[10], HiFi-GAN [14] 和 MelGAN [15]。为实现歌声合成（SVS），将歌词和旋律信息共同作为声学建模网络的输入 [16; 17; 18; 19],, 并采用与文本转语音类似的声码器最终生成音频信号。

有监督歌声合成面临的一个主要问题是需要大型标注数据集，其中包含音频以及指明内容和旋律的对应文本稿。尽管已有一些公开的歌声合成数据集，例如 OpenCpop [20] 和 VocalSet [21], , 构建此类数据集仍需投入大量人力进行录制和标注。这不仅阻碍了为任意个人灵活构建语音模型，也由于邀请歌手录音的成本，限制了对高水平专业歌手演唱技巧的模拟。为摆脱对标注的依赖，可以使用歌声转换（SVC）对音频进行风格迁移，其中音色被定义为“风格”。已有许多方法 [22; 23; 24; 25; 26] 被提出，通过无监督学习解耦内容和音色信息，然后将音色替换为目标说话人的音色。然而，由于这些方法仅将音色替换为目标歌手的音色，因此难以建模该歌手独特的演唱技巧，也不易精确控制其他多样化的演唱特征。

在本文中，我们提出了一种以完全无监督方式将语音数字化的新框架。所提出的方法仅依靠任意内容和语言的音频录音，即可构建灵活的语音模型，并能够控制演唱音高、旋律和歌词等详细特征。通过这种方式，所提出的方法甚至可以从说话语音数据集中生成歌声。

我们开发了一种基于变分自编码器（VAE）的框架，该框架利用一组预训练模型将音频编码为代表歌声不同因素的各种隐藏嵌入，并进一步将这些嵌入解码为原始音频。通过操控代表不同因素的隐藏嵌入，可以控制生成的歌声，还可以通过在音色之间进行插值来进一步生成新的虚拟歌手。此外，通过在大规模数据上进行训练，该模型还能够学习建模其他独特的技能，包括歌手的口音和情感表达。我们针对各种任务在不同数据集上开展了实验，结果证明了所提出方法的有效性。所提出的方法也是制作人工智能合唱团的关键技术，该合唱团于 2022 [27]月在香港助力了人机共生交响乐团。

本文其余部分安排如下。第 2 节回顾相关工作。第 3 节将介绍所提出的无监督语音建模框架，包括编码器和解码器模型，以及端到端适配。此外，还将介绍如何操控表征以实现可控歌唱。第 4 节将说明如何基于所提出的框架生成由人工智能合唱团。第 5 节将展示并讨论实验结果。

2 相关工作

2.1 语音合成

将人声数字化的一种传统方式是语音合成，即文本转语音（TTS），它根据输入文本生成自然语音。当前最先进的文本转语音系统基于神经网络来建模自然语音复杂的动态特性，并使用成对音频和文本的数据集以监督学习方式训练模型。文本转语音的流程通常由声学模型和声码器组成：声学模型将文本信息转换为音频梅尔频谱图，声码器则进一步生成可听波形。声学模型主要致力于从文本中学习低层级的语音表示，而声码器则旨在以高保真度生成超长信号（例如，每秒 24,000 个采样点） [11]。

早期基于深度神经网络的方法通常使用循环神经网络对语音的时间依赖关系进行建模，例如 Tacotron [4] 和 DurIAN [6], , 并以自回归的方式生成梅尔频谱图。随后，人们提出了基于 Transformer的方法，包括 [28; 29; 8; 9], , 通过注意力机制中的并行计算显著加快了生成过程。变分自编码器（VAE）也被用于声学建模，典型方法包括 GMVAE-Tacotron

[30] 和变分自编码器文本转语音 [31]。最近，基于扩散的声学模型（例如 [32； 33], ）得到进一步发展，以提升合成语音的质量。将声学模型用于歌声合成（SVS）时 [16；17；18； 19], , 主要的改动是明确给出歌词的音高和时长，而不是像语音合成那样由独立模块进行估计。

声码器通常被视为一个序列到序列建模问题。一项重要工作是 WaveNet [10], ，它使用自回归（AR）卷积神经网络进行序列建模。Parallel WaveNet [11] 进一步使用逆自回归流（IAF），从预训练的 WaveNet 教师模型中提取知识，大幅提升了推理速度。人们还提出了基于 GAN 的方法来改善语音质量，包括 MelGAN [15], 、GAN-TTS [34],和 HiFi-GAN [14]。扩散模型进一步提升了语音质量，典型工作包括 SpecGrad [35] 和 DiffWave [36]。微分信号处理也被用于设计声码器，例如 Source-Filter HiFi-GAN [37]和 SawSing [38]。分别优化声学建模和声码器有助于提升模型训练的稳定性，而且通用声码器还可以训练为不同音色合成语音。不过，人们已经开发出许多端到端方法，直接根据文本输入生成音频波形。著名工作包括 Char2Wav [39], 、ClariNet [40], 和 FastSpeech 2s[9； 41], 这些方法通常将声学建模和声码器结合到一个大型编码器-解码器框架中。

2.2 语音转换

要利用语音合成将人声数字化，主要障碍之一是对规模庞大且经过仔细标注的数据集的高度依赖。数据集包含有关说话内容的文本信息；对于歌声合成，还必须精确提供歌词中每个词的时长和音高。生成高质量音频需要数小时的训练数据。由于专业的语音录制和标注成本高昂，这一要求大大增加了普通人训练语音模型的难度。对于歌声合成，生成的语音模型在演唱技巧方面也受到限制，因为几乎不可能邀请许多流行歌星参与录制。

通过将计算机视觉领域广泛研究的风格迁移 [42; 43; 44; 45； 46], 应用于音频，可以提出语音转换（VC）和歌声转换（SVC）问题：将音色视为音频的风格，并将其余信息视为内容。受 CycleGAN [45] 和 StarGAN [46] 等图像处理方法的启发，无监督学习可以从不同人群的非平行数据中解耦隐藏的音频表征。此外，通过改变音色嵌入即可实现语音转换，由此产生了 CycleGAN-VC [47], 、StarGAN-VC [48], 和 StarGANv2-VC[49]。研究重点主要在于将语言学特征与说话人表征解耦，例如基于 VAE 的方法 [50； 51], 、基于音素后验图（PPG）的方法 [52; 53], 以及基于向量量化（ VQ ）的方法 [54; 26； 55]。尽管这些方法可以将生成音频的音色转换为目标音色，却无法充分表达人声的其他独特特征，例如口音、情感和节奏。另一方面，尽管已有许多无监督方法用于解耦音频表征，但要明确解耦不同特征仍然具有挑战性。

3 语音数字化的拟议框架

本文提出一种新方法，无需标注即可根据任何人的录音将其声音数字化。录音内容和语言不限，而且该方法仅凭说话录音就能预测歌声。如图1所示，我们构建了一个无监督的基于变分自编码器的框架。类似于人类运用已有技能学习新任务，我们充分利用一组预训练音频模型的能力，将声音的不同特征解耦。与文本转语音的流程类似，所提出的框架由一个声学模型和一个声码器组成：声学模型将不同的控制因素映射到梅尔频谱图，声码器则将梅尔频谱图转换为音频波形。在分别训练声学模型和声码器后，我们进一步采用端到端训练对这两个网络进行微调。下文将详细介绍所提出的框架。

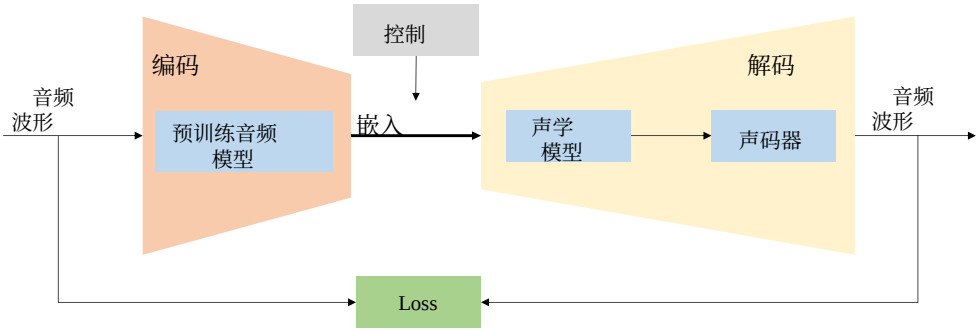


图 1：整体框架示意图。

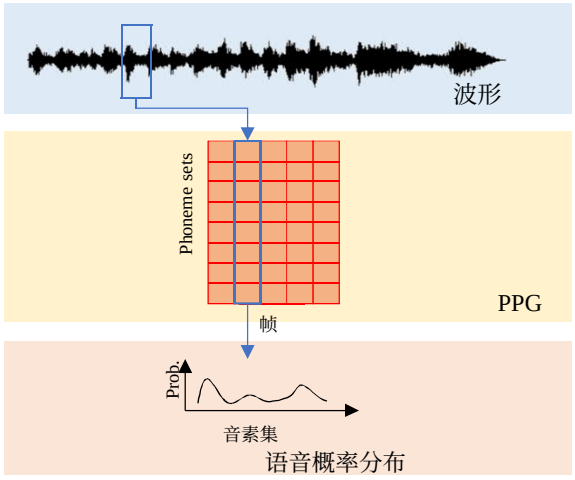


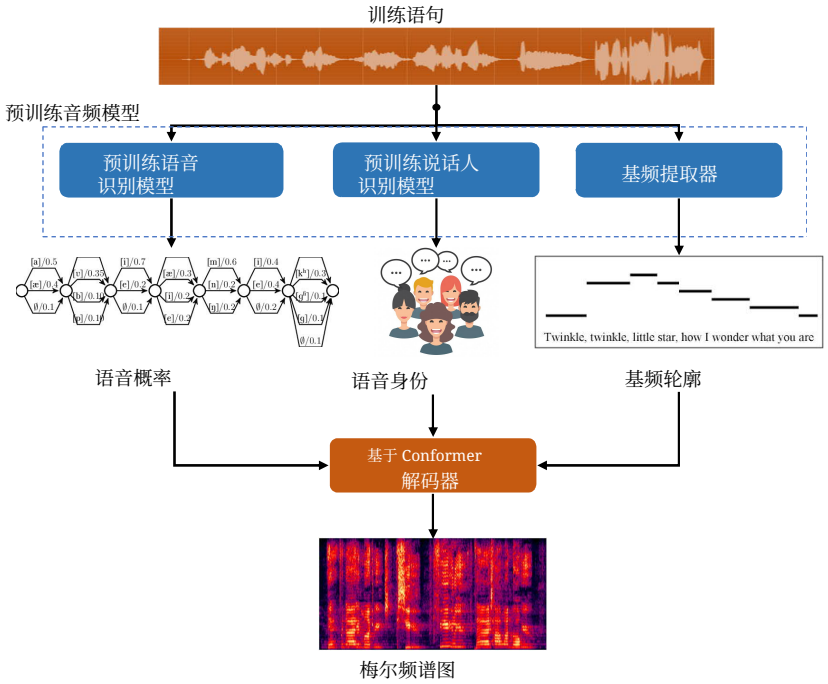
图 2：PPG 示意图。二维 PPG 的每一列都表示不同音素的概率分布。

3.1 音频模型预训练

为了识别细致的音频特征，我们旨在弄清楚：a）音频内容是什么，b）说话或唱歌的人是谁，以及 c）说话或唱歌的旋律是什么。我们注意到，目标说话人的训练数据中并没有音频和文本信息配对的数据。因此，我们不采用端到端无监督学习来解耦音频表征，而是依靠在其他标注数据集上训练的模型来识别区别性特征。这样一来，便无需依赖目标说话人的标注数据集，同时也能生成判别性嵌入。

在大规模自动语音识别数据集上训练的自动语音识别（ASR）模型能够识别内容信息。我们使用 Wenetspeech 数据集 [56], 进行训练，该数据集包含超过 10,000 小时经过准确标注的普通话语音；并采用基于 U2++network [58] 的预训练模型 [57] 生成音频内容嵌入，它是 Conformer[59] 编码块的输出。如图 2 所示，音频内容嵌入本质上是二维 PPG，每一列表示当前帧对应不同音素的概率分布。将目标说话人未标注的音频输入预训练的自动语音识别模型后，可用 PPG 表示其内容信息。自动语音识别处理的帧长为 25 ms，帧移为 10 ms。假设话语包含 T 帧，为加快解码，在自动语音识别解码时以 3 为下采样因子，得到 $T/3$ 个 PPG 帧。随后将生成的 PPG 上采样 3 倍，使其帧数与原始帧数一致。

同样，使用 VoxCeleb2 [60] 数据集进行说话人识别，以获取说话人身份。我们采用在 SpeechBrain [62] 上预训练的 ECAPA-TDNN 模型 [61] 来生成



Figu图 3：声学模型示意图，该模型将从预训练中提取的嵌入转换为
ed audi
将模型映射到输入信号的梅尔频谱图。

说话人嵌入，即模型的输出。每条话语仅获得一个说话人嵌入，并对同一说话人的所有话语的嵌入取平均值，最终表示其音色。随后将得到的嵌入重复 T 次，以表示每一帧的身份。

我们最终通过估计有声语音中的基频轮廓，识别出语音和歌声的旋律。我们利用预训练的 CREPE 模型 [63]，所得的T帧话语的基频轮廓是一个 $T \times 1$ 信号，表示每一帧的音高。

3.2 声学模型

借助预训练音频模型，我们可以将音频信号转换为一组无需任何标注的嵌入，用于表示不同的特征。随后进一步训练声学模型，将获得的嵌入转换为梅尔频谱图，之后将利用该梅尔频谱图合成可听波形。由于梅尔频谱图包含的信息比预训练模型生成的嵌入丰富得多，声学模型能够学习表示说话人的独特性，从而将该人的声音数字化。此外，嵌入还为控制生成的梅尔频谱图提供了接口。

所提出声学模型的示意图如图3所示。给定表示内容、身份和旋律信息的提取嵌入，我们设计了一个基于 Conformer 的解码网络，用于重建输入语音的梅尔频谱图。梅尔频谱图可以通过在短时傅里叶变换（STFT）域中对输入信号应用梅尔滤波器组直接计算得到。因此，声学模型可视为自动编码器，训练过程可以完全无监督地进行。

基于 Conformer 的解码器的细节如图4所示。在图4 (a) 中，三个线性层首先分别处理不同的音频嵌入，以确保所有嵌入具有相同的维度。随后，将投影后的嵌入相加，并通过一个线性层进一步变换，然后输入 Conformer 块。声学模型最终预测生成这些嵌入的话语的梅尔频谱图；与 [8], 类似，模型使用 PostNet 作为辅助网络来提升预测性能。

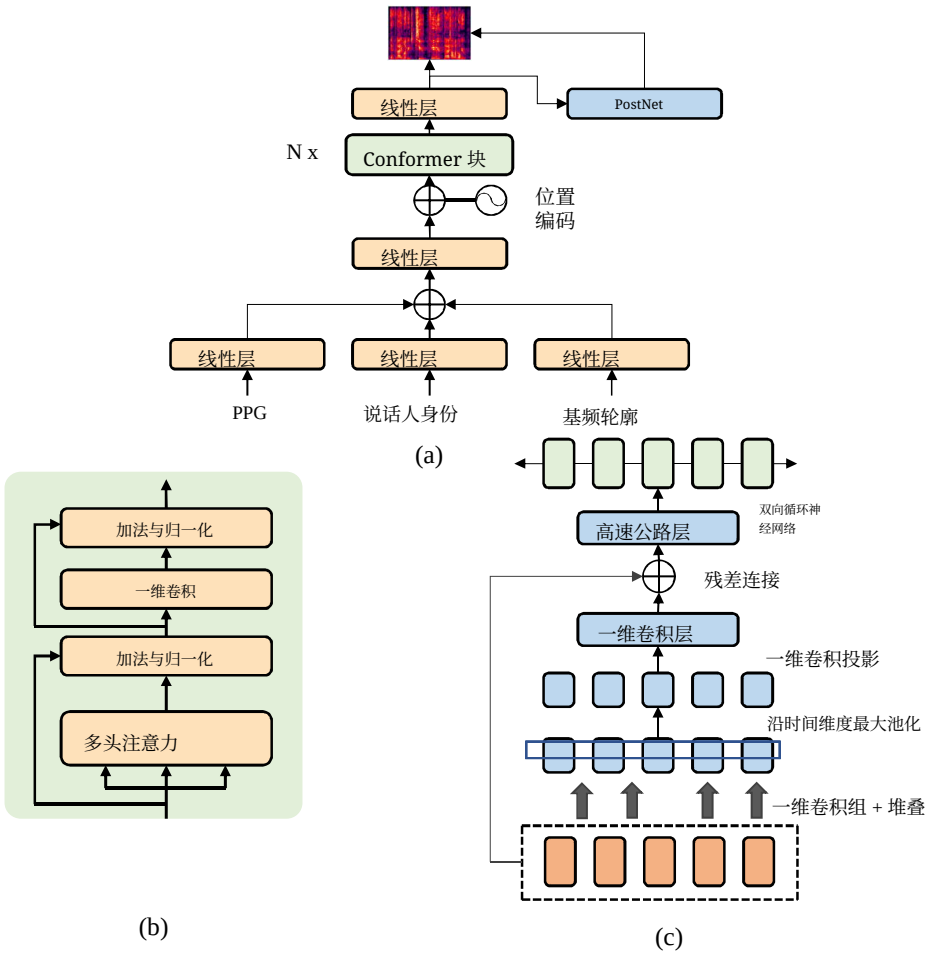


图4：声学模型中基于 Conformer 的解码器细节。(a) 整体结构；(b) Conformer 块；(c) PostNet。

Conformer 块广泛用于自动语音识别和文本转语音 [59；8； 64], ，它基本上是在 Transformer 中集成了一个卷积块，其结构如图4 (b) 所示。PostNet 的结构如图4 (c) 所示，它采用带有跳跃连接的卷积循环神经网络（CRNN）来建模梅尔频谱图的时间和结构信息。

训练期间，仅使用 L_1 和 L_2范数的组合对基于 Conformer 的解码器进行优化：

$$
\begin{aligned}
&L(\mathbf{S}_{\text{linear}}, \mathbf{S}_{\text{PostNet}}, \mathbf{S}_{\text{gt}}) \\
&= ||\mathbf{S}_{\text{linear}} - \mathbf{S}_{\text{gt}}||_1 + ||\mathbf{S}_{\text{linear}} - \mathbf{S}_{\text{gt}}||_2 + ||\mathbf{S}_{\text{PostNet}} - \mathbf{S}_{\text{gt}}||_1 + ||\mathbf{S}_{\text{PostNet}} - \mathbf{S}_{\text{gt}}||_2,
\end{aligned} \tag{1}
$$

whe其中$\mathbf{S}_{\text{linear}}$和$\mathbf{S}_{\text{PostNet}}$ 分别是最后一个线性层和 PostNet 输出的梅尔频谱图 y, and $\mathbf{S}_{\text{gt}}$ 是真实梅尔频谱图。

3.3 声码器

梅尔频谱图随后通过声码器转换为音频波形。这里采用广泛使用的 HiFi-GAN [14] 声码器，它使用多尺度和多周期判别器来确保生成波形的保真度。如图 5 所示，对于每段话语，首先计算真实梅尔频谱图，然后将其输入声码器以生成原始信号。

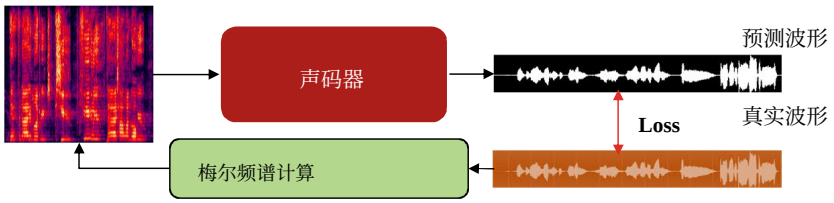


图5：声码器训练示意图。

3.4 端到端训练

我们注意到，在实际应用中，将使用声学模型预测的梅尔频谱图，而非真实值来生成音频信号。然而，预测的梅尔频谱图与真实值之间始终会存在误差，最终会影响生成音频的质量。在图1中，整个框架旨在对目标人物的音频波形执行端到端 VAE。因此，将对声学模型和声码器进行端到端训练，以进一步微调这两个模型。

在端到端训练中，声学模型与声码器结合，将预训练音频模型提取的嵌入直接转换为原始波形；更新权重时，通过冻结另一个模型交替优化二者。使用 HiFi-GAN 的损失函数优化整个端到端框架。

3.5 可控音频生成

给定声学模型和声码器，如图1所示，可以通过操控输入声学模型的音频嵌入来生成可控音频信号，甚至可以将说话录音转换为歌声。

通过对具有目标内容的现有语音录音（例如真实或合成的朗读语音，或歌曲的人声轨）应用自动语音识别模型，可以获得表示音频内容的音素后验概率。可以进行强制对齐 (FA) [56]，以确定每个音素的区间，并对与各音素对应的音素后验概率进行重采样，从而调整至目标时长。

基频轮廓也可以从现有音频中提取，无论该音频是说话还是演唱，并且可以进一步以可解释的方式进行编辑。如果训练数据中的音高变化很大，即使其中只包含说话数据，模型也能学会如何生成演唱旋律中目标音高对应的声音，从而合成歌声。

我们可以为每个人生成一个个性化模型，对应于音频形式数字孪生体；也可以生成一个能够改变音色以创造虚拟人的通用模型。通过训练一对专用声学模型和声码器，可以充分建模个人声音的独特性，例如音色、口音，以及说话和演唱时细微的节奏控制。在这种情况下，图3所示框架中的说话人身份嵌入可以固定为零向量或移除。若要生成可变音色，可以联合使用不同人的音频数据训练声学模型，此时可以操控说话人身份嵌入来控制音色。还可以训练通用声码器，将任意梅尔频谱图转换为可听信号。

4 案例研究：人工智能合唱团生成

本节介绍如何基于所提出的方法，生成由数百名虚拟歌手组成的人工智能合唱团。该方法曾助力 2022年7月 香港的人机共生交响乐团。与对单个人的声音进行数字化不同，生成合唱团会带来一个问题：生成的歌声在趋同性与多样性之间如何权衡。

为了获得令人满意的合唱效果，真人歌手必须在音色、节奏和表现力方面进行高度协调的演唱。值得注意的是，相同声音的组合并不会产生合唱效果，合唱效果实际上源于精心……

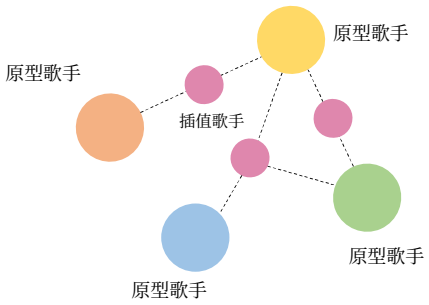


图 6：从原型歌手生成插值虚拟歌手。

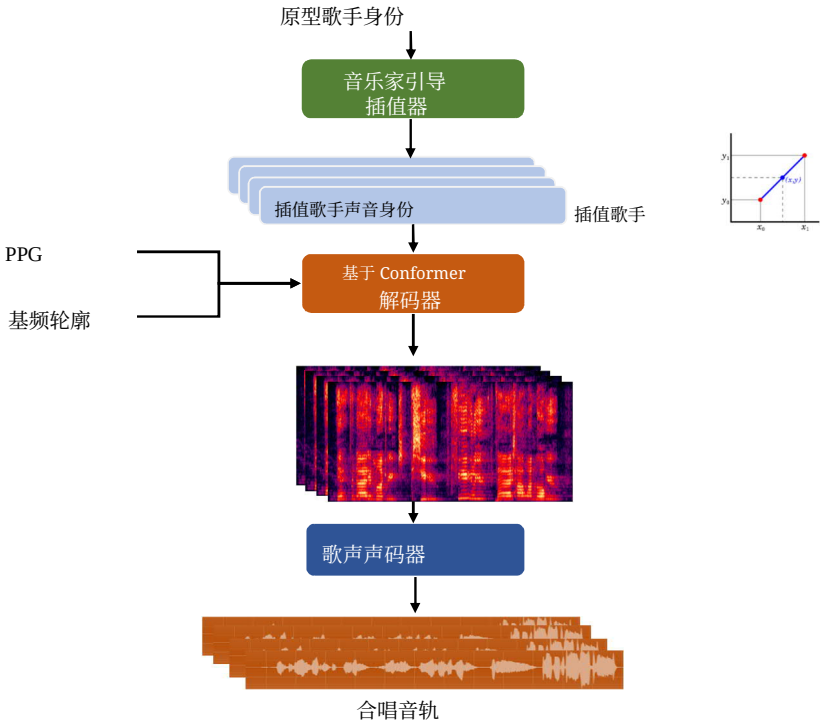


图 7：通过音色插值实现人工智能合唱团生成。

精心塑造集体演出中每位歌手在音色、节奏和表现力方面的多样性。这带来了一个多声部歌声联合生成的控制问题：必须对歌手歌声的一致性与合唱团音色的多样性进行联合优化。

我们开发了一种两阶段方法，来打造由数百位虚拟歌手组成的人工智能合唱团。首先，在第一阶段生成几位音色相似的“原型”歌手。然后，在第二阶段，如图 6 所示，通过对这些原型歌手的音色进行插值，生成数百位新的虚拟歌手。通过精细控制每位虚拟歌手的节奏和音高，即可获得和谐合唱效果。

4.1 原型歌手

在获得干净的人声录音后，通过对现有歌手的声音进行数字化，生成原型歌手。实际上，尽管网上有许多歌手的公有领域歌曲，但其中的人声信号都与伴奏混合在一起。为了收集大规模训练数据，我们使用预训练的语音分离模型 Demucs [65], 从原始声音中提取人声。利用从多位原型歌手提取的人声录音，我们训练了一个通用模型，该模型

依靠说话人嵌入来控制生成歌声的音色。为了生成合唱团，我们生成了八位音色相似的原型歌手，这些歌手由人工评估者精心挑选。

4.2 插值歌手

原型歌手可视为现有歌手的音频形式数字孪生体，而通过音色插值可以创建新的虚拟歌手。如图 7 所示，通过在原型歌手之间进行线性插值，可生成不同新虚拟歌手的说话人嵌入，并利用这些嵌入生成梅尔频谱图。

所有不同的虚拟歌手都使用相同的音素后验概率和基频轮廓嵌入，因此生成的歌声具有相同的歌词和节奏。最后，使用在包含所有歌手的数据集上训练的通用声码器，将生成的梅尔频谱图转换为音频波形。总共生成了 320 位虚拟歌手，最终组成一个合唱团。

由于使用了音色相似的原型歌手，尽管所有插值虚拟歌手的音色各不相同，但彼此仍然相似，从而确保合唱团能够协同演唱。人工评估者在组合虚拟歌手以组成合唱团的过程中也发挥着至关重要的作用。将大量虚拟歌手组合起来组成合唱团，似乎很简单。然而，人工评估者必须评估合唱团的音色融合效果是否理想，并调整原型歌手所占的比例。

5 实验

5.1 数据集

我们使用三个不同的数据集，分别评估歌声生成、语音转歌声和人工智能合唱团生成的表现。

OpenCpop。 我们采用公开可用的高质量普通话歌唱语料库 OpenCpop [20], , 以检验所提方法生成高质量歌声的能力。该语料库包含由一位专业女歌手录制的 100 首无伴奏歌曲，音频被切分为 3,756 段话语，总时长为 5。2 小时。尽管原始数据集中包含音符和音素信息，所提方法仅使用音频波形来实现歌手声音的数字化。

口语音频。 为了测试语音转歌声的表现，我们根据一位男性同事在日常 Zoom 会议中的录音，自行构建了一个时长为 3。7 小时的口语音频数据集。值得注意的是，目前已有许多高质量的公开数据集（例如用于朗读语音合成的 LibriTTS [66]）。然而，使用从 Zoom 会议中收集的数据，有助于检验利用日常普通质量的语音实现声音数字化的可行性。为便于讨论，在后续实验中，我们将该数据集称为“口语”。
多位歌手的音频。为了生成合唱团，我们进一步从网上（YouTube、Spotify 等）收集了八位专业歌手的歌曲，其中包括四位男歌手和四位女歌手。我们为每位歌手收集了近 4 小时的数据，并使用 Demucs [65] isused 从所有歌曲中提取人声轨。下文将这个数据集称为“Multi-Sing”。

对于所有数据集，音频都被转换为单声道，采样率为 24 kHz。进行无监督训练不需要其他信息。我们通过根据 VocalSet [21], 中的“straight”片段合成歌声来测试所得模型，并确保训练数据集与测试话语的性别相匹配。

5.2 实现

声学模型。声学模型的结构如图4所示。音素后验概率、说话人身份和基频轮廓嵌入的维度分别为 320、256 和 1，且均通过线性层转换为 320 维。声学模型包含 6 个 Conformer 块，每个 Conformer 块使用 2 个注意力头。训练目标为 80 维梅尔频谱图，分析帧长为 25 ms，帧移为 10 ms。Conformer 的最大长度为 1000 帧，在帧移为 10 ms 时对应 10 s。该模型是】【。

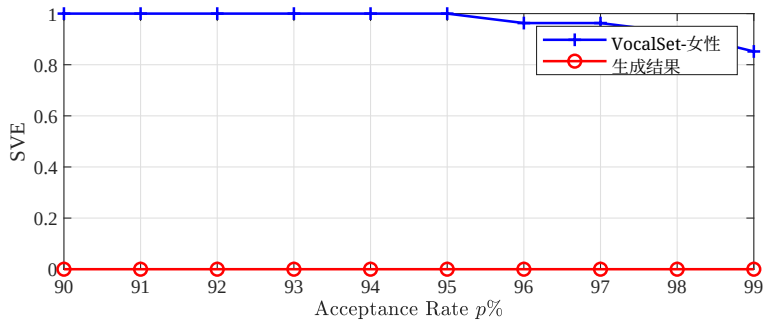


图 8：OpenCpop 数据集的 SVE 随接受率变化的情况。接受率用于确定说话人验证的阈值。

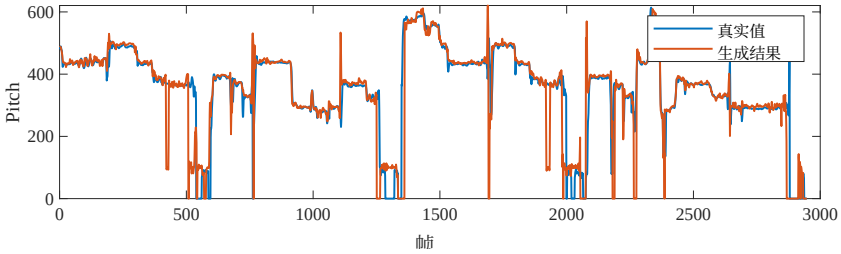


图 9：OpenCpop 歌手生成的演唱音频中基频轮廓与真实值的比较。

以批次大小 24、学习率为 1×10^{-3}、权重衰减为 1×10^{-6} 进行训练。训练共进行 2000 个周期。

由于训练数据集中只包含一个人，因此在 OpenCpop 和 Speaking 数据集上训练得到的模型依赖于说话人。在 Multi-Sing 数据集上训练时，可以通过说话人身份嵌入控制输出音频的音色。

声码器。 为从 80 维梅尔频谱图生成采样率为 24 kHz 的音频波形，HiFi-GAN 生成器中包含 4 个基于转置卷积的上采样模块，上采样率分别为 $\{8, 8, 2, 2\}$，上采样卷积核大小分别为 $\{16, 16, 4, 4\}$。序列到序列建模的片段大小为 8192 个采样点。训练时批量大小设为 32，学习率为 2×10^{-4}，模型训练了 200，000 步。
与声学模型类似，我们针对 OpenCpop 和 Speaking 数据集分别训练了说话人专属模型，并使用 Multi-Sing 数据集中的所有数据训练了一个通用声码器。
端到端训练。 针对每个数据集训练好声学模型和声码器后，我们最终进行端到端训练，进一步微调这两个模型。我们首先冻结声学模型，更新声码器 5，000 步；然后冻结所得声码器，训练声学模型 100 个 epoch。此过程重复 5 次。

5.3 结果

歌声生成。 在本实验中，我们根据 OpenCpop 歌手的演唱录音对她进行数字化。

我们首先检查生成音频的说话人身份相似度。使用预训练的 ECAPA- 时延神经网络模型 [61], , 计算 OpenCpop 数据集中所有话语的说话人嵌入，然后取平均值，得到 OpenCpop 歌手的整体说话人嵌入，在此记作$\mathbf{S}_{\text{cpop}}$。下文将使用$\mathbf{S}_{\text{cpop}}$ 作为锚点，并根据阈值检查生成的音频是否与 OpenCpop 歌手相似。对于 OpenCpop 中的所有话语，计算其说话人嵌入与$\mathbf{S}_{\text{cpop}}$ 之间的余弦相似度，然后确定阈值

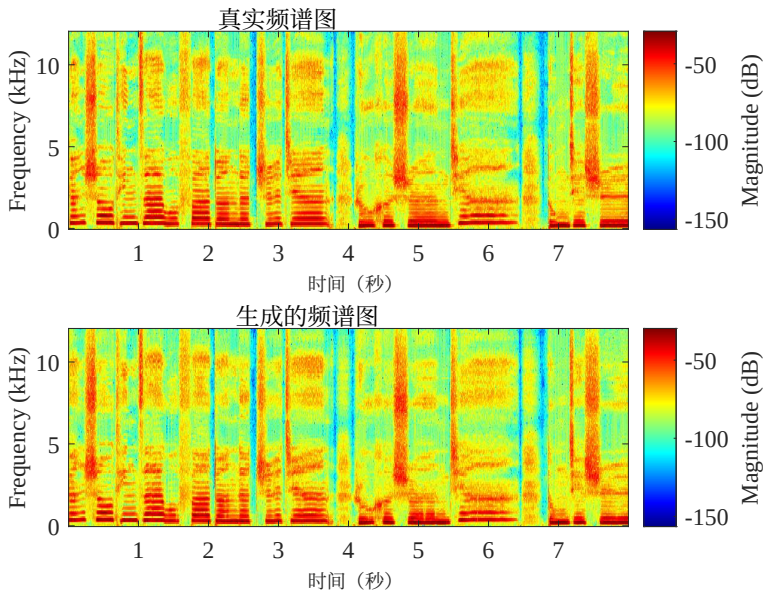
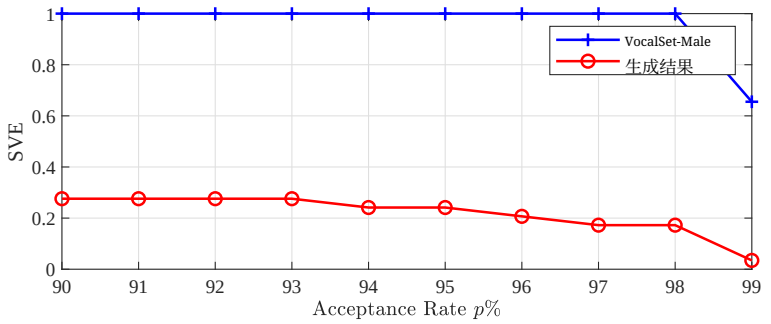


图 10：OpenCpop 歌手的真实值与生成信号的线性频谱图。



Fig图 11：“Speaking”数据集的 SVE 随接受率变化的曲线。接受率 e is u用于确定说话人验证的阈值。

is d该阈值通过对相似度分数排序后，选取“ p%”的话语来确定。可以看出， e hig越高 p，阈值越低。

如上文所述，使用 VocalSet 中女歌手的 25 个“直声”片段生成 25 个话语。计算生成音频与 Scpop 之间的余弦相似度，并通过将相似度与阈值进行比较，得到说话人验证错误率（SVE）。作为对比，还计算了 VocalSet 原始片段的相同指标。图 8 展示了生成音频和 VocalSet 音频的 SVE 随接受率p%变化的情况。在所有接受率下，生成音频的 SVE 均为零，而 VocalSet 音频的 SVE 接近一，这表明所提出的方法能够有效学习 OpenCpop 歌手的音色。

随后，我们评估生成音频的音高准确度。图 9 展示了一个示例，其中包含生成音频的基频轮廓以及来自 VocalSet 的输入音频的基频轮廓。将基频轮廓作为声学模型的控制因素后，生成音频应当遵循指定的旋律。结果表明，几乎所有情况下，生成音频都能产生所需的音高，但在起始和结束处存在一些误差。

我们进一步在图10中展示了生成音频与真实值频谱图的一个示例。生成音频的获取过程是：首先从真实值音频中提取嵌入，然后使用未经修改的嵌入重建波形信号。我们可以观察到

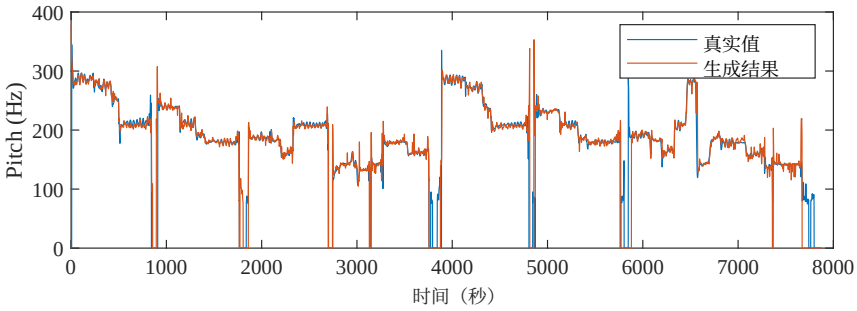


图 12：男性说话人从说话转为演唱时，生成音频中的演唱音高基频轮廓与真实值的对比。

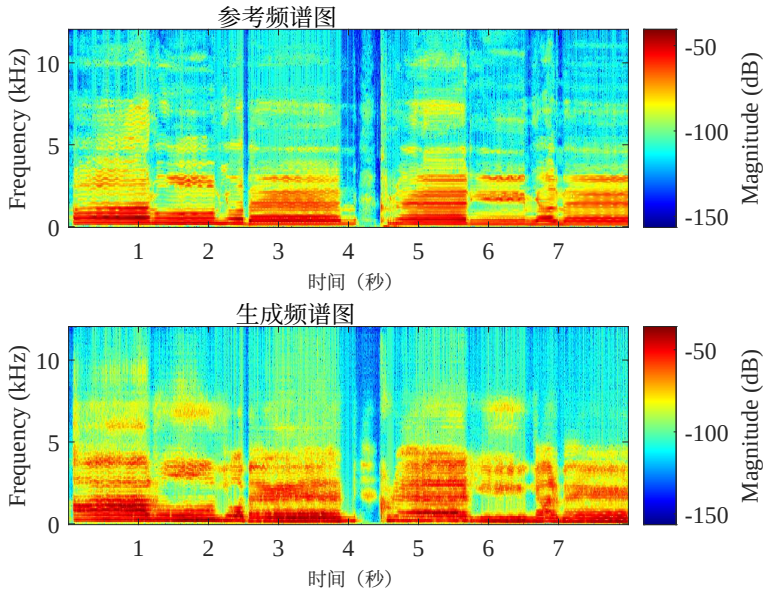


图 13：“Speaking”数据集的参考信号与生成信号的线性频谱图。生成了歌声。值得注意的是，参考音频与生成音频对应的是不同的说话人。

生成音频能够有效捕捉原始歌声音频的时频特征，表明该方法具备生成高质量音频输出的能力。

从说话到歌唱。使用“说话”数据集，我们进一步评估所提出的方法能否仅根据只包含说话语音的数据集生成歌声。与 OpenCpop 数据集的实验类似，我们使用来自 VocalSet 男歌手的 29 段“直声”片段作为模型的输入信号。结果如图11所示。我们可以看到，按照说话人嵌入相似度剔除 10% 的训练数据时，近 27% 的生成音频被分类为与目标说话人不同。将接受率提高至 99% 后，几乎所有生成样本都被判定为目标说话人，这表明生成音频与平均说话人嵌入之间的最大差异与训练集相当。我们还注意到，所有情况下的 SVE 值都明显高于 OpenCpop 的情况，说明仅使用说话数据生成歌声更具挑战性。

图12比较了生成的演唱音频与真实值的基频轮廓。可以看出，尽管模型仅使用说话数据集进行训练，训练得到的模型仍能以很高的精度生成演唱基频轮廓。图13比较了来自 VocalSet 的参考音频与生成音频的频谱图。可以发现，尽管生成的信号能够有效跟随参考信号的旋律和节奏，但两者的频率

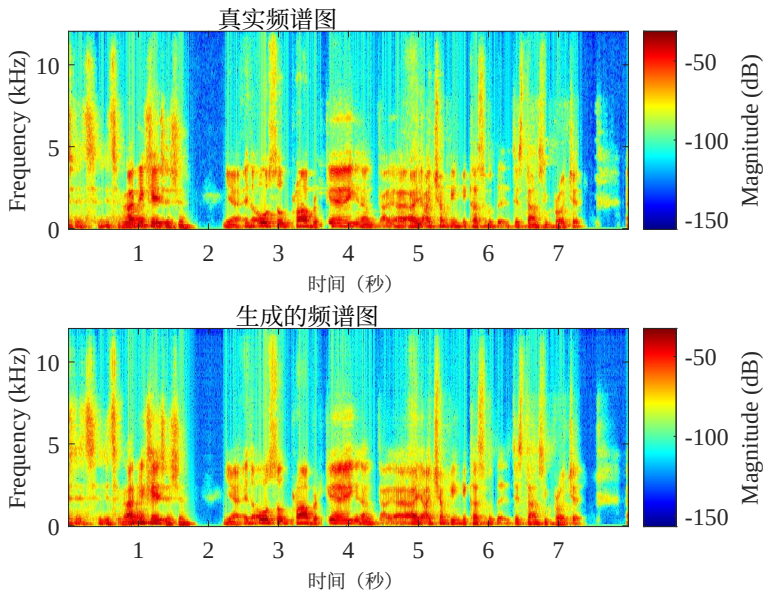


图 14：“说话”数据集的真实值信号和生成信号的线性频谱图。重建了说话语音。请注意，真实值音频和生成音频对应同一位说话人。

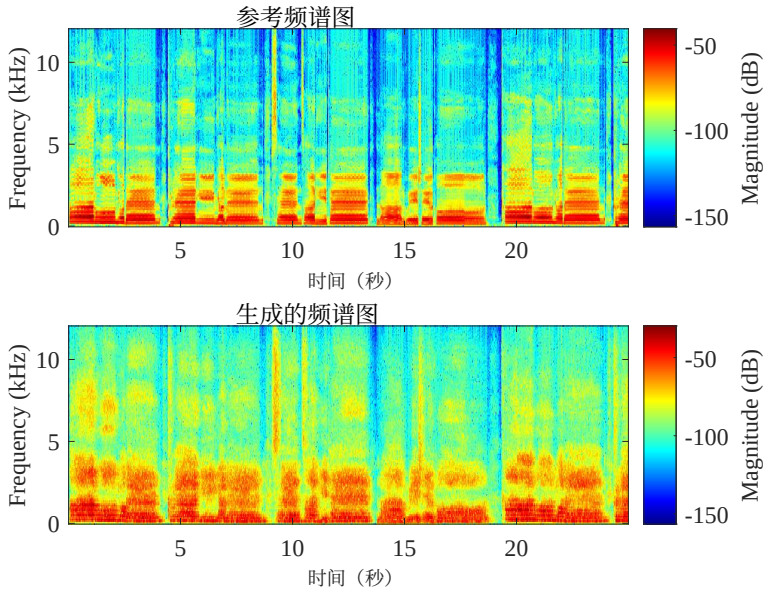


图 15：参考合唱团与由 120 位男歌手组成的生成合唱团的线性频谱图。

分布有所不同。这进而表明，生成的音频与参考音频呈现出不同的共振峰，说明说话人身份已成功修改。如图14所示，在重建训练说话人的说话音频时，所提出的方法仅凭提取的嵌入即可实现准确的重建效果。

合唱团生成。最后，我们评估所提出的合唱团生成方法的性能，该合唱团由数十至数百名虚拟歌手组成。通过身份插值，可以利用原型歌手生成新的虚拟歌手，而原型歌手是使用在线收集的数据训练而成的。我们注意到，每位歌手的歌声表现都与 OpenCpop 歌手相似，因此可以得出相同的结论；因此，对每段生成音频的评估是

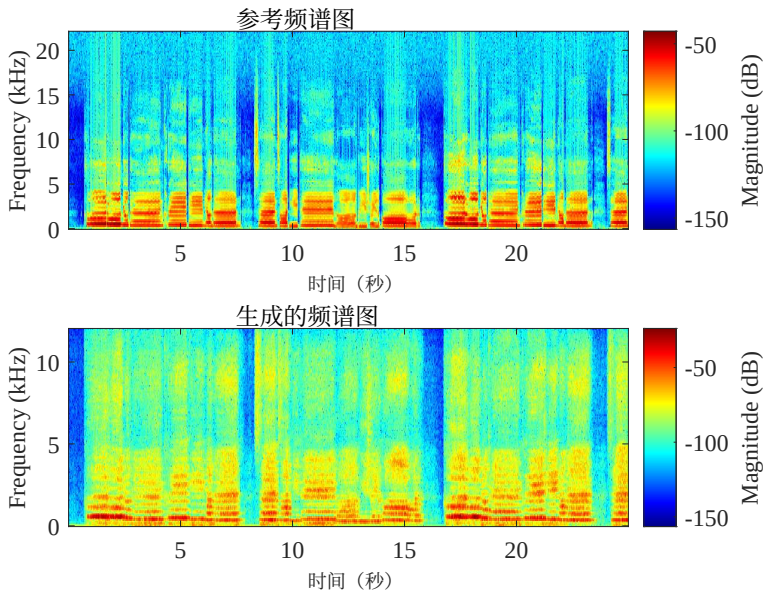


图 16：由 120 位女歌手组成的合唱团的参考线性频谱图和生成的线性频谱图。

此处未予展示。在图1 5中，我们展示了一个由160名男性虚拟歌手组成的人工智能合唱团。我们注意到，由于每位歌手都具有独特的共振峰，音高谐波在合唱团的混合信号中变得难以区分，从而产生了合唱效果。从图1 6所示的女性合唱团中也可以观察到类似的结果。由于女性音色具有更高的多样性，其频率分布范围比男性合唱团宽得多。

6 结论

本文提出了一种用于无监督语音模型的全新框架，能够创建数字歌唱人类。我们的方法依赖于变分自编码器（VAE），将录音编码为代表歌声不同因素的各种隐藏嵌入，随后可对这些嵌入进行操控，以控制音高、旋律和歌词等不同的演唱特征。通过在大规模数据上进行训练，该方法还能够学习建模其他独特技能，包括歌手的口音和情感表达。在不同数据集上的实验结果证明了所提方法的有效性。

参考文献

[1] J. Thyssen、H. Nielsen 和 S. D. Hansen，《语音编码中的非线性短期预测》，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，1994年4月，第185–188页。[2] J. D. Markel 和 A. H. Gray, Jr.，《语音线性预测》。施普林格出版社，1976年。[3] L. R. Rabiner 和 R. W. Schafer 编，《数字信号处理理论与应用》。培生，2010年。[4] Y. Wang、R. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio 等，《Tacotron：迈向端到端语音合成》，arXiv 预印本 arXiv:1703.10135，2017年。[5] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，《通过以梅尔频谱图预测结果为条件的 WaveNet 实现自然文本转语音合成》，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，2018年，第4779–4783页。[6] C. Yu、H. Lu、N. Hu、M. Yu、C. Weng、K. Xu、P. Liu、D. Tuo、S. Kang、G. Lei、D. Su 和 D. Yu，《DurIAN：时长感知注意力网络用于语音合成》，载于国际语音通信协会会议（INTERSPEECH）论文集，2020年。[7] S. Vasquez 和 M. Lewis，《MelNet：频域音频生成模型》，arXiv 预印本 arXiv:1906.01083，2019年。[8] Y. Ren、Y. Ruan、X. Tan、T. Qin、S. Zhao、Z. Zhao 和 T. Liu，《FastSpeech：快速、稳健且可控的文本转语音》，载于神经信息处理系统大会（NeurIPS）论文集，2019年，第3165–3174页。[9] Y. Ren、C. Hu、X. Tan、T. Qin、S. Zhao、Z. Zhao 和 T. Liu，《FastSpeech 2：快速且高质量的端到端文本转语音》，载于国际学习表征会议（ICLR）论文集，2021年。[10] A. van den Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. W. Senior 和 K. Kavukcuoglu，《WaveNet：原始音频生成模型》，载于国际语音通信协会语音合成研讨会，2016年。[11] A. van den Oord、Y. Li、I. Babuschkin、K. Simonyan、O. Vinyals、K. Kavukcuoglu、G. van den Driessche、E. Lockhart、L. C. Cobo、F. Stimberg、N. Casagrande、D. Grewe、S. Noury、S. Dieleman、E. Elsen、N. Kalchbrenner、H. Zen、A. Graves、H. King、T. Walters、D. Belov 和 D. Hassabis，《Parallel WaveNet：快速高保真语音合成》，载于国际机器学习大会（ICML）论文集，2018年，第3915–3923页。[12] S. Mehri、K. Kumar、I. Gulrajani、R. Kumar、S. Jain、J. Sotelo、A. C. Courville 和 Y. Bengio，《SamplerRNN：无条件端到端神经音频生成模型》，载于国际学习表征会议（ICLR）论文集，2017年。[13] J. Valin 和 J. Skoglund，《LPCNet：通过线性预测改进神经语音合成》，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集。IEEE，2019年，第5891–5895页。[14] J. Kong、J. Kim 和 J. Bae，《HiFi-GAN：用于高效高保真语音合成的生成对抗网络》，载于神经信息处理系统大会（NeurIPS）论文集，2020年。[15] K. Kumar、R. Kumar、T. de Boissiere、L. Gestin、W. Z. Teoh、J. Sotelo、A. de Brébisson、Y. Bengio 和 A. C. Courville，《MelGAN：用于条件波形合成的生成对抗网络》，载于神经信息处理系统大会（NeurIPS）论文集，2019年。[16] P. Lu、J. Wu、J. Luan、X. Tan 和 L. Zhou，《小冰歌唱：高质量集成式歌声合成系统》，载于国际语音通信协会会议（INTERSPEECH）论文集，2020年。[17] Y. Song、W. Song、W. Zhang、Z. Zhang、D. Zeng、Z. Liu 和 Y. Yu，《结合颤音建模和潜在能量表征的歌声合成》，载于 IEEE 多媒体信号处理国际研讨会（MMSP）论文集，2022年，第1–6页。[18] J. Liu、C. Li、Y. Ren、F. Chen 和 Z. Zhao，《DiffSinger：通过浅层扩散机制实现歌声合成》，载于人工智能促进协会会议（AAAI）论文集，2022年，第11 020–11 028页。[19] R. Huang、C. Cui、F. Chen、Y. Ren、J. Liu、Z. Zhao、B. Huai 和 Z. Wang，《SingGAN：用于高保真歌声生成的生成对抗网络》，载于 ACM 多媒体国际会议（ACM MM）论文集，2022年，第2525–2535页。[20] Y. Wang、X. Wang、P. Zhu、J. Wu、H. Li、H. Xue、Y. Zhang、L. Xie 和 M. Bi，《OpenCpop：用于歌声合成的高质量开源中文流行歌曲语料库》，载于国际语音通信协会会议（INTERSPEECH）论文集，2022年，第4242–4246页。

[21] J. Wilkins、P. Seetharaman、A. Wahl 和 B. Pardo，《Vocalset：歌声数据集》，载于 国际音乐信息检索学会会议论文集（*ISMIR*），2018年。[22] J. Chou、C. Yeh、H. Lee 和 L. Lee，《通过对抗式学习解耦音频表征实现无需平行数据的多目标语音转换》，载于 国际语音通信协会会议论文集（*INTERSPEECH*），2018年，第501–505页。[23] J. Chou 和 H. Lee，《通过实例归一化分离说话人表征与内容表征，实现单样本语音转换》，载于 国际语音通信协会会议论文集（*INTERSPEECH*），2019年，第664–668页。[24] Y. Wang、D. Stanton、Y. Zhang、R. J. Skerry-Ryan、E. Battenberg、J. Shor、Y. Xiao、Y. Jia、F. Ren 和 R. 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》，载于 国际机器学习大会论文集（*ICML*），2018年，第5167–5176页。[25] H. Lu、Z. Wu、D. Dai、R. Li、S. Kang、J. Jia 和 H. Meng，《利用全局说话人嵌入实现单样本语音转换》，载于 国际语音通信协会会议论文集（*INTERSPEECH*），G. Kubin 和 Z. Kacic 编，2019年，第669–673页。[26] D. Wang、L. Deng、Y. T. Yeung、X. Chen、X. Liu 和 H. Meng，《VQMIVC：基于向量量化和互信息的无监督语音表征解耦，用于单样本语音转换》，载于 国际语音通信协会会议论文集（*INTERSPEECH*），2021年，第1344–1348页。[27] Y. Guo、Q. Liu、J. Chen、W. Xue、H. Jensen、F. Rosas、J. Shaw、X. Wu、J. Zhang 和 J. Xu，《通往未来共生式创造力的路径》，*arXiv*预印本 *arXiv:2209.02388*，2022年。[28] N. Li、S. Liu、Y. Liu、S. Zhao 和 M. Liu，《基于 Transformer 网络的神经语音合成》，载于 人工智能促进协会人工智能会议论文集（*AAAI*），2019年。[29] M. Chen、X. Tan、Y. Ren、J. Xu、H. Sun、S. Zhao 和 T. Qin，《Multispeech：基于 Transformer 的多说话人文本转语音》，载于国际语音通信协会会议论文集（*INTERSPEECH*），2020年。[30] W. Hsu、Y. Zhang、R. J. Weiss、H. Zen、Y. Wu、Y. Wang、Y. Cao、Y. Jia、Z. Chen、J. Shen、P. Nguyen 和 R. Pang，《用于可控语音合成的分层生成建模》，载于 国际学习表征会议论文集（*ICLR*），2019年。[31] Y. Zhang、S. Pan、L. He 和 Z. Ling，《学习用于端到端语音合成中风格控制与迁移的潜在表征》，载于 *IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2019年。[32] M. Jeong、H. Kim、S. J. Cheon、B. J. Choi 和 N. S. Kim，《Diff-tts：用于文本转语音的去噪扩散模型》，载于国际语音通信协会会议论文集（*INTERSPEECH*），2021年，第3605–3609页。[33] V. Popov、I. Vovk、V. Gogoryan、T. Sadekova 和 M. A. Kudinov，《Grad-tts：用于文本转语音的扩散概率模型》，载于国际机器学习大会论文集（*ICML*），2021年，第8599–8608页。[34] M. Binkowski、J. Donahue、S. Dieleman、A. Clark、E. Elsen、N. Casagrande、L. C. Cobo 和 K. Simonyan，《利用对抗网络实现高保真语音合成》，载于 国际学习表征会议论文集（*ICLR*），2020年。[35] Y. Koizumi、H. Zen、K. Yatabe、N. Chen 和 M. Bacchiani，《Specgrad：基于扩散概率模型、具有自适应噪声频谱整形的神经声码器》，载于 国际语音通信协会会议论文集（*INTERSPEECH*），2022年，第803–807页。[36] Z. Kong、W. Ping、J. Huang、K. Zhao 和 B. Catanzaro，《Diffwave：用于音频合成的通用扩散模型》，载于 国际学习表征会议论文集（*ICLR*），2021年。[37] R. Yoneyama、Y. Wu 和 T. Toda，《Source-Filter HiFi-GAN：快速且可控制音高的高保真神经声码器》，*arXiv.2210.15533*，2022年。[38] D. Wu、W. Hsiao、F. Yang、O. Friedman、W. Jackson、S. Bruzenak、Y. Liu 和 Y. Yang，《基于 DDSP 的歌声声码器：一种新的减法式合成器及全面评估》，*arXiv.2208.04756*，2022年。[39] J. Sotelo、S. Mehri、K. Kumar、J. F. Santos、K. Kastner、A. C. Courville 和 Y. Bengio，《Char2Wav：端到端语音合成》，载于国际学习表征会议论文集（*ICLR*），2017年。[40] W. Ping、K. Peng 和 J. Chen，《ClariNet：端到端文本转语音中的并行波形生成》，载于 国际学习表征会议论文集（*ICLR*），2019年。

[41] D. J. Lim, S. Jung, 1927 F. Kim, “JETS：联合训练 FastSpeech2 和 HiFi-GAN，实现端到端文本转语音”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2022年，第 21–25 页。[42] L. A. Gatys、A. S. Ecker 和 M. Bethge，“艺术风格的神经算法”，*arXiv.1508.06576*，2015年。[43] ——，“使用卷积神经网络进行图像风格迁移”，发表于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），2016年。[44] Y. Li、N. Wang、J. Liu 和 X. Hou，“揭开神经风格迁移的奥秘”，发表于国际人工智能联合会议论文集（*IJCAI*），2017年，第 2230–2236 页。[45] J. Zhu、T. Park、P. Isola 和 A. A. Efros，“使用循环一致对抗网络进行非配对图像到图像转换”，发表于*IEEE*国际计算机视觉会议论文集（*ICCV*），2017年，第 2242–2251 页。[46] Y. Choi、M. Choi、M. Kim、J. Ha、S. Kim 和 J. Choo，“StarGAN：用于多域图像到图像转换的统一生成对抗网络”，发表于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），2018年，第 8789–8797 页。[47] T. Kaneko 和 H. Kameoka，“CycleGAN-VC：使用循环一致对抗网络进行非平行语音转换”，发表于欧洲信号处理会议论文集（*EUSIPCO*），2018年，第 2100–2104 页。[48] H. Kameoka、T. Kaneko、K. Tanaka 和 N. Hojo，“StarGAN-VC：使用星形生成对抗网络进行非平行多对多语音转换”，2018年，第 266–273 页。[49] Y. A. Li、A. Zare 和 N. Mesgarani，“StarGANv2-VC：一种用于自然音质语音转换的多样化无监督非平行框架”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2021年。[50] H. Lu、D. Wang、X. Wu、Z. Wu、X. Liu 和 H. Meng，“利用 β-VAE进行单样本跨语言语音转换的解耦语音表征学习”，发表于*IEEE*口语语言技术研讨会（*SLT*），2022年，第 814–821 页。[51] J. Williams、Y. Zhao、E. Cooper 和 J. Yamagishi，“在半监督VQ-VAE范式下学习解耦的音素和说话人表征”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2021年，第 7053–7057 页。[52] Z. Li、B. Tang、X. Yin、Y. Wan、L. Xu、C. Shen 和 Z. Ma，“基于PPG的歌声转换与对抗表征学习”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2021年，第 7073–7077 页。[53] L. Sun、K. Li、H. Wang、S. Kang 和 H. M. Meng，“无需平行数据训练的多对一语音转换音素后验图”，发表于国际多媒体与博览会会议论文集（*ICME*），2016年。[54] H. Tang、X. Zhang、J. Wang、N. Cheng 和 J. Xiao，“AVQVC：通过应用对比学习和向量量化实现单样本语音转换”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2022年，第 4613–4617 页。[55] D. Wu 和 H. Lee，“通过向量量化实现单样本语音转换”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2020年。[56] B. Zhang、H. Lv、P. Guo、Q. Shao、C. Yang、L. Xie、X. Xu、H. Bu、X. Chen、C. Zeng、D. Wu 和 Z. Peng，“WENETSPEECH：用于语音识别的 10000+ 小时多领域普通话语料库”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2022年，第 6182–6186 页。[57] “WeNet中的预训练模型”，https://github.com/wenet-e2e/wenet/blob/main/docs/pretrained_models.md，2022年。[58] B. Zhang、D. Wu、Z. Peng、X. Song、Z. Yao、H. Lv、L. Xie、C. Yang、F. Pan 和 J. Niu，“Wenet 2.0：更高效的端到端语音识别工具包”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2022年，第 1661–1665 页。[59] A. Gulati、J. Qin、C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 和 R. Pang，“Conformer：用于语音识别的卷积增强型Transformer”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2020年，第 5036–5040 页。[60] J. S. Chung、A. Nagrani 和 A. Zisserman，“VoxCeleb2：深度说话人识别”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2018年，第 1086–1090 页。

[61] B. Desplanques、J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：时延神经网络说话人验证中的强调通道注意力、传播与聚合”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2020年，第 3830–3834 页。[62] M. Ravanelli、T. Parcollet、P. Plantinga、A. Rouhe、S. Cornell、L. Lugosch、C. Subakan、N. Dawalatabad、A. Heba、J. Zhong、J. Chou、S. Yeh、S. Fu、C. Liao、E. Rastorgueva、F. Grondin、W. Aris、H. Na、Y. Gao、R. D. Mori 和 Y. Bengio，“SpeechBrain：通用语音工具包”，*arXiv*预印本：*arXiv.2106 04624*，2021。.[63] J. W. Kim、J. Salamon、P. Li 和 J. P. Bello，“CREPE：用于基音估计的卷积表示”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2018年，第 161–165 页。[64] O. Chang、H. Liao、D. Serdyuk、A. Shah 和 O. Siohan，“视觉语音识别只需Conformer”，*arXiv*预印本 *arXiv.2302.10915*，2023年。[65] A. Défossez、N. Usunier、L. Bottou 和 F. R. Bach，“Demucs：利用额外无标签数据混音的音乐源深度提取器”，*arXiv*预印本 *arXiv.1909.01174*，2019年。[66] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，“LibriTTS：从LibriSpeech衍生的文本转语音语料库”，发表于国际语音通信协会会议论文集（*INTERSPEECH*），2019年，第 1526–1530 页。