面向印度语言语音合成的快速、小型混合 HMM-HiFi-GAN 系统

Sudhanshu S rivastava[1], *Ishika Gupta*[1],*Anusha Prakash*[2],*Jom Kuriakose*[1],*Hema A. Murthy* [1]

[1]计算机科学与工程系，印度理工学院马德拉斯分校[2]电气工程系，印度理工学院马德拉斯分校

srivastava.rishabh4321@gmail.com, ishika@cse.iitm.ac.in, anushaprakash90@gmail.com
jom@cse.iitm.ac.in, hema@cse.iitm.ac.in

arXiv:2302.06227v1 [eess.AS] 13 Feb 2023

摘要

基于隐马尔可夫模型（HMM）的文本转语音（HTS）在说话风格方面具有灵活性，训练和语音合成速度快，且计算开销较低。即使在低资源场景下，HTS 也能表现出良好的效果。其主要缺点是，与端到端系统相比，语音质量较差。本文提出一种混合方法，将基于 HMM 的特征生成与基于神经网络的 HiFi-GAN 声码器相结合，以提高 HTS 的语音合成质量。HTS 使用高分辨率梅尔频谱图进行训练，而非传统的梅尔广义系数（MGC）；再将与输入文本对应的输出梅尔频谱图输入一个基于印度语言训练的 HiFi-GAN 声码器，从而生成自然度媲美端到端系统的语音，这一点也得到了差异平均意见分和成对比较测试结果的证实。

索引词— 语音合成（会话式）、HiFi-GAN、基于 HMM 的语音合成、端到端系统、混合文本转语音

1. 引言

最先进的端到端系统能够生成高质量语音，但需要大量计算资源和数据，模型占用空间也很大。单元选择语音合成系统速度快，生成的语音质量高，但拼接处可能不够平滑，且占用空间较大。本文旨在提升基于隐马尔可夫模型（HMM）的语音合成系统（HTS）的质量 [1, 2]。基于 HTS 的语音合成系统具有占用空间小、在低资源场景下鲁棒性强的优势。HTS 需要使用 STRAIGHT 声码器 [3]（专有）才能生成高质量语音。本文尝试结合 HTS 技术与 HiFi-GAN 声码器，以显著降低计算和内存开销，同时生成高质量语音。序列到序列（S2S）自回归模型，例如 Tacotron2 [4] 和 Transformer TTS [5] 模型，与 WaveNet [6]和 WaveGlow [7], 等声码器结合后，能够生成高质量语音。尽管质量有了显著提升，但质量与计算复杂度之间仍需权衡。高质量语音仅限于朗读语音，其中的句子语法正确。由于端到端系统使用整个句子的上下文，因此无法扩展到对话类文本，而且常会出现漏词及其他瑕疵 [8]。尽管非自回归的 FastSpeech[9] 很少出现漏词，但在低资源场景下表现仍然不佳。基于 HTS 的合成系统中的子词模型对应于五音素。得益于基于树的聚类，未登录词也可以合成。尽管如此，基于 HTS 的系统合成的语音质量较差。我们此前的工作 [10] 表明，合成语音的质量主要取决于梅尔频谱图的保真度。混合文本转语音方法结合了两个或更多文本转语音框架。当前工作受到 [10], 的启发，该研究将 HTS 合成的语音转换为 WaveGlow 声码器所需的梅尔频谱图。在[10], 中，从 HTS 生成的音频（用于训练文本）和真实语音（GT）音频中提取梅尔频谱图。针对每个梅尔滤波器系数，分别估计 HTS 源音频以及真实原始音频的一维直方图，并对源直方图和目标直方图进行直方图均衡化。测试时，首先使用 HTS 合成语音，并将其转换为一系列梅尔滤波器系数。随后使用查找表，将 HTS 生成的梅尔滤波器系数替换为真实语音的梅尔滤波器系数。然后使用 WaveGlow 声码器合成语音。研究还尝试直接使用 HTS 生成的 MGC 训练 WaveGlow 声码器，但输出音色不佳。HTS 生成的 MGC 缺乏所需的频谱分辨率。在当前工作中，我们首先使用高分辨率梅尔滤波器训练 HTS 系统。我们用 HiFi-GAN 声码器替换原有声码器，因为它在参数选择方面具有更大的灵活性[1]。

近期的混合系统包括 [12, 13, 14]。在 [12] 中，采用 USS 和端到端合成语音中最适合的多音节片段。通过所提出的混合系统，选择合适的片段

[1]HiFi-GAN 是一种基于生成对抗网络（GAN）的声码器，语音合成质量优于 WaveNet 等自回归模型。由于采用非自回归架构，其推理速度相对较快，且不需要 GPU。 [11]

时使用 USS 单元或端到端单元。 [14] 和 [15] 将传统方法与 Tacotron2 等基于神经网络的架构相结合。 [15]将 Tacotron 编码器输出的向量作为音素表示，并使用它们之间的欧几里得距离进行单元选择； [14] 则根据选择成本，在端到端语境中合成额外的音素级单元。 [13] 使用隐半马尔可夫模型（HSMM）对对齐进行潜在表示，以减少暴露偏差，即训练（教师强制）与推理（自由运行）之间的不匹配。[16] 是近期一项关于对话语音合成的研究，通过使用对话语音增强训练数据，改善了对话语音的韵律。

在本研究中，我们提出了一种结合 HTS 和 HiFi-GAN 的方法。基线系统包括传统的 HEQ，以及搭配 HiFi-GAN 声码器的 Fastspeech2。HEQ 模型与我们先前的工作 [10], 类似，区别在于将 WaveGlow 声码器替换为 HiFi-GAN 声码器。在所提出的方法中，HTS 系统基于梅尔频谱图进行训练。该模型支持无GPU推理，总占用空间为 ∼ 72MB（包括声码器），因此易于部署。研究发现，即使仅使用 1 小时的标注数据进行训练，该模型也能生成高质量语音。本文其余部分安排如下：第 2 节回顾相关研究；第 3 节介绍所提出的方法；第 4 节讨论实验及结果；第 5 节总结全文。

2. 相关研究

2.1. HTS 系统

HTS 基于语音的声源-滤波器模型 [17],，其中频谱参数和激励参数经过卷积后生成语音。梅尔广义倒谱（ MGC）和对数基频（lf0、lδ f0、l$\delta\delta$ f0；f0 为基频）分别是从音频文件中提取的频谱参数和激励参数。HTS流程由训练阶段和测试阶段组成。

2.1.1. 文本转语音合成系统的训练阶段基于 HMM 的语音合成使用少量标注数据来引导音素模型的初始化，或者采用平坦初始化，即假设给定话语中的所有音素（如文本所示）时长相同。通过迭代嵌入式重估来校正边界并重新估计音素模型。HTS 使用带标注/对齐的音频和文本数据进行训练。训练隐马尔可夫模型，以对上下文相关的五音素（前置 + 2后置 +当前音素）单元建模。采用基于决策树的聚类对广泛的上下文标签进行建模，从而使系统能够稳健地处理未见过的序列。

2.1.2. 使用生成的模型合成语音测试文本通过统一解析器 [18]解析为标签，然后将与这些标签对应的隐马尔可夫模型拼接起来——接。HTS 生成频谱和激励参数。可以使用梅尔对数谱近似（MLSA） [19],STRAIGHT [3]或其他声码器。

2.2. HiFi-GAN

HiFi-GAN [20] 是一种非自回归的先进神经声码器。它包含一个生成器和两个判别器，采用对抗方式进行训练。除了 MelGAN [21] 提出的用于建模连续依赖关系和长期依赖关系的多尺度判别器外，它还采用了由多个子判别器组成的多周期判别器，以处理输入音频中部分周期信号。此外，GAN目标函数中还加入了新的梅尔频谱图损失（原始波形与生成波形的梅尔频谱图之间的 L1 距离）。

3. 提出的方法

在所提出的工作中，我们在两个方面有别于现有方法：a) 虽然初始分段采用了平坦初始化，但边界会使用一种语言无关信号处理方法进行修正，该方法主要利用音节的声学特性 [22, 23]。UTF-8 文本会使用 [18]转换为标签序列（包括音素和音节标签）。b) 传统的 HTS 训练使用梅尔广义倒谱系数，而我们提出的方法使用高分辨率梅尔滤波器组系数。基于群时延（GD）的分段利用了这样一种特性：每个音节的音节核都是一个高能量区域，对应一个元音；其音节起始和音节韵尾则由一系列辅音构成（C*VC*）。隐马尔可夫模型无法很好地建模边界，但在与文本强制对齐时，可以给出正确数量的边界。由于 GD 分段主要基于信号处理，因此可能产生额外的伪边界。我们利用 GD 处理得到的精确边界来修正隐马尔可夫模型边界。基于 GD 的处理无法为擦音和塞擦音提供精确边界，因此使用子带谱通量（SBSF）修正擦音和塞擦音的边界[23]。这样便可将语音信号分割为一系列音节。嵌入式再估计在音节级执行。由于音节的平均时长约为 130 毫秒，因此可以得到精确的音素边界。相比之下，在句子级执行嵌入式再估计时，时长最长可达 15 秒，导致音素边界较差，如图1所示。图1展示了采用此方法获得的一段语音及其边界 [24]。在图1中，音节 *khoj* 和 *kar* 之间的边界使用 SBSF 的 GD 进行修正，而音节*kar* 和*taa* 之间的边界则使用 GD 进行修正。如果第一个音节的末尾音素不是擦音或鼻音，且第二个音节的起始音素不是擦音、塞擦音、鼻音或半元音，则使用短时能量（STE）修正两个音节之间的边界。如果第一个音节的末尾或第二个音节的起始音素（但不能两者同时）是擦音或塞擦音，则使用 SBSF 修正两个音节之间的边界。声码器

HTS 中使用的声码器包括 MLSA [19] 和 STRAIGHT [3], , 它们使用 MGC 和生成的基频来合成语音。在当前工作中，我们用端到端合成器所使用的声码器替换这些声码器。在端到端框架中，声码器包括 WaveNet [6], 、WaveGlow [7], 、HiFi-GAN 和 CARGAN [25]。值得注意的是，所有这些声码器的频谱分辨率都远高于传统 HTS 系统所用的分辨率（传统 HTS 系统通常使用 34 个梅尔滤波器）。WaveNet、WaveGlow 和 HiFi-GAN 默认使用 80 个梅尔滤波器。因此，我们使用 80 个和 120 个梅尔滤波器训练 HTS 系统。图2展示了训练流程图

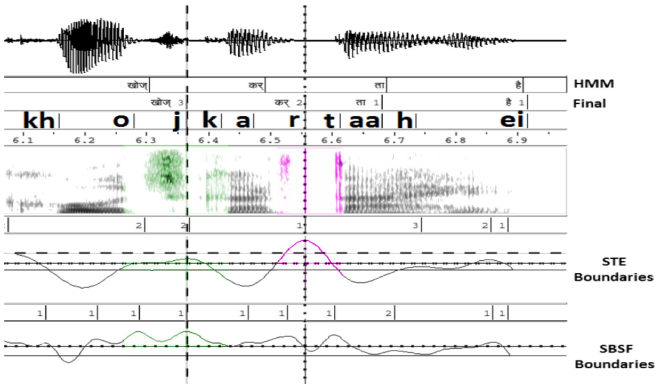


图 1。混合分段。粉色圆点表示短时能量，绿色圆点表示由 SBSF 校正的边界。经作者许可，本图转载自 [24] 。

以及所提出系统的语音合成阶段。改进后的 HTS（基于梅尔频谱图的 HTS）和传统的 HiFi-GAN 分别称为 HTS (M) 和 HiFi-GAN（C）。在语音合成过程中，测试句子输入 HTS (M) 系统，由该系统生成相应的梅尔频谱图。随后将梅尔频谱图输入 HiFi-GAN（C）声码器，以生成最终的语音输出。在此方法中，我们还使用 80 和 120 个梅尔滤波器进行了实验。

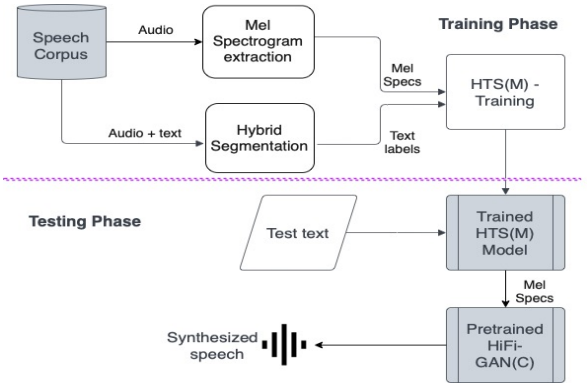


图 2。所提系统训练阶段和测试阶段的流程图。

4. 实验与结果

本节将介绍所使用的数据集、开展的各项实验及其结果。

4.1. 所用数据集

所提出的系统在印地语和卡纳达语这两种语言上进行了测试。HTS 和 HiFi-GAN 模块分别使用 8.5 小时的男性印地语和男性卡纳达语数据集进行训练。数据集来自 IndicTTS 数据库 [26]。IndicTTS 是一个开源文本转语音数据库，录音在录音棚环境中以 48 kHz 采集。测试时，我们使用来自“面向青年有志者的学习网络”（SWAYAM）平台 [27] 的译文文本。原始 SWAYAM 讲座为英语。先进的自动语音识别（ASR）系统 [28]生成转录文本，再通过机器翻译（MT）系统将其译为印地语和卡纳达语 [29]。生成的文本具有对话性质。此外，我们还合成了通用领域的句子，以测试所提出系统的可扩展性。

4.2. 实验

将传统的 HTS+HEQ 系统（如 [10] 中所述）作为基线系统。我们还训练了一个采用 HiFi-GAN 声码器的 FastSpeech 系统，以比较语音质量。FastSpeech 模型使用 ESPnet工具包 [30] 进行训练。对齐取自 tacotron2[4]教师模型。我们还评估了所提系统在低资源场景下的质量，即仅使用 1 小时的数据进行训练。在这种低资源场景下，虽然合成质量会有所下降，但仍可训练 HTS。然而，由于数据不足，FastSpeech 模型未能得到良好训练。通常，梅尔频谱图使用 80 个梅尔滤波器生成。由于我们在 HiFi-GAN 声码器中不使用 HTS 生成的音高，因此训练时采用 120 个梅尔滤波器。

4.3. 评估

通过退化平均意见分（DMOS）测试评估各种系统的性能。在DMOS测试中，该语言的母语者听取合成音频，并以1至5分的等级对话语进行评分，其中1分表示较差，5分表示具有人类语音特征。测试中加入了GT话语作为参考。话语以随机顺序呈现给评估者。最终得分根据GT话语的得分进行归一化后报告。在所提出的混合系统与HEQ混合系统之间还进行了成对比较测试（PC测试）。在该测试中，评估者听取由两个系统生成的相同合成话语，并以偏好形式给出评分。表1展示了印地语和卡纳达语系统的DMOS结果。在DMOS测试中，15名（印地语）和10名（卡纳达语）听众对总共20个话语（每个系统8个 + 4 GT）进行了评估。最后一行显示了不包括声码器大小在内的模型占用空间。HEQ

表 1。比较 Fastspeech（FS）和所提出方法生成的合成话语的差异平均意见分评估结果

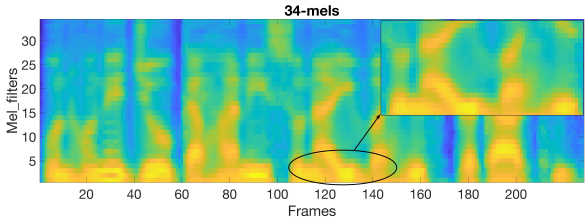
语言	FS	所提出的方法
印地语	4.32	**3.99**
卡纳达语	3.01	**3.22**
占用空间（MB）（不含声码器）	146	**15.9**

表 2。成对比较测试结果：HEQ系统与所提出的混合系统（偏好率，%）

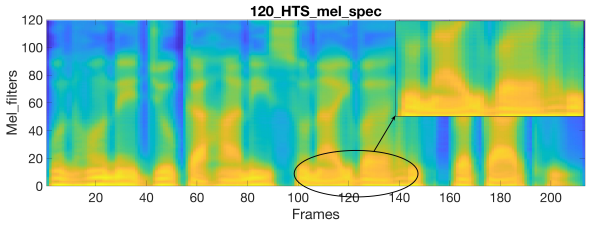
语言	HEQ	所提出的方法	相同
印地语	17.5	**60**	22.5
卡纳达语	20	**63.33**	16.66

模型占用空间为 ∼ 17.5 MB。所提出的系统在卡纳达语上的表现优于FS。达罗毗荼语具有显著的黏着性，这可能导致对齐质量较差，进而影响语音合成性能。表2展示了印地语和卡纳达语系统的PC测试结果。15名（印地语）和10名（卡纳达语）母语者评估了由上述系统合成的10对话语。此处[链接] 包含1小时HTS（梅尔频谱）、8.5小时80维梅尔滤波器组HTS以及传统HTS生成的话语样本，并附有上述模型。

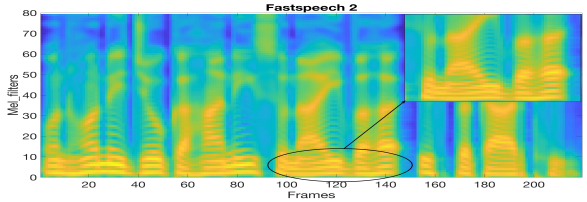
4.4. 结果与讨论



(a) 34 个梅尔滤波器组



(b) 120 个梅尔滤波器组（所提出的方法）



(c) 80 个梅尔滤波器组（传统 Fastspeech）

图 3。不同系统的梅尔频谱图。从表 1 和表 2 可以看出，所提出的系统优于基线 HEQ系统，其性能与 Fastspeech 相当。然而，在低资源场景下，Fastspeech 模型无法完成训练，而所提出的方法能够生成具有合理可懂度的语音。 [链接] 展示了时长为 1 小时的话语示例。此外，在低资源环境中，如果比较单个 CPU 核心上对包含 8 个单词的示例句子进行梅尔频谱图提取所需的时间，可以发现 FS 在用户模式下消耗的 CPU 时间为 3.591 秒，相较于 HTS系统的 0.488 秒明显更高。仅使用 CPU 训练 HiFi-GAN 的速度较慢。但是，其推理过程与 GPU 无关。这 [链接]包含了在 IndicTTS 数据库上从头训练的声码器。图 3 展示了不同系统的梅尔频谱图（34 个梅尔滤波器，即传统 HTS 中使用的滤波器，如图 3(a) 所示）。图 3(b) 中生成的频谱图与 Fastspeech2 生成的频谱图更为接近。

5. 结论

所提出的系统（HTS+HiFi-GAN）结合了基于 HMM 的语音合成与 HiFi-GAN 声码器的优势：前者能够对词汇表之外的词进行建模，后者则能够快速推理并生成高质量语音。所提出系统最重要的优势是占用空间小。该系统中的梅尔频谱图生成器大小为 ∼16 MB，而基于神经网络的 Fastspeech 模型则为 >100 MB 至 GB。HiFi-GAN 声码器占用 ∼ 56 MB，且推理无需 GPU。此外，即使仅使用 1 小时的转录数据训练 HTS 系统，该模型也能生成清晰易懂的结果；而端到端系统则无法完成训练。从占用空间小、推理速度快的角度来看，这种系统颇具吸引力，并且可以移植到智能手机上。

致谢

本研究作为项目“印度语言语音技术”（SP21221960CSMEIT0- 03119）的一部分开展，该项目由电子与信息技术部（MeitY）资助。

6. 参考文献

[1] H Zen、K Tokuda 和 A W Black，“统计参数语音合成”，*SpeechCommunication*，第 51 卷第 3 期，第 1039–1064 页，2009 年 11 月。

[2] Simon King，“统计参数语音合成初学者指南”，语音技术研究中心，爱丁堡大学，英国，2010 年。

[3]Hideki Kawahara 等，《使用基音自适应时频平滑和基于瞬时频率的 f0 提取重构语音表示：声音中重复结构的可能作用》，*Speech Commun.*，第 187–207 页，1999 年 4 月。

[4] J. Shen 等，《通过在梅尔频谱图预测上对 WaveNet 进行条件化实现自然文本转语音合成》，载于*2018 (ICASSP)*，2018 年，第 4779–4783 页。

[5] Naihan Li 等人，《基于 Transformer 网络的神经语音合成》，载于第三十三届 *AAAI 2019* 会议论文集，2019，AAAI’19/IAAI’19/EAAI’19。

[6] A¨aron van den Oord 等人，《WaveNet：一种原始音频生成模型》，载于*SSW*，2016。

[7]Ryan Prenger 等人，《WaveGlow：一种用于语音合成的基于流的生成网络》，载于*(ICASSP)*，2019，第 3617–3621 页。

[8] Bhagyashree Mukherjee 等人，《对话语音分析及其在语音适配中的应用》，2021 年 12 月，第 765–772 页。[9] Yi Ren 等人，*FastSpeech*：快速、稳健且可控的文本转语音，Curran Associates Inc.，美国纽约州雷德胡克，2019。[10]Mano Kumar M、Sudhanshu Srivastava、Anusha Prakash 和 Hema Murthy，《一种基于混合 HMM-WaveGlow 的文本转语音合成器：采用直方图均衡化的低资源印度语言语音合成》，2020 年 8 月。

[11] Ian Goodfellow等人，“生成对抗网络”，载于神经信息处理系统进展。2014 年，第 27 卷，Curran Associates, Inc.

[12] Sudhanshu Srivastava 和 Hema A Murthy，“面向印度语言的 USS 引导式端到端语音合成”，载于 *2022IEEESPCOM*，2022 年，第 1–5 页。

[13] Takato Fujimoto 等人，“基于隐半马尔可夫模型结构化注意力的自回归变分自编码器语音合成”，载于 *ICASSP2022 -2022*，2022 年，第 7462–7466 页。

[14] Xiao Zhou 等人，“基于 UnitNet 的混合语音合成”，2021 年。

[15] X. Zhou、Z. Ling 和 L. Dai，“利用序列到序列声学模型提取单元嵌入以进行单元选择语音合成”，载于*ICASSP2020*，2020 年。

[16] Johannah O’Mahony、Catherine Lai 和 Simon King，“将对话语音与朗读语音相结合，以改善文本转语音合成的韵律”，载于《*Proc. Interspeech 2022*》，2022 年，第 3388–3392 页。

[17] Gunnar Fant，《语音产生声学理论：基于俄语发音的 *X*射线研究进行的计算》，De Gruyter Mouton，2012 年。

[18] Arun Baby 等人，“用于开发印度语言文本转语音合成器的统一解析器”，发表于国际文本、语音与对话会议，2016年。

[19] Satoshi Imai 等人，“用于语音合成的梅尔对数谱近似（MLSA）滤波器”，发表于日本电子与通信（第一部分：通信），1983年。

[20] 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”，载于《*HiFi-GAN*：用于高效、高保真语音合成的生成对抗网络》，收录于《*2020* 年神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，2020 年，第 33 卷，第 17022–17033 页，Curran Associates, Inc.

[21] Kundan Kumar 等人，*MelGAN*：用于条件波形合成的生成对抗网络。

[22] Arun Baby 等人，《结合信号处理线索与深度学习技术，对印度语言文本转语音合成进行音素切分》，载于 *INTERSPEECH*，2017 年，第 3817–3821 页。[23] S.A. Shanmugam 和 H. Murthy，《基于群延迟处理和 HMM 嵌入式再估计的语音切分混合方法》，*INTER-SPEECH*，2014 年。

[24] Arun Baby、Jeena J. Prakash、Aswin Shanmugam Subramanian 和 Hema A. Murthy，《频谱线索对印度语言语音合成器自动语音切分的重要性》，《Speech Communication》，第 123 卷，第 10–25 页，2020 年。

[25] Max Morrison 等人，《用于条件波形合成的分块自回归生成对抗网络》，*arXiv* 预印本 *arXiv:2110.10139*，2021年。

[26] Arun Baby 等人，《印度语言资源》，载于社区协作式语言资源建设（国际文本、语音与对话会议），2016年，第37–43页。

[27] “SWAYAM”，https://swayam.gov.in/。

[28] “IITM语音识别系统”，访问日期：2022年3月29日。

[29] Vandan Mujadia 和 Dipti SharmaMisra，《SWAYAM课程的多语言翻译转录》，即将发表，2022年。

[30] Tomoki Hayashi 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》，2019年10月。