arXiv:2304.04618v1 [cs.SD] 10 Apr 2023

增强型 SP EECH 到语音翻译，支持多个文本转语音目标 S

Jiatong Shi[1], *Yun Tang*[2], *A nn Lee*[2], *Hirofumi Inaguma*[2],*Changhan Wang*[2],*Juan Pino*[2],*Shinji Watanabe*[1]

[1] 卡内基梅隆大学 [2] MetaAI{jiatongs,swatanbe}
@cs.cmu.edu, {yuntang,annl}@meta.com

摘要

众所周知，由于现有的源语和目标语语音平行数据有限，直接语音到语音翻译（S2ST）模型通常面临数据稀缺问题。因此，为了训练直接 S2ST 系统，以往的研究通常利用文本转语音（TTS）系统，通过扩增语音到文本翻译（S2TT）数据来生成目标语言的样本。然而，合成的目标语音会如何影响 S2ST 模型，目前仍缺乏研究。本文分析了改变合成目标语音对直接 S2ST 模型的影响。我们发现，简单地结合来自不同文本转语音系统的目标语音，有望提升 S2ST 的性能。在此基础上，我们还提出了一种多任务框架，利用来自不同文本转语音系统的多个目标共同优化 S2ST 系统。大量实验表明，在 Fisher 西班牙语-英语数据集上，我们提出的框架始终优于基线方法（提升 2.8 双语评估替补分）。

索引词— 语音到语音翻译、文本转语音增强、离散单元

1. 引言

语音到语音翻译（S2ST）旨在将源语言的语音翻译成目标语言的语音 [1]。传统的级联语音到语音翻译模型将任务分解为三个组成部分，包括自动语音识别（ASR）、机器翻译（MT）和文本转语音（TTS） [2, 3]。此外，部分先前研究采用端到端语音到文本翻译（S2TT）来取代 ASR 和 MT。然而，这会给后续应用带来较高的计算成本和推理延迟。为缓解这一问题，近期研究聚焦于构建无需三个独立模块的直接语音到语音翻译模型 [4–12]。

直接语音到语音翻译模型的训练不可避免地需要大量平行语音到语音翻译语料库，而这些语料库比传统的级联方法更难获取 [6]。为缓解这一问题并实现语音到语音翻译模型的训练，先前的工作将文本转语音系统纳入其中，以构建用于语音到语音翻译的数据集 [4–11]。几乎所有已发表的语音到语音翻译数据集都扩展自语音到文本语料库，其中用于语音到语音翻译的目标语音由文本转语音系统合成 [13–15]。在合成用于语音到语音翻译的目标语音时，先前工作的研究人员通常会选择特定的文本转语音系统。例如，在 [14], 中，他们使用了 Non-attentive Tacotron（NAT）的一个变体 [16], , 而在 [15], 中，他们采用了 Fastspeech2 [17]。据我们所知，目前尚未有研究探讨不同的合成目标语音将如何影响语音到语音翻译建模。

为填补上述研究空白，本文聚焦于不同文本转语音系统生成语音的影响。我们发现，仅使用来自多个文本转语音系统的训练数据即可提- 高语音到语音翻译的性能。为进一步利用共享知-在多个文本转语音系统之间的优势基础上，我们进一步提出了一种框架，联合优化以不同文本转语音系统为目标的语音到语音翻译系统。结果表明，与基线模型相比，我们提出的方法能够显著提升语音到语音翻译系统的性能。具体而言，在 Fisher 西班牙语-英语数据集 [18] 上，与采用单一文本转语音目标的最佳基线系统相比，我们提出的框架使 BLEU 分数提高了 2.8 分。本工作的贡献可概括如下：

- 我们首先研究了不同文本转语音系统对语音到语音翻译目标合成语音的影响。
- 我们提出了一种多任务框架，融合来自不同文本转语音数据的知识；实验结果表明，该框架带来了合理的性能提升。

2. 方法

本节首先回顾本研究的背景，包括基于离散单元的语音到语音翻译系统，以及本文使用的各种文本转语音系统。随后，我们介绍所提出的框架，该框架用于融合来自不同文本转语音系统的目标语音知识。

2.1. 背景

基于离散单元的语音到语音翻译：语音自监督学习（SSL）模型在各种任务中都展现出卓越的性能 [19,20]。值得注意的是，这些模型也适用于合成任务 [21–23]。应用 SSL 表征的一种常见策略是通过聚类方法将其离散化为语音单元 [21, 24]。以往研究表明，离散单元能够将语言内容与其他声学属性（例如说话人身份或韵律信息）分离，从而更容易直接从语音中学习语言信息 [21]。因此，Lee等人提出了一种直接的语音到语音翻译模型，以离散语音单元作为系统的预测目标 [7]。他们的实验还表明，该模型优于基于 Translatotron 的方法[4, 5] ，且性能与级联式语音到语音翻译系统相当[2, 3, 25]。其系统中的离散单元由预训练 HuBERT 模型表征上的 K-Means 聚类生成 [26]。

文本转语音系统：如第 1 节所述，以往研究通常采用文本转语音系统为语音到语音翻译生成目标语音。Translatotron 系列研究主要采用自回归（AR）文本转语音系统（即 NAT） [4, 5, 14], , 也有其他研究采用非自回归（NAR）文本转语音，例如 Fastspeech2 [15]。这些模型都是文本转梅尔频谱模型，会将文本转换为梅尔频谱图，因此需要额外的声码器来获取语音波形。声码器的选择也各不相同，包括非参数的 Griffin-Lim 和神经声码器。

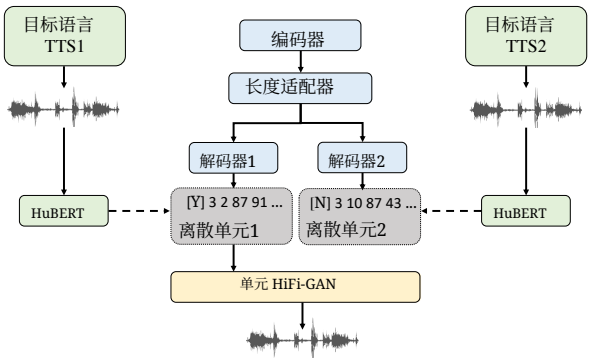

图 1。我们提出的采用多个文本转语音目标的语音到语音翻译模型框架。蓝色模块表示语音到语音翻译建模；绿色模块表示用于生成目标离散单元的模块；灰色模块表示语音到语音翻译模型的目标，其中第一个 token 用于预测推理时使用哪种文本转语音目标；黄色模块表示模型推理。详情见第 2.2 节。

在这项工作中，我们选择了三种文本转语音模型：Tacotron2（TT2）[27]，、Fastspeech2（FS2）[17]，和 VITS [28]。Tacotron2 是一种经典的自回归文本转梅尔频谱模型，而 Fastspeech2 是一种典型的非自回归文本转梅尔频谱模型。与其他模型（文本转梅尔频谱模型 + 声码器）不同，VITS 直接对从文本到波形（文本到波形）的过程进行建模，无需额外的声码器。对于文本转梅尔频谱模型（即 TT2 和 FS2），我们采用三种不同的声码器进行研究：Parallel WaveGAN（PWG）[29]，、Hifi-GAN（HFG）[30]，和 StyleMel-GAN（SMG）[31]。我们还通过在推理时调整速度因子，研究了时长控制对非自回归模型的影响。

在第一组实验中，我们评估了采用不同目标文本转语音语音的语音到语音翻译模型。随后，我们进一步开展实验，将不同文本转语音系统合成的语音进行组合。本研究仅关注单说话人文本转语音系统。

总体工作流程：基于前文讨论的相关工作，构建语音到语音翻译系统的总体工作流程如下：(1) 目标语音合成：使用文本转语音模型合成目标语音，该模型可以是声学模型与声码器，也可以是直接的文本到波形模型。(2) 离散单元提取：使用 HuBERT 模型通过聚类将合成的目标语音转换为离散单元。(3) 语音到语音翻译系统训练：以源语音为输入、目标离散单元为输出训练语音到语音翻译模型。(4) 推理：推理时，语音到语音翻译模型将源语音转换为离散单元序列。然后，使用基于单元的声码器生成最终的波形语音。

2.2. 所提出的框架

如 [21, 24]，所示，从语音 SSL 表征中提取的语音离散单元有望将语言、韵律和说话人相关信息解耦。然而，这种方法同时仍然存在噪声。例如，[12]，作者展示了，不同说话人说出的同一句话可能会产生不同的语音离散单元序列。当

相同的句子由不同的文本转语音系统生成。为验证我们的假设，我们测量了在 LJSpeech [32]。[1] 上训练的不同文本转语音系统之间 HuBERT 单元分布的皮尔逊相关系数。如图2 所示，这清楚地表明，尽管具有相同的语言学来源并在相同语料库上训练，不同的文本转语音系统仍然存在差异。数据包括

[1]HuBERT 和文本转语音系统的配置将在第 3.2 节中详细讨论。

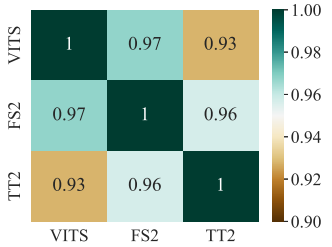

图 2。不同文本转语音系统之间的皮尔逊相关系数：单元分布是使用 Fisher 西班牙语-英语数据集 [18]的开发集收集的。对于 FS2 和 TT2 的波形生成，我们均使用 HiFi-GAN 声码器。

Fisher 西班牙语-英语语料库的开发集 [18]。另一方面，给定目标语言中的相同话语，合成语音应具有相同的语言内容。因此，假设在使用相同文本从不同的文本转语音系统生成语音时，所提取的单元之间可能具有类似的共识，这是合理的。

基于上述假设，我们提出了图 1 所示的框架。该框架基于 [7] 提出的模型，同时还专门设计用于捕获不同文本转语音系统之间语言信息的高层共识。具体而言，我们为不同文本转语音系统生成的语音离散单元添加了独立的解码器分支。为简洁起见，图 1 展示了包含两个目标的情况；但由于其并行特性，该框架很容易扩展到三个或更多目标。

我们并非直接并行预测单元，而是在**目标** 单元序列开头附加一个特殊标记，如图 1 中的灰色方块所示。该特殊标记用于指示合成语音的质量，可在模型推理时用来选择更好的输出。具体而言，我们首先计算不同文本转语音系统中每个语句的句子级字符错误率（CER）。然后，在训练阶段，我们将标记 [Y] 分配给候选目标语音离散单元中 CER 最低的文本转语音系统，并将标记 [N] 分配给其他系统。[2] 在推理时，我们计算所有解码器预测出第一个特殊标记 [Y] 的概率，并选择概率最高的解码器继续生成序列。需要注意的是，由于特殊标记位于序列开头，如果其概率低于其他分支，推理过程便无需自回归地生成后续标记。因此，与不含多个文本转语音目标的基线系统相比，推理时的额外搜索负担并不大。

由于离散单元中存在噪声，[12]探索了类似的方法。作者提出了一种说话人归一化方法，将不同说话人的单元归一化到参考说话人。然而，在使用不同文本转语音系统的情况下，由于这些系统合成的文本相同，因此很难确定应将哪个系统作为参考。

3. 实验

3.1. 数据集

对于语音到语音翻译（S2ST），我们使用 Fisher 西班牙语-英语数据集 [18]，，该数据集也广泛用于以往的语音到语音翻译研究 [7,9]。我们将英语文本转语音系统应用于根据英语文本合成目标语音，以用于训练和验证。对于文本转语音系统的训练，我们使用 LJSpeech，这是一个时长为 24 小时的单说话人语料库 [32]。

[2]对于并列的情况，我们将 [Y] 赋予所有字符错误率（CER）最佳的系统。

3.2. 实验设置

3.2.1. 模型架构

语音到单元翻译模型（S2UT）：我们遵循 [11], 中所述的 S2 UT 模型更新版本，该版本是对 [7] 的扩展。在语音离散单元生成方面，我们采用 [11, 12] 发布的预训练多语言 HuBERT、K-Means 聚类和基于单元的 HFG 声码器。离散单元的词汇表大小为 1,000，对应于 K-Means 模型中的聚类数量。在 S2UT 模型训练期间，通过重复池化对生成的离散单元进行缩减。另一方面，缩减后的单元通过与基于单元的 HFG 联合训练的时长预测器恢复为其原始长度。与 [11], 相同，编码器初始化自基于 Conformer 的 wav2vec 2.0，而解码器则初始化自 [33] 中发布的 mBART 解码器 [11]。为保持不同设置之间的一致性，我们不调整超参数，而是对不同系统采用 [11] 中的设置。

TTS模型：如第2.1 节所述，我们采用了三种不同的文本转语音系统（即TT2、FS2 和 VITS）以及三种不同的声码器（即 PWG、HFG 和 SWG）。为保证实验的可复现性，所有模型均来自 ESPnet-TTS [34, 35], 这一用于文本转语音的开源框架中公开提供的检查点。

3.2.2. 训练与解码

对于 S2UT 模型的训练，我们使用 AdamW 优化器 [36]，学习率为 0.0005。调度器采用预热策略，从 1e-7 的学习率开始，并在 20k 步时达到最大学习率。我们每 120 步累积一次梯度，以模拟较大的批量大小；事实证明，这对语音到语音翻译学习有效。在训练过程中，我们遵循 [11], 中提出的“LNA-D”策略，该策略不会对预训练的 mBART 解码器的所有参数进行微调，而只微调 Layer-Norm 和自注意力参数。对于 S2UT 的解码，我们采用波束搜索，波束大小为 10。

对于非自回归 TTS 模型，可以通过速度因子调节时长，从而控制生成时长。在我们的实验中，除了第 2.1 节讨论的 TTS 模型外，我们还采用了不同的速度因子，包括 0.95、1.0 和 1.05。

3.2.3. 实验设计

实验通常包括三个方面：
单一 TTS 系统：为了比较不同 TTS 系统的效果，我们直接使用由单一 TTS 系统生成的离散单元训练 S2UT 模型。

TTS 系统的简单组合：如第 2.2 节所述，多个 TTS 系统有望通过对单元序列中的噪声进行归一化来改善语音到语音翻译。为了系统地研究不同系统的影响，我们根据不同 TTS 系统的建模特性开展实验，包括自回归与非自回归、不同的声码器、不同的速度因子，以及文本到梅尔频谱与文本到波形。

用于多个文本转语音目标的多任务框架：我们遵循文本转语音系统简单组合中的实验设计，但如第 2.2 节所述，将数据组合改为多任务训练方式。我们使用基于wav2vec2.0 的自动语音识别模型来计算第 2.2 节中提到的字符错误率。由于训练具有三个以上额外分支的模型需要大量 GPU显存，我们将文本转语音系统的数量限制为少于四个。

表 1。不同文本转语音系统上的语音到语音翻译性能。“数据标识”列表示由文本转语音系统生成的目标语音单元。字符错误率和双语评估替补的计算方法如第 3.2.4 节所述。声学模型（声学模型）和声码器见第 2.1 节。

数据标识	AM	声码器	字符错误率(↓)	双语评估替补（↑）
A	TT2	PWG	9.1	37.3
B		HFG	8.9	37.3
C		SWG	8.7	37.3
Avg.	TT2	/	8.9	37.3
D	FS2	PWG	9.4	37.5
E		HFG	**8.3**	37.6
F		SWG	8.7	37.5
Avg.	FS2	/	8.8	37.5
G	VITS		8.4	**38.3**

3.2.4. 评估指标

首先，通过将合成语音输入自动语音识别模型，并计算自动语音识别预测结果与参考文本之间的字符错误率（CER），来评估文本转语音的质量。对于自动语音识别模型，我们采用基于 wav2vec 2.0[3] 训练的开源自动语音识别模型。为评估翻译质量，我们首先使用相同的自动语音识别模型获取语音到语音翻译系统的转写结果（即预测单元 + code-HFG 声码器），然后使用 SacreBLEU [37] 计算其与参考文本之间的 BLEU 分数。需要注意的是，参考文本也经过了分词处理，并转换为不含标点符号的小写形式，以便计算 BLEU 分数。

3.3. 结果与讨论

我们使用由单一 TTS 系统生成的目标语音离散单元开展实验。最佳系统采用 VITS 合成语音（数据**G**），结果如表 1 所示。我们发现，使用不同声码器对模型 **A**-**C** 和模型 **D**-**F**进行处理，对语音到语音翻译性能没有显著影响。然而，声学模型的差异可能会影响语音到语音翻译结果：最佳系统使用 VITS 合成的目标语音，最差系统则使用 TT2 合成的目标语音。我们还发现，各 TTS 声学模型的字符错误率大致与其语音到语音翻译性能相关。

表 1 中一个有趣的发现是，尽管不同声码器的系统具有不同的自动语音识别模型字符错误率，但它们取得了相近的性能。我们推测，这是因为 HuBERT 单元有助于在语音合成中归一化声码器差异。为验证这一假设，我们对使用不同声码器的 FS2 上的 HuBERT 单元分布进行了皮尔逊系数分析，发现皮尔逊得分均高于 0.98。

此外，我们发现，对于口语词语（例如“HMM”和“HUM”），VITS 的字符错误率通常高于 TT2 和 FS2。这可能是由于用于 TTS 训练（即 LJSpeech）和 ASR 训练（即 Librispeech）的朗读语音，与用于语音到语音翻译训练（即 Fisher）的会话语音之间存在数据领域不匹配。

表2展示了不同速度因子对非自回归文本转语音系统的影响。结果表明，较小的速度因子可以改- 善语音到语音翻译系统的性能。然而，当测-

[3]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

表 2。不同速度因子下的语音到语音翻译性能。默认情况下，Fast-speech2 (FS) 模型与 HFG 声码器结合用于文本转语音。TTS 模型介绍见第 2.1 节。

数据标识	TTS	速度因子	字符错误率(↓)	双语评估替补(↑)
H	FS2	0.95	8.7	38.0
E		1.0	8.3	37.5
I		1.05	9.5	37.4
J	VITS	0.95	8.6	**38.7**
G		1.0	8.4	38.3
K		1.05	8.5	38.3

表3。多个文本转语音目标上的语音到语音翻译性能。我们按照第3.2.3节列出的类别进行实验。“多任务”列中带有 ✓ 的模型采用第2.2节提出的框架进行训练。

类别	Data	多任务	双语评估替补(↑)
最佳单一文本转语音系统	**J**	/	38.7
TT2 + FS2	**B** + **E**	✗	37.0
		✓	37.6
TT2 + 声码器	**甲** + **乙** + **丙**	✗	37.3
		✓	37.3
FS2 + 声码器	**D** + **E** + **F**	✗	37.7
		✓	37.9
FS2 + 速度因子	**H** + **E** + **I**	✗	38.8
		✓	39.7
VITS + 速度因子	**J** + **G** + **K**	✗	39.7
		✓	**41.5**
VITS + TT2	**B** + **G**	✗	37.2
		✓	38.4
VITS + FS2	**E** + **G**	✗	39.6
		✓	40.5
VITS + TT2 + FS2	**B** + **E** + **G**	✗	38.5
		✓	40.1

测量合成语音的字符错误率时，使用默认速度因子 1.0 更有利。结合表 1 和表 2 的结果，我们发现，速度因子为 0.95 的 VITS 能够为构建基于单元的语音到语音翻译系统提供质量最佳的合成数据。因此，我们在以下组合实验中将这一数值作为参考。

表3展示了先前实验中研究的不同文本转语音系统**A**-**K**组合后的结果。采用“✗”的实验是指直接合并不同文本转语音系统的数据进行训练的方法，而采用“✓”的实验则基于我们在第 2.2 节提出的方法。比较表 1 和表3中的模型可以发现，直接合并不同文本转语音系统的数据通常能提升语音到语音翻译的效果。例如，合并表3中的**E** 和**G** 的数据后，双语评估替补得分为 39.6；而在表 1 中，仅使用**B** 或**G**进行训练时，得分分别为 37.6 和 38.3。值得注意的是，有些情况下，简单合并数据并不能提升语音到语音翻译的效果（例如**A** + **B** + **C** 与**A**、**B**和**C**相比）。同时，与不采用多目标训练的模型相比

表4。来自所提出框架不同分支的语音到语音翻译性能。BLEU差值是与对应的单一 TTS 系统上训练的系统之间的绝对差值。详情见第3.3节。

类别	分支	双语评估替补(↑)	BLEU差值
VITS + TT2 + FS2	**B**	37.9	+0.6
	E	38.8	+1.2
	G	38.9	+0.6
	B + **E** + **G**	40.1	/
VITS + FS2	**E**	38.8	+1.2
	G	39.0	+0.7
	E + **G**	40.5	/
VITS + 速度因子	**J**	39.5	+0.8
	G	39.1	+0.8
	K	39.0	+0.7
	J + **G** + **K**	41.5	/

采用第 2.2 节提出的框架可以获得更好的结果。

如第 2.2 节所述，在推理过程中给定一个话语，所提出的框架首先预测每个解码器分支的第一个标记（例如，[Y] 或 [N]）。随后，它选择标记 [Y] 概率最高的解码器分支，作为该话语所使用的解码器分支。在表4中，我们针对基于特殊标记的推理进行了消融实验。具体而言，我们报告了所提出的框架中每个解码器分支的语音到语音翻译性能，并将结果与仅基于相应文本转语音目标训练的语音到语音翻译系统进行比较。我们选择性能最高的三个语音到语音翻译系统进行比较。结果表明，与仅使用单个文本转语音系统训练的系统相比（例如，表 1 中使用**E** 或**G** 训练的模型，以及表 2 中使用**J**、**G**、**K** 训练的模型），所有分支的语音到语音翻译性能均有所提升。同时，这也证明了所提出推理过程的有效性，因为使用我们所提出方法组合的分支仍然优于单分支性能。

4. 结论

本研究首先探讨了使用不同文本转语音系统生成的不同目标语音所产生的影响。实验表明，简单地结合文本转语音系统生成的目标语音，就有助于语音到语音翻译的学习，尤其是语音到单元模型（S2UT）。基于这些发现，我们提出了一种将多个文本转语音目标整合到语音到语音翻译建模中的新框架。实验结果表明，我们所提出的框架能够持续提升最佳基线语音到语音翻译系统的性能，双语评估替补分数提高了 2.8 分。

5. 致谢

本研究得到 Meta AI SRA 资助。Jiatong Shi 和 Shinji Watanabe 部分经费来自 Bridges 系统 [38]，该系统由美国国家科学基金会（NSF）资助，奖项编号为 ACI-1445606，托管于匹兹堡超级计算中心（PSC）。

6. 参考文献

[1] Enrique Vidal，《有限状态语音到语音翻译》，载于*ICASSP*，1997年。[2] Shigeki Matsuda、XinhuiHu、Yoshinori Shiga 等，《多语言语音到语音翻译系统：Voicetra》，载于*ICMDM*，2013年。[3] Quoc Truong Do、TomokiToda、Graham Neubig等，《在语音到语音翻译中保留词级重音》，*TASLP*，2016年。[4] YeJia、Ron J Weiss、FadiBiadsy 等，《基于序列到序列模型的直接语音到语音翻译》，*Interspeech*，2019年。[5] Ye Jia、Michelle Tadmor Ramanovich、Tal Remez 等，《Translatotron 2：具有语音保留功能的高质量直接语音到语音翻译》，载于*ICML*，2022年。[6] Ye Jia、YifanDing、Ankur Bapna 等，《利用无监督和弱监督数据改进直接语音到语音翻译》，载于 *Interspeech*，2022年，第 1721–1725 页。[7] Ann Lee、Peng-Jen Chen、Changhan Wang 等，《使用离散单元进行直接语音到语音翻译》，载于*ACL*，2022年。[8]Takatomo Kano、Sakriani Sakti 和 Satoshi Nakamura，《基于 Transformer 和转码器的直接语音到语音翻译》，载于*SLT*，2021年。[9] ChenZhang、XuTan、Yi Ren 等，《Uwspeech：面向无文字语言的语音到语音翻译》，载于*AAAI*，2021年。[10] Xutai Ma、HongyuGong、Danni Liu 等，《直接同步语音到语音翻译》，*arXiv* 预印本*arXiv:2110.08250*，2021年。[11] Sravya Popuri、Peng-Jen Chen、ChanghanWang等，《使用自监督预训练和数据增强的增强型直接语音到语音翻译》，载于 *Interspeech*，2022年，第 5195–5199 页。[12] Ann Lee、HongyuGong、Paul-AmbroiseDuquenne 等，《基于真实数据的无文本语音到语音翻译》，载于 *NAACL*，2022年，第 860–872 页。[13] Genichiro Kikui、EiichiroSumita、Toshiyuki Takezawa 等，《创建语音到语音翻译语料库》，载于 *Eurospeech*，2003年。[14] Ye Jia、Michelle TadmorRamanovich 等，《CVSS语料库与超多语言语音到语音翻译》，载于*LREC*，2022年，第 6691–6703 页。[15] PedroJeuris和Jan Niehues，《Libris2s：德英语音到语音翻译语料库》，载于*LREC*，2022年，第 928–935 页。[16] Jonathan Shen、Ye Jia、MikeChrzanowski等，《非注意力 Tacotron：包含无监督时长建模的稳健且可控的神经文本转语音合成》，*arXiv* 预印本*arXiv:2010.04301*，2020年。[17] Yi Ren、Chenxu Hu、Xu Tan 等，《Fastspeech 2：快速且高质量的端到端文本转语音》，载于*ICLR*，2020年。[18] Matt Post、Gaurav Kumar、AdamLopez等，《Fisher 和 CALLHOME 西班牙语–英语语音翻译》，*LDC2014T23*。网络下载。费城：语言数据联盟。

[19] Shu-wen Yang、Po-Han Chi、Yung-Sung Chuang 等人，“SUPERB：语音处理通用性能基准”，载于*Proc. Interspeech 2021*，2021年，第1194–1198页。[20] Hsiang-Sheng Tsai、Heng-Jui Chang、Wen-Chin Huang 等人，“SUPERB-SG：面向语义与生成能力的增强型语音处理通用性能基准”，载于*ACL*，2022年，第8479–8492页。[21] Adam Polyak、Yossi Adi、Jade Copet 等人，“基于离散解耦自监督表征的语音重合成”，载于*Interspeech*，2021年。[22] Wen-Chin Huang、Shu-Wen Yang、Tomoki Hayashi 等人，“S3PRL-VC：基于自监督语音表征的开源语音转换框架”，载于*ICASSP*，2022年。[23] Tomoki Hayashi 和 Shinji Watanabe，“Discretalk：将文本转语音视为机器翻译问题”，*arXiv preprint arXiv:2005.05525*，2020年。[24] Kushal Lakhotia、Eugene Kharitonov、Wei-Ning Hsu 等人，“基于原始音频的生成式口语语言建模”，*TACL*，2021年。[25] PD Aguero、Jordi Adell 和 Antonio Bonafonte，“语音到语音翻译的韵律生成”，载于*ICASSP*，2006年。[26] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai 等人，“HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习”，*TASLP*，2021年。[27] Jonathan Shen、Ruoming Pang、Ron J Weiss 等人，“通过以梅尔频谱图预测为条件的 Wavenet 实现自然文本转语音合成”，载于*ICASSP*，2018年。[28] J aehyeon Kim、Jungil Kong 和 Juhee Son，“用于端到端文本转语音的对抗学习条件变分自编码器”，载于*ICML*，2021年。[29] Ryuichi Yamamoto、Eunwoo Song 和 Jae-Min Kim，“Parallel WaveGAN：一种基于多分辨率频谱图生成对抗网络的快速波形生成模型”，载于*ICASSP*，2020年。[30] 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”，*NeurIPS*，2020年。[31] Ahmed Mustafa、Nicola Pia 和 Guillaume Fuchs，“Stylemelgan：一种采用时间自适应归一化的高效高保真对抗声码器”，载于*ICASSP*，2021年。[32] Keith Ito 和 Linda Johnson，“LJ语音数据集”，2017年。[33] Yinhan Liu、Jiatao Gu、Naman Goyal 等人，“面向神经机器翻译的多语言去噪预训练”，*TACL*，2020年。[34] Tomoki Hayashi、Ryuichi Yamamoto、Katsuki Inoue 等人，“ESPnet-TTS：统一、可复现且可集成的开源端到端文本转语音工具包”，载于*ICASSP*，2020年。[35] Tomoki Hayashi、Ryuichi Yamamoto、Takenori Yoshimura 等人，“ESPnet2-TTS：拓展文本转语音研究的边界”，*arXivpreprintarXiv:2110.07840*，2021年。[36] Ilya Loshchilov 和 Frank Hutter，“解耦权重衰减正则化”，载于*ICLR*，2018年。[37] Matt Post，“呼吁清晰报告 BLEU 分数”，载于机器翻译会议，2018年。[38] Nicholas A Nystrom、Michael J Levine、Ralph Z Roskies 等人，“Bridges：面向新兴社群和数据分析的灵活高性能计算资源”，载于*XSEDE*，2015年。

，2014年。