arXiv:2304.00714v1 [eess.AS] 3 Apr 2023

# 集成 表现力语音合成的韵律预测

*Tian Huey Teh*[†⋆] *Vivian Hu*[†] *Devang S Ram Mohan*[†] *Zack Hodari*[†] *Christopher G. R. Wallis*[†]
*Tom´as G´omez Ibarrondo*[†] *Alexandra Torresquintero*[†] *James Leoni*[†] *Mark Gales*[†‡] *Simon King*[†§]

[†] Papercup Technologies Ltd.，英国
[‡] 剑桥大学，英国 [§] 爱丁堡大学，英国 tehtianhuey@gmail.com vivian@papercup.com

**摘要**

使用丰富多样的韵律生成表现力语音，仍然是文本转语音面临的挑战。大多数工作都集中于复杂的神经网络架构，旨在更好地建模数据分布。然而，评估通常发现，对于所有输入文本，没有任何单一模型都能获得优先选择。这表明，可以将一种此前很少用于文本转语音的方法应用于此：模型集成。

我们将集成学习应用于韵律预测。通过改变模型架构或模型参数值，我们构建了简单的韵律预测器集成。

为了在执行文本转语音时自动从集成中选择模型，我们提出了一种新颖且计算开销极低的基于方差的准则。我们证明，即使是规模很小的韵律预测器集成也能产生有用的多样性；结合所提出的选择标准后，其性能优于集成中的任何单个模型。

索引词— 文本转语音、韵律预测、集成方法

## 1. 引言

文本转语音（TTS）系统已经能够生成自然度近乎媲美人类的合成语音。因此，研究重点转向生成多样且富有表现力、超越“平均韵律”的语音 [1]。通常，人们会提出一种新的挑战者模型，通过利用语言和上下文信息 [2], 、使用参考音频 [3], 或采用新的架构和损失函数 [4], 来证明其性能优于现有模型。

然而，听众几乎不会在所有 情况下都更偏好新提出的模型。有些 时候，听众仍然更偏好现有模型。因此，我们有理由以结合多个模型优势的方式使用它们。

这一思路已广为人知，称为集成学习。集成学习算法在处理两个子问题 [5]时采用不同的方法：i)集成生成，即用于构建多个基学习器的方法；ii)集成整合，即用于整合其预测结果的方法，可以通过组合这些结果（例如取平均值），也可以从中进行选择。集成学习已广泛应用于自动语音识别：系统组合会构建多个错误具有互补性的系统，然后使用投票 [6] 或基于似然的重评分 [7]来整合其预测结果。

本文通过构建韵律预测器集成，以提高合成语音的韵律多样性和表现力。理论上，单个生成模型应当能够对数据分布进行建模，从而生成多样化的样本。然而，实践证明这很难实现。由于韵律具有多面性和概率性，不同模型受各自归纳偏置以及模型初始化和训练过程中固有随机性的影响，最终会学到数据分布的不同方面。集成是解决这些问题的一种便捷实用的方法 [8]。

在集成生成方面，我们研究了两种类型：同构集成由架构相同、仅参数值不同的学习器组成；异构集成则由架构各异的学习器组成。为展示这一思路，我们将集成规模限定为最小值，即由两个基学习器组成。在集成整合方面，我们从各学习器的预测结果中进行选择，而非将它们合并（例如取平均值），以避免加剧“平均韵律”效应 [1]。

据我们所知，这是首次将带有选择方法的集成学习用于提升语音合成的表现力。我们的贡献在于证明：

i) 仅由两个韵律预测器组成的最小集成，就能提供足够明显的多样性。[1] ii) 自动准则能够从集成中进行选择，使其表现优于任何单个模型。iii) 我们基于 $F_0$ 方差提出的最佳准则，能够将与人工上限之间的差距缩小 31%。

[⋆] 在 Papercup Technologies Ltd. 任职期间完成的工作。

[1] 样例：https://research.papercup.com/samples/ensemble-prosody-prediction

©2023年 IEEE。允许将本材料用于个人用途，对于任何其他用途，包括为广告或宣传目的重印/再版本材料、创作新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分，均必须获得 IEEE 的许可。

## 2. 相关工作

### 2.1. 语音合成中的生成式集成

关于文本转语音集成学习的先前研究很少。[9]通过对令牌级集成进行知识蒸馏，提高字素到音素（G2P）转换的准确率。他们使用由 Transformer、双向长短期记忆网络和卷积序列到序列模型组成的异构集成，并对预测的概率分布取平均。由于 G2P 的输出是离散的，对预测取平均既恰当又易于实现。

[10] 与我们的工作最为接近。他们使用浅层神经网络集成的加权平均预测来建模 $F_0$ 和时长轮廓。他们通过数据交叉验证构建集成，将训练数据划分为六个不同的子集，并在每个子集上训练八个全连接网络，这些网络在层数、隐藏维度和输入方面各不相同。我们的工作与之不同之处在于：我们关注表现力而非准确率，使用深度神经网络生成集成，并在集成整合中采用选择而非组合。

### 2.2. 过度平滑与平均韵律

神经语音合成中的“平均韵律”问题，与隐马尔可夫模型语音合成中的过度平滑现象类似：统计处理常常导致生成的参数轨迹过于平滑，使语音听起来含混不清。 [11]提出在参数生成过程的似然目标函数中加入全局方差（GV）惩罚项，其中 GV 用于衡量一个话语中频谱参数和 $F_0$ 参数的方差。他们发现，GV 与过度平滑效应呈负相关，而且在主观评估中，该惩罚项提升了自然度。

基于此，我们探讨频谱全局方差和 $F_0$ 方差能否作为从韵律预测器集成中进行选择的标准。

### 2.3. 语调

$F_0$ 或许是建模多样化且富有表现力的语音时最显著的声学特征。[1, 12, 13]使用基于变分自编码器的方法，通过连续或离散潜变量来建模语调（$F_0$ 轨迹）。在这些研究中，必须从波形中提取 $F_0$ 。然而，显式声学特征建模再度兴起，在这类方法中，模型会显式预测 $F_0$ 等声学特征，并通过声学特征预测器 （AFP）实现。在推理过程中，这些声学特征在前向传播时得到预测，然后传递给
用于梅尔频谱生成的解码器。由于此类模型生成的语音，其声学特征与模型内部预测的值相匹配 [15, 17], , 我们提出一种利用这些明确预测的$F_0$ 值来衡量 $F_0$ 方差的方法。

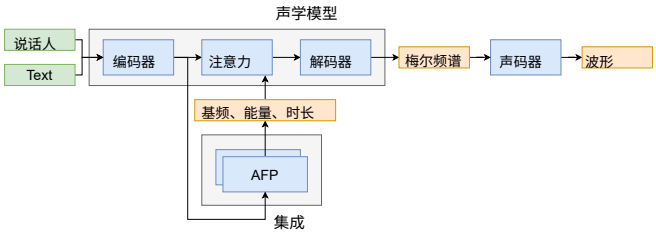



**图 1**。我们研究声学特征预测器（AFPs）的集成，同时保持声学模型和声码器不变。

## 3. 集成设置

### 3.1. 模型集成

我们在声学特征预测器集成上展示了所提出的方法。所有声学特征预测器均以声学模型的编码器输出为输入，并在音素级预测三种说话人归一化韵律声学相关量： $F_0$、能量和时长。

我们采用两种不同的声学特征架构 P:

i) 一种在 [17] 中描述的**循环**神经网络。该网络由 4 个双向长短期记忆网络层组成，维度分别为 = 64、64、32、32，用于将编码器输出映射为一系列隐藏状态。随后，这些状态经过一个 16 维全连接层、一个双曲正切激活函数和一个 3 维投影层。

ii) 一种基于 FastPitch 时序预测器的**卷积**神经网络 [16]。该网络由 2 个堆叠模块组成，每个模块包含一个核大小为 3、滤波器大小为 256 的一维卷积层、一个 ReLU 非线性函数、层归一化和随机失活（p=0.1）。随后，将该堆叠结构的输出输入最终投影层，映射至 3 维。

对于每种声学特征预测器架构，我们训练了 2 个版本，它们的差异仅在于参数值（通过使用不同的随机种子进行训练）。基于这 4 个检查点，我们构建了 3 个各含 2 个成员的集成。其中两个是同质集成，分别包含单一架构（**RNN-2** 和 **CONV-2**）的两个检查点。第三个是异质集成，包含每种架构各一个随机选取的检查点（**RNN-CONV**）。我们假设，异质集成可能产生更多样化的韵律预测，因而比同质集成更能受益于集成选择。

图1展示了声学模型中的声学特征预测器集成。该声学模型是多说话人版本的 Tacotron 2 [18]，与循环声学特征预测器结合后便成为 Ctrl-P [17]。推理时，声学特征预测器的输出会与编码器输出拼接，经过注意力处理和解码后生成梅尔频谱图，并将其输入独立的 WaveRNN声码器 [19]。

### 3.2. 选择标准

我们所有的候选选择标准都用于衡量模型对待合成话语的预测方差。每项标准都会选出方差最高的合成版本：

i) **全局方差**：随后 [11], ，我们从声学模型输出的梅尔频谱图中提取前 25 个梅尔频率倒谱系数（MFCCs）（使用 librosa），计算它们在整个话语中的逐参数平均方差，再将这些数值相加，得到该话语的全局方差；

ii) **WAV-F0** 是从合成语音波形中使用 RAPT [20]以 10 毫秒帧移估算的 $F_0$ 在所有有声帧上的方差；

iii) **AFP-F0** 是 AFP 预测的 $F_0$ 值在所有有声音素上的方差。

AFP-F0、全局方差和 WAV-F0 分别根据声学特征预测器、声学模型和声码器的输出计算。AFP-F0 只需对编码器和声学特征预测器进行一次前向传播，此时选择标准便可从集成成员提出的多个演绎版本中进行选择。全局方差还需要针对每个候选演绎版本对解码器进行一次前向传播。WAV-F0 则还需要使用声码器生成波形，同样要针对每个候选演绎版本进行。由于 AFP-F0 基于网络中最早预测的输出，因此其计算量比全局方差少 10 倍，比 WAV-F0 少 1000 多倍。

## 4. 实验设置

### 4.1. 训练

所有模型均在一个专有的墨西哥西班牙语表现力语料库上进行训练，该语料库共包含 32 位说话人的 38 小时语音。配音演员受指示以多种风格说话，例如“快乐”“愤怒”“讲故事”等，因此基频、能量和时长变化范围广泛。我们采用了两步训练流程 [17]：先将声学模型训练 20 万次迭代，然后冻结。随后，对每个声学特征预测器（第 3.1 节）训练 40 万次迭代，使其相对于从强制对齐的参考语音中提取的真实声学特征值，将 L2损失降至最低。所有阶段均使用 Adam 优化器，批量大小为 16。所有实验使用相同的 WaveRNN 检查点；该模型在教师强制模式下，使用来自相同架构声学模型的真实声学特征值生成的梅尔频谱图训练了 300 万次迭代。

### 4.2. 测试材料

我们使用了一个包含 30 个句子的测试集，这些句子是从视频合集（我们的目标应用场景）中随机选出的，并经人工专家转录、翻译和检查，以确保语义和语言准确性。这些句子未出现在训练数据中。

我们提出的方法旨在处理高度表现力语音，因此我们选择了 1 名男性和 1 名女性说话人，他们在训练数据中涵盖了最广泛的风格变化之一（38 种不同风格）。对于每位说话人，我们针对每个测试句子，为第 3.1 节中的 3 个模型集成分别生成一对合成版本（每个集成总计 60 对样本），用于听力测试。

| 集成 | 听者 1 | 听者 2 | 听者 3 |
|---|---|---|---|
| CONV-2 | 22% | 40% | 8% |
| RNN-2 | 33% | 57% | 25% |
| RNN-CONV | 58% | 70% | 35% |

**表1**。每个集成中被评为不同的话语对所占的百分比如下。三位专家听众均认为异构集成（RNN-CONV）的多样性最高。

### 4.3. 感知集成多样性

集成多样性是实现任何改进的必要前提。我们邀请了3位定期接触合成语音、且具有翻译或语言学背景的专家听众对此进行了评估。对于每个集成，我们展示了所有话语对，并询问：“A和B听起来是否不同？如果不同，请描述它们有何差异。”可选答案为“是”和“否”，并提供一个可选的自由文本字段。

### 4.4. 听众偏好

根据上述回答，我们确定了被评为最多样化的集成：这是我们用于基于集成的文本转语音系统的最佳候选方案。我们进行了第二次听测，以衡量听众对集成中各个模型的偏好，采用 A/B 测试。我们通过 Prolific 平台众包招募了 30 名以墨西哥西班牙语为母语的说话人，并用西班牙语询问他们：“请选择您更喜欢的版本：A 还是 B？”，同时提供“A”“B”和“无法决定”三个选项。

### 4.5. 选择标准准确率

随后，我们可以分析这些偏好评分，以衡量选择标准利用集成多样性的能力。对于 3 项评判标准，我们将 900 项选择（30 个句子 × × 30 名听众）中每位听众的个体偏好与相应评判标准选出的合成版本进行比较，并据此计算准确率 。如果两者一致，就认为该评判标准准确。共有 292 项回答（16.2%）标记为“无法决定”，因此未纳入准确率计算。

我们还构建了一个人工**ORACLE**：为每个话语选择听众最偏好的版本（即，在30位听众的评分中出现频率最高的标签）。如果“未决定”是最常见的选择，则排除该话语的所有评分。ORACLE以大众的多数投票作为选择标准，代表了在本次听力测试中任何选择标准或单个模型所能达到的最高性能。如果所有听众始终意见一致，ORACLE的准确率将达到100%。

## 5. 结果

### 5.1. 感知到的集成多样性

表1报告了各位专家听众感知到的平均多样性。感知到的多样性差异很大，但结果始终表明，异构集成模型RNN-CONV能够生成最多样的韵律预测。

表2总结了描述每个集成中不同合成版本之间感知差异的自由文本评论，我们为这些评论标注了关键词。集成多样性主要源于表现力、语调和重音，而我们预计这些因素与选择标准相关。“清晰度”和“语速/停顿”经常被同时提及，较快的语速会降低清晰度。由于这些韵律方面与时长有关，我们所探讨的标准无法充分捕捉它们；未来的标准可以将其纳入考量。因此，我们预计当前的选择标准能够利用集成中部分但并非全部感知到的多样性。

### 5.2. 听众偏好

图2展示了我们最多样化集成中各个声学特征预测器（**RNN**、**CONV**）的偏好结果。正如预期，没有任何一个模型总是更受偏好。两种模型的表现都处于随机水平：也就是说，听众偏好每种模型的概率大约各占一半，这表明可以结合它们各自的优势。我们将 RNN（因其略微更受偏好）作为非集成基线，并将所有集成系统与之进行比较。

### 5.3. 选择标准准确率

图3报告了第4.5节所定义的准确率。图中以垂直虚线标示非集成基线 RNN 的平均偏好度。我们进行了双侧费 Fisher 精确检验，以确定哪些准确率与 RNN 存在显著差异（$p \leq 0.05$），并采用 Holm-Bonferroni 校正。

最上方的三条柱形显示了自动选择准则 GV、WAV-F0 和 AFP-F0 的结果。AFP-F0 的准确率显著高于 RNN：自动选择

| 关键词 | CONV-2 | RNN-2 | RNN-CONV |
|---|---|---|---|
| 表现力 | 7 | 11 | 27 |
| 语调 | 5 | 20 | 24 |
| 重音 | 14 | 17 | 20 |
| 清晰度 | 3 | 9 | 16 |
| 语速/停顿 | 9 | 14 | 25 |
| 自然度 | 3 | 9 | 13 |
| 其他 | 3 | 4 | 3 |

**表2**。话语对之间的差异通常与表现力、语调和重音有关。该表显示了专家听众在描述各个集成中感知到的差异时，其描述性评论中标注的关键词数量。

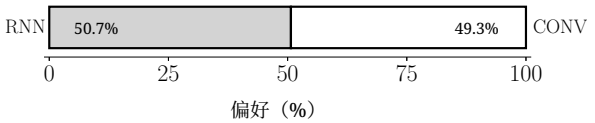



**图2**。我们多样性最高的集成中的两个独立声学特征预测器均处于随机水平（双侧二项检验，$p \leq 0.05$)

韵律预测器集成中的成员，能够胜过其中的单个成员。WAV-F0 衡量的属性与 AFP-F0 相似，但与 RNN 相比，差异在统计上并不显著（未校正的 p 值：0.094）。出乎意料的是，GV 显著更差于 RNN：也就是说，选择 GV最低的集成成员会是更好的评判标准。GV 显然从生成的梅尔频谱图中测量到了有用的信息，但具体测量的是什么还有待未来研究。

最后一根柱表示第4.5节定义的人工 ORACLE 准确率上限；由于听众的偏好当然无法完全一致，该值低于100%。AFP-F0 缩小了 RNN 与 ORACLE 之间 31% 的可用性能差距，表明自动评判标准能够利用集成中的多样性。

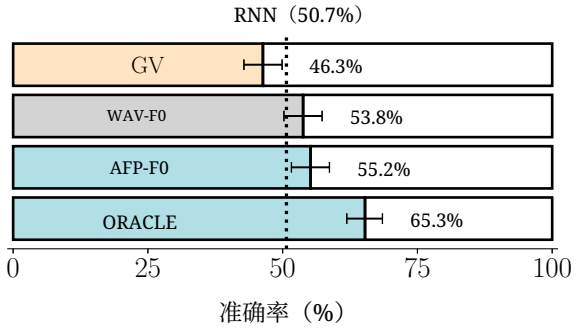



**图 3**。我们提出的评判标准（AFP-F0）比单一模型 RNN 基线更准确地预测听众偏好（虚线表示图 2 中的偏好）。准确率以 95% 置信区间表示；蓝色和橙色表示具有统计显著性。

## 6. 结论与未来工作

我们证明，韵律预测器集成能够产生多样性，而自动选择标准可以利用这种多样性，因此其表现优于任何单个模型。我们以最小的集成规模 2 展示了这一优势；采用了简单的集成生成方法（声学模型中同一组件的两种不同架构），以及简单且计算高效的 $F_0$ 方差选择标准。未来还有很大空间可进一步探索规模更大、更多样化的集成，以及更复杂的选择标准，尤其是在时长方面。

## 7. 参考文献


[1] Zack Hodari、Oliver Watts 和 Simon King，“利用生成建模为语音合成生成多样化的语调”，载于*SSW*论文集，2019年，第239–244页。[2]Peter Makarov、Syed Ammar Abbas、Mateusz Łajszczak、Arnaud Joly、Sri Karlapati、Alexis Moinet、Thomas Drugman 和 Penny Karanasou，“采用富有表现力且连贯的韵律实现简单有效的多句文本转语音”，载于*Interspeech*论文集，2022年，第3368–3372页。[3] Yi Ren、Ming Lei、Zhiying Huang、Shiliang Zhang、Qian Chen、Zhijie Yan 和 Zhou Zhao，“Prosospeech：通过文本转语音中的量化向量预训练增强韵律”，载于*ICASSP*论文集，2022年，第7577–7581页。[4] Yi Ren、Jinglin Liu 和 Zhou Zhao，“PortaSpeech：便携且高质量的生成式文本转语音”，载于*NeurIPS*论文集，2021年，第34卷，第13963–13974页。[5] Jo˜ao Moreira、Carlos Soares、Al´ıpio Jorge 和 Jorge Sousa，“回归集成方法综述”，《ACM计算调查》，第45卷，第10:1–10:40页，2012年11月。[6] J.G. Fiscus，“用于降低词错误率的后处理系统：识别器输出投票错误削减（ROVER）”，载于*ASRU*论文集，1997年，第347–354页。[7]Gunnar Evermann 和 P.C. Woodland，“后验概率解码、置信度估计与系统组合”，载于语音转录研讨会论文集，2000年，第27卷，第78–81页。[8] Taiga Abe、E. Kelly Buchanan、Geoff Pleiss、Richard Zemel 和 John P. Cunningham，“深度集成有效，但是否必要？”，*arXiv:2202.06985*，2022年。[9] Hao Sun、Xu Tan、Jun-Wei Gan、Hongzhi Liu、Sheng Zhao、Tao Qin 和 Tie-Yan Liu，“用于字素到音素转换的词元级集成蒸馏”，载于*Interspeech*论文集，2019年，第2115–2119页。

[10] Harald Romsdorfer，《用于语音韵律控制的加权神经网络集成模型》，载于*Interspeech*论文集，2009年，第492–495页。

[11] Tomoki Toda 和 Keiichi Tokuda，《一种考虑全局方差的基于隐马尔可夫模型的语音合成语音参数生成算法》，电子信息通信学会信息与系统汇刊，第E90-D卷，第5期，第816–824页，2007年5月。

[12] Zack Hodari、Catherine Lai 和 Simon King，《使用无监督离散基频表示感知语音合成中的韵律变化》，载于*Speech Prosody* 论文集，2020年。

[13] Xin Wang、Shinji Takaki、Junichi Yamagishi、Simon King 和 Keiichi Tokuda，《用于统计参数语音合成的向量量化变分自编码器（VQ-VAE）自回归神经$f_0$ 模型》，*IEEE/ACM* 音频、语音与语言处理汇刊，第28卷，第157–170页，2020年。[14] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速高质量端到端文本转语音》，载于*ICLR* 论文集，2021年。[15] Tuomo Raitio、Ramya Rasipuram 和 Dan Castellani，《使用直观韵律特征实现可控神经文本转语音合成》，载于*Interspeech* 论文集，2020年，第4432–4436页。[16] Adrian Lancucki，《FastPitch：基于基频预测的并行文本转语音》，载于*ICASSP* 论文集，2021年。[17] Devang S Ram Mohan、Vivian Hu、Tian Huey Teh、Alexandra Torresquintero、Christopher G R Wallis、Marlene Staib、Lorenzo Foglianti、Jiameng Gao 和 Simon King，《Ctrl-P：语音合成中韵律变化的时间控制》，载于*Interspeech* 论文集，2021年。[18]Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan 等，《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* 论文集，2018年，第4779–4783页。[19] Nal Kalchbrenner、Erich Elsen、Karen Simonyan、Seb Noury、Norman Casagrande、Edward Lockhart、Florian Stimberg、A¨aron van den Oord、Sander Dieleman 和 Koray Kavukcuoglu，《高效神经音频合成》，载于*ICML* 论文集，2018年，第2415–2424页。[20] David Talkin，《一种稳健的基频跟踪算法（RAPT）》，载于《Speech Coding and Synthesis》，W. B. Kleijn 和 K. K. Paliwal 编，第497–18页。Elsevier Science B.V.，1995年。