《数学》

文章深度说话人嵌入架构中的残余信息

Adriana Stan

克卢日-纳波卡技术大学通信系，400114，克卢日-纳波卡，罗马尼亚；adriana.stan@com.utcluj.ro

摘要： 说话人嵌入旨在从语音信号中提取具有代表性的向量表示，使该表征仅与说话人身份相关。说话人嵌入通常用于对不同说话人进行分类和区分。然而，目前尚无客观指标能够评估说话人嵌入将说话人身份与其他语音特征解耦的能力。这意味着说话人嵌入并不理想，高度依赖训练语料库，并且仍包含与语言内容、录音条件或话语说话风格等因素相关的一定程度的残余信息。本文对使用近期一些性能优异的深度神经网络（DNN）架构提取的六组说话人嵌入进行了分析，尤其考察了这些嵌入在多大程度上能够真正将说话人身份与语音信号解耦。为正确评估这些架构，本文使用了一个大型多说话人平行语音数据集。该数据集包含46名说话人，他们朗读相同的一组提示语，录音地点为专业录音棚或各自的家庭环境。分析考察了在不同嵌入集合上计算的说话人内及说话人间相似度指标，并分析简单的分类和回归方法能否从说话人嵌入中提取多种残余信息因素。结果表明，所分析嵌入的判别能力非常强，但在所有分析的架构中，表征中仍存在残余信息，表现为其与录音条件、语言内容和话语时长高度相关。然而，我们表明，这种相关性虽不理想，但在下游任务中仍可能具有实用价值。说话人嵌入的低维投影在说话人内数据聚类和话语异常值检测方面表现出相似的行为模式。

arXiv:2302.02742v1 [eess.AS] 6 Feb 2023

check for updates

C引文： Stan, A. 残余I信息：深度说话人E嵌入架构。*M*数学**2022**，*1*，0。

https://doi.org/

A学术编辑：陈博豪

R收稿日期：2022年9月16日

A接受日期：2022年10月19日

P发表日期：

P

出版者说明： MDPI 对已发表地图中的管辖权主张及机构隶属关系保持中立。

版权： © 2022 作者所有。许可方为MDPI，bas，瑞士。本文是根据知识共享署名（CC BY）许可条款和条件发布的开放获取文章（https://creativecommons.org/licenses/by/ 4.0/）。

关键词： 说话人嵌入；X向量；深度表征；深度嵌入；说话人解缠；说话人识别；残余信息；神经网络架构；深度学习；人工智能

数学学科分类： 68T01；68T07

1. 引言

录制的语音本质上是一种复杂的信号，其中包含与语言内容、韵律或风格因素（如节奏和语调）以及说话人特征（如生理特征、性别、族裔或社会背景）相关的信息。人类能够将几乎所有这些因素分离出来并提取其抽象表征，也能够在来自不同来源的口语数据中复现并识别相似的模式。语音信号中一个具有众多下游应用的重要组成部分，便与说话人身份的分离有关。对说话人身份进行如此准确的表征，将使其在说话人识别与身份验证、文本转语音合成和语音克隆等任务中极具价值 [1],匿名化或生成新的、未曾出现过的说话人身份 [2]。目前已有大量已发表的研究专注于说话人区分，即判断两个或更多声学

信号是否属于同一说话人身份。然而，就为生成过程（如文本转语音或语音克隆）准确表征说话人而言，迄今尚无已发表的方法能够准确地只表征说话人身份，同时忽略声学信号的其他相关因素，例如录音条件或语言内容，尽管许多此类应用会将得到的表征用作输入或条件。因此，有必要分析当前的说话人表征对说话人身份的建模效果。

因此，本文的重点是 выяс明各种深度说话人嵌入中包含多少残余信息 （即与说话人身份无关的信息）。研究选取了六种开源、易于使用且 readily available 的实现。这些实现针对说话人分类和说话人日志任务报告了当前最先进的结果。本研究开展的第一项评估旨在直接比较这些架构在其预期用途（即说话人区分）方面的性能。在一个大型多说话人平行数据集上，计算了等错误率（EER）以及说话人间和说话人内相似度度量。在评估的第二步中，根据其中包含的残余信息量，对这些架构生成的说话人嵌入进行了分析，包括话语时长、信噪比、语言内容和录音条件等。研究采用简单的分类和回归算法，尝试从这些表征中提取相关信息。结果表明，在很大程度上，这些嵌入高度依赖于上述因素，因此说话人身份并未真正实现解耦。在这些初步结果的基础上，我们进一步探究了这些残余信息是否仍可用于下游任务。研究将生成的说话人嵌入绘制为低维表示，以验证它们在聚类不同录音会话和背景条件、以及区分话语离群点和异常说话人方面是否表现出相似模式。这类信息可用于为文本转语音合成系统或多说话人系统中的数据增强过程选择合适的说话人和一组样本。

本文的**贡献**可总结如下：

- 直接比较了六种最新的说话人嵌入深度神经网络的判别特性和生成特性；
- 分析基于一个平行数据集展开，该数据集包含 46 名说话人朗读相同的提示语；
- 评估了六种架构的等错误率（EER）以及说话人间和说话人内相似度度量；
- 采用决策树和轻量级梯度提升机算法，评估嵌入中所含残余信息的数量；
- 从离群点检测和说话人内数据聚类的角度，评估了六种架构的嵌入空间的低维 t-SNE 表征。

本文结构安排如下：第2 节介绍了以往一些关于开发高精度说话人嵌入的研究，以及这些嵌入在语音克隆和文本转语音合成系统中的应用。第3 节描述了本研究采用的音频数据和说话人嵌入架构。评估结果见第4节，结论与讨论则在第5节中介绍。

2. 相关工作

长期以来，说话人识别一直是研究界关注的重点，因为它对于身份识别、身份验证和说话人日志任务至关重要。说话人身份识别是指从一组预定义说话人中确定某段语音话语的身份。说话人验证旨在预测两个话语是否属于同一位说话人，而说话人分离则旨在分离较长

音频片段中出现的不同说话人身份，并将每个音频片段分配给相应的说话人。尽管说话人识别的这三个子任务在原理上似乎有所不同，但它们都包含一个共同环节：提取能够准确刻画各个说话人身份的数值表征。

早期的一些说话人识别方法基于频谱分析和模板匹配，通常使用梅尔频率倒谱系数 [3,4], 线性预测系数（LPC） [5,6] 或感知线性预测（PLP）系数 [7,8]。从 20 世纪 90 年代开始，高斯混合模型（GMM） [9,10] 逐渐得到广泛应用。这些模型针对数据集中的每位说话人，估计各种语音信号表示的统计特征。通过计算目标说话人与高斯混合模型集合中各概率分布函数之间的距离，进行身份验证或识别。随着通用背景模型（UBM） [11] 和支持向量机（SVM） [12], 的引入，说话人识别方法的性能持续提升。然而，评估通常是在规模较小、经过筛选且干净的数据集上进行，而且往往是在文本相关场景中进行。随后，人们应用降维技术，以提取感兴趣说话人之间差异最大的变化轴。在这一领域，主成分分析（PCA） [13] 和 i向量 [14]迅速受到青睐。

与许多其他应用领域一样，说话人识别方面的重大进展源于深度学习架构的引入。这类架构能够从大量口语数据中受益，抽象出语音信号所包含的信息。迈向基于深度神经网络的表示的第一步，只是用深度架构的后验概率取代基于高斯混合模型的后验概率 [15]。与传统说话人识别模型中的主成分分析技术类似，基于深度神经网络的瓶颈特征逐渐流行起来 [16], ，并被称为 d向量，在帧级别提取。d向量属于更广泛的一类基于深度神经网络的表示，即嵌入。X向量也是嵌入，由时延神经网络（TDNN）在片段级别提取 [17–19] ，并已成为说话人识别应用的标准方法。说话人识别中采用的其他深度架构还包括 RawNet [20,21] 和 ResNet [22–24]。这类嵌入以端到端方式提取，也就是说，网络既负责找到合适的表征，也负责确定与说话人相关任务有关的最终决策。以往的方法采用概率线性判别分析或余弦相似度来估计输出表征之间的相似性或差异性。近期一些研究甚至尝试将其他基于语音的神经表征应用于说话人识别 [25]。关于基于深度学习的说话人嵌入表示的扩展综述见 [26,27]。

随着越来越多的方法陆续发表，人们需要一个通用评估基准，以便准确比较各方法的结果。为此，组织了多场说话人识别研讨会和挑战赛，例如 NIST说话人识别评测（https://www.nist.gov/itl/iad/mig/speaker-recognition，访问日期：2022年10月22日）、Odyssey（http://www.odyssey2022.org/，访问日期：2022年10月22日）以及 VoxSRC。（https://www.robots.ox.ac.uk/~vgg/data/voxceleb/interspeech2022.html，访问日期：2022年10月22日）。在 2022 年 VoxSRC 挑战赛中，设有两个与说话人验证（开放集和封闭集）及说话人分离相关的主要赛道。表现最佳的系统采用了 ResNet 和 ECAPA-TDNN 架构，并结合了通过自监督学习（SSL）获得的音频信号表征 [28–31]。

尽管上述方法仅针对说话人识别、验证和说话人日志应用，但其研究发现也可应用于其他语音相关任务。其中最重要且应用最广泛的任务之一是语音合成和语音克隆。从使用大量说话人训练的网络中提取的说话人嵌入，可用于对多说话人合成模型进行条件控制。借助外部学习得到的嵌入，模型能够对未见过的说话人进行快速或零样本自适应 [32–36]。而

正是对于这些任务，更深入地分析嵌入准确性才尤为重要。

3. 实验设置

3.1. 语音数据

说话人嵌入评估中一个棘手的问题是，不同说话人的口语数据可能存在差异。这意味着每位说话人的话语子集的数量和语言内容可能会影响结果。因此，本研究使用了现有规模最大的平行口语数据集之一。该数据集是 SWARA 语料库的扩展版本 [37]。语料库的初始版本（下文称为 SWARA1.0）包含在专业录音棚录制的 18 位说话人。每位说话人朗读了 921 至 1493 条话语。该数据集最近新增了 28 位说话人，我们将这一子集称为 SWARA2.0。然而，由于2019冠状病毒病疫情，录音是在说话人的家庭环境中使用半专业设备完成的。说话人朗读了 1597 至 1797 条话语。由于 SWARA2.0 是在家庭环境下录制的，我们预计背景噪声和混响会影响从该数据集中提取的嵌入的表现。

在 SWARA1.0 和 SWARA2.0 子集中，说话人收到的朗读文本提示相同。然而，由于缺乏控制，尤其是在 SWARA2.0 场景中，46 位说话人之间真正平行的话语只有 712 条。这意味着其余话语中，说话人要么相对于提示语进行了删除、插入或替换，要么根本没有录制某些提示语。合并后的数据集中共有 24 位女性说话人和 22 位男性说话人，包含 32,752 条话语，总时长为 38 小时 29 分钟。所有数据均重采样至 16 kHz，并裁去了开头和结尾的静音片段。

3.2. 说话人嵌入网络

许多研究致力于提取用于表征说话人特征的深度学习表征。当然，这些研究大多旨在区分不同说话人，并准确完成识别和说话人日志任务。在我们的评估中，我们选取了基于DNN的架构，这些架构均为开源、易于获取和使用，并且提供优质的预训练模型。我们选择了以下架构：

(1) **Pyannote** (https://github.com/pyannote/pyannote-audio，访问日期：2022年10月22日）是一个以 Python 编写的开源说话人分离工具包 [38,39]。它采用 SincNet [40] 架构，后接一系列时延神经网络层和一个平均池化层。它还包含 Speech Brain 和 NeMo Titanet 架构的实现，但本研究中并未使用这些架构。

(2) **Speech Brain 的** [41] 说话人身份验证网络。（https://github.com/speechbrain/speechbrain，访问日期：2022年10月22日）基于 ECAPA-TDNN [42] 模型。该模型包含一系列卷积块和残差块，并使用加性间隔Softmax损失作为训练目标函数。说话人嵌入通过注意力统计池化生成。

(3) **Clova AI** [22,23] 采用类 ResNet 架构。，（https://github.com/clovaai/voxceleb_训练器，访问日期：2022年10月22日）同样会对帧级表示取平均，得到最终嵌入。区别在于目标函数的变化以及训练数据增强的使用。

(4) **NeMo Titanet** (所有 NeMo 架构的代码库可在此处获取：. https://github.com/NVIDIA/NeMo/，访问日期：2022年10月22日）采用 Titanet 架构 [43]，该架构基于 ContextNet [44]。该模型使用带有压缩与激励层的一维深度可分离卷积。输出嵌入通过对中间变长表示的统计量取平均获得。

(5) **NeMo SpeakerNet** 使用自动语音识别架构的编码器，即 QuartzNet [45]，，作为高级特征提取器，并在池化层中对这些特征取平均，以捕获与时间无关的说话人特征。

(6) **NeMo ECAPA-TDNN** 与 Speech Brain 架构相似，并采用 ECAPA-TDNN 结构 [46]。不同之处在于，NeMo 的实现没有使用残差块，而是使用膨胀率为 1 的组卷积块。

由于我们希望在语音合成系统中使用这些嵌入，因此没有对现有的预训练模型进行微调。微调意味着每当使用新的数据集进行合成时，都需要重新训练嵌入网络，这并不可行。因此，我们研究这些架构的原始表现，即按照作者提供给广大研究界的方式使用它们。

4. 评估

我们的评估基于一系列分析；我们认为，这些分析对于在说话人识别或判别应用之外的下游任务中使用说话人嵌入具有重要意义。以下小节将介绍各评估场景的结果。

4.1. 等错误率、说话人内与说话人间相似度

由于说话人嵌入源自旨在进行说话人识别的神经网络架构，因此属于面向判别的任务，其性能通常以等错误率（EER）进行评估。EER 定义为 ROC 曲线上虚警率（FAR）等于漏检率（FRR）的点。计算 EER 的阈值通常基于说话人嵌入对之间的余弦相似度，其定义如下：

$$\cos(\mathbf{e}_1, \mathbf{e}_2) = \frac{\mathbf{e}_1 \mathbf{e}_2}{\|\mathbf{e}_1\| \|\mathbf{e}_2\|} = \frac{\sum_{i=1}^{n} \mathbf{e}_{1,i} \mathbf{e}_{2,i}}{\sqrt{\sum_{i=1}^{n} (\mathbf{e}_{1,i})^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_{2,i})^2}} \tag{1}$$

由于我们使用的是一个新的语音数据集，因此首先评估所选架构的目标性能至关重要。因此，我们分析的第一步是计算各架构、说话人及说话人集合的等错误率结果。我们随机选取了 46,000 对话语。选取这些随机配对时，确保每位说话人出现在 1,000 对中，其中 200 对为同一说话人的配对。等错误率结果如表 1所示。我们对全部 46,000 对进行评估，并分别评估仅包含女性说话人、男性说话人、SWARA1.0 子集中的说话人以及 SWARA2.0 子集中的说话人样本的配对。可以看出，所有架构的等错误率均低于 1%，且女性说话人比男性说话人更容易区分。SWARA1.0 中的说话人与 SWARA2.0 中的说话人相比，情况也相同。NeMo Titanet 架构取得了最佳结果。

表1. 不同说话人嵌入架构和说话人子集的 EER 值。箭头表示最佳性能的方向。最佳结果以粗体突出显示。

架构	All↓	女性↓	男性↓	SWARA1.0↓	SWARA2.0↓
Pyannote	0.040	0.055	0.039	0.024	0.047
Speech Brain	0.025	0.027	0.031	0.011	0.031
Clova AI	0.055	0.060	0.081	0.031	0.073
NeMo Titanet	**0.018**	**0.014**	**0.027**	**0.005**	**0.024**
NeMo SpeakerNet	0.039	0.045	0.051	0.024	0.048
NeMo ECAPA-TDNN	0.032	0.035	0.041	0.023	0.038

对于每位说话人，NeMo Titanet 也是性能最佳的架构（见图 1）。不过，说话人bvl、mgl 和 pbl 的等错误率显著高于其他说话人。分析这些说话人的录音时，我们注意到，他们的背景条件确实比其他说话人差得多，而且在某些情况下，波形分段发生在话语结束之后或开始之前。

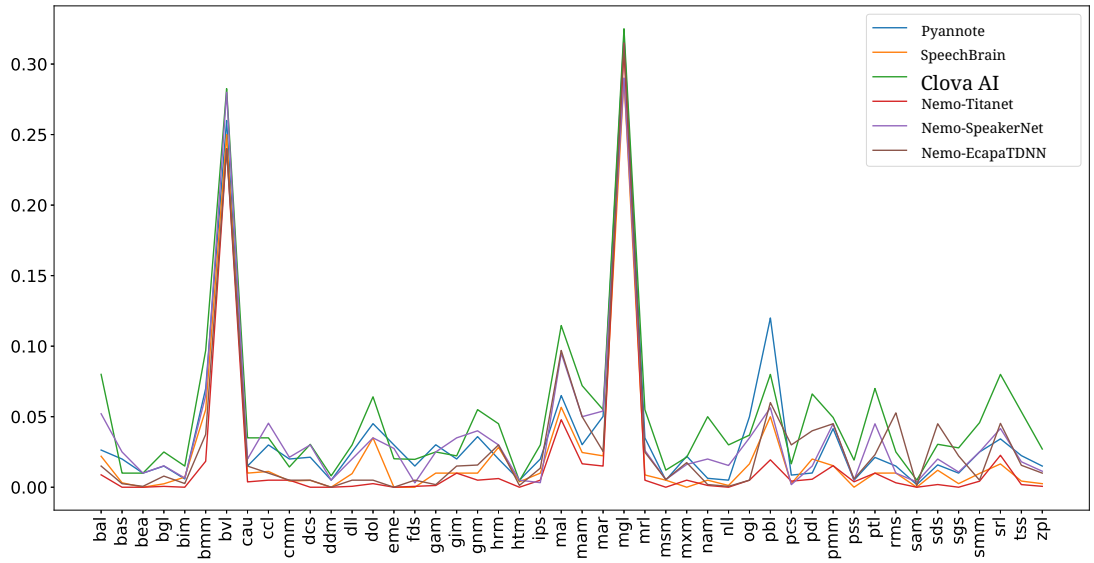


图 1. 各嵌入架构下不同说话人的等错误率。

等错误率给出了使虚警率等于漏检率的最佳阈值，但并未详细说明表征的准确率。对于下游任务，说话人间和说话人内的相似度度量会更有参考价值。这意味着，说话人之间的区分能力可以体现为较高的说话人内相似度和较低的说话人间相似度，而中间值理想情况下应代表感知上相似的说话人。图2展示了说话人内相似度值。相似度是通过计算同一说话人的所有话语对的相似度，并取其平均值而得出的。可以看出，Clova AI 架构的说话人内相似度最高，而说话人bvl、mgl和pbl 的说话人内相似度仍然最低。表现最差的架构是 Pyannote。表2中，我们对上述分数进行了系统级平均，并按不同的说话人子集分别计算平均值。Clova AI、女性和 SWARA1.0 说话人的说话人内相似度最高。

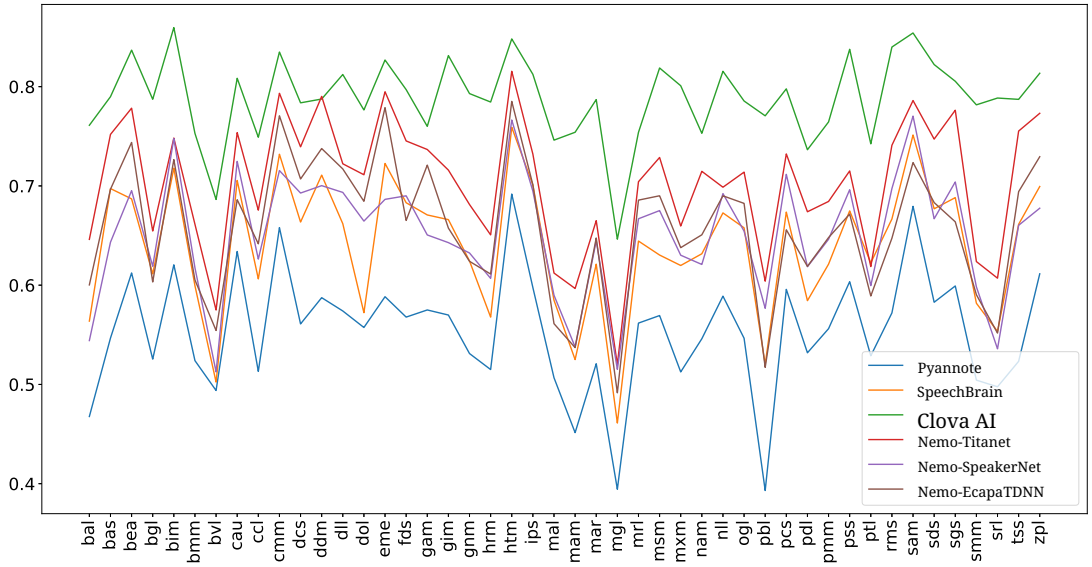


图2. 各说话人及嵌入架构的说话人内余弦相似度。

表2. 不同说话人嵌入架构及不同说话人子集的平均**说话人内** 余弦相似度。箭头表示最佳表现的方向。最佳结果以粗体标出。

架构	All↑	女性↑	男性↑	Swara1.0↑	Swara2.0↑
Pyannote	0.554	0.557	0.550	0.589	0.531
Speech Brain	0.640	0.651	0.629	0.686	0.610
Clova AI	**0.788**	**0.790**	**0.786**	**0.810**	**0.774**
NeMo Titanet	0.702	0.711	0.695	0.750	0.672
NeMo SpeakerNet	0.651	0.658	0.644	0.693	0.623
NeMo ECAPA-TDNN	0.658	0.670	0.647	0.696	0.635

对于说话人间相似度，我们使用了用于评估等错误率的 46,000 对随机话语，并仅选取说话人身份不同的话语对。随后，我们计算每对说话人之间余弦相似度的平均值。在这种情况下，我们希望各架构的相似度值尽可能低，从而最大限度地发挥表征的判别特征。表3中，我们列出了这些结果，并按架构和说话人子集分别取平均。尽管 NeMo Titanet 在之前的任务中表现似乎最佳，但就判别能力而言，NeMo ECAPA-TDNN 的效率最高（男性说话人之间的情况除外，此时 Pyannote 表现最佳）。图3展示了 NeMo ECAPA-TDNN 架构的说话人间相似度矩阵。根据这些分数，最相似的说话人是`htm` 和`mar`，相似度为 0.42；其次是`cmm` 和`cau` ，相似度为 0.40。这些配对均为女性说话人，其声音在感知上确实相似。

本节中，我们考察了评估说话人嵌入架构的常见指标，同时致力于提取可能对下游任务有用的额外信息。例如，在数据有限的情况下，可以利用说话人间相似度来训练语音合成系统，并通过使用另一位声音相似的说话人的语音进行数据增强。

表3. 不同说话人嵌入架构和不同说话人子集之间的平均**说话人间**余弦相似度。箭头标示最佳表现的方向。最佳结果以粗体突出显示。

架构	All↓	女性↓	男性↓	Swara1.0↓	Swara2.0↓
Pyannote	0.127	0.195	**0.129**	0.139	0.127
Speech Brain	0.122	0.188	0.143	0.139	0.118
Clova AI	0.133	0.188	0.157	0.142	0.135
NeMo Titanet	0.132	0.187	0.156	0.143	0.135
NeMo SpeakerNet	0.195	0.272	0.226	0.199	0.198
NeMo ECAPA-TDNN	**0.107**	**0.151**	0.140	**0.136**	**0.110**

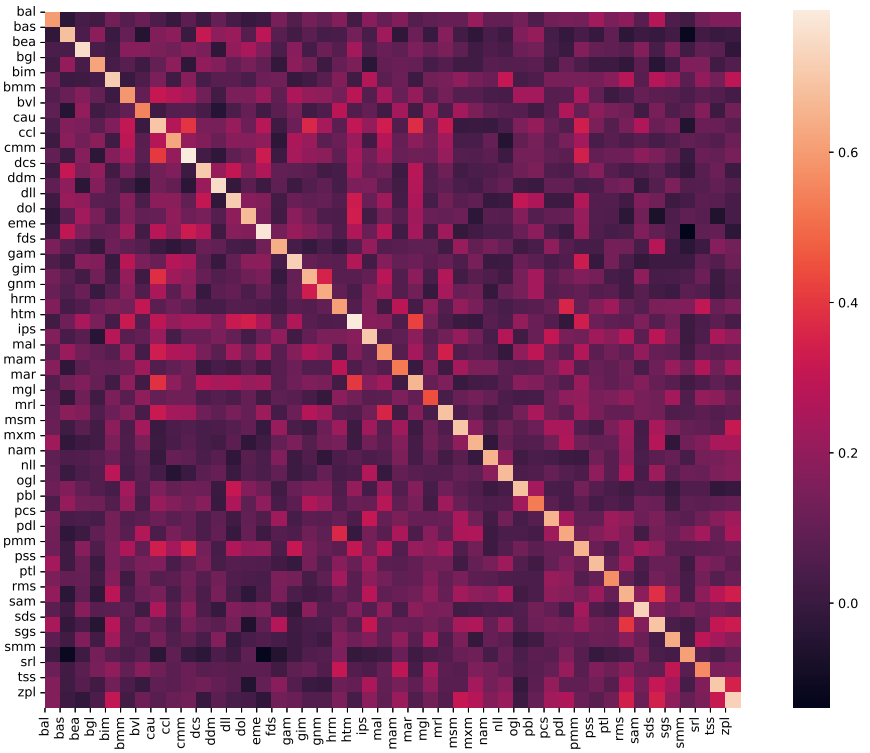


图3. NeMo ECAPA-TDNN 架构的说话人间相似度矩阵。

4.2. 说话人身份解耦

前一组结果显然旨在为说话人分类任务提供最佳表征。然而，对于说话人嵌入的许多下游应用而言，这还不够。语音合成系统尤其如此：在这类系统中，嵌入作为额外输入，而与语言内容和韵律模式相关的信息才是主要输入。因此，如果嵌入中包含与语音其他方面相关的信息，就可能在训练过程中产生不良影响和偏差。基于上述论述，本节将探讨说话人嵌入中存在多少残余信息，以及说话人身份是否真正与其他所有语音因素解耦。

判断是否存在残余信息的一种简单方法，是查看简单的机器学习算法能否从嵌入中提取这些信息。我们根据任务采用两种不同的算法：用于分类任务的决策树

任务和用于回归任务的 LightGBM [47] 。选择这两种算法，是因为它们属于最简单但功能强大的传统机器学习方法，且其结果在不同任务和数据集之间具有可比性。所有任务均使用相同的参数，并将语音数据集随机划分为 80% 的训练集和 20% 的测试集。

表 4 显示了说话人性别和说话人身份分类的 F1分数结果。对于这两个目标，预期结果应较高，因为嵌入应包含这些信息。表中显示的另外两个目标是文本长度（以字符数表示）和录音条件。录音条件被编码为针对两个子集 SWARA1.0 和 SWARA2.0 的二分类。对于这两列，预测准确率应有限，因为这两个目标的信息不应能轻易从嵌入中提取。文本长度的结果符合预期，无法从嵌入中提取话语中的字符数。然而，录音条件列中出现了一个有趣的结果：预测准确率相当高。这意味着，家庭录音中存在的额外频谱伪影确实会影响所有嵌入架构生成的最终嵌入。

表4. 各种分类任务的F1分数。箭头表示最佳性能的方向。最佳结果以粗体突出显示。

架构	说话人身份↑	性别↑	话语时长↓	录音条件↓
Pyannote	0.76	0.94	0.016	**0.87**
Speech Brain	0.84	0.96	0.011	0.90
Clova AI	0.85	0.98	0.015	0.92
NeMo Titanet	**0.90**	**0.97**	**0.010**	0.95
NeMo SpeakerNet	0.85	0.96	0.014	0.91
NeMo ECAPA-TDNN	0.87	0.96	0.015	0.95

对于回归任务，我们考察话语时长（以秒为单位）、信噪比（SNR）和语言内容。SNR 使用 WADA [48] 算法计算。对于文本内容，我们只考察话语 ID，并假设相同语言内容的说话人嵌入会具有相似特征。LightGBM 的准确率通过预测值与目标值之间的斯皮尔曼等级相关系数（SRCC）来衡量。表 5 展示了结果，但结果并不理想。所有架构与评估目标之间都表现出较高的相关系数（>0.7）。这意味着这类信息并未从最终的说话人嵌入中得到有效解耦。

另一个单独的回归任务考察了对话语级别平均基频的预测（表中最后一列5）。相关性相当高，六种架构的平均均方误差为 12 Hz。

表5. 基于 LightGBM 的各类回归任务 SRCC 结果。箭头标示最佳性能的方向。最佳结果以粗体突出显示。

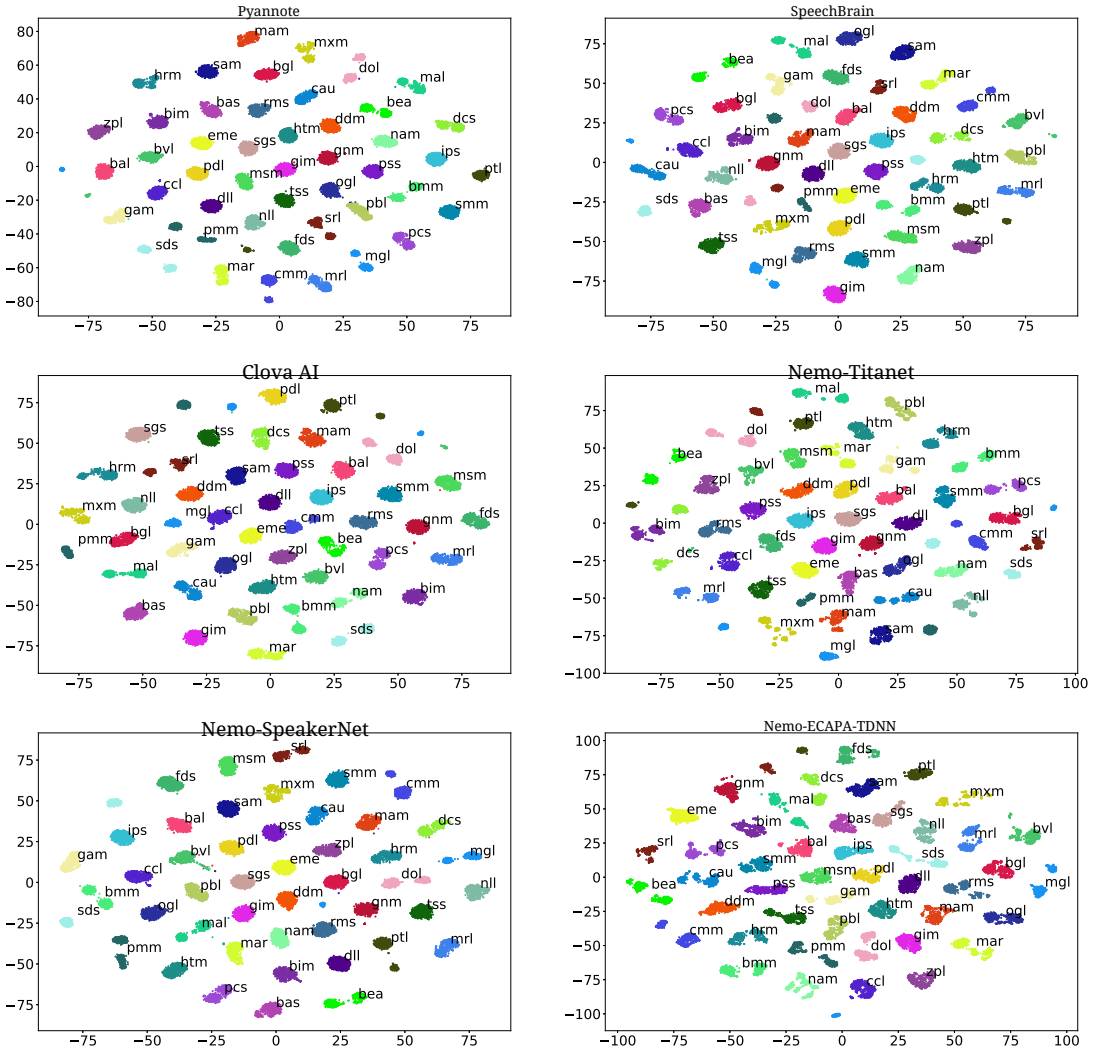
架构	话语时长↓	SNR↓	语言内容↓	F0↑
Pyannote	0.830	0.771	0.723	0.958
Speech Brain	0.734	0.749	0.742	0.959
Clova AI	0.750	0.791	0.734	**0.976**
NeMo Titanet	**0.704**	**0.747**	0.798	0.964
NeMo SpeakerNet	0.796	0.758	**0.709**	0.958
NeMo ECAPA-TDNN	0.862	0.787	0.775	0.962

本节结果表明，尽管嵌入架构的任务是尽可能准确地表征说话人身份，但其中仍然存在残余信息，因此还需要开展更多工作，以找到更适合下游任务的表征。在下一节中，我们将探讨如何通过可视化嵌入的表示，利用这些残余信息来助益其他任务。

4.3. 可视化表示

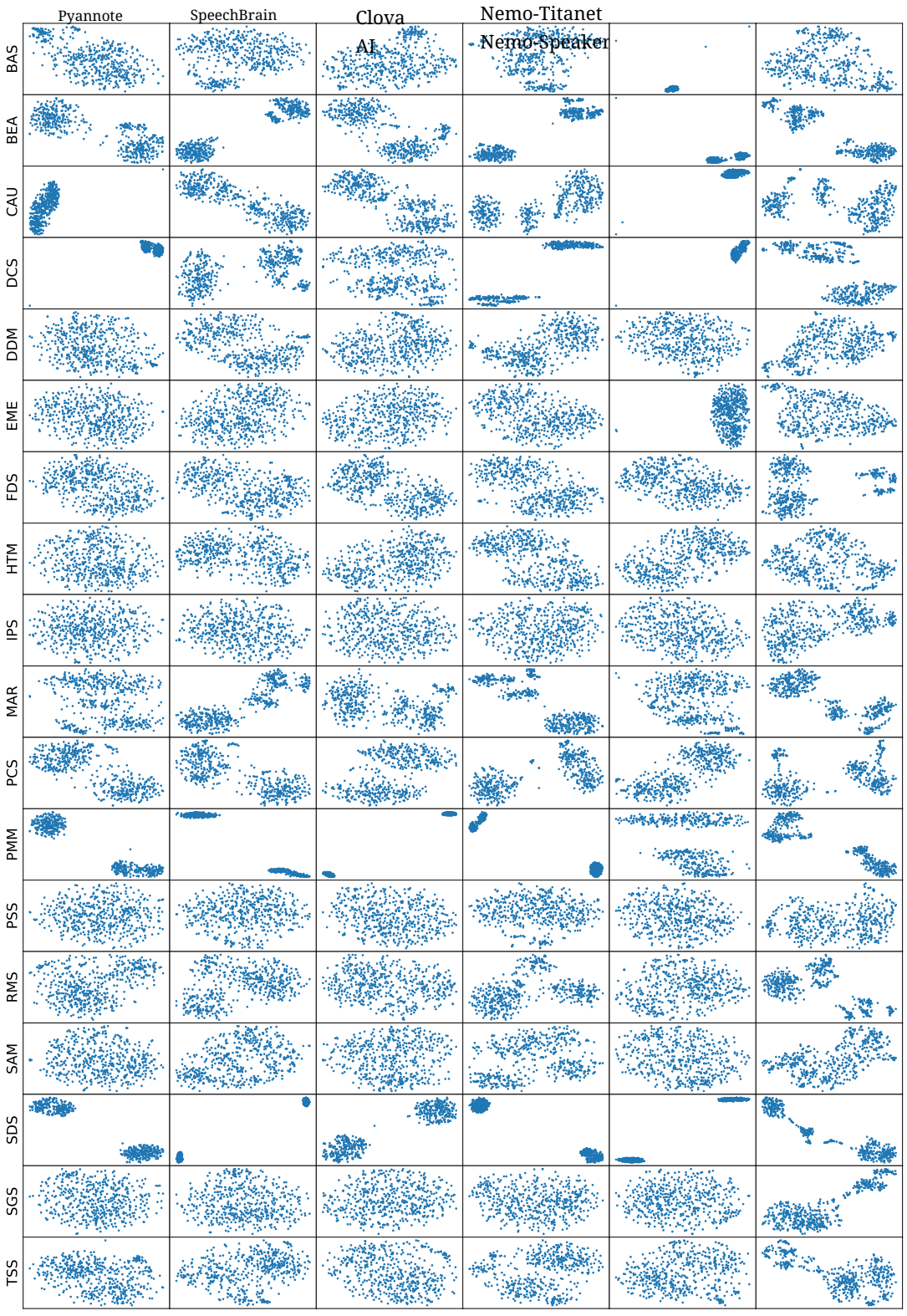
对高维数据进行可视化检查并不可行。然而，在大多数情况下，相比单纯的百分比和数字，数据的可视化表示能为开发者提供更多信息。因此，我们采用了一种 t-SNE 降维技术 [49]，并将从六种架构中获得的说话人嵌入绘制在二维空间中。该算法应用于每种架构各自的全部嵌入，运行了 1000 步，并将困惑度设为 30。图 4 展示了各架构对应的 t-SNE 图。可以看出，除少数例外，说话人都聚类得较为明显。此外，似乎没有任何一种说话人嵌入架构在说话人分组方面表现更佳。部分说话人的嵌入似乎形成了子簇，少数情况下还能观察到离群点。通过检查单独的语音样本，我们发现离群点通常对应短语音，这进一步证实了此前关于嵌入与音频时长之间相关性的结果。对于这些子簇，大多数说话人至少通过两次录音会话录制了全部提示语。不同会话之间的背景噪声、与麦克风的距离、说话节奏和发声力度各不相同。形成的 t-SNE 簇确实与不同的录音会话相吻合。可以看出，这些簇在包含家庭录音的 SWARA2.0 子集中更为常见。在不同的嵌入架构中，所有架构似乎都能够检测到这些子簇，而 NeMo SpeakerNet 和 Pyannote 似乎更易受音频时长影响，且比其他系统出现更多离群点。这意味着，这两种架构更适合用于对说话人数据集进行预处理，以供下游应用使用。

然而，当我们放大查看这些低维可视化时，会发现一些有趣的模式。图 5 和 6分别展示了两个数据子集（即 SWARA1.0 和 SWARA2.0）的说话人级 t-SNE 表征。



Figu图 4. 从不同架构中提取的说话人嵌入的 t-SNE 表征

s.



Fi图 5. 不同嵌入架构的说话人级 t-SNE 图—**SWARA1.0** 子集 t.

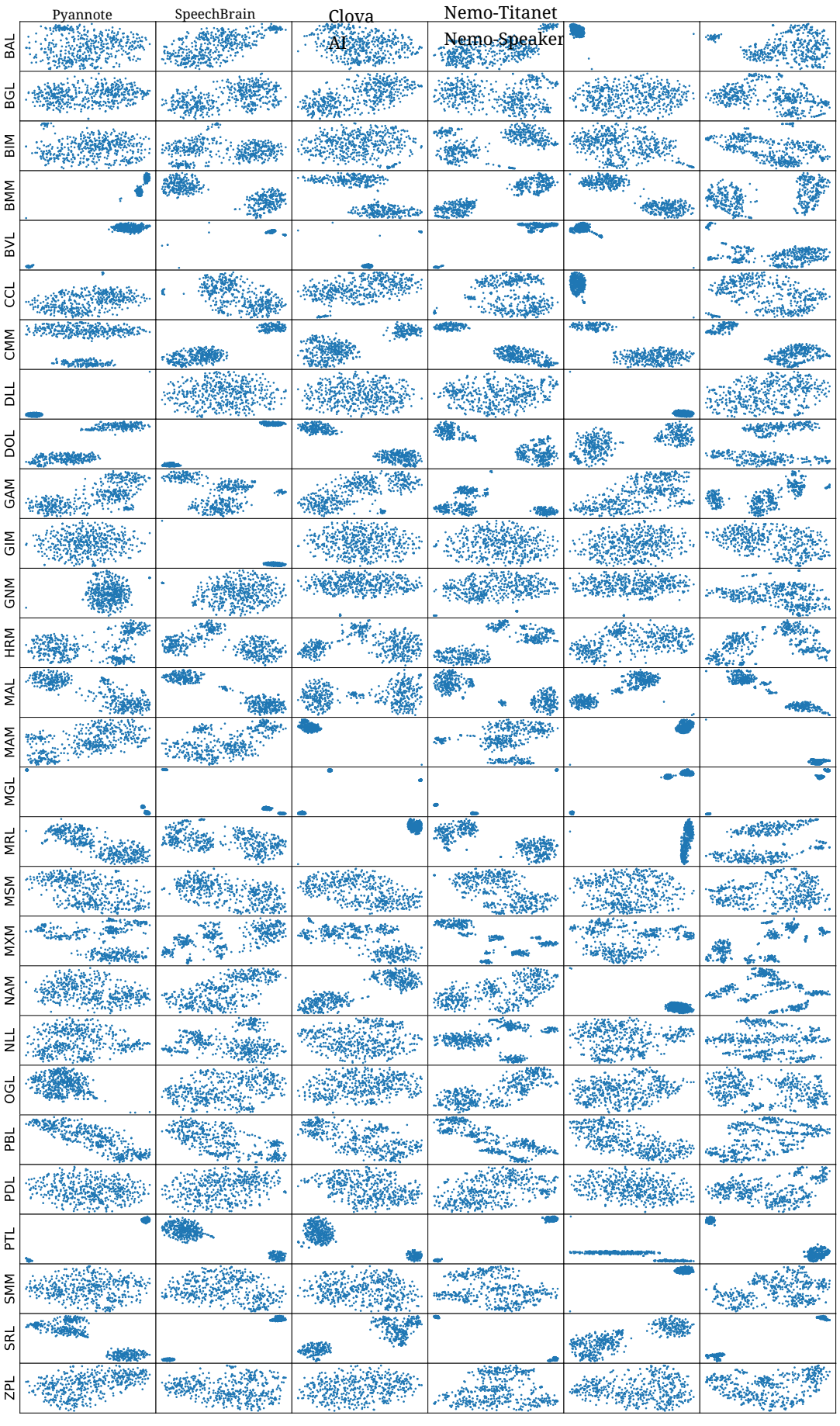


图6。 不同嵌入架构的说话人级 t-SNE 图——**SWARA2.0** 子集。

5. 结论与未来工作

本文尝试评估一些最新且性能优异的深度说话人嵌入，既考察它们在预期用途（即说话人区分）方面的表现，也考察其不足之处，即嵌入中存在的残余信息。所选架构包括 Pyannote、Speech Brain、Clova AI、NeMo Titanet、NeMo SpeakerNet 和 NeMo ECAPA-TDNN，并在一个大型多说话人平行数据集上进行评估，该数据集包含超过 38 小时的口语数据。

在第一组实验中，我们根据等错误率以及说话人间和说话人内相似度度量对这些架构进行了评估。就等错误率而言，NeMo Titanet 的区分能力最佳。然而，就说话人内相似度而言，最擅长将说话人聚类的架构是 Clova AI，而 NeMo ECAPA-TDNN 在最大化说话人表征之间的距离方面表现最佳。这组评估还考察了音频数据的不同子集，即男性与女性，以及录音室录音与家庭录音子集。结果显示，女性说话人和录音室录制的说话人具有较低的等错误率和较高的说话人内余弦相似度。此外，男性说话人和家庭录制的说话人表现出更大的说话人间聚类距离。

第二组实验测量了六组说话人嵌入中存在的残余信息量。研究采用了简单的分类和回归算法。当嵌入中存在不同的语音因素时，这些算法应能取得较高的准确率。考察的因素包括：以字符数和信号时长表示的话语时长、录音条件、信噪比以及语言内容。六种架构均与信号时长和录音条件（包括信噪比）表现出较高的相关性。Pyannote 架构中与录音条件相关的残余信息最少。就话语时长和信噪比而言，基于 NeMo Titanet 的嵌入与这些因素的相关性较低，而 NeMo SpeakerNet 嵌入与话语语言内容的相关系数最小。然而，六种架构之间的差异很小，因此我们认为，至少到目前为止，它们都尚未真正获得解耦说话人表征。

鉴于嵌入中残余信息存在性的实验结果，第三组实验探究了如何在说话人嵌入的进一步下游应用中利用这些残余因素。研究绘制了六组嵌入的低维 t-SNE 表征。就全局说话人表征而言，除 NeMo ECAPA-TDNN 外，所有架构的表现都相近，均形成了结构良好的簇；NeMo ECAPA-TDNN 的簇则具有更大的分布范围。当进一步放大观察说话人层面的这些 t-SNE 表征时，所有深度表征都呈现出与不同录音会话相关的子簇，以及与短语音相关的离群话语。这意味着，嵌入投影中存在的这些信息实际上可以用于文本转语音或语音克隆应用中的数据选择过程。可以从训练集中移除离群话语，并进一步整理表现不佳的说话人数据集，或将其彻底移除。这实际上是扩展本文所述工作的下一步计划之一。我们计划考察不同的文本转语音架构如何受到某些说话人变异性的影响，以及移除话语离群点是否能提升输出合成语音的性能。这项工作的另一项重要结果涉及使用基于嵌入的相似说话人进行文本转语音系统中的数据增强，即使用相对于目标说话人最相似的说话人，确实能够提升所得系统的自然度和说话人相似度。

此外，鉴于说话人嵌入网络现已可用，我们计划将本研究的发现应用于多说话人文本转语音合成系统的训练任务，并确定将这些嵌入输入合成

网络的最高效方式，同时验证合成输出会如何影响这些嵌入，以及如何调整它们以更好地表征不同的说话人身份。

资助： 本文的研究工作部分由 Zevo Technology 根据编号为 21439 的资助协议提供资助。文章处理费由克卢日-纳波卡技术大学资助。

机构审查委员会声明： 不适用

知情同意声明： 不适用。

数据可用性声明： 本文使用的处理后的数据集和实验流程可向作者索取。

利益冲突： 作者声明不存在利益冲突。

参考文献

1. Jia, Y.; Zhang, Y.; Weiss, R.J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Lopez-Moreno, I.; 等. 从说话人验证到多说话人文本转语音合成的迁移学习。*CoRR***2018**, abs/1806.04558. http://xxx.lanl.gov/abs/1806.04558. 2. Stanton, D.; Shannon, M.; Mariooryad, S.; Skerry-Ryan, R.; Battenberg, E.; Bagby, T.; Kao, D. 说话人生成。ICASSP 2022 - 2022 IEEE声学、语音与信号处理国际会议（ICASSP），2022，页码 7897–7901。doi:10.1109/ICASSP43922.2022.9747345. 3. Luck, J.E. 使用倒谱测量进行自动说话人验证。美国声学学会期刊 **1969**, *46*, 1026–1032. 4. Furui, S. 用于自动说话人验证的倒谱分析技术。*IEEE*声学、语音与信号处理汇刊 **1981**, *29*, 254–272. doi:10.1109/TASSP.1981.1163530. 5. Ye, W.; Wu, D.; Nucci, A. 基于高斯混合模型的说话人识别实验研究。见 SPIE—国际光学工程学会会议论文集，美国佛罗里达州奥兰多，2010年4月28日。doi:10.1117/12.849201. 6. Chandra, M.; Nandi, P.; Kumari, A.; Mishra, S. 用于说话人身份识别的基于谱减法的特征。智能系统与计算进展 **2015**,*328*, 529–536. doi:10.1007/978-3-319-12012-6_58. 7. McLaren, M.; Scheffer, N.; Graciarena, M.; Ferrer, L.; Lei, Y. 通过多系统融合提升说话人身份识别对严重信道受损语音的鲁棒性。见2013年IEEE声学、语音与信号处理国际会议论文集，加拿大不列颠哥伦比亚省温哥华，2013年5月26日至31日；页码 6773–6777。doi:10.1109/ICASSP.2013.6638973. 8. Plchot, O.; Matsoukas, S.; Matejka, P.; Dehak, N.; Ma, J.; Cumani, S.; Glembek, O.; Hermansky, H.; Mallidi, S.; Mesgarani, N.; 等. 为DARPA RATS项目开发说话人身份识别系统。见2013年IEEE声学、语音与信号处理国际会议论文集，加拿大不列颠哥伦比亚省温哥华，2013年5月26日至31日；页码 6768–6772，doi:10.1109/ICASSP.2013.6638972. 9. Reynolds, D.A. 一种用于文本无关说话人识别的高斯混合建模方法。博士论文，佐治亚理工学院，美国佐治亚州亚特兰大，1992年。10. Reynolds, D.; Rose, R. 使用高斯混合说话人模型进行鲁棒的文本无关说话人识别。*IEEE*语音与音频处理汇刊 **1995**, *3*, 72–83. doi:10.1109/89.365379. 11. Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. 使用自适应高斯混合模型进行说话人验证。数字信号处理 **2000**,*10*, 19–41. https://doi.org/10.1006/dspr.1999.0361. 12. Campbell, W.; Sturim, D.; Reynolds, D. 使用高斯混合模型超向量进行说话人验证的支持向量机。*IEEE*信号处理快报 **2006**,*13*, 308–311, doi:10.1109/LSP.2006.870086. 13. Zhang, W.; Yang, Y.; Wu, Z. 利用主成分分析分类器进行说话人识别。见国际神经网络联合会议论文集，美国俄勒冈州波特兰，2003年7月20日至24日；第1卷，页码 820–823。doi:10.1109/IJCNN.2003.1223488. 14. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. 用于说话人验证的前端因子分析。*IEEE*音频、语音与语言处理汇刊 **2011**, *19*, 788–798. doi:10.1109/TASL.2010.2064307. 15. Chen, K.; Salman, A. 利用深度神经网络架构学习说话人特定特征。*IEEE*神经网络汇刊 **2011**,*22*, 1744–1756. doi:10.1109/TNN.2011.2167240. 16. Hinton, G.E.; Salakhutdinov, R. 使用神经网络降低数据维度。科学**2006**, *313*, 504–507. 17. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X向量：用于说话人识别的鲁棒深度神经网络嵌入。见2018年IEEE声学、语音与信号处理国际会议（ICASSP）论文集，加拿大艾伯塔省卡尔加里，2018年4月15日至20日；页码 5329–5333。doi:10.1109/ICASSP.2018.8461375. 18. Garcia-Romero, D.; Snyder, D.; Sell, G.; McCree, A.; Povey, D.; Khudanpur, S. 使用完整长度录音对用于说话人识别的x向量深度神经网络进行精炼。见INTERSPEECH 2019论文集，奥地利格拉茨，2019年9月15日至19日。19. You, L.; Guo, W.; Dai, L.; Du, J. 基于高阶统计量的多任务学习，用于基于x向量的文本无关说话人验证。*arXiv***2019**, arXiv:1903.12058.

20. Jung, J.W.; Heo, H.S.; Kim, J.H.; Shim, H.J.; Yu, H.J. RawNet：使用原始波形进行文本无关说话人验证的先进端到端深度神经网络。*arXiv*2019, arXiv:1904.08104. 21. Jung, J.W.; Kim, S.B.; Shim, H.J.; Kim, J.H.; Yu, H.J. 改进的RawNet：利用特征图缩放和原始波形进行文本无关说话人验证。*arXiv*2020, arXiv:2004.00526. 22. Chung, J.S.; Huh, J.; Mun, S.; Lee, M.; Heo, H.S.; Choe, S.; Ham, C.; Jung, S.; Lee, B.J.; Han, I. 为说话人识别中的度量学习辩护。*arXiv*2020, arXiv:2003.11982. 23. Kwon, Y.; Heo, H.S.; Lee, B.J.; Chung, J.S. 说话人识别的方方面面：来自VoxSRC 2020的经验。见ICASSP 2021—2021 IEEE声学、语音与信号处理国际会议（ICASSP）论文集，加拿大安大略省多伦多，2021年6月6日至11日。24. Thienpondt, J.; Desplanques, B.; Demuynck, K. IDLAB VoxCeleb说话人识别挑战赛2020系统说明。*arXiv*2020, arXiv:2010.12468. 25. Vaessen, N.; Van Leeuwen, D.A. 微调用于说话人识别的Wav2Vec2。见ICASSP 2022—2022 IEEE声学、语音与信号处理国际会议（ICASSP）论文集，新加坡，2022年5月23日至27日；页码 7967–7971。doi:10.1109/ICASSP43922.2022.9746952. 26. Bai, Z.; Zhang, X.L. 基于深度学习的说话人识别：综述。神经网络 **2021**, *140*, 65–99. https://doi.org/10.1016/j.n

eunet.20

27. Ohi, A.Q.; Mridha, M.F.; Hamid, M.A.; Monowar, M.M. 深度说话人识别：流程、进展与挑战。IEEE Access 2021, 9, 89619–89643. doi:10.1109/ACCESS.2021.3090109. 28. Makarov, R.; Torgashov, N.; Alenin, A.; Yakovlev, I.; Okhotnikov, A. ID R&D 针对VoxCeleb说话人识别挑战赛2022 的系统描述；ID R&D Inc.：纽约州纽约，美国，2022. 29. Cai, Q.; Hong, G.; Ye, Z.; Li, X.; Li, H. Kriston AI 在 VoxCeleb说话人识别挑战赛2022 中的系统。arXiv 2022, arXiv:2209.11433. 30. Suh, S.; Park, S. ReturnZero 在 VoxCeleb说话人识别挑战赛2022 中的系统。arXiv 2022, arXiv:2209.10147. 31. Zhao, Z.; Li, Z.; Wang, W.; Zhang, P. 用于 VoxCeleb说话人识别挑战赛2022 的 HCCL 系统，2022. 在线可用：https://www.robots.ox.ac.uk/~vgg/data/voxceleb/data_workshop_2022/reports/zzdddz_report.pdf（访问日期：2022年10月22日）。32. Casanova, E.; Shulby, C.; Gölge, E.; Müller, N.M.; de Oliveira, F.S.; Junior, A.C.; Soares, A.d.S.; Aluisio, S.M.; Ponti, M.A. SC-GlowTTS：一种高效的零样本多说话人文本转语音模型。arXiv 2021, arXiv:2104.05557. 33. Cooper, E.; Lai, C.I.; Yasuda, Y.; Fang, F.; Wang, X.; Chen, N.; Yamagishi, J. 基于先进神经说话人嵌入的零样本多说话人文本转语音。载于 ICASSP 2020—2020 IEEE声学、语音与信号处理国际会议论文集（ICASSP），西班牙巴塞罗那，2020年5月4日至8日；第6184–6188页。doi:10.1109/ICASSP40776.2020.9054535. 34. Zhou, Y.; Tian, X.; Li, H. 用于跨语言个性化语音生成的语言无关说话人嵌入。IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 3427–3439. doi:10.1109/TASLP.2021.3125142. 35. Hu, Q.; Bleisch, T.; Petkov, P.; Raitio, T.; Marchi, E.; Lakshminarasimhan, V. 气声与隆巴德语音神经合成。载于2021 IEEE口语语言技术研讨会（SLT）论文集，中国深圳，2021年1月19日至22日；第454–461页。doi:10.1109/SLT48900.2021.9383454. 36. Zhang, M.; Zhou, Y.; Zhao, L.; Li, H. 利用非平行训练数据，将语音合成迁移学习至语音转换。IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 1290–1302. doi:10.1109/TASLP.2021.3066047. 37. Stan, A.; Dinescu, F.; Tiple, C.; Meza, S.; Orza, B.; Chirila, M.; Giurgiu, M. SWARA语音语料库：大型罗马尼亚语平行朗读语音数据集。载于第9届语音技术与人机对话会议（SpeD）论文集，罗马尼亚布加勒斯特，2017年7月6日至9日。38. Bredin, H.; Yin, R.; Coria, J.M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; Gill, M.P. Pyannote.audio：用于说话人分离的神经构建模块。载于 ICASSP 2020（IEEE声学、语音与信号处理国际会议）论文集，西班牙巴塞罗那，2020年5月4日至8日。39. Bredin, H.; Laurent, A. 用于重叠感知再分割的端到端说话人分段。arXiv 2021, arXiv:2104.04045. 40. Ravanelli, M.; Bengio, Y. 使用 SincNet 从原始波形进行说话人识别。载于2018 IEEE口语语言技术研讨会（SLT）论文集，希腊雅典，2018年12月18日至21日。doi:10.48550/ARXIV.1808.00158. 41. Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; 等。SpeechBrain：通用语音工具包。arXiv 2021, arXiv:2106.04624. 42. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN：基于时延神经网络的强调通道注意力、传播与聚合，用于说话人验证。载于 Interspeech 2020 论文集，中国上海，2020年，第3830–3834页。doi:10.21437/Interspeech.2020-2650. 43. Koluguri, N.R.; Park, T.; Ginsburg, B. TitaNet：采用一维深度可分离卷积和全局上下文的说话人表征神经模型。载于 ICASSP 2022—2022 IEEE声学、语音与信号处理国际会议（ICASSP）论文集，新加坡，2022年5月23日至27日。

44. Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.C.; Qin, J.; Gulati, A.; Pang, R.; Wu, Y. ContextNet：利用全局上下文改进自动语音识别的卷积神经网络。*arXiv* **2020**，arXiv:2005.03191. 45. Koluguri, N.R.; Li, J.; Lavrukhin, V.; Ginsburg, B. SpeakerNet：用于文本无关说话人识别与验证的一维深度可分离卷积网络。*arXiv* **2020**，arXiv:2010.12653. 46. Dawalatabad, N.; Ravanelli, M.; Grondin, F.; Thienpondt, J.; Desplanques, B.; Na, H. 用于说话人分离的 ECAPA-TDNN嵌入。*arXiv* **2021**，arXiv:2104.01466. 47. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM：一种高效的梯度提升决策树。载于第*31*届国际神经信息处理系统会议论文集，*NIPS'17*；Curran Associates Inc.：美国纽约州雷德胡克，2017；第3149–3157页。48. Kim, C.; Stern, R. 基于波形幅度分布分析的稳健信噪比估计。载于 Interspeech 2008（含 SST 08）论文集，澳大利亚布里斯班，2008年9月22日至26日；第2598–2601页。doi:10.21437/Interspeech.2008-644. 49. van der Maaten, L.; Hinton, G.E. 使用 t-SNE 可视化高维数据。机器学习研究杂志 **2008**，*9*，2579–2605。