arXiv:2603.05091v1 [eess.AS] 5 Mar 2026

采用紧凑且具可解释性的免训练声学参数进行嗓音音色属性检测

Aemon Yat Fei Chiu ‡ *Yujia Xiao* ‡ *Qiuqiang Kong* ◇ *Tan Lee* ‡, ◇

‡ 电子工程系，香港中文大学，中国香港◇ 数据科学学院，香港中文大学，中国深圳{aemon.yf.chiu,yujiaxiao}@link.cuhk.edu.hk, {qqkong,tanlee}@ee.cuhk.edu.hk

摘要

嗓音音色属性检测（vTAD）旨在确定不同语音话语之间音色属性的相对强度。嗓音音色是语音感知中至关重要却又固有复杂的组成部分。尽管深度神经网络（DNN）嵌入在说话人建模中表现出色，但它们通常作为黑箱表征，物理可解释性有限且计算成本高。本文研究了一种用于vTAD 的紧凑声学参数集。该参数集捕捉了重要的声学测量指标及其时变特性，而研究发现这些因素对该任务至关重要。尽管结构简单，该声学参数集仍具有竞争力，其性能优于传统倒谱特征和有监督深度神经网络嵌入，并接近最先进的自监督模型。更重要的是，所研究的参数集无需可训练参数，计算开销可忽略不计，并为分析人类音色感知背后的物理特征提供了明确的可解释性。**索引词**：嗓音音色、声学参数、可解释性、说话人特征

1. 引言

嗓音音色是人类语音的重要组成部分，却也是一个难以理解的维度，可用于识别和区分说话者 [1, 2]。它就像说话者的“听觉面孔” [3]。嗓音音色传达与性别、年龄和生理特征相关的稳定个人特征，也传达情绪和健康状况等动态状态 [4]。根据 [5], , 音色被定义为一种属性，它“使听者能够判断两个不相同的声音在以相同响度、音高、空间位置和时长呈现时，彼此不同”，并且“通常用定性形容词（例如明亮或沉闷）”来描述。这凸显了音色的复杂性，也表明嗓音音色分析依赖主观的自然语言描述 [2, 6]。

在 [7], 中，研究者提出了嗓音音色属性检测（vTAD）框架，用于探究并解释嗓音音色与语音声学之间的关系。在该框架中，音色属性由一组嗓音音色语言描述词定义，如表1所示。vTAD 系统使用 Diff-Net 测量来自不同说话者的两段语音话语之间某一音色属性的相对强度。真实标签由人类感知得出 [8]。vTAD 任务利用一个大规模语音数据集，该数据集提供语音话语与嗓音音色主观描述之间的直接映射。其目标是评估最新说话人嵌入模型在人类专家感知声音音色属性方面的表现。

vTAD 的任务与说话人验证（SV）类似，如图1所示。通用说话人验证系统使用说话人嵌入——

表1： *vTAD [7, 8]*中使用的描述符集。比例 表示数据集中每个描述符所占的百分比（%）。

描述符	比例	描述符	比例
明亮	17.10	Thin	13.03
粗糙	11.62	Slim	11.31
Low	7.43	Pure	5.48
Rich	4.71	富有磁性	3.64
浑浊的	3.59	沙哑	3.32
圆润	2.48	Flat	2.15
尖细（女性特有）	2.08	干瘪	1.74
沉闷	1.44	Soft	0.82
通透	0.66	沙哑（男性特有）	0.59

这些嵌入会纠缠语音变异中的多个因素，例如内容、音色和韵律 [9, 10]。vTAD 专注于评估潜在说话人嵌入如何专门编码语音中的嗓音音色信息。尽管使用大量数据训练的说话人嵌入系统在 vTAD 中表现出色 [11–17], , 但这些系统需要消耗大量计算能力来提取高维嵌入，而且缺乏对有意义的声音分析至关重要的可解释性。可解释性对于构建可靠的语音人工智能系统至关重要，在法证和法律场景等真实应用中发挥着重要作用。

在本研究中，我们重新审视并探究了一组用于vTAD的声学参数。该集合包含13个与语音产生相关的特征及其时变特性，构成一个26维向量。与在大规模语音数据上训练、维度达数百甚至数千的深度神经网络学习得到的说话人嵌入相比，这一26维声学参数集在特征提取时无需GPU加速，却能在该任务上取得相当的性能，并提供明确的可解释性。Anal-

实验结果分析表明，语音的时间动态特征在区分音色属性方面发挥着重要作用。本文分析了各项声学特征的贡献，以深入理解可解释语音音色分析与说话人嵌入设计。

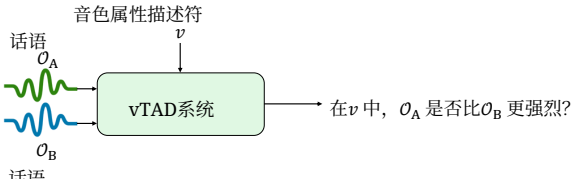


图 1： *vTAD [7]* 的定义。

2. 相关工作

早期的音色研究主要关注其物理相关因素 [18, 19]。后续研究则转向关注音色的感知层面 [20, 21]。尽管研究者尝试构建语音音色空间 [1]，并开发用于音色分类的声学参数集 [6, 23–25]，但音色至今仍缺乏清晰且全面的定义 [22],。

当前的 vTAD 方法主要由数据驱动的基于深度神经网络的说话人嵌入模型和分类器主导 [11]。Chiu 等人 [14] 将 WavLM-Large [26] 用作说话人编码器。Song 等人 [15] 研究了改进的说话人嵌入融合方法。其他研究 [16, 17] 则专注于改进下游 Diff-Net，以获得更好的分类性能。从大量语音数据中学习得到的高维嵌入缺乏可解释性。它们通过黑箱模型分析，能够很好地区分声音音色，但无法对音色为何相似或不同提供任何有意义的解释。

随着近几十年来人工智能和深度神经网络的快速发展，人们提出了各种说话人嵌入和表征，用于在潜在空间中表示独特的说话人 [12, 26–30]。学习得到的嵌入通常会将多个说话人相关成分纠缠在一起，包括音色、韵律和口音 [10, 31]。已有研究尝试对这些成分进行显式解耦，并对人类语音的音色进行显式建模 [9, 13]。

Liu 等人 [32] 研究了人类感知、基于深度神经网络的说话人验证系统，以及用于语音相似度测量的一组声学参数之间的一致性。研究发现，在进行二元同异判断时，声学参数与人类感知具有中等程度的一致性，但在评定说话人相似度时，与人类判断的一致性有限。说话人验证系统仅在宽泛的整体语音层面上与人类感知一致。尽管两组特征内部相关性很高 [32]，但这两种自动系统都无法充分捕捉人类听觉判断中的细微语音细节。

本研究将关注重点从一般的说话人验证转向更具体的 vTAD 任务。研究采用了一个包含细粒度专家标注的大型数据集，以及用于描述音色属性的受控自然语言框架。在本研究中，专家标注被视为真实标注。通过将计算特征的输出与专家
标签。

3. 用于 vTAD 的声学参数

3.1. 任务定义

在 vTAD 任务中，说话人A和说话人B的一对语音话语 O_A 和 O_B 与音色描述符 $v \in V$一同输入检测系统。系统会生成一个范围为 $[0, 1]$的预测分数。分数接近 0 表示我们有较高把握认为 O_A 在音色属性 v上的表现更突出，反之亦然 [7, 33]。图2展示了 vTAD 系统的详细工

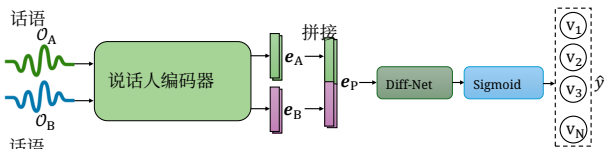


图2： *vTAD* 的整体工作流程 *[7]*。

作流程。通过冻结的说话人编码器，分别从 O_A和 O_B 中提取说话人嵌入 e_A和 e_B，并将其拼接为 e_P，再输入 Diff-Net，以获得音色属性 v_1 至 $v_N \in V$对应的预测分数 $\hat{y}$。

3.2. 声学参数

在所提出的系统中，说话人编码器提取 13 个基础声学特征及其各自的变异系数（CoV），这些特征源自 [6, 24, 25]。这些声学参数汇总于 [6, 24]。基础参数包括基频（F_0）、前四个共振峰频率（F_1、 F_2、 F_3 和 F_4）、共振峰离散度、四项谐波频谱形状测量指标（$H_1^* - H_2^*$、 $H_2^* - H_4^*$、 $H_4^* - H_{2\text{kHz}}^*$ 和 $H_{2\text{kHz}}^* - H_{5\text{kHz}}$），以及三项非谐波声源指标：倒谱峰值突出度（CPP）、均方根（RMS）能量和次谐波与谐波比（SHR）。这 13 个参数及其变异系数组成一个 26 维特征向量。

本文保留了 [24] 中总结的核心声学定义，同时对提取方法进行了若干改进，以增强鲁棒性并实现动态对齐。本研究使用 Praat-Parselmouth 工具 [34] 提取声学参数。原始声学测量采用 10 毫秒的时间步长，以捕捉高度动态的嗓音音色。通过自相关提取 F_0，基频下限设为 75 Hz。使用 Burg算法估计共振峰及其带宽，最多估计五个共振峰，频率上限为 5.5 kHz。共振峰离散度在计算上定义为 F_4 与 F_1 之间的频率差除以三。

为计算频谱和能量特征，我们对每个浊音帧采用固定的 40 毫秒分析窗。在这些片段中计算均方根能量，并提取频谱倾斜度量，随后校正共振峰频率和带宽重叠造成的影响。根据片段的幅度谱计算 SHR，并通过稳健的抛物线功率倒谱图方法提取 CPP。提取 SHR 和 CPP 时均使用 Praat-Parselmouth 的 [34]内置函数。

提取这些原始时间序列指标后，直接计算所有有效浊音帧的全局均值和全局变异系数，从而为每个音频文件构建最终的紧凑型 26 维话语级表示。

3.3. 下游分类器

我们采用一个简单的 Diff-Net 作为下游分类器并对其进行训练，以执行 vTAD。该网络由两个全连接层（FC）组成，中间设有批归一化（BN）层、ReLU 激活函数和丢弃层。

4. 实验设置

4.1. 数据集

本研究使用了 VCTK-RVA 数据集 [8],，该数据集源自 VCTK 数据集 [35]，并由语音专家对音色属性强度进行了人工标注。该数据集中，共有 6,038 对同一性别的说话人对，涉及 101 位说话人（40 位男性和 61 位女性）。每对说话人都标注了 1 到 3 种音色属性的强度 [7, 8]。

建立训练集和测试集，用于训练 Diff-Net 并评估 vTAD 系统。最初，训练集

* 对共振峰对谐波振幅的影响进行校正

包含来自 78 位说话人（29 位男性说话人和 49 位女性说话人）的 136,320 条话语 [7]。在本研究中，从训练集中移除了包含样本 p341/p341 101（一个静音文件）的话语对，最终得到 136,316 个有效话语对。采用 [7] 提供的未见过的测试集，确保其中的说话人与训练集中的说话人没有重叠。测试集包含来自 23 位说话人（11 位男性和 12 位女性）的 91,600 个话语对。

在数据预处理方面，音频文件直接从 [35] 获取，仅使用 mic1.flac 文件。数据从 FLAC 转换为 WAV 格式，并从 48 kHz 降采样至 16 kHz，不做进一步修改，严格遵循 [7, 8] 中的设置。

4.2. 基线系统

本研究评估了五种说话人嵌入和倒谱特征，将其作为基线说话人编码器：ECAPA-TDNN [12], FA-Codec [13], 梅尔频率倒谱系数（MFCC） [1], 以及 WavLM [26]。

• ECAPA-TDNN [12]是一种广受欢迎的监督式说话人嵌入模型，专为说话人验证训练。FA-Codec 是 NaturalSpeech 中的说话人编码器 3 [13], , 这是一种零样本文本转语音合成系统，能够以高度自然且相似的效果复现说话人的声音。1024 通道 ECAPA-TDNN[12] 和 FA-Codec 的音色分支 [13] 是 vTAD任务 [7]的两个主要基线。

• 梅尔频率倒谱系数是常用的音频信号特征，广泛应用于包括语音识别和说话人识别在内的各种下游任务。LFC 是一种“稻草人”表示，它在 [1]中用其他尺度替代了梅尔频率倒谱系数中的梅尔频率尺度。梅尔频率倒谱系数和 LFC 均在一项关于语音音色潜在特征的早期研究中进行了测试，并在此处使用遵循 [1]的实现进行评估。

• WavLM [26] 是一种大型语音自监督学习模型，由一系列 Transformer 层组成， [36] 在 960 至 94,000 小时的语音数据上进行训练。Chiu 等人 [14] 使用注意力统计池化（ ASTP） [37] 聚合 WavLM-Large [26]所有 Transformer 层的输出，以获得最终的说话人嵌入。

WavLM 模型的 Diff-Net 包含四个模块，每个模块在最终的全连接层之前由全连接层、批归一化、ReLU 和丢弃层组成， [14], 而其他系统则严格遵循第 3.3 节所述的设置。

5. 结果与分析

5.1. 性能分析

表 2 展示了声学参数集在 vTAD 任务上与若干基线系统的性能，评估指标为准确率和等错误率。采用 ASTP-L（一种多层输出聚合方法）的 WavLM-Large 达到了最先进（SOTA）水平。26 维声学参数集的准确率为 82.87%，等错误率为 17.21%。其性能优于 WavLM 模型（ WavLM-Base、WavLM-Base+ 和 WavLM-Large）的最终表征、梅尔频率倒谱系数（MFCC）及其变体 LFC，以及专门的有监督说话人嵌入模型，即 ECAPA-TDNN 和 FA-Codec 的音色嵌入。此外，其性能与 SOTA 非常接近。

说话人嵌入的结果与先前研究发现一致 [7] ，其中 FA-Codec 在未见数据上表现出更强的泛化能力，

表 2： 各系统在 *vTAD* 上的性能，以准确率（*Acc*）和等错误率（*EER*）衡量。

模型	**准确率（%）↑**	**等错误率（%）↓**
ECAPA-TDNN	70.37	28.67
FA-Codec	79.32	20.60
MFCC	68.72	31.15
LFC	80.32	19.41
WavLM-Base	75.76	23.89
WavLM-Base+	78.02	22.23
WavLM-Large	80.08	19.21
搭配 ASTP-***L*** [14]	**83.13**	**16.87**
声学参数	82.87	17.21

而 ECAPA-TDNN 在已见测试数据上表现更好，因为这些说话人已出现在训练集中。然而，对两种常见音频特征 MFCC 和 LFC 的比较，得出了与先前研究不同的结果 [1], ；先前研究认为 MFCC 是进行感知音色分析的更佳特征。表 2 显示，在 vTAD 任务中，LFC 的表现显著优于 MFCC。由于梅尔刻度会大幅压缩高频信息，因此可能会掩盖对感知特定音色属性至关重要的高频非谐波能量和谱噪声 [24]。LFC 通过保留整个频谱的线性分辨率，更有效地捕捉这些高频细节；这与线性频率刻度在说话人识别任务中更稳健的研究发现相符 [38]。

表 2 还表明，训练后的模型可以从更大的训练数据规模和模型规模中受益，这一点从 WavLM-Base、WavLM-Base+ 和 WavLM-Large 之间的性能差异中可见。WavLM-Base 由 12 个 Transformer 层组成，具有 768 个隐藏状态，并使用 960 小时的语音数据进行训练。WavLM-Base+ 与 WavLM-Base 采用相同的模型架构，但使用 94,000 小时的语音数据进行训练；WavLM-Large 也采用类似的训练数据规模，由 24 个 Transformer 层组成，具有 1,024 个隐藏状态。结果还表明，使用 WavLM-Large 的中间表示可以进一步显著提升性能。

5.2. 可解释性

该声学参数集的根本优势在于其物理可解释性。图 3 展示了从训练好的 Diff-Net 中直接提取的特征重要性权重，该网络额外加入了一个简单的可学习权重层。CPP、能量、F_0、SHR 的均值以及 F_1 的变异系数，是区分音色属性最重要的正向指标。 F_0 代表声带的实际振动速率，而 CPP 则直接衡量周期性和谐波丰富度的程度；二者都是感知嗓音音色的基础。SHR 均值具有较大的权重，表明周期倍增的存在也可能是 vTAD 的重要指标。共振峰频率 F_1 通常与元音身份的关系更为密切，但研究也显示，它在音色检测中发挥着重要作用。频谱斜率的动态变化（例如 $H_2^* - H_4^*$、 $H_4^* - H_{2\text{kHz}}^*$ 和 $H_1^* - H_2^*$ 的变异系数）与气声或听觉上的明亮度有关，是网络中最重要的负权重。这表明高频非谐波能量的时间变异性具有区分不同说话人感知特征的能力。

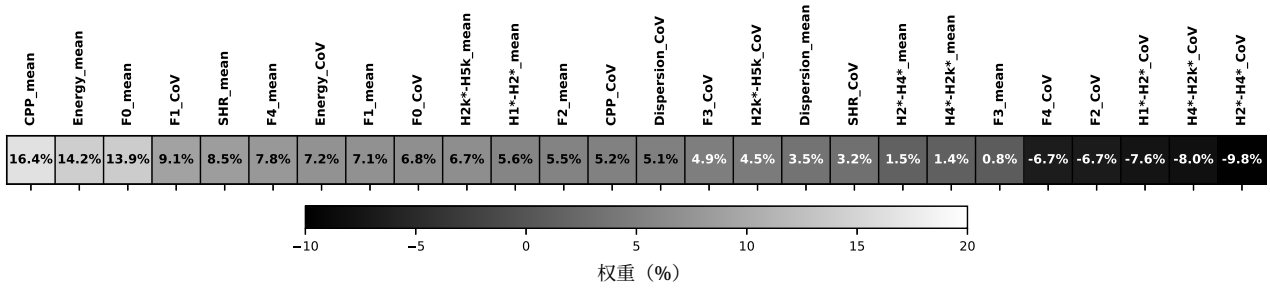


图3： 采用声学参数集的 *Diff-Net* 特征权重。

这与 [24] 的结论一致：高频谱形状和谱噪声的变化对于区分说话人的音色至关重要。然而， [24] 认为 F_0 在语音空间中的重要性不高，而本研究发现 F_0 是一项主要的区分特征，这与近期说话人建模研究的结论相符；这些研究通常将 F_0 视为说话人特定属性的基础要素 [6,10,39,40]。

传统基线倒谱特征（如梅尔频率倒谱系数和 LFC 系数）通过对滤波器组应用离散余弦变换得到。这种变换会将谱包络去相关为数学基函数，无法映射回可解释的物理特征。高维的基于深度神经网络的模型（包括 ECAPA-TDNN、FA-Codec 和 WavLM）以黑箱方式运行，并将各种语音成分交织到抽象的潜在空间中。因此，其输出嵌入通常缺乏可解释性。值得注意的是，这些说话人嵌入模型通常采用帧平均，从而消除了已被证明对区分音色属性至关重要的时变特性。即使采用 ASTP 等广泛使用、能够捕捉方差的技术，这些动态信息也会被投射到抽象的潜在空间中，并与均值高度交织

5.3. 效率

除了具有可解释性，声学参数集的计算效率也很高，且无需使用图形处理器。如表3所示，声学参数集的提取过程不需要任何可训练参数，处理一秒语音仅需 17.85 M 次每秒浮点运算（FLOPs），即可为每段话语输出一个 26 维向量。相比之下，训练好的深度神经网络模型需要 10 到 300 M 个可训练参数，每秒语音处理量为 80 M 到 25 G FLOPs，输出的嵌入维度从 192 到 1024（按帧）不等（为获得最佳性能，还需聚合 WavLM-Large 全部 25 层的表征，从而进一步扩展维度）。嵌入规模的差异表明，与基于深度神经网络的方法相比，针对声学参数集训练下游 Diff-Net 也更加高效，如表4所示。虽然梅尔频率倒谱系数和 LFC 的提取及下游训练成本略低，但其性能较差，且可解释性相对不足。

6. 结论

本研究表明，一组紧凑且可解释的 26 维声学参数能够有效捕捉难以捉摸的嗓音音色特征，成为一种强大的.....

表3： 各系统提取说话人嵌入的计算成本，按模型参数量（参数）、每秒浮点运算次数（*FLOPs*）以及每条话语的维度（*dim*）衡量。

模型	参数量	每秒浮点运算次数	每条语句的维度
ECAPA-TDNN	17.24 M	84.33 M	192
FA-Codec	103.61 M	3.57G	256
MFCC	0	5.41 M	13
LFC	0	4.46 M	13
WavLM-Base	94.70 M	7.25 G	13 × [***F***, 768]
WavLM-Base+	94.70 M	7.25 G	13 × [***F***, 768]
WavLM-Large	316.62 M	25.88 G	25 × [***F***, 1,024]
声学参数	0	17.85 M	26

表4：各系统 *Diff-Net* 训练过程的计算成本，以模型参数（*params*）和每对话语（*uttr*）对的 *FLOPs* 衡量。

模型	参数量	每对语句的 FLOPs
ECAPA-TDNN	70.31 k	140.80 k
FA-Codec	53.92 k	108.03 k
MFCC	7.84 k	15.36 k
LFC	7.84 k	15.36 k
WavLM-Base	1.03 M	2.06 M
WavLM-Base+	1.03 M	2.06 M
WavLM-Large	1.29 M	2.59 M
搭配 ASTP-***L*** [14]	10.74 M	844.36 M
声学参数	11.17 k	22.02 k

在 vTAD 任务中，它优于复杂的高维 DNN 嵌入。该组具有物理依据的参数在嗓音音色属性检测任务中达到 82.87% 的准确率，优于广泛使用的有监督说话人嵌入和传统倒谱特征。此外，其性能与最先进的 WavLM 模型相当，同时具有更高的可解释性，且所需计算资源显著更少。研究发现，语音中的时间变化对音色感知至关重要，但 DNN说话人嵌入并未对其进行显式建模。这些发现表明，将可解释的声学知识融入现代人工智能系统，是实现可解释、高效说话人特质分析的一条有前景的途径。

7. 生成式人工智能使用声明

作者声明，生成式人工智能工具仅用于编辑和润色本稿件，包括纠正语法错误、优化句子结构以及提升英文文本的整体清晰度。所有作者（包括共同作者）均对稿件中呈现的原创研究、数据分析和技术内容承担全部责任。未使用任何生成式人工智能工具撰写稿件的任何重要部分。

8. 参考文献

[1] H. Terasawa、M. Slaney 和 J. Berger，“语音的音色空间”，载于*Interspeech 2005*论文集，2005年，第1729–1732页。[2] J. Kreiman 和 D. Sidtis，声音研究基础：声音产生与感知的跨学科方法。马尔登，马萨诸塞州：Wiley-Blackwell，2011年。[3] P. Belin、S. Fecteau 和 C. B´edard，“思考声音：声音感知的神经相关性”，*Trends in Cognitive Sciences*，第8卷，第3期，第129–135页，2004年3月。[4] J. Laver，嗓音质量的语音学描述，剑桥语言学研究丛书第31卷。剑桥：剑桥大学出版社，1980年。[5]A.S.A.，“标准术语：音色”，https://asastandards.org/terms/timbre/。[6] J. Kreiman，“嗓音质量传递的信息”，*The Journal of the Acoustical Society of America*，第155卷，第2期，第1264–1271页，2024年2月。[7] J. He、Z. Sheng、L. Chen、K. A. Lee 和 Z.-H. Ling，“介绍嗓音音色属性检测”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2505.09661[8] Z.-Y. Sheng、L.-J. Liu、Y. Ai、J. Pan 和 Z.-H. Ling，“利用文本提示编辑声音属性”，*IEEE Transactions on Audio, SpeechandLanguage Processing*，第33卷，第1641–1652页，2025年。[9] H. Lu、X. Wu、Z. Wu 和 H. Meng，“Speechtriplenet：面向内容、音色和韵律的端到端解耦语音表征学习”，载于第*31*届*ACM*国际多媒体会议论文集，2023年，第2829–2837页。[10] A. Y. F. Chiu、K. C. Fung、R. T. Y. Li、J. Li 和 T. Lee，“对自监督语音表征中说话人特定属性的大规模探测分析”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2501.05310[11] L. Chen、J. He、Z. Sheng、K. A. Lee 和 Z.-H. Ling，“首届嗓音音色属性检测挑战赛”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2509.06635 [12] B. Desplanques、J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：基于TDNN的说话人验证中的强调式通道注意力、传播与聚合”，载于*Interspeech 2020*，2020年，第3830–3834页。[13] Z. Ju、Y. Wang、K. Shen、X. Tan、D. Xin、D. Yang、Y. Liu、Y. Leng、K. Song、S. Tang、Z. Wu、T. Qin、X.-Y. Li、W. Ye、S. Zhang、J. Bian、L. He、J. Li 和 S. Zhao，“NaturalSpeech 3：采用因子化编解码器和扩散模型的零样本语音合成”，载于第*41*届国际机器学习会议论文集，ICML’24系列，JMLR.org，2024年。[14] A. Y. F. Chiu、J. Li、Y. Tian、G. Zhang 和 T. Lee，“参加NCMMSC 2025 VTAD挑战赛的CUHK-EE系统”，载于人机语音交流，J. Jia、Z. Wu、L. Gao、G. Huang 和 Y. Li 编。新加坡：Springer Nature Singapore，2026年，第290–298页。[15] Z. Song 和 L. He，“介绍用于嗓音音色属性检测的判别式说话人嵌入”，载于人机语音交流，J. Jia、Z. Wu、L. Gao、G. Huang 和 Y. Li 编。新加坡：Springer Nature Singapore，2026年，第447–458页。[16] M. Deng、J. Wang、Y. Lang、S. Zhao 和 X. Xie，“融合预训练说话人特征提取器与性别先验特征的音色属性判别系统”，载于《人机语音交流》，J. Jia、Z. Wu、L. Gao、G. Huang 和 Y. Li 编。新加坡：Springer Nature Singapore，2026年，第470–481页。[17]Z. Wu、J. Fang、Y. Tang、Y. Zheng、Y. Wang 和 H. Fei，“Qvtad：用于嗓音音色属性检测的差分相对属性学习”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2508.15931[18] H. von Helmholtz，“论音调感觉：音乐理论的生理学基础”，《自然》，第12卷，第308期，第449–452页，1875年9月。 [在线]。获取地址：https://doi.org/10.1038/ 012449a0[19] C. Stumpf，《语音》。海德堡：Springer Berlin，1926年。[20]S. McAdams，“声源与事件的识别”，载于《声音思维：人类听觉的认知心理学》。牛津大学出版社，1993年4月。 [在线]。获取地址：https://doi.org/10.1093/acprof:oso/9780198522577.003.0006[21] S. Handel，“第12章——音色感知与听觉对象识别”，载于《听觉》，《感知与认知手册》系列，B. C. Moore 编。圣迭戈：学术出版社，1995年，第425–461页。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/B9780125056267500145[22]Y. Wei、L. Gan 和 X. Huang，“音色感知神经认知研究综述”，《心理学前沿》，第13卷（2022年），2022年。 [在线]。获取地址：https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.869475[23] J. Va˘nkov´a 和 R. Skarnitzl，“表征短时嗓音质量的参数在说话人内部及说话人之间的变异性”，载于Speech Prosody 2014论文集，2014年，第1081–1085页。[24] Y. Lee、P. Keating 和 J. Kreiman，“说话人内部及说话人之间的嗓音声学变异”，The Journal of the Acoustical Society of America，第146卷，第3期，第1568–1579页，2019年9月。[在线]。获取地址：https://doi.org/10.1121/1.5125134[25] J. Kreiman、Y. Lee、M. Garellek、R. Samlan 和 B. R. Gerratt，“心理声学嗓音质量模型的验证”，The Journal of the Acoustical Society of America，第 149， 1， 457 465, 卷。期。01 2021 [ ] 10 1121 10 0003331 页码。–。在线。获取地址：https://doi.org/ 。/。[26] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao、J. Wu、L. Zhou、S. Ren、Y. Qian、Y. Qian、J. Wu、X. Zeng、M.和 Yu、F. Wei，“WavLM：面向全栈语音处理的大规模自监督预训练”，《IEEE信号处理专题期刊》，第16卷，第6期，第1505–1518页，2022年。[27] D. Snyder、D. Garcia-Romero、D. Povey 和 S. Khudanpur，“用于文本无关说话人验证的深度神经网络嵌入”，载于Interspeech 2017论文集，2017年，第999–1003页。[28] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，“X向量：用于说话人识别的鲁棒深度神经网络嵌入”，载于2018年IEEE声学、语音与信号处理国际会议论文集，2018年，第5329–5333页。[29] H. Zeinali、S. Wang、A. Silnova、P. Mat˘ejka 和 O. Plchot，“BUT系统在VoxCeleb 2019说话人识别挑战赛中的系统描述”，2019年。 [在线]。获取地址：https://arxiv.org/abs/1910. 12592[30] H. Wang、S. Zheng、Y. Chen、L. Cheng 和 Q. Chen，“Cam++：一种利用上下文感知掩蔽进行说话人验证的快速高效网络”，载于Interspeech 2023论文集，2023年，第5301–5305页。[31] C. Luu、S. Renals 和 P. Bell，“利用解耦说话人表征研究说话人属性对说话人可分性的贡献”，载于Interspeech 2022论文集，2022年，第610–614页。[32] S. Liu、M. Babel 和 J. Zhu，“通过声学、人类感知和深度神经网络（DNN）说话人验证系统比较声音相似度”，载于Interspeech 2024论文集

[33] Z. Sheng、J. He、L. Chen、K. A. Lee 和 Z.-H. Ling，《嗓音音色属性检测2025年挑战赛评估方案》，2025年。[在线]。可访问：https://arxiv.org/abs/2505.09382[34]Y. Jadoul、B. Thompson 和 B. de Boer，《介绍Parselmouth：面向Praat的Python接口》，《语音学杂志》，第71卷，第1–15页，2018年。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S0095447017301389[35] J. Yamagishi、C. Veaux 和 K. MacDonald，《CSTR VCTK语料库：用于CSTR语音克隆工具包的英语多说话人语料库（版本0.92）》，2019年。 [在线]。可访问：https: //doi.org/10.7488/ds/2645[36] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin，《注意力就是你所需要的一切》，载于神经信息处理系统进展，I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，第30卷。Curran Associates, Inc.，2017年。 [在线]。可访问：https://proceedings.neurips.cc/paper files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf[37] K. Okabe、T. Koshinaka 和 K. Shinoda，《用于深度说话人嵌入的注意力统计池化》，载于*Interspeech 2018*年会议论文集，2018年，第2252–2256页。[38] X. Zhou、D. Garcia-Romero、R. Duraiswami、C. Espy-Wilson 和 S. Shamma，《用于说话人识别的线性倒谱系数与梅尔频率倒谱系数比较》，载于*2011*年*IEEE*自动语音识别与理解研讨会论文集，2011年，第559–564页。[39] S. Ji、J. Zuo、M. Fang、Z. Jiang、F. Chen、X. Duan、B. Huai 和 Z. Zhao，《TextrolSpeech：采用编解码器语言文本转语音模型的文本风格控制语音语料库》，载于*2024*年*IEEE*声学、语音与信号处理国际会议论文集，2024年，第10 301–10 305页。[40] Z. Jin、J. Jia、Q. Wang、K. Li、S. Zhou、S. Zhou、X. Qin 和 Z. Wu，《SpeechCraft：带有自然语言描述的细粒度表现性语音数据集》，载于第*32*届*ACM*国际多媒体会议论文集，2024年，第1255–1264页。