受 KA 表示定理启发的语音音色属性检测模型

彭　程[1]，姜　林[1,2]，彭　博[1]，陈颖超[1]，邓赛男[1]
（1 湖南工商大学人工智能与先进计算学院（湘江书院），长沙 410205;
2 湘江实验室，长沙 410205）

摘 要：针对现有语音音色属性检测模型的特征检测模块多采用“特征拼接+浅层分类器”范式，存在缺乏上下文依赖建模、难以关注判别性特征以及特征融合过程过于简单的问题。本文提出一种受 Kolmogorov-Arnold（KA）表示定理启发的语音音色属性检测模型。该方法聚焦属性检测模块的优化，在检测模块中引入 Transformer 结构替代传统 MLP 分类器，利用其多头注意力机制增强上下文依赖建模能力并实现对判别性特征的动态关注；同时提出 KA 启发的特征融合模块（KAfusion），通过内层函数（InnerFunction）建模单个说话人特征的内部交互关系，外层函数（OutterFunction）捕捉多说话人特征间的交互关系，实现对音色属性的融合表征。实验结果表明，相比基线模型，检测准确率和等错误率上分别提升了 4.13%和 3.76%。

关键词：语音音色属性检测；特征融合；Kolmogorov-Arnold 表示定理；Transformer 模型
中图分类号：TP393.1　　**文献标志码**：A

语音音色作为说话人声音的独特属性，具有生物特征唯一性，在语音识别[1]、语音合成[2]和情感计算[3]等领域具有重要的研究价值。随着多场景语音交互需求的快速增长，对各类语音音色属性进行特征表达，并准确检测音色属性强弱等特征具有重要的应用意义。

语音音色属性检测通常由特征提取模块和属性检测模块组成。特征提取模块负责将原始语音信号建模为通用的语音嵌入表达，常用的方法包括基于梅尔频率倒谱系数（MFCC）的统计建模[4]、深度神经网络（如 TDNN[5]、ResNet[6]）的端到端特征学习，以及自监督预训练模型（如 Wav2Vec 2.0[7]）的嵌入提取。属性检测模块则对语音嵌入进行属性分类或比较，传统方法主要依赖于多层感知机（MLP）或支持向量机（SVM）等浅层分类器[8]。然而，这些方法在建模复杂音色属性关系时性能有限。本文聚焦属性检测模块的设计，旨在通过引入更高效的建模机制提升音色属性检测的性能。

现有的属性检测模块通常由分类器构成，如：多层感知机（MLP），通过全连接层实现特征的非线性变换与分类决策[9]；支持向量机，利用核函数将特征映射到高维空间进行线性划分[10]；图神经网络，通过图结构建模特征节点间的拓扑关系[11]；时延神经网络，采用时序滑动窗口机制捕获语音特征的动态演变[12]；卷积神经网络，通过局部感受野提取具有平移不变性的特征表示[13]。上述属性检测模块对多特征处理时，多采用“特征拼接+浅层分类器”这一极简范式构成：首先对原始语音信号经预训练说话人编码器获取语音特征嵌入向量，然后按通道维度直接拼接，随后使用 MLP 或 SVM 完成二分类（即属性强弱判断）。该类方法尽管能够完成基本的属性比较任务，但在训练阶段仅依赖交叉熵损失对末端分类器参数进行更新，存在以下关键问题：一是缺乏对上下文依赖关系的建模能力，难以捕捉音色属性的全局相关性；二是未能有效关注对判别任务更重要的音色属性特征；三是在处理多说话人音色属性比较时，特征融合过程过于简单，既未考虑单个特征内部的交互关系，也未建模不同特征之间的相互影响，这限制了模型在细粒度音色属性检测中的性能。

针对上述问题，引入 Transformer[14]建模特征全局依赖关系，受 KA 表示定理[15]启发，采用层级函数结构实现特征内外交互的信息融合，提高属性检测模块性能，主要贡献如下：

（1）在检测模块中引入 Transformer 结构替代传统 MLP 分类器，通过多头注意力机制增强上下文依赖建模能力，并实现对判别性特征的动态关注。

（2）提出 KA 表示定理启发的特征融合模块，人特征的内部交互关系，通过外层函数（OutterFunction）通过内层函数（InnerFunction）建模单个

收稿日期：2025-07-28
基金项目：湘江实验室重大项目（No.23XJ01003、No.23XJ01009）；湖南省教育厅科学研究重点项目（No.22A0441）；大学生创新创业训练计划（S202410554039）
作者简介：彭程（2006－），男，本科生。
通信作者：姜林，教授，E-mail: jlcdf@163.com

属性的融合表征。实验结果表明，所提方法在音色属性检测任务上显著优于现有基线模型。

1 本文问题

1.1 问题定义

本文语音音色属性检测任务属于比较型二分类问题。在测试集中的说话人已在训练集中出现的设定下，系统需判断给定属性下两个语音片段的音色强度差异。设输入为两个语音段O_A与O_B，属性维度为 v，输出为标签$y \in [0,1]$，其中$y=1$ 表示O_A的音色属性 v 强于O_B，否则为 0。任务需从音频$\{O_A, O_B\}$对中建模细粒度音色属性的强弱比较关系。强弱比较假设由算法函数$\mathcal{F}(\langle O_A, O_B\rangle|v;\theta)$确定，其中$\theta$是算法参数集，问题定义如图 1 所示。

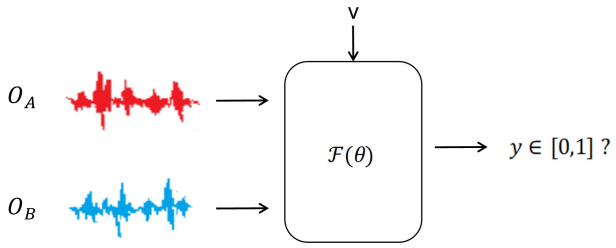

图 1 问题定义

1.2 模型架构

本文提出的属性检测模型如图 2 所示。包括三个模块：Speaker Encoder、KAfusion 和 Transformer。语音对$\langle O_A, O_B\rangle$通过预训练 Speaker Encoder 提取对应的语音嵌入向量$e_A \in \mathbb{R}^n$和$e_B \in \mathbb{R}^n$，n表示特征维度；将语音嵌入向量输入到核心模块 KAfusion，该模块对不同说话人特征进行交互建模，获取融合特征；最后将融合特征输入到 Transformer 模块进行全局建模，并突出重要特征。模型通过全连接层进行分类得到音色属性强弱类别。

本文提出的属性检测模型如图 2 所示。包括三个模块：Speaker Encoder、KAfusion 和 Transformer。语音对$\langle O_A, O_B\rangle$通过预训练 Speaker Encoder 提取对应的语音嵌入向量$e_A \in \mathbb{R}^n$和$e_B \in \mathbb{R}^n$，n表示特征维度；将语音嵌入向量输入到核心模块 KAfusion，该模块对不同说话人特征进行交互建模，获取融合特征；最后将融合特征输入到 Transformer 模块进行全局建模，并突出重要特征。模型通过全连接层进行分类得到音色属性强弱类别。

(1) **Speaker Encoder**

Speaker Encoder 部分采用冻结参数的 ECAPA-TDNN[16]预训练模型作为基础特征提取器，该模型能够捕捉语音中的说话人特征与 timbre 相关差异，其高质量嵌入为后续属性建模提供语义支撑。

（2）KAfusion

KAfusion 模块是本文的关键创新模块，其核心目标是通过内外层函数建模单特征的内部交互和不同特征的交叉交互。该模型具有结构可分解、属性可解释的优点，形成音色属性的中间表示，详细设计过程见 1.3 节。

（3）Transformer

Transformer 模块利用多头注意力机制，在语音特征表示空间中挖掘各音色属性之间的依赖关系与对比线索，强化整体的判别建模能力。

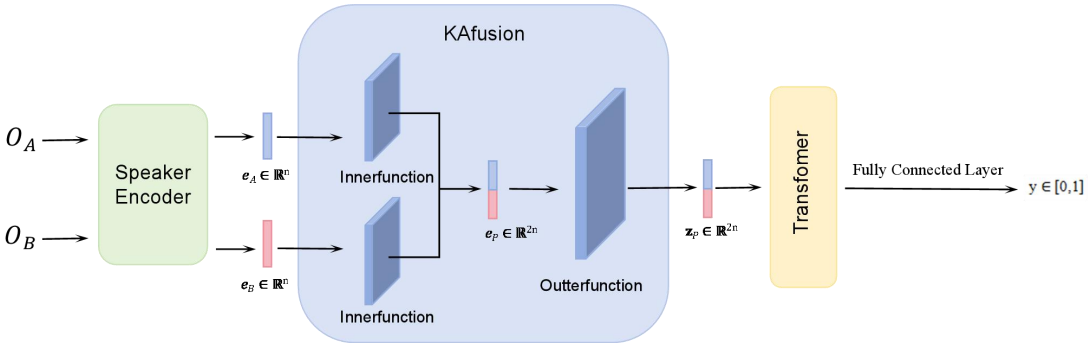

图 2 模型框架

1.3 KAfusion 模块

近年来，KA 表示定理的理论框架为特征融合网络设计提供了全新启发。它在动态卷积机制[17]，语音分割模型[18]等领域取得了全新进展，将 KA 表示定理与特征融合机制结合，有望突破现有方法在特征交互融合和模型适应性方面的瓶颈。KA 表示定义将高给函数分解为单变量函数组合，提示了

深度网络通过分层非线性变换与逼近复杂映射的本质能力，这与特征融合中层组级组合具有理论同构性。KA 表示定理严格地证明了：对于任意定义在紧致域上的多元连续函数$f(x)$，存在一组有限个一元连续函数ϕ_q和$\psi_{q,p}$，使得该函数可被分解为嵌套的两层函数结构，如下所示：

$$f(x_1, x_2, ..., x_n) = \sum_{q=1}^{Q} \phi_q \left(\sum_{p=1}^{n} \psi_{q,p}(x_p) \right). \quad (1)$$

其中$\psi_{q,p}(\cdot)$表示第 i 路径作用于第 j 维输入的可学习一元映射函数，$\Phi_q(\cdot)$表示路径级别的非线性聚合函数，该定理揭示了任意复杂映射函数在理论上的“可分解性”，即通过有限次一元非线性映射即可对其进行逼近[19]。受 KA 表示定理启发，本文将复杂的语音特征交互过程划分为两个阶段：个体内部属性特征交互与跨个体特征交互。具体而言，在语音嵌入中，内层函数$\psi_{q,p}(\cdot)$可类比为针对单个说话人语音特征的非线性映射，用于捕捉其内部属性之间的复杂非线性交互关系；而外层函数$\Phi_q(\cdot)$则负责将两个说话人处理后的拼接特征进行融合交互，提炼出在某一特定音色属性维度上的相对差异性表达。基于此启示，本文提出的 KAfusion 模块分内层非线性变换与外层组合交互映射两阶段实现语音特征交互建模过程，具体实现如图 3 所示。

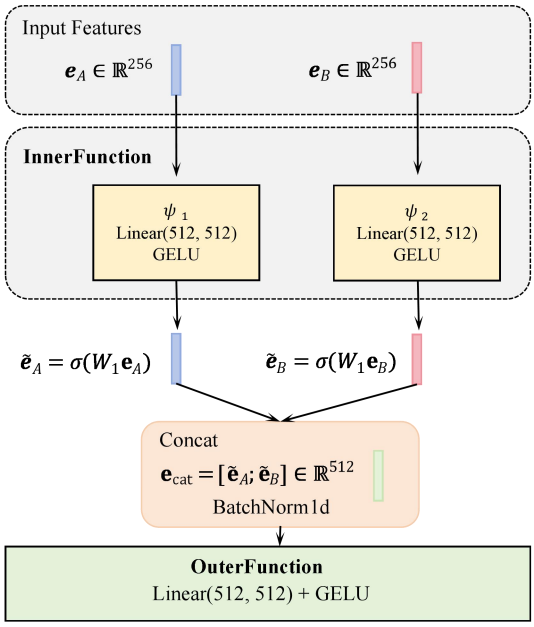

图 3 KAfusion 模块设计

在内层变换阶段，首先对两个说话人各自的语音嵌入特征$e_A \in \mathbb{R}^{256}, e_B \in \mathbb{R}^{256}$分别输入 Innerfunction 模块进行个体特征交互。该模块内部由全连接层 Linear(256,256)与 GELU 激活函数组成，用于模拟 KA 理论中的一元非线性函数$\psi(\cdot)$。该阶段实现了对单个说话人语音音色特征内部交互关系的建模，输出经过非线性映射的嵌入表示：

$$\tilde{e}_A = \sigma(W_1 e_A), \quad \tilde{e}_B = \sigma(W_1 e_B). \quad (2)$$

其中σ为 GELU 激活函数，W_1为内层线性变换权重。随后，将$\tilde{e}_A = \sigma(W_1 e_A)$与 $\tilde{e}_B = \sigma(W_1 e_B)$拼接成组合特征$e_{cat} = [\tilde{e}_A; \tilde{e}_B] \in \mathbb{R}^{512}$，并通过 BatchNorm 1d 对其进行归一化处理，提升模型训练的稳定性与特征表示的一致性。

在外层组合阶段，归一化后的拼接向量作为 Outterfunction 的输入，模拟$\phi(\cdot)$对两个说话人的音色属性特征进行融合交互建模。该模块同样由 Linear(512,512)与 GELU 激活构成，其作用是进一步提取联合语音对中属性感知维度上的深层交互特征，从而增强模型对于音色属性的判别能力。最终输出为$z_{KA} = \phi(e_{cat}) \in \mathbb{R}^{512}$，作为后续 Transformer 模块的输入。

整个 KAfusion 模块以非结构化、属性驱动的方式实现了语音对之间音色属性的显式交互表示，兼顾了特征间的非线性建模与表达紧凑性。相比于传统的直接拼接或线性投影方法，KAfusion 更贴合音色属性本身的复杂语义结构，有助于提升模型在多标签音色比较任务中的判别性能。

2 实验及数据集分析

2.1 数据集介绍

本研究采用 VCTK-RVA[20]数据集进行模型训练与评估。该数据集基于 VCTK 语音语料，新增了人工标注的声纹属性强度差异。每条标注由一对说话人在特定音色维度（如“明亮”“低沉”“磁性”等）上的主观强度比较组成，形式为有序对〈 SpeakerA, SpeakerB，u〉，其中 $u \in \{0,1\}$，u=1 表示 SpeakerB 在该属性上相较 SpeakerA 更为显著。数据集中共包含 101 名说话人，共构建 6038 对带音色标签的说话人配对，每对配对最多包含 1 至 3 个音色维度。

本文聚焦于可见测试任务（Seen Track），测试说话人全部来自训练集，但测试所用语音片段与训练无重合，且说话人配对亦为全新组合。该设置用于评估模型在已知说话人上的音色属性迁移与判别能力。

训练集中包含 29 名男性和 49 名女性说话人，分别标注了 17 个通用及性别特有的音色维度（如“干哑”仅用于男性，“尖锐”仅用于女性）。不同属性下说话人配对数量不均，例如“明亮”属性下女性训练集中配对数为 428 对，而“干净”属性下仅为 1

96 对。在测试集中，每个说话人配对随机选取各 20 个语音段，构成共 400 个语音段配对。

2.2 实验设置与评价指标

实验在 Ubuntu 22.04 操作系统下进行，硬件平台为 16GB Tesla V100 GPU。使用 Python 3.8 和 PyTorch 1.12.1 构建实验框架，训练共 30 个 epoch，batch size 设置为 128，学习率为 5e-4。损失函数采用二元交叉熵（BCE）。

模型评估指标为准确率（ACC）和等错误率（EER），并在各属性维度上取平均以衡量整体性能。计算公式如下：

$$\text{Avg ACC}(\%)=\frac{\sum_{i=1}^{N}\text{ACC}_i}{N}. \tag{2}$$

$$\text{Avg EER}=\frac{\sum_{i=1}^{N}\text{EER}_i}{N}. \tag{3}$$

其中公式(2),(3)中的 N 代表类别数。

2.3 对比实验

本文选择 ECAPA-TDNN[16]+MLP 和 FACodec[21]+MLP 作为基线模型，旨在对比其在音色属性迁移与判别任务中的表现。ECAPA-TDNN 作为一种经典的基于 TDNN 的模型，广泛应用于说话人验证任务[22][23][24]。它通过强调时间序列数据的局部特征建模能力，能够有效提取说话人的音色特征。FACodec 作为一种音频编解码器，采用因式分解的方式将语音信号分解为多个子空间，能够有效提取音色信息。本文选择这两种模型作为基线，是因为它们分别代表了不同的音频处理方法，且在音色属性任务中均有较好的效果。

表 1 性能对比

模型	Seen 测试集准确率	Seen 测试集错误率
ECAPA-TDNN+MLP	93.825%	5.94%
facodec+MLP	86.085%	13.405%
ECAPA-TDNN+KA+Transformer（本文方法）	97.955%	2.18%

实验结果表明（表 1），所提方法在测试集上取得了良好的性能。在 Seen 测试集上的准确率与错误率的表现明显优于传统的 ECAPA-TDNN+MLP 模型，相比 ECAPA-TDNN+MLP 方法，准确率与错误率分别提升了 4.13%和 3.76%，相比于 FACodec+MLP,准确率与错误率分别提升了 11.87%和 11.225%。

为了更加直观展示本文方法在模型上的贡献，本文绘制了如图 4-8 所示的柱状图，分别展现出测试集中不同性别的说话人音色识别的准确率和等错误率，不同性别的音色识别准确率，不同性别的音色识别等错误率。其中图 4 左侧是基于准确率的对比，右侧是基于等错误率的对比。图 5，图 6，图 7，图 8 分别代表女性不同音色特征识别准确率，男性不同音色特征识别准确率，女性不同音色特征识别等错误率，男性不同音色特征识别等错误率。

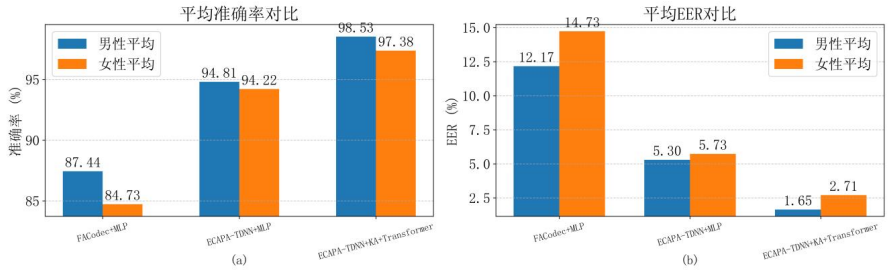

图 4 性别平均指标对比

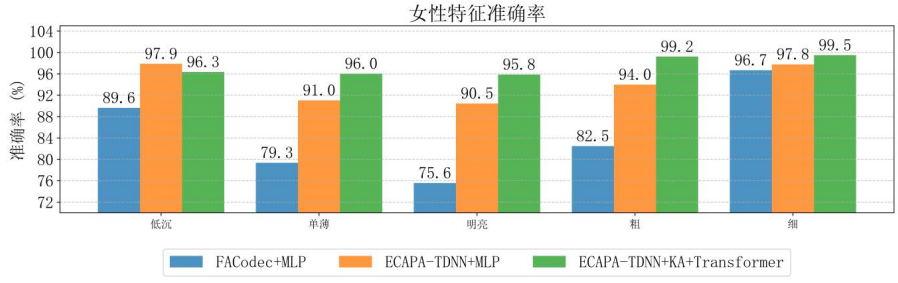

图 5 女性不同音色特征识别准确率对比

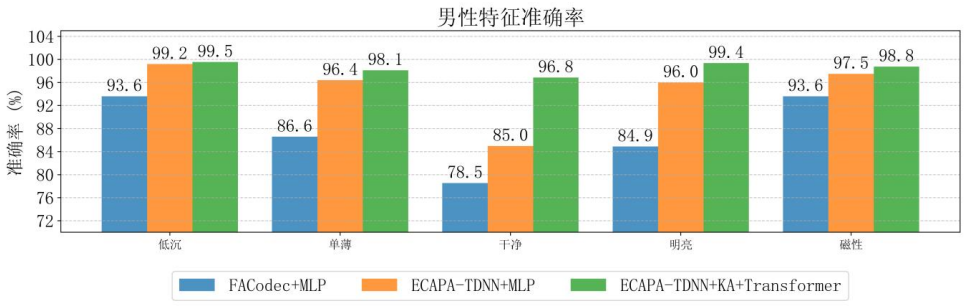

图 6 男性不同音色特征识别准确率对比

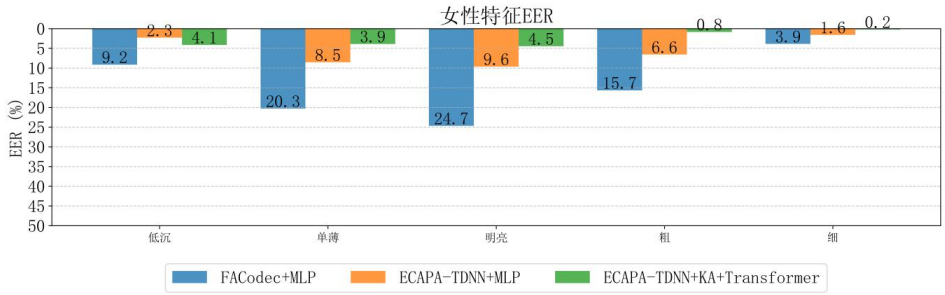

图 7 女性不同音色特征识别等错误率对比

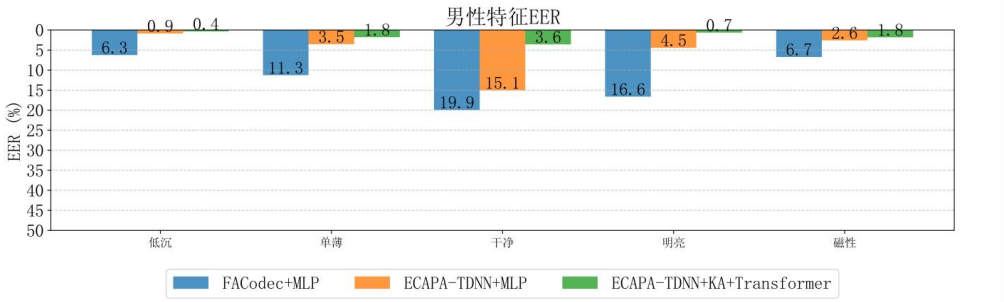

图 8 男性不同音色特征识别等错误率对比

由图 4-6 得出，在不同模型结构下性别识别任务的性能差异显著，且 ECAPA-TDNN 结合 KA 模块与 Transformer 结构（即 ECAPA-TDNN+KA+Transformer）整体表现最为优异。在性别平均指标对比图中，该模型在男女两个群体上都取得了最高的平均准确率（男性 98.53%，女性 97.38%）和最低的平均 EER（男性 1.65%，女性 2.71%），明显优于 FACodec+MLP 与 ECAPA-TDNN+MLP 的基线结构。进一步从特征层面的 EER 对比图可以看出，ECAPA-TDNN+KA+Transformer 在多种语音特征（如低沉、单薄、干净、明亮等）下都显著降低了错误识别率，特别是在 FACodec+MLP 存在显著性能瓶颈的特征上，该组合结构表现出更强的鲁棒性和适应性。而在特征准确率的对比中，ECAPA-TDNN+KA+Transformer 同样在大多数特征上均获得最高的准确率，例如女性低沉语音的识别准确率达到 96.3%，男性明亮语音准确率高达 99.4%，均远超其他模型结构。综合来看，融合 KA 模块与 Transformer 的注意力机制能够更有效地建模语音信号的复杂时序与特征分布，从而显著提升在性别识别任务中的泛化能力与判别性能。

2.4 消融实验

为全面验证所提出的 ECAPA-TDNN+KA+Transformer 模型各组件的有效性，本文设计了一系列的消融实验，分析每个组件对于整体性能的贡献。表 4 展示了消融实验的详细结果，包括准确率和错误率两个指标。

消融实验结果表明（表 2），基线模型 ECAPA-TDNN+MLP 已展现出较高的性能（准确率为 93.825%，错误率为 5.94%）。引入 KA 特征融合模块后，准确率提升 3.845%，错误率降低 3.565%，充分验证了该模块在增强特征交互与表示能力方面的有效性。进一步地，单独引入 Transformer 模块

后，准确率提升 2.75%，错误率降低 2.395%，表明 Transformer 在 ECAPA-TDNN 编码器中补充了对上下文信息的建模能力，有效缓解了原始模型在处理复杂上下文信息时的局限性。

表 2 不同模型组合的性能对比

模型组合	Seen 测试集准确率	Seen 测试集错误率	相比基线模型（ACC）	相比基线模型（EER）
ECAPA-TDNN+MLP	93.825%	5.94%	/	/
ECAPA-TDNN+KA+MLP	97.67%	2.375%	3.845%↑	-3.565%↑
ECAPA-TDNN+Transformer	96.575%	3.545%	2.75%↑	-2.395%↑
ECAPA-TDNN+KA+Transformer	97.955%	2.18%	4.13%↑	-3.76%↑
FACodec+MLP	86.085%	13.405%	/	/
FACodec+KA+MLP	88.87%	10.91%	2.785%↑	-2.495%↑
FACodec+Transformer	79.885%	18.76%	-6.2%↓	5.355%↓
FACodec+KA+Transformer	89.455%	10.8%	3.37%↑	-2.605%↑

此外，结合 KA 和 Transformer 模块后的 ECAPA-TDNN+KA+Transformer 模型，准确率进一步提升至 97.955%，错误率降至 2.18%，相比基线模型 ECAPA-TDNN+MLP，准确率提升了 4.13%，错误率下降了 3.76%。这表明 KA 与 Transformer 的联合使用能有效增强模型的表达能力，并在提高准确率的同时显著降低错误率，进一步验证了组合模块的优势。

针对 FACodec+Transformer 组合导致性能下降的现象，表明 FACodec 编码器提取的音色表示中缺乏有效的交互信息，难以激发 Transformer 注意力机制的优势。通过引入 KA 特征融合模块，能够挖掘并增强音色特征之间的交互关系，从而使 Transformer 得以利用其注意力机制，有效聚焦于与音色强度差异相关的关键特征区域，进一步发挥其上下文建模能力，提升整体模型性能。

对于 ECAPA-TDNN+Transformer 的性能提升，ECAPA-TDNN 本身具有处理时间序列数据的优势，它通过不同的卷积层捕捉音频信号的局部特征。然而，传统的 ECAPA-TDNN 模型存在局限性，因为它的时间上下文建模能力有限，通常只能考虑较小范围的时间步。结合 Transformer 后，利用其注意力机制来捕捉更长时间范围的依赖关系。Transformer 的全局上下文建模能力，使得模型能够学习到更丰富的时间特征，从而增强了对长时间依赖的建模能力[25]。

3 结论

针对现有语音音色属性检测模型在特征融合过程中存在的维度交互缺失与全局依赖建模薄弱问题，本文提出基于 KA 表示定理的特征融合模型，通过构建 KAfusion 模块实现对单个说话人特征内部交互和多说话人特征间交互关系的显式建模，并引入 Transformer 编码器有效捕捉音色属性的全局依赖关系。实验验证表明，KAfusion 模块相较传统特征融合策略表现出更优的结构表征能力，对比实验证实该模型在音色属性检测任务中优于主流基线方法，消融实验验证了特征融合与全局建模组件的协同有效性。总体而言，该方法通过特征交互的显式建模与全局依赖的协同优化，实现了音色属性检测性能的有效提升。然而，在多属性并行计算场景下仍存在效率优化空间，未来将聚焦轻量化架构设计与跨语种自适应学习策略，进一步提升模型在复杂声学环境中的鲁棒性。

参考文献（References）

[1] Liu C, Zhang J, Zhang T, et al. Detecting voice cloning attacks via timbre watermarking[J]. arXiv preprint arXiv: 2312.03410, 2023.

[2] Guo M, Wang J, Liu C, et al. Innovative Speaker-Adaptive Style Transfer VAE-WadaIN for Enhanced Voice Conversion in Intelligent Speech Processing[C]//2024 4th International Symposium on Computer Technology and Information Science (ISCTIS). IEEE, 2024: 919-925.

[3] Adagale S S, Gupta P, Sharma R P. Multiple Acoustic Feature-Based Speech Emotion Recognition for Sentiment Analysis[C]//2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS). IEEE, 2024: 1298-1303.

[4] Leu F Y, Lin G L. An MFCC-based speaker identifica

tion system[C]//2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA). IEEE, 2017: 1055-1062.

[5] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[M]//Backpropagation. Psychology Press, 2013: 35-61.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[7] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.

[8] Zarrouk E, Ben Ayed Y, Gargouri F. Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study[J]. International Journal of Speech Technology, 2014, 17(3): 223-233.

[9] Hu H, Zahorian S A. A neural network based nonlinear feature transformation for speech recognition[C]//INTERSPEECH. 2008: 1533-1536.

[10]Smith N, Gales M. Speech recognition using SVMs[J]. Advances in neural information processing systems, 2001, 14.

[11]Li Y, Wang Y, Yang X, et al. Speech emotion recognition based on Graph-LSTM neural network[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023 (1): 40.

[12]Wang D, Ding Y, Zhao Q, et al. ECAPA-TDNN Based Depression Detection from Clinical Speech[C]//Interspeech. 2022: 3333-3337.

[13]Hema C, Marquez F P G. Emotional speech recognition using cnn and deep learning techniques[J]. Applied Acoustics, 2023, 211: 109492.

[14]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[15]KOLMOGOROV A N. On the representation of continuous functions of several variables by superpositions of continuous functions of one variable and addition [J]. Doklady Akademii Nauk SSSR, 1957, 114(5): 953-956.

[16]Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification[J]. arXiv preprint arXiv:2005.07143, 2020.

[17]Chen Y, Dai X, Liu M, et al. Dynamic con- volution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020:11030-11039.

[18]Li Y, Song L, Chen Y, et al. Learning dyna- mic routing for semantic segmentation[C]//Pr- oceedings of the IEEE/CVF conference on c- omputer vision and pattern recognition. 2020:8553-8562.

[19]ARNOLD V I. On functions of three variables [J]. Doklady Akademii Nauk SSSR, 1957, 114(4): 679-681.

[20]Sheng Z Y, Liu L J, Ai Y, et al. Voice attribute editing with text prompt[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025.

[21]Ju Z, Wang Y, Shen K, et al. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models[C],International Conference on Machine Learning. PMLR, 2024: 22605-22623.

[22]Z. Zhao, Z. Li, W. Wang and P. Zhang, "PCF: ECAPA-TDNN with Progressive Channel Fusion for Speaker Verification," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5

[23]S. Han, Y. Ahn, K. Kang and J. W. Shin, "Short-Segment Speaker Verification Using ECAPA-TDNN with Multi-Resolution Encoder," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5

[24]J. Thienpondt and K. Demuynck, ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings,"2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 2023, pp. 1-8

[25]F. Wang, Z. Song, H. Jiang and B. Xu, "MACCIF-ECAPA-TDNN: Multi Aspect Aggregation of Channel and Context Interdependence Features in ECAPA-TDNN-Based Speaker Verification," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 2021, pp. 214-219

Speech timbre attribute detection model inspired by KA repre sentation theorem

PENG Cheng[1], JIANG Lin[1,2], PENG Bo[1], CHENG Ying Chao[1], DENG Sai Nan[1]

(1. School of Artificial Intelligence and Advanced Computing, Hunan University of Technology and Business, Changsha 410205, China;

2. Xiang Jiang Laboratory, Changsha 410205, China)

Abstract: For the existing voice timbre attribute detection models, the feature detection module mostly ad opts the "feature splicing shallow classifier" paradigm, which has the problems of lack of context-depend ent modeling, difficulty in focusing on discriminant features, and too simple feature fusion process. In thi s paper, a speech timbre attribute detection model inspired by the Kolmogorov-Arnold (KA) representatio n theorem is proposed. The method focuses on the optimization of the attribute detection module, introdu ces the Transformer structure to replace the traditional MLP classifier in the detection module, and uses its multi-head attention mechanism to enhance the context-dependent modeling ability and realize the dyn amic attention of discriminant features. At the same time, a KA-inspired feature fusion module (KAfusion) is proposed, which models the internal interaction relationship of a single speaker feature through the in ner function (InnerFunction), and the outer function (OutterFunction) captures the interaction relationship between multiple speaker features to realize the fusion representation of timbre attributes. Experimental re sults show that the proposed method is significantly better than the existing baseline model in the detecti on of timbre attributes.

Key words: speech timbre attribute detection; feature fusion; Kolmogorov-Arnold represents the theorem; Transformer model