说话人嵌入编码了什么?

Shuai Wang、*Yanmin Qian*、*Kai Yu*

上海市教育委员会智能交互与认知工程重点实验室、智能交互与认知工程语音实验室，计算机科学与工程系，脑科学与技术研究中心，上海交通大学，中国上海

wsstriving@gmail.com, yanminqian@sjtu.edu.cn, kai.yu@sjtu.edu.cn

arXiv:2512.18286v1 [eess.AS] 20 Dec 2025

摘要

开发良好的说话人嵌入已在语音领域引起广泛关注，诸如*i*-vector 和*d*-vector 等表征在多种任务中展现出卓越性能。尽管这些方法已被广泛采用，一个根本性问题仍未得到充分探究：这些嵌入表征实际上编码了哪些属性？为弥补这一空白，我们对三种代表性的说话人嵌入方法进行了全面分析：*i*-vector、*d*-vector 和基于 RNN/LSTM 的序列向量（*s*-vector）。通过精心设计的分类任务，我们系统地考察了它们在多个维度上的编码能力，包括说话人身份、性别、语速、文本内容、词序和信道信息。分析揭示了各类嵌入的不同优势与局限：*i*-vector 擅长说话人区分，但编码的序列信息有限；*s*-vector 能有效捕捉文本内容和词序，但难以表征说话人身份；*d*-vector 表现较为均衡，但在平均化过程中会丢失序列信息。基于这些发现，我们提出了一种新颖的多任务学习框架，将*i*-vector 与*s*-vector 相结合，得到一种新的说话人嵌入（*i*-*s*-vector），兼具二者的互补优势。RSR2015 上的实验结果表明，在内容不匹配试验中，与*i*-vector 基线相比，所提出的*i*-*s*-vector 可使等错误率降低 50% 以上，验证了该方法的有效性。

索引词：说话人识别、说话人嵌入、*i*-向量、*d*-向量、*s*-向量、表征学习

1. 引言

说话人识别旨在通过声音特征识别个人，过去几十年来取得了显著进展。根据语音内容是否受限，说话人识别系统可分为文本相关和文本无关两类。两种范式都得到了广泛研究，近年来性能也因此显著提升。

说话人识别的发展历程中有几个重要的里程碑。自从 [1], 采用带最大后验自适应的高斯混合模型-通用背景模型（GMM-UBM-MAP）框架以来，该框架多年来一直主导着这一领域。在这种方法中，说话人模型是通过最大后验自适应从通用背景模型（UBM）中导出的。然而，GMM-UBM-MAP 在注册过程中会受到数据稀疏性问题的影响，因为只有部分 UBM 参数得到了有效自适应 [1, 2]。为了解决这一局限，研究人员探索了通过全局变换关联不同高斯分量的方法，由此提出了超向量表示 [3, 4], ，与支持向量机结合后取得了有竞争力的结果。

随着联合因子分析（JFA）的提出，定长向量表征的研究逐渐兴起。该方法在超向量空间中分别为说话人因子和信道因子建模，并将其置于不同的子空间。随后，*i*-向量方法 [6], 对其进行了简化，通过对单一总变异子空间建模，成为说话人识别领域事实上的标准。

深度学习在语音识别中的成功 [7] 推动了其在说话人识别中的应用。早期研究将深度神经网络集成到*i*向量框架中，以基于深度神经网络的声学模型取代高斯混合模型 [8, 9]。随后出现了其他方法，包括基于深度神经网络的瓶颈特征提取 [10, 11, 12, 13] 和直接的说话人表征学习[14, 15, 16, 12]。其中，*d*向量 [14] 已成为一种重要的代表方法。近年来，研究人员开始探索使用循环神经网络，尤其是长短期记忆（LSTM）网络，来获取话语级表征 [17, 18, 19]。

尽管说话人嵌入方法层出不穷，且已展现出显著效果 [6, 14, 16, 17, 18], ，一个根本性问题仍未得到充分解答：这些嵌入究竟编码了哪些属性？了解不同嵌入的编码特性，对于为特定应用选择合适的方法以及开发更优的表征至关重要。受自然语言处理领域句子嵌入细粒度分析方法的启发 [20], ，我们提出了一个系统性的研究框架，用于分析说话人嵌入。通过精心设计的分类任务，我们揭示了以下关键见解：

- *i*-向量展现出卓越的说话人区分能力，因此非常适用于说话人身份任务。虽然它在很大程度上能够编码口语词项，但无法捕捉词序等序列信息。
- 基于 LSTM/RNN 的序列向量（*s*-向量）擅长编码文本内容和词序信息，但与 *i*-向量相比，在说话人身份建模方面效果较差。
- *d*向量在多项属性上表现均衡，但由于其平均运算而丢失了序列信息。

本研究得到上海市扬帆计划项目（编号：16YF1405300）、国家自然科学基金委员会项目（编号：61573241 和 61603252）以及中国上海交通大学交叉学科计划（14JCZ03）的资助。实验在上海交通大学的 PI 超级计算机上完成。

{v1>精心设计的*i*向量与*s*向量组合，可以发挥二者的互补优势，从而生成更强大的说话人嵌入。

本研究的主要贡献有三点：(1) 提出一个系统化框架，用于分析不同说话人嵌入中编码了哪些属性；(2) 对*i*-vector、*d*-vector 和*s*-vector 进行多维度的全面实证比较分析；(3) 提出一种新颖的多任务学习架构，将*i*-vector 与*s*-vector 相结合，在文本相关说话人验证任务中显著提升性能。

本文其余部分安排如下。第 2 节介绍本文研究的三种说话人嵌入方法。第 3 节介绍我们的分析方法。第 4 节展示实验结果并进行详细分析。第 5 节介绍我们提出的*i*-*s*-vector 框架，并展示其在 RSR2015 数据集上的结果。第 6 节总结全文。

2. 说话人嵌入方法

本节概述本文分析的三种说话人嵌入方法：*i*-vector、*d*-vector 和*s*-vector。

2.1. *i*-向量

i-vector 框架 [6]将依赖于说话人和会话的超向量$\mathbf{M}$ （由通用背景模型推导而来）表示为低维子空间中的线性变换：

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

其中，$\mathbf{m}$ 是与说话人和会话无关的超向量（通常为通用背景模型的均值超向量），$\mathbf{T}$是一个低秩总变异矩阵，用于捕捉说话人和会话的变异性，而$\mathbf{w}$ 是一个潜向量。$\mathbf{w}$ 在给定观测到的声学特征时的后验均值定义为*i*-向量。这种紧凑表示（通常为 400–600 维）因其有效性和稳健性，已成为说话人识别的标准。

2.2. *d*-向量

d-向量方法 [14] 利用深度神经网络学习具有说话人区分性的表示。首先，以帧级声学特征为输入，通过监督学习训练深度神经网络来对说话人进行分类。训练完成后，从深度神经网络的最后一个隐藏层提取帧级向量。话语级*d*-向量通过对整个话语中的这些帧级表示取平均获得。虽然这种平均操作能够生成定长表示，但也不可避免地丢弃了时间顺序信息。图 1 展示了*d*-向量提取的模型架构。

2.3. 说话人向量

*s*向量（序列向量）方法将说话人识别视为序列建模问题，利用循环神经网络，特别是 LSTM网络，捕捉时间依赖关系。如 [18], 所述，序列级表示可以通过多种策略获得：最后时间步嵌入、基于注意力的嵌入或平均嵌入。在本研究中，我们采用最后时间步方法，即将最后一个时间步的隐藏状态作为序列

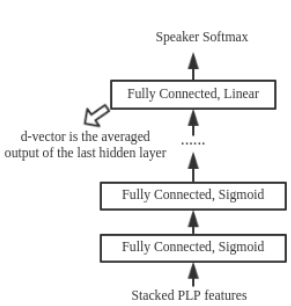

图1： *d-vector*提取

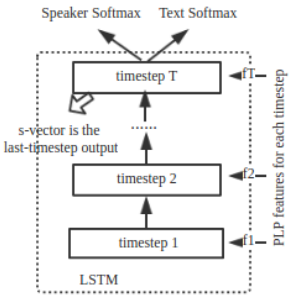

图2： *s*向量提取

表示（*s*向量）。这种方法能够自然地保留序列信息，因此非常适合捕捉词序和时间模式。然而，由于话语级训练样本有限，训练用于说话人识别的 LSTM模型可能颇具挑战。为此，我们采用了与 [16], 类似的多任务学习框架，同时训练模型预测说话人身份和文本内容。模型架构如图2所示。

3. 分析方法

我们的分析方法基于一个基本原则：**如果某种属性被编码在说话人嵌入中，那么就应当能够训练分类器，根据该嵌入预测这一属性，而分类准确率则反映了该属性被编码的程度**。这一原则使我们能够系统地研究不同嵌入表征所捕获的信息。

给定从话语中提取的说话人嵌入，我们构建分类任务，以探究特定属性。我们重点关注三大类属性：

- **说话人相关属性**：说话人身份、性别和语速。

- **文本相关属性**：口语词项（词语出现情况）、词序和话语长度。
- **信道相关属性**：手机身份和录音信道特征。

我们分析了三种说话人嵌入类型：*i*-vector、*d*-vector 和 *s*-vector。针对每种嵌入类型和每项属性，我们提取嵌入表征、构建带标签的数据集，并训练分类器来预测目标属性。为确保分类性能反映嵌入本身固有的编码能力，而非分类器的复杂程度，我们在所有实验中均采用一个简单的多层感知机（MLP），其包含单个隐藏层和 ReLU激活。这样设计可以将性能差异归因于嵌入本身，而非分类器的复杂度。

3.1. 预测任务

基于上述方法，我们设计了八项预测任务，以探查说话人嵌入所编码的不同属性：

- **说话人身份任务**：此任务评估说话人嵌入对说话人身份的编码程度，而这正是说话人识别的核心目标。评估集包含 106 位不同的说话人，构成一个 106 类分类问题。

• **语音文本任务**：此任务评估说话人嵌入是否捕获了有关口语文本内容的信息。数据集包含 30 个不同的句子，因此形成一个 30 类句子分类任务。

• **口语词项任务**：该任务旨在探究嵌入表征是否编码了话语中出现哪些特定词语的信息。数据集包含 147 个不同的词语。我们针对每个词语训练一个二元分类器（逻辑回归），以预测该词语是否出现在话语中。若对应话语中的所有词语均被正确识别，则认为该嵌入表征分类正确。

• **词序任务**：该任务评估嵌入表征捕捉序列信息，尤其是词序的能力。我们将其构建为一个“片段顺序”预测任务：给定两个话语 $\mathbf{u}_1$ 和 $\mathbf{u}_2$，以及它们的拼接序列 $\mathbf{s}$，目标是根据它们的说话人嵌入 $\mathbf{se}_{u1}$、$\mathbf{se}_{u2}$ 和 $\mathbf{se}_s$，预测 $\mathbf{u}_1$ 在 $\mathbf{s}$ 中出现在 $\mathbf{u}_2$ 之前还是之后。这被建模为一个二元分类任务，输入为三个嵌入表征的拼接。通过颠倒 $\mathbf{u}_1$ 和 $\mathbf{u}_2$ 在 $\mathbf{s}$ 中的顺序，生成正负样本。

• **话语长度任务**：此任务用于衡量嵌入表征是否编码了话语时长信息。根据话语长度（经过语音活动检测后），我们将样本分为四类：1-3 秒、4-6 秒、7-9 秒和 10-12 秒（保留间隔以便更好地区分）。通过拼接较短的话语来生成较长的话语。此任务训练一个四分类分类器。

• **信道任务**：此任务用于研究嵌入表征是否捕捉了信道相关信息，而这可能是非期望变异的来源。该数据集包含来自 6 种不同手机的录音，对应 6 个不同的信道。训练一个六分类分类器来预测录音信道。

• **说话人性别任务**：此任务用于评估嵌入表征是否编码了说话人性别信息。训练一个二元分类器，从说话人嵌入中预测性别。

• **语速任务**：此任务用于评估嵌入表征是否捕捉了语速信息。原始数据集仅包含正常语速的语音。我们通过对话语进行时间伸缩，生成速度为原始速度的 $0.5\times$ 和 $1.5\times$ 的额外样本，从而创建慢速、正常和快速三类。此任务训练一个三分类分类器。

4. 实验与分析

4.1. 实验设置

所有实验均在 RSR2015 第1部分数据集[21], 上进行，该数据集包含 300 名说话人（143 名女性和 157 名男性）。每位说话人朗读 30 个从 TIMIT 数据库 [22], 中选取的固定句子，每个句子有 9 次录音会话，以确保涵盖所有英语音素。平均录音时长为 3.20 秒。该数据集分为三个子集：背景集（*bkg*）、开发集（*dev*）和评估集（*eval*）。

表 1：*RSR2015 第 1 部分的子集定义*

子集	# 女性说话人	# 男性说话人	# 总计
bkg	47	50	97
dev	47	50	97
eval	49	57	106

bkg子集 用于训练所有嵌入提取器：高斯混合模型-通用背景模型和*i*-向量提取器、*d*-向量深度神经网络提取器，以及*s*-向量长短期记忆网络提取器。所有实验均使用39维感知线性预测特征（13维静态特征，包含差分系数和二阶差分系数）。*eval* 子集仅用于第3节所述的预测任务。

各嵌入方法的实现细节如下：

• ***i*-向量**：*i*-向量系统由一个包含 1024 个分量的高斯混合模型-通用背景模型以及维度各异的*i*-向量提取器组成。整个流程使用 Kaldi [23] 实现。

• ***d*-向量**：深度神经网络架构包括：(1) 一个包含 429 个神经元的输入层（39 维特征，带有左右上下文帧）；(2) 5 个隐藏层，每层各有 1024 个神经元；(3) 一个包含 $|vector|$ 个神经元的提取层（其中 $|vector|$ 是所需的嵌入维度）；以及 (4) 一个包含 97 个神经元的输出 soft- max 层（对应*bkg* 集合中的每个说话人）。整个网络均采用 Sigmoid 激活函数。网络使用受限玻尔兹曼机（RBM）初始化，然后通过结合反向传播的随机梯度下降（SGD）进行微调。训练完成后，对整个话语中从最后一个隐藏层提取的帧级表示取平均，得到*d*-向量。

• ***s*-vector**：LSTM架构最初由一个39维输入层、一个标准LSTM隐藏层和一个包含97个神经元（每位说话人对应一个神经元）的Softmax输出层组成。通过改变隐藏层中的LSTM单元数量，我们可以提取不同维度的*s*-vectors。然而，按话语训练LSTM模型会导致训练样本有限，进而造成泛化能力较差。为了解决这一问题，我们采用了与 [16], 类似的多任务学习方法，模型设有两个输出层：一个用于说话人分类（97个神经元），另一个用于文本分类（30个神经元）。这一多任务框架提供了额外的监督信号，并改善了模型训练。LSTM使用RMSprop算法进行优化 [24]。

在所有实验图中，标记为*i*-*s*-vector（基于单向长短期记忆网络）的绿线代表新提出的说话人嵌入，将在第5节中介绍。

4.2. 说话人身份任务

说话人区分能力是说话人识别系统的基础。如图3所示，*i*-vector 的表现显著优于另外两种嵌入方法，取得了最高的分类准确率。相比之下，*s*-vector 在这项任务中的表现最差。这种性能差距可归因于长短期记忆网络/循环神经网络模型在话语级运行，因此与基于帧的方法相比，训练样本更少，进而削弱了其说话人区分泛化能力。

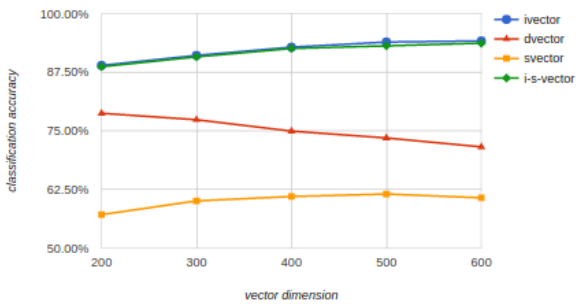

图3：说话人身份任务

4.3. 语音文本任务

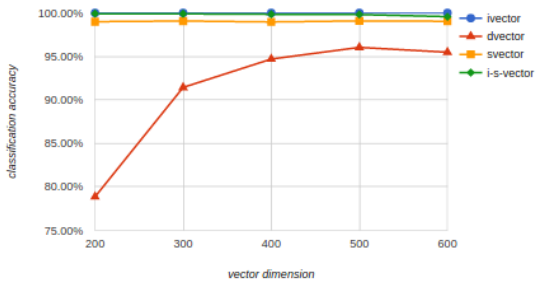

图 4： 语音文本任务

如图4所示，*i*-vector 和 *s*-vector 在语音文本分类任务中均表现出色，准确率都接近 100%。有趣的是，尽管 *d*-vector 对话语级表示取平均，按理说可能会丢失文本信息，但其准确率仍接近 95%。为了更好地理解这一现象并探究其内在机制，我们通过口语词项任务、词序任务和话语长度任务开展了进一步分析。

4.3.1. 口语词项任务

该任务用于评估说话人嵌入是否编码了话语中出现的具体词语信息。如图 5 所示，当嵌入维度超过 300 时，*s*-vector 的表现优于 *i*-vector，体现了其捕获词级信息的卓越能力。相比之下，*d*-vector 无法编码词语信息，也不能可靠地预测口语词项。这符合预期，因为提取 *d*-vector 时的平均操作会去除大部分词级判别信息。

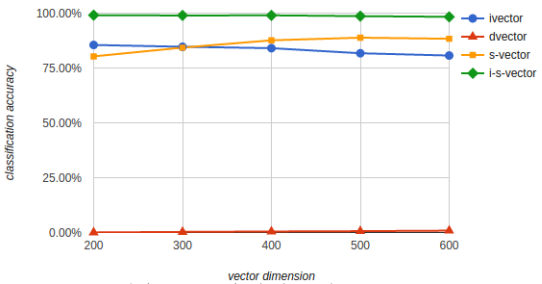

图 5： 语音内容任务

4.3.2. 词序任务

图 6 比较了不同说话人嵌入表征在词序任务上的表现。由于这是一个二元分类任务，随机基线为 50%。正如预期，*d*-vector 无法保留任何顺序信息，表现处于基线水平。*i*-vector 在该任务上的表现也不佳，证实了它无法捕捉序列信息。相比之下，*s*-vector 表现出色，准确率接近 100%。这些结果与各方法的架构特征相符：*i*-vector 和 *d*-vector 不会显式建模时间顺序，而循环网络则通过其循环结构天然地捕捉序列依赖关系。

4.3.3. 话语长度任务

这项任务评估嵌入表征是否编码了有关话语时长的信息。将长度分为四个类别时，随

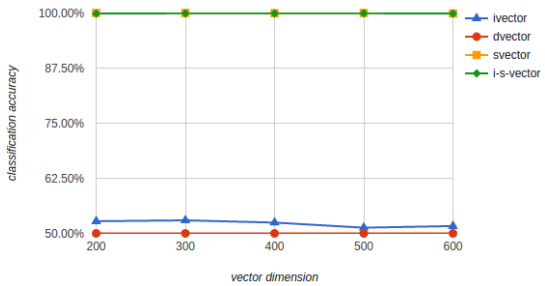

图 6： 词序任务

机基线准确率为 25%。如图 7 所示，正如预期，*s*-vector 和*i*-vector 的表现均优于*d*-vector。这表明这些方法保留了部分时长信息，而*d*-vector 的平均操作在很大程度上消除了这类信息。

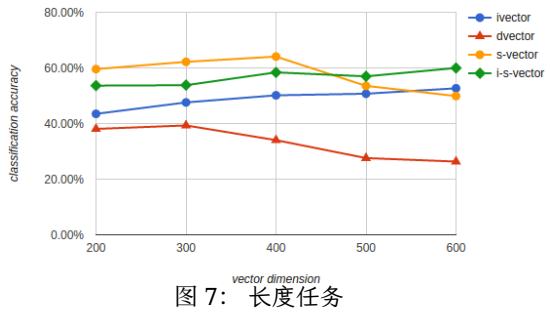

图 7： 长度任务

4.4. 信道任务

RSR2015 数据集包含来自 6 种不同手机的录音，对应 6 个不同的信道。如图 8 所示，三种嵌入类型的预测准确率均显著高于随机基线（16.7%）。这表明这些嵌入表征会无意中编码与信道相关的信息，这可能会带来问题，因为它们会因此对信道失配和录音条件敏感。这一发现凸显了在实际应用中部署这些嵌入表征时采用信道补偿技术的重要性。

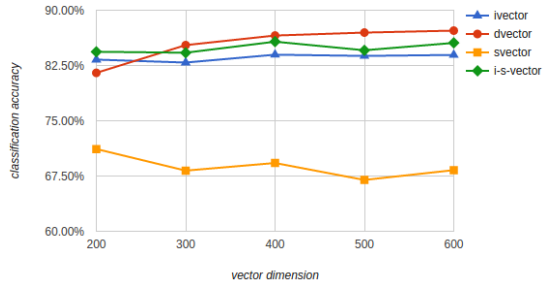

图 8：信道任务

4.5. 说话人性别与语速任务

在说话人性别任务中，三种嵌入类型的准确率都很高（超过 97%），其中 *d*-vector 接近 100%。这表明，性别信息在所有嵌入类型中都有很强的编码，这可能是因为男性和女性声音之间存在根本性的声学差异。

在语速预测方面，*i*-vector、*d*-vector 和 *s*-vector 的准确率分别为 90%、95% 和 77%。*d*-vector 在这项任务中表现最佳，可能是因为帧级深度神经网络表示比话语级聚合更有效地捕捉了时间动态。

5. 结合 *i*-向量和 *s*-向量

我们的分析表明，不同的说话人嵌入表征各有优势，且彼此互补：*i*-向量在说话人区分方面表现出色，而*s*-向量则更擅长编码文本内容和序列信息。对于文本相关说话人验证（TDSV）而言，说话人区分信息和文本信息都至关重要。因此，我们决定开发一个统一框架，结合这两种方法的优势。

我们提出了一种多任务学习框架，将*i*-向量和*s*-向量整合为一种统一表征，称为*i-s*-向量。其核心创新在于训练一个长短期记忆网络（LSTM）模型，在训练过程中同时利用这两种表征。具体而言，我们将最后时刻隐藏输出（*s*-向量）与相应的*i*-向量拼接起来，再将组合后的表征输入说话人分类 Softmax 层。整个架构采用多任务学习进行端到端优化，使模型能够同时预测说话人身份和文本内容。与简单拼接不同，这种方法能够让网络在训练过程中学习最优的特征交互方式。该架构如图 9 所示。

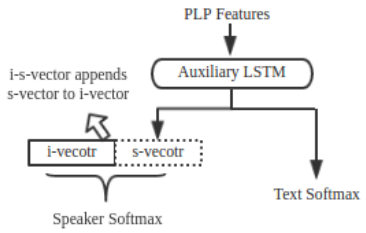

图 9： 用于文本相关说话人验证的 *i-s-vector* 框架

如图 3 至图 7 所示，在几乎所有分析任务中，所提出的 *i-s*-vector 都取得了有竞争力或更优的表现，展现出其有效结合*i*-vector 和*s*-vector互补优势的能力。

5.1. RSR2015 第 1 部分上的评估集

我们在 RSR2015 第1部分数据集上评估所提出的 *i-s*-vector 在文本相关说话人验证任务中的表现。根据不同的冒名者情景，定义了三种测试条件：（I）内容不匹配（说话人正确，文本错误）；（II）说话人不匹配（说话人错误，文本正确）；（III）两者均不匹配（说话人错误，文本错误）。结果见表 2，评分指标为余弦相似度。

表 2： *RSR2015* 第 *1* 部分上文本相关说话人验证的等错误率（%）对比

向量/条件	I	II	III
i-向量	0.35	**1.13**	0.06
i-向量 +*s*-向量	0.28	**1.13**	0.03
i-s-向量（长短期记忆网络）	0.17	1.98	0.03
i-s-向量（双向长短期记忆网络）	**0.11**	1.72	**0.02**

第一行展示了*i*向量基线，其性能与文献中报告的最先进水平相当 [21, 25]。第二行展示了直接拼接*i*向量和*s*向量的结果。最后两行分别展示了所提出的*i-s*向量框架采用单向和双向长短期记忆网络的结果。

主要观察结果：(1) 与内容不匹配条件（I 和 III）下的*i*-向量基线相比，所提出的*i-s*-向量使等错误率降低了 50% 以上，证明了其捕捉文本相关信息的有效性。(2) 使用双向长短期记忆网络（BLSTM）可进一步提升性能，因为它利用了前向和后向时间上下文。(3) 在条件 II（说话人不匹配但文本正确）下，性能有所下降，这表明对文本信息的关注增加可能会略微削弱纯说话人区分能力。然而，对于内容匹配至关重要的文本相关场景而言，这种权衡是有益的。(4) 所提出的多任务学习框架优于简单拼接，验证了联合优化的重要性。这些结果证实，*i-s*-向量特别适用于“说话人相关的内容匹配问题” [25], ，而这正是文本相关说话人验证的核心。

6. 结论

本文全面分析了三种重要的说话人嵌入方法：*i*-vector、*d*-vector 和基于 LSTM/RNN 的序列向量（*s*-vector）。我们通过系统设计的分类任务，考察了每种嵌入类型编码了哪些属性。分析揭示了以下几项重要发现：

- *i*-vector 在说话人区分方面表现出色，因此非常适用于说话人身份任务。高斯混合模型-通用背景模型的音素建模使 *i*-vector 能够编码语音内容信息，但由于其特征袋的性质，它无法捕捉词序。
- 基于 LSTM 的 *s*-vector 擅长编码语音内容和词序信息，因此对于需要序列理解的任务尤其有价值，例如随机数字说话人识别。
- *d*-vector 在多种属性上表现均衡，但平均操作会使其丢失序列信息，从而限制了它在顺序敏感任务中的有效性。
- 所有嵌入类型都会在一定程度上无意中编码信道信息，这凸显了在实际部署中采用信道补偿技术的重要性。
- *i*向量与*s*向量各自的优势互补，通过精心设计的多任务学习框架可以有效结合。与*i*向量基线相比，所提出的*i-s*向量在 RSR2015 第 1 部分的内容不匹配试验中使等错误率降低了 50% 以上，彰显了理解并利用嵌入表征特性的实际价值。

我们的工作提供了相关见解，可指导针对特定应用选择和设计说话人嵌入，并展示了理解嵌入表征特性如何通过有原则的组合策略带来更优的表征。

7. 参考文献

[1] D. A. Reynolds、T. F. Quatieri 和 R. B. Dunn，“使用自适应高斯混合模型进行说话人验证”，数字信号处理，第 10 卷，第 1–3 期，第 19–41 页，2000 年。[2] S. Shum，“ 说话人日志的无监督方法”，博士论文，麻省理工学院，2011 年。

[3] W. M. Campbell、D. E. Sturim 和 D. A. Reynolds，“使用 GMM 超向量进行说话人验证的支持向量机”，*IEEE* 信号处理快报，第 13 卷，第 5 期，第 308–311 页，2006 年。

[4] W. M. Campbell、D. E. Sturim、D. A. Reynolds 和 A. Solomonoff，“基于支持向量机的说话人验证：使用 GMM 超向量核与 NAP 可变性补偿”，载于声学、语音与信号处理，*2006*。*ICASSP 2006* 论文集。*2006 IEEE* 国际会议，第 1 卷。IEEE，2006 年，第 I–I 页。

[5] P. Kenny，“说话人与会话可变性的联合因子分析：理论与算法”，*CRIM*，蒙特利尔（报告）*CRIM-06/08-13*，2005。[6] N. Dehak、P. J. Kenny、R. Dehak、P. Dumouchel 和 P. Ouellet，“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，第 19 卷第 4 期，第 788–798 页，2011。[7] G. Hinton、L. Deng、D. Yu、G. E. Dahl、A.-r. Mohamed、N. Jaitly、A. Senior、V. Vanhoucke、P. Nguyen、T. N. Sainath 等，“语音识别中的声学建模深度神经网络：四个研究小组的共同观点”，*IEEE* 信号处理杂志，第 29 卷第 6 期，第 82–97 页，2012。[8] Y. Lei、N. Scheffer、L. Ferrer 和 M. McLaren，“使用音素感知深度神经网络进行说话人识别的新方案”，载于声学、语音与信号处理（*ICASSP*），*2014* 年 *IEEE* 国际会议。IEEE，2014，第 1695–1699 页。[9] P. Kenny、V. Gupta、T. Stafylakis、P. Ouellet 和 J. Alam，“用于提取说话人识别 Baum-Welch 统计量的深度神经网络”，载于*Odyssey* 会议论文集，2014，第 293–298 页。[10] T. Fu、Y. Qian、Y. Liu 和 K. Yu，“用于文本相关说话人验证的串联深度特征”，载于 *INTERSPEECH*，2014，第 1327–1331 页。[11] Y. Liu、Y. Qian、N. Chen、T. Fu、Y. Zhang 和 K. Yu，“用于文本相关说话人验证的深度特征”，*Speech Communication*，第 73 卷，第 1–13 页，2015。[12] Y. Tian、M. Cai、L. He 和 J. Liu，“瓶颈特征与多语言深度神经网络在说话人验证中的研究”，载于 *Interspeech*，2015，第 1151–1155 页。[13] F. Richardson、D. Reynolds 和 N. Dehak，“用于说话人与语言识别的统一深度神经网络”，*arXiv* 预印本 *arXiv:1504.00923*，2015。[14] E. Variani、X. Lei、E. McDermott、I. L. Moreno 和 J. Gonzalez-Dominguez，“用于小型化文本相关说话人验证的深度神经网络”，载于声学、语音与信号处理（*ICASSP*），*2014* 年 *IEEE* 国际会议。IEEE，2014，第 4052–4056 页。[15] L. Li、Y. Lin、Z. Zhang 和 D. Wang，“改进的文本相关说话人识别深度说话人特征学习”，载于信号与信息处理协会（*APSIPA*）*2015* 年亚太峰会与会议。IEEE，2015，第 426–429 页。[16] N. Chen、Y. Qian 和 K. Yu，“用于文本相关说话人验证的多任务学习”，载于国际语音通信协会第十六届年会，2015。[17] G. Heigold、I. Moreno、S. Bengio 和 N. Shazeer，“端到端文本相关说话人验证”，载于声学、语音与信号处理（*ICASSP*），*2016* 年 *IEEE* 国际会议。IEEE，2016，第 5115–5119 页。[18] G. Bhattacharya、J. Alam、T. Stafylakis 和 P. Kenny，“基于深度神经网络的文本相关说话人验证”。[19] S. Hochreiter 和 J. Schmidhuber，“长短期记忆”，神经计算，第 9 卷第 8 期，第 1735–1780 页，1997。[20] Y. Adi、E. Kermany、Y. Belinkov、O. Lavi 和 Y. Goldberg，“使用辅助预测任务对句子嵌入进行细粒度分析”，*arXiv* 预印本 *arXiv:1608.04207*，2016。

[21] A. Larcher、K. A. Lee、B. Ma 和 H. Li，“文本相关说话人验证：分类器、数据库与 RSR2015”，*Speech Communication*，第 60 卷，第 56–77 页，2014。[22] J. S. Garofolo、L. F. Lamel、W. M. Fisher、J. G. Fiscus 和 D. S. Pallett，“美国国防高级研究计划局 TIMIT 音频-音系连续语音语料库只读光盘。美国国家标准与技术研究院语音光盘 1-1.1”，*NASA STI/Recon* 技术报告 *n*，第 93 卷，1993。[23] D. Povey、A. Ghoshal、G. Boulianne、L. Burget、O. Glembek、N. Goel、M. Hannemann、P. Motlicek、Y. Qian、P. Schwarz等，“Kaldi语音识别工具包”，载于*IEEE 2011* 年自动语音识别与理解研讨会，编号 EPFL-CONF-192584。IEEE信号处理学会，2011。[24] T. Tieleman 和 G. Hinton，“第 6.5 讲——RMSprop：用其近期幅度的运行平均值除以梯度”，*Coursera*：机器学习神经网络，第 4 卷第 2 期，2012。[25] S. Dey、S. Madikeri、M. Ferras 和 P. Motlicek，“用于文本相关说话人验证的深度神经网络后验概率”，载于声学、语音与信号处理（*ICASSP*），*2016* 年 *IEEE* 国际会议。IEEE，2016，第 5050–5054 页。