arXiv:2301.06277v1 [cs.SD] 16 Jan 2023

利用稀疏 LDA 变换说话人嵌入改进目标说话人提取

Kai Liu、Xucheng Wan、Ziqing Du 和 Huan Zhou

华为技术有限公司人工智能应用研究中心，中国深圳
{liukai89,wanxucheng,duziqing1,zhou.huan}@huawei.com

摘要。 作为语音分离的一种实用替代方案，目标说话人提取（TSE）旨在利用从说话人处提取的额外说话人提示，提取目标说话人的语音。其主要挑战在于如何恰当地提取并利用说话人提示，以提升提取语音的质量。现有大多数目标说话人提取研究所采用的提示提取方法，是直接使用从预训练说话人验证模型中提取的判别性说话人嵌入。尽管较强的说话人区分能力是说话人验证任务最理想的特性之一，但我们认为，这种特性对于目标说话人提取而言可能过于复杂。本研究提出，对于目标说话人提取来说，具有明确类别可分性的简化说话人提示或许更为适用。为验证这一设想，我们引入了多种形式的说话人提示，包括朴素说话人嵌入（如 x-vector 和 xi-vector），以及通过稀疏 LDA 变换生成的新型说话人嵌入。我们将这些说话人提示与 SepFormer（一种最先进的语音分离模型）结合，构建了相应的目标说话人提取模型，并在基准数据集 WSJ0-2mix 上评估其性能。实验结果验证了该设想的有效性和泛化能力，SI-SDRi 相对提升最高可达 9.9%。此外，我们表现最佳的 TSE 系统取得了 19.4 dB 的 SI-SDRi 和 3.78 的 PESQ，显著优于当前最先进的系统，并在 WSJ0-2mix 上实现了迄今报道的最佳目标说话人提取结果。

关键词： 目标说话人提取 · 说话人嵌入 · 线性判别分析。

1 引言

在真实的多说话人通信场景中，例如人际对话或多方会议，感兴趣说话人的语音受到其他人干扰是常见情况。这种受干扰的语音混合信号可能会让听者听到部分可辨的语音，或对下游应用（如语音识别和说话人分离）产生不利影响。如何从复杂的听觉场景中获取所需语音，一直是计算听觉场景分析领域的一项长期挑战 [1]。随着深度神经网络（DNN）的出现，人们提出了两种基于 DNN 的方法——语音分离（SS）和目标说话人提取（TSE）算法，以不同的目标解决这一问题。

说话人分离算法的目标是从混合信号中恢复所有单个说话人的语音。给定目标语音、混合信号以及说话人总数，说话人分离可以构建为一个监督学习问题。大多数说话人分离算法会估计时频分离掩码或滤波器，并对其进行优化，以尽量减小信号重建损失。凭借大量训练数据，这种方法能够实现出色的分离性能。据报道，在应用最广泛的数据集 WSJ0-2mix [2], 上，SepFormer [3], 作为表现最佳的说话人分离模型之一，在尺度不变信号失真比提升（SI-SDRi）[4]方面取得了接近该数据集性能上限的表现（22.3 dB 对比 23.1 dB）。

另一种方法是，目标说话人提取旨在提取目标说话人的语音，同时忽略其他人的语音，因此更适合实际部署。作为语音分离的一种特殊情况，目标说话人提取只输出单个语音，并将额外的说话人识别信息（本文称为说话人提示）作为辅助输入。

辅助输入有多种形式，最常见的是目标说话人预先录制的注册语音。然而，将表现良好的语音分离模型扩展为仍能保持高性能的目标说话人提取模型并非易事。其核心挑战主要在于如何恰当地提取并利用说话人提示，从而提升所提取语音的质量。本文关注的挑战是如何恰当地提取说话人提示，使其能够捕捉目标说话人的语音特征并提升所提取语音的质量。

从以往来看，先前的目标说话人提取研究提出了两种主流的说话人提示提取策略。一种直接的方法是利用深度说话人嵌入作为说话人提示，通过预训练的深度神经网络从注册语音中提取这些嵌入。这些深度神经网络专为说话人验证（SV）任务精心设计，即判断测试语音中的说话人身份是否与注册说话人一致。这些网络通常在大规模数据集上训练，数据集中的录音涵盖各种声学条件。因此，所学习的深度说话人嵌入作为可变长度语音的定维表示，具有说话人区分能力，即同一说话人内部差异较小，而不同说话人之间差异较大。近年来提出的目标说话人提取模型采用了许多在说话人验证中广泛使用的深度说话人嵌入。具体而言，d-vector 用于 VoiceFilter [5], VoiceFilter-lite [6] 和 pDCCNR [7]；i-vector 用于一系列基于Speaekerbeam的目标说话人提取模型 [8,9]；基于深度 ResNet 的说话人嵌入用于一种基于注意力的目标说话人提取模型 [10]；而 x-vector 则用于近期的双路径滤波网络 [11]。

尽管大多数目标说话人提取模型都采用了深度说话人嵌入，但有人认为，这些嵌入是针对说话人验证单独优化的，因此可能并非目标说话人提取的最优选择。受此观点启发，另一种策略是从辅助说话人编码器中提取线索。通常，该说话人编码器会与目标说话人提取模型通过多任务学习联合训练，采用多类分类损失（在 [8,12,13] 中）或度量学习损失（在 [14] 中）。

尽管付出了这些努力，最先进的目标说话人提取模型的性能仍不及单说话人模型。在广泛采用的 WSJ0-2mix

数据集上，最佳目标说话人提取模型的性能为 18.6 dB，远低于最先进水平的单说话人模型（22.3 dB）。这与通常的预期相悖：目标说话人提取模型的整体性能上限可能高于单说话人模型，因为该网络只需专注于重构目标说话人。

我们发现，目前的说话人线索提取方法，无论是源自深度说话人嵌入，还是通过联合优化的说话人编码器提取，设计时都主要关注可区分性。然而，众所周知，说话人嵌入特征分散在高维特征空间中（通常高于 128 维）。在这种情况下，类别可分性（衡量嵌入区分效果的指标）不可避免地会受到各类别内部嵌入方差的影响。受此启发，我们假设，类别可分性明确的说话人提示或许更适用于目标说话人提取。这是一个简单而新颖的想法，此前的目标说话人提取研究尚未对此进行探索。

为此，本文对所提出的方法进行了初步探究，研究了稀疏 LDA 变换方案。具体而言，我们首先将语音分离模型骨干网络与 x-vector 说话人提示或一种新提出的说话人嵌入 xi-vector（通过在说话人嵌入神经网络中引入不确定性建模而设计）相结合，构建目标说话人提取基线。随后，我们将上述说话人提示替换为其 LDA 变换版本，期望后者凭借更强的分类能力促进说话人提取。最后，为缩小 WSJ0-2mix 上最先进水平的目标说话人提取与语音分离之间的性能差距，我们所有的目标说话人提取模型均采用 SepFormer 这一最先进的语音分离模型作为骨干网络。

本文其余部分安排如下。第 2 节简要介绍与 SepFormer 和 xi-vector 相关的研究工作。第 3 节介绍我们提出的方法。第 4 节和第 5 节报告并分析实验结果。最后，我们在第 6 节总结全文。

2 相关工作

我们首先简要概述 SepFormer [3], ，它是我们提出的目标说话人提取模型的骨干网络。SepFormer 是一种基于 Transformer 的单通道语音分离模型。它基于学习域掩蔽方法，整体架构由编码器、解码器和掩码模块组成。该模型按以下流程处理混合语音输入：编码器使用卷积层将输入转换为类似 STFT 的表示；随后，掩码模块根据该表示估计掩码；最后，解码器使用转置卷积层，将掩码与表示的乘积转换为时域分离源。

该流程的核心模块是掩码模块，由多个 Transformer 组成，并采用双路径处理，通过多尺度方法学习依赖关系。具体而言，首先将混合语音帧分割为相互重叠的分块。这些堆叠的分块首先沿着

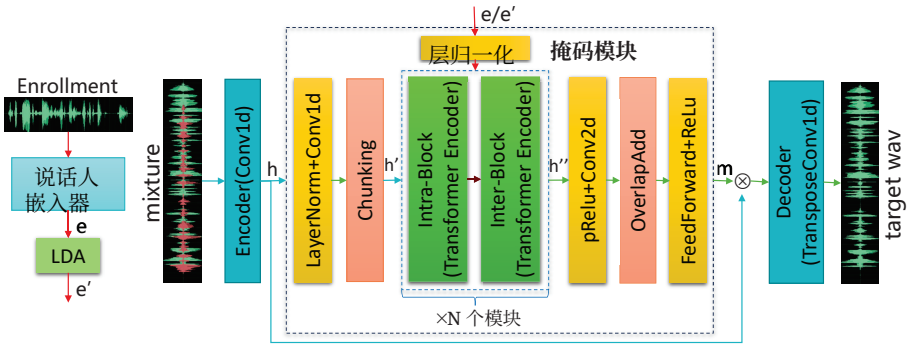

图 1. 所提目标说话人提取架构。其中， h 表示学习得到的语音特征， m表示混合语音的目标声源掩码， e/e' 分别表示原始说话人嵌入和经线性判别分析变换后的嵌入。

华为专有，限量分发

分块维度独立处理（块内处理），以捕捉短期依赖；随后转置维度，并跨分块执行序列处理（块间处理），以汇总所有分块的信息。块内和块间处理可以迭代执行并重复多次（N 次）。最终汇总的信息经过线性层和非线性激活处理，再通过重叠相加操作生成多个掩码（对于 WSJ0-2mix，则生成两个掩码）。

SepFormer 展现出出色的分离性能，并在一系列基准数据集（如 WSJ0-2mix[1]和 WSJ0-3mix[2], ）上取得了最先进水平的 SI-SDRi。因此，我们选择它作为下文所提出的目标说话人提取模型的骨干网络。此外，为了提升分离性能，我们对原始 SepFormer 进行了少量修改：将 Intra-InterT 处理重复次数从 $N = 2$ 次增加到 $N = 4$ 次，并将 Transformer 编码器的层数从 8 层简化为 4 层。

3 所提方法

在本节中，我们提出几种目标说话人提取模型，通过将不同的说话人提示与 SepFormer 的分离骨干网络相结合来构建。这些模型采用相同的架构，如图 1 所示。与 SepFormer 骨干网络不同，目标说话人提取模型在掩码模块中引入了额外的说话人提示 e/e' ，并且仅为目标说话人生成一个掩码。更多架构细节将在后文的实验部分介绍。

[1] https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix [2] https://paperswithcode.com/sota/speech-separation-on-wsj0-3mix

3.1 使用 xi-vector 的目标说话人提取模型

如前所述，i-vector、d-vector、x-vector 等常见的深度说话人嵌入已广泛用作目标说话人提取任务的说话人提示。近期提出的 xi-vector 嵌入 [15] 将用于衡量输入变化性的帧不确定性度量纳入深度说话人表征学习。换言之，它结合了 i-vector 的生成式特性与 x-vector 的说话人区分能力。

鉴于能够考虑逐帧不确定性，所提出的 xi-vector 嵌入表现出更强的抗扰动鲁棒性

通过显式处理输入特征的不确定性，我们预计 xi-vector 相较于其前身（例如 x-vector）能够更好地抵抗扰动，因此有助于目标说话人提取。因此，本文提出使用 xi-vector 作为说话人提示，而此前的目标说话人提取研究尚未对此进行探索。

为了寻找在 SepFormer 中利用这一新说话人提示的合适方法，我们开展了一些初步实验。根据实验结果，我们通过经验发现，最好在 Intra-Transformer 和 Inter-Transformer 模块的第 1 层将说话人提示与 SepFormer 集成（如图 2 所示）。这意味着，xi-vector 会在 8 个不同位置与我们的 SepFormer（由 $N = 4$ 个 Intra-InterT 处理过程组成）反复拼接。

同时，在掩码模块中，我们探索并比较了多种融合配置，以有效地将说话人提示与 IntraT 或 InterT 模块的输入相结合，例如基于拼接、基于乘积、基于注意力的缩放、基于 FiLM 以及基于交叉注意力的融合。其中，表现最佳的基于交叉注意力的融合技术 [16] 被应用于我们的目标说话人提取模型中，说话人提示同时影响注意力权重和融合偏置。此类说话人与语音融合过程（如图 2 中橙色部分所示）可表示为：

$$A_{pc} = f_{weight}(e, h', \theta_{Q_e,K_e,V_e}, \theta_{Q_{h'},K_{h'},V_{h'}})$$
$$f_{pc} = A_{pc}(e\theta_{V_e} + h\theta_{V_{h'}})$$

其中， e 表示说话人提示， h' 表示 IntraT 或 InterT 模块的输入，$\theta_{Q,K,V}$ 为注意力参数， f_{weight} 为权重函数， A_{pc} 为注意力权重， f_{pc} 为通过并行拼接得到的融合输出。

3.2 基于稀疏 LDA 变换嵌入的目标说话人提取模型

如前所述，我们假设具有较强分类能力的说话人提示可能有利于目标说话人提取。为验证这一点，一种简单的方法是对原始说话人嵌入执行 LDA（线性判别分析）[17] 。作为一种经典的降维技术，LDA 通过着重最大化各个 M 类别之间的可分性，将原始数据投影到一个维度更低的子空间中（其维度严格小于类别数 M）。

通常，类别可分性准则不容易计算。因此，在 LDA 中，我们假设数据服从高斯分布。在这一假设下，通过不相关的线性判别向量最大化类间方差与类内方差之比，即可获得最优的类别分离效果。

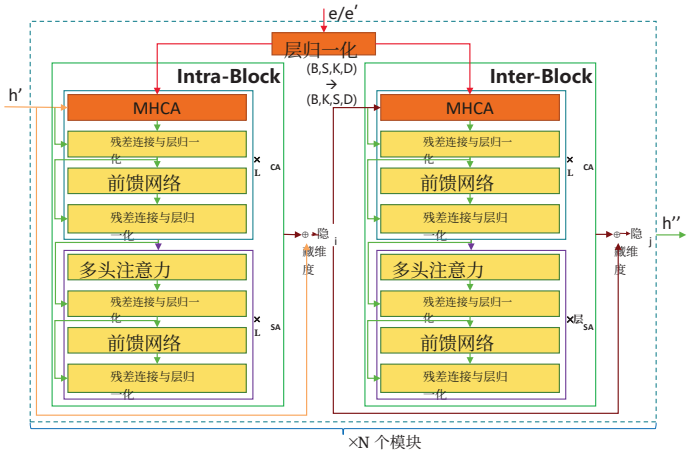

图 2. 采用我们提出的说话人与语音融合策略的掩码模块。这里，MHCA 表示多头交叉注意力。 B=批大小， S=分块数， K=每个分块中的帧数， D=每帧的语音特征维度。 L_{CA}和 L_{SA} 分别表示带交叉注意力和自注意力的 Transformer 编码器层数。 $\bigoplus$ 表示相加。

Huawei 专有，限制分发

需要注意的是，将嵌入维度降至 $M-1$ 不会损失类别可分性。对于用于目标说话人提取的说话人提示，我们认为可分性和泛化能力都很重要。因此，我们提议采用从一个小子空间中提取的稀疏 LDA 变换嵌入（该子空间由前 l 个判别方向张成，维度为 l，且$l < M-1$）。换句话说，我们只选择重要的判别方向，同时舍弃其他无关或冗余的方向，以牺牲部分数据方差为代价。由此得到的新说话人提示能够在可分性和泛化能力之间取得平衡：舍弃少量不需要的信息，同时尽可能保留可分性。

为了首次应用这一思路，我们在第 3.1 节提出的目标说话人提取模型基础上进行扩展，在说话人与语音融合过程之前，额外使用线性判别分析（LDA）对朴素的说话人提示（即 x-vector 或 xi-vector）进行变换。

4 实验设置

4.1 数据集

数据集 WSJ0-2mix [2] 是一个广泛使用的语音分离数据集，其中的单通道混合信号由《华尔街日报》（WSJ0）语料库中的两位说话人的语句随机混合而成，采样率为 8 kHz。语句以 0 dB 至 5 dB 的不同信噪比（SNR）进行混合。

为研究我们提出的方法的有效性，我们按照脚本[3] 基于 WSJ0-2mix 构建了一个目标说话人提取数据集。该数据集共包含训练集（20000 条语句）、验证集（5000 条语句）和测试集（3000 条语句）。训练集和验证集中的所有语句均来自 101 位说话人，而测试集中的语句来自另外 18 位说话人。测试集中辅助语音的平均时长为 7.3 秒。

为了构建说话人嵌入数据集，我们将 WSJ0 官方训练集si_tr_s 按照 0.95:0.05 的划分比例分为训练集和验证集。测试集由合并后的 WSJ0 开发集 si_dt_05和评估集 si_et_05构成。对于合并数据集中出现的每位说话人，我们随机选取 200 条语音作为正语音样本，并从其他说话人中再选取 200 条语音作为负语音样本。

4.2 实现细节

我们提出的 TSE系统采用 PyTorch 实现，架构详情如下：

– 编码器模块：编码器中使用了一维卷积层，卷积核大小为 16，步幅为 8，输入和输出维度分别为 1 和 256；– 掩码模块：掩码模块主要由分块、重叠相加、块内-块间 Transformer 模块和二维卷积组成。我们设置每个分块中的帧数为 $K=250$ ，并设置语音特征 $D=256$。对于每个块内和块间 Transformer 模块，交叉注意力层和自注意力层的数量分别设为 1 和 3。与 SepFormer 骨干网络不同，由于 TSE 任务只需提取一位说话人的语音，因此我们将二维卷积的输出通道设为 256，而不是 512。– 解码器模块：作为编码器的逆过程，解码器模块使用一维转置卷积层。其卷积核大小和步幅与编码器模块相同，输入和输出维度分别为 256 和 1。

输入混合语音和辅助说话人提示后，我们的目标说话人提取系统在尺度不变信号失真比损失的监督下进行训练。受内存限制，我们将批次大小设为 1。我们的目标说话人提取系统使用 Adam 优化器进行训练，并采用 $ReduceLROnPlateau$ 学习率调度器。前 20 个训练轮次的初始学习率设为 $1.5e-4$ ，之后若验证集上的损失连续 2 个训练轮次未改善，则将学习率降至原先的一半。

在 LDA 的实现中，我们采用 *scikit-learn* 库的实现，并选择特征分解作为求解器。我们为 components 参数设置了少量维度，其余参数均保持默认值。该库还提供解释方差比这一指标，用于表示每个成分所解释的方差百分比。

[3] https://github.com/xuchenglin28/speaker_extraction_SpEx/tree/master/data/ wsj0_2mix

4.3 评估指标

按照大多数现有技术所采用的相同评估方法和指标，我们使用 SDRi、SI-SDRi 等重建指标以及 PESQ 语音质量指标来评估目标说话人提取系统。对于这些指标，数值越高表示性能越好。本研究报告所有指标，以便与更多现有技术进行充分的性能比较。此外，我们还使用标准指标最小检测代价函数（minDCF）和等错误率（EER）来评估说话人嵌入。一般来说，EER 和 minDCF 越低，嵌入的可分辨性能越好。

5 结果与讨论

5.1 目标说话人提取基线

首先，我们通过复现作者提供的代码[4]构建语音分离骨干系统。复现的系统性能与原始 SepFormer [3], 相当，但 SI-SDRi 略有下降。原始 SepFormer 使用两个模块时宣称的性能为 22.3 dB，我们复现的系统达到 21.6 dB，而采用四个模块的改进版 SepFormer 则进一步将性能提升至 21.9 dB。下文中，我们所有的目标说话人提取系统均基于改进版 SepFormer 的骨干网络实现。

基于该骨干网络，我们首先构建了目标说话人提取基线。该系统采用广泛使用的 x-vector 作为说话人提示，因此称为 x-TSE 系统。我们的 x-vector 模型按照 SpeechBrain 配方[5]进行训练。为与现有技术公平比较，x-vector 的训练集仅限于 WSJ0-2mix 的训练数据和开发数据，共包含 101 位说话人，且与测试集完全不重叠。

同样，在相同的训练设置下，我们将 x-vector 中的统计池化替换为高斯后验推断，从而训练 xi-vector 模型。对于 xi-vector，我们仅提取后验均值向量（ϕ）。

说话人嵌入	维度	等错误率(%)	最小检测代价函数
x-vector	512	3.58	0.364
xi-vector	512	3.31	0.370

表 1. 我们训练的说话人嵌入的性能比较

表 1 对比了两种嵌入的相应性能。对比结果表明，xi-vector 在等错误率方面有所改善，但最小检测代价函数略有下降。此外，通过 t-SNE 图可视化 512 维嵌入，图 3 显示 xi-vector 的聚类确实比 x-vector 更清晰

[4] https://github.com/speechbrain/speechbrain/tree/develop/recipes/WSJ0Mix/separation [5] https://github.com/speechbrain/speechbrain

尽管两种嵌入中的所有样本彼此间隔都较远。此外，还应指出，训练数据中的说话人数

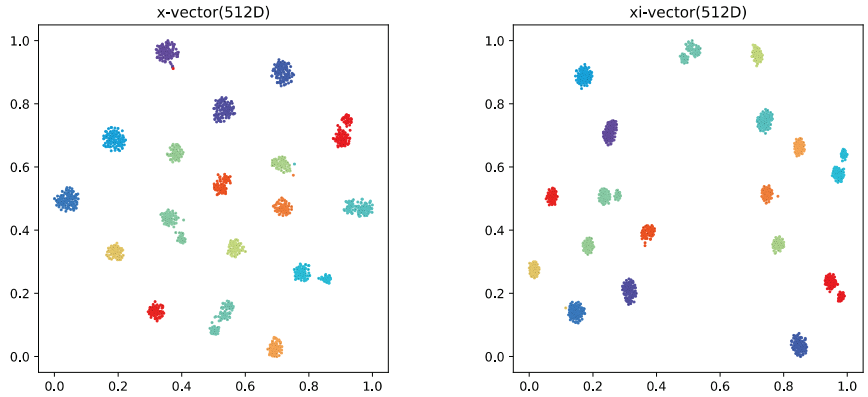

图 3. 使用 t-SNE 对测试集中的说话人嵌入进行可视化的图

在 WSJ0-2mix 中为 101，远少于标准说话人识别任务中的数量（例如，VoxCeleb1 中超过 1k，VoxCeleb2 中为 6k）。

5.2 提出的目标说话人提取系统

除基线外，我们还按照提出的方法实现了一系列目标说话人提取系统。这些系统采用图 1 所示的相同目标说话人提取架构，但使用三种形式的说话人提示：i）xi-vector；ii）经线性判别分析变换的 xi-vector；以及 iii）经线性判别分析变换的 x-vector。因此，这些系统分别命名为 xi-TSE、xi-LDA-TSE 和 x-LDA-TSE。

提出的目标说话人提取系统与基线的比较 首先，将提出的三种目标说话人提取系统与基线 x-TSE 进行比较。此外，考虑到 LDA变换的维度灵活性，对于采用经 LDA变换嵌入的目标说话人提取系统，我们考察了两种不同的维度设置。所有六种目标说话人提取系统的实验结果列于表 2。此处，我们为基线结果加下划线，并将所有系统中的最佳结果加粗。此外，还报告等错误率和最小检测代价函数，以评估相应说话人提示的可分辨性能。

根据表 2 所列结果，我们得出以下关键观察和结论：

a) 在基线上采用 LDA变换，显著提升了基线性能，这验证了我们第二种方案的有效性；b) xi-TSE也呈现出类似的表现，这表明所提出的 LDA变换方案可能具有通用性，可应用于其他说话人嵌入；

	系统	SI-SDRi（dB）	PESQ	最小检测代价函数	等错误率（%）
基线	x-TSE	17.1	3.59	0.36	3.58
我们提出的TSE系统	x-LDA-TSE（64D）	18.5	3.70	0.39	2.94
	x-LDA-TSE（32D）	17.9	3.67	**0.32**	2.50
	xi-TSE	18.0	3.65	0.37	3.31
	xi-LDA-TSE（64D）	18.2	3.67	0.43	2.75
	xi-LDA-TSE（32D）	**18.8**	**3.71**	0.45	2.78

表2。所提出的目标说话人提取系统与基线在 WSJ0-2mix 测试集上的比较

c) TSE系统的性能对 LDA变换的维度设置较为敏感；最佳维度可通过经验确定；d) xi-TSE在 SI-SDRi 和 PESQ 两项指标上均优于基线，这符合我们先前的预期，即 xi-vector 的鲁棒性有助于目标说话人提取；e) 等错误率（最小检测代价函数）最低的 TSE系统并未取得最佳的目标说话人提取性能。这再次表明，说话人验证与目标说话人提取之间可能存在目标不匹配。

鉴于上述性能提升，后续实验中仅进一步探究 xi-vector。

	目标说话人提取系统	训练集说话人提示	SDRi（dB）	SI-SDRi（dB）	PESQ
混合语音	-	-	0	-0.001	2.01
现有技术	SpEx[18]	✓	16.3	15.8	-
	Spex+[12]	✓	17.2	16.9	-
	Speakerbeam [8]	✓	9.7	-	-
	SpeakerBeam + DC [8]	✓	10.9	-	-
	DPRNN-Spe-IRA[19]	✗	17.6	17.3	3.43
	WASE [20]	✓	17.0	-	-
	SpExsc[16]	✓	18.6	18.4	-
	SpExpc[16]	✓	18.8	18.6	-
ours	xi-TSE	✓	18.6	18.0	3.65
	xi-LDA-TSE（32D）	✓	19.2	18.8	3.71
	xi-TSE + DM	✓	19.7	19.1	3.75
	xi-LDA-TSE（32D） + DM	✓	**19.9**	**19.4**	**3.78**

表 3. 在 WSJ0- 2mix 测试集上，将所提出的系统与现有技术进行比较

所提出的目标说话人提取系统与现有技术的比较 现在，我们将所提出的目标说话人提取系统与 WSJ0-2mix 上八种表现最优的最先进水平系统进行比较。整体比较结果见表 3。表格上半部分列出了直接引用自文献的已发表结果，最后四行则对应我们所提系统的实验结果。

此处 – 表示原论文未报告相应结果。为确保比较公平，同时报告了说话人提示所用的训练数据：表示使用的数据集与我们的基线模型完全相同，表示使用了某个大型数据集。

从上述结果可以观察到，我们的 xi-LDA-TSE（32D）系统优于所有现有方法，有时优势还十分显著。特别是，据作者所知，SpExpc 是现有工作中最具竞争力的方法。我们的 xi-LDA-TSE（32D）系统在所有评估指标上都稳定地超越了它。此外，动态混合（DM） [21] 数据增强通过即时创建新的混合语音，带来了显著的性能提升。表 3 的最后一个要点是，我们性能最佳的系统（结合 xi-LDA-TSE（32D）和 DM）实现了 WSJ0-2mix 上已报道的最佳 TSE 性能。

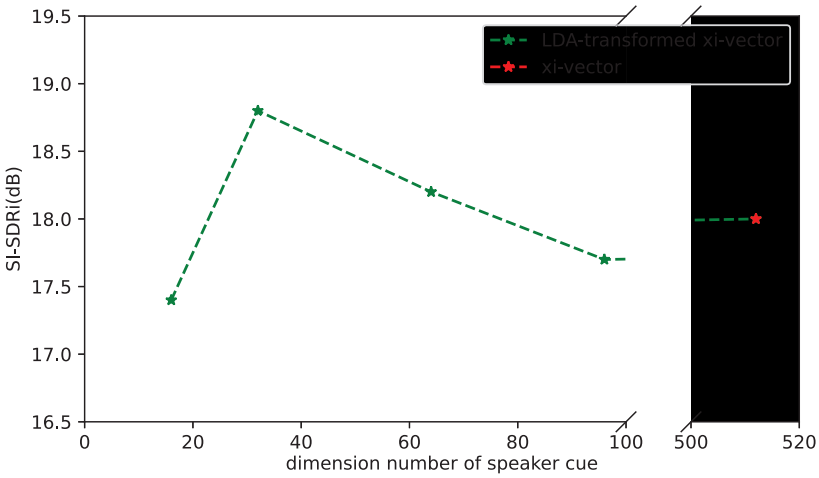

图 4. xi-vector 的 LDA 维度与相应 TSE 在 SI-SDRi 上性能之间的关系

LDA 变换说话人提示的最佳维度如表 2 所示，与说话人提示相关的 LDA 维度会影响相应 TSE 系统的性能。因此，最后我们通过测试更多不同维度的说话人提示，研究维度所带来的影响。图 4 展示了 SI-SDRi 随 LDA 定义的维度数量变化的情况。从图中可以看出，基于 xi-LDA 的说话人提示的最佳维度为 32。根据 *sklearn* 库提供的指标 解释方差比，我们发现这种说话人提示保留了 xi-vector 中约 82% 的总变异。

6 结论

本研究探讨了目标说话人提取中的说话人提示问题。与大多数使用说话人嵌入、主要关注可区分性的现有 TSE 研究不同，我们提出提取类别可分性明确的说话人提示。具体而言，我们采用稀疏 LDA 变换说话人嵌入作为新的说话人提示。为验证这一方案的有效性，我们基于相同架构构建了多个 TSE 模型，这些模型的区别仅在于所使用的说话人提示。基准数据集 WSJ0-2mix 上的实验结果证实了该方案的有效性，表明 LDA 变换的说话人提示带来了显著的性能提升。此外，在采用不同说话人嵌入的 TSE 模型中，该方案均能持续带来性能提升，这表明该方案不依赖于嵌入形式。针对最佳 LDA 维度的进一步实验表明，兼顾可分性与泛化能力的恰当说话人提示能够带来更好的 TSE 性能。此外，我们最佳的 TSE 系统优于现有最先进水平的 TSE 系统，并在 WSJ0-2mix 上实现了 SI-SDRi 和 PESQ 的新纪录。我们认为这些实证结果颇具启发性。未来的研究将评估该方案在更大规模语音分离数据集上的表现，并进一步加以改进，以便实际部署。

参考文献

1. Wang, D. 和 Brown, G.J.：计算听觉场景分析：原理、算法与应用。**计算听觉场景分析：原理、算法与应用**，1–395（2006） 2. J. R. Hershey、Z. Chen、J. Le Roux 和 S. Watanabe：深度聚类：用于分割与分离的判别式嵌入，载于：ICASSP 2016，第 31–35 页（2016） 3. Subakan, Cem、Ravanelli, Mirco、Cornell, Samuele、Bronzi, Mirko 和 Zhong, Jianyuan：注意力就是你所需要的一切：语音分离。载于：ICASSP 2021，第 21–25 页（2021） 4. Shahar Lutati、Eliya Nachmani 和 Lior Wolf：SepIt：逼近单通道语音分离的性能上限，载于：Interspeech 2022 论文集，第 5323–5327 页（2022） 5. Quan Wang、Hannah Muckenhirn、Kevin Wilson、Prashant Sridhar、Zelin Wu、John R. Hershey、Rif A. Saurous、Ron J. Weiss、Ye Jia 和 Ignacio Lopez Moreno：VoiceFilter：通过说话人条件频谱掩蔽实现目标语音分离，载于：Interspeech 2019 论文集，第 2728–2732 页（2019） 6. Quan Wang、Ignacio Lopez Moreno、Mert Saglam、Kevin Wilson、Alan Chiao、Renjie Liu、Yanzhang He、Wei Li、Jason Pelecanos、Marily Nika 和 Alexander Gruenstein：VoiceFilter-Lite：用于端侧语音识别的流式目标语音分离，载于：Interspeech 2020 论文集，第 2677–2681 页（2020） 7. Eskimez, Sefik Emre、Yoshioka, Takuya、Wang, Huaming、Wang, Xiaofei、Chen, Zhuo 和 Huang, Xuedong：个性化语音增强：新模型与全面评估。载于：ICASSP 2022，第 356–360 页（2022）

8. Žmolíková, Kateřina、Delcroix, Marc、Kinoshita, Keisuke、Ochiai, Tsubasa、Nakatani, Tomohiro、Burget, Lukáš 和 Černocký：SpeakerBeam：用于语音混合信号中目标说话人提取的说话人感知神经网络，**IEEE《信号处理精选主题期刊》**，第 13 卷，第 800–814 页（2019） 9. Xu, Chenglin、Rao, Wei、Chng, Eng Siong 和 Li, Haizhou：采用幅度与时域频谱近似损失优化说话人提取神经网络，载于：ICASSP 2019，第 6990–6994 页（2019） 10. Tingle Li、Qingjian Lin、Yuanyuan Bao 和 Ming Li：Atss-Net：通过基于注意力的神经网络实现目标说话人分离，载于：Interspeech 2020 论文集，第 1411–1415 页（2020） 11. Fan-Lin Wang、Yu-Huai Peng、Hung-Shin Lee 和 Hsin-Min Wang：双路径滤波器网络：面向语音分离的说话人感知建模，载于：Interspeech 2021 论文集，第 3061–3065 页（2021） 12. Meng Ge、Chenglin Xu、Longbiao Wang、Eng Siong Chng、Jianwu Dang 和 Haizhou Li：SpEx+：完整的时域说话人提取网络，载于：Interspeech 2020，第 1406–1410 页（2020） 13. Delcroix, Marc、Ochiai, Tsubasa、Zmolikova, Katerina、Kinoshita, Keisuke、Tawara, Naohiro、Nakatani, Tomohiro 和 Araki, Shoko：利用时域 Speakerbeam 提升目标语音提取中的说话人区分能力，载于：ICASSP 2020，第 691–695 页（2020） 14. Zifeng Zhao、Dongchao Yang、Rongzhi Gu、Haoran Zhang 和 Yuexian Zou：端到端说话人提取中的目标混淆：分析与应对方法，载于：Interspeech 2022 论文集，第 5333–5337 页（2022） 15. K. A. Lee、Q. Wang 和 T. Koshinaka：用于说话人验证的 Xi-Vector 嵌入，**IEEE《信号处理快报》**，第 28 卷，第 1385–1389 页（2021） 16. Wupeng Wang、Chenglin Xu、Meng Ge 和 Haizhou Li：采用说话人-语音交叉注意力网络的神经说话人提取，载于：Interspeech 2021 论文集，第 3535–3539 页（2021） 17. A. M. Martinez 和 A. C. Kak：主成分分析与线性判别分析，**IEEE《模式分析与机器智能汇刊》**，第 23 卷，第 228–233 页（2001） 18. Chenglin Xu、Wei Rao、Eng Siong Chng 和 Haizhou Li：SpEx：多尺度时域说话人提取网络，**IEEE/ACM《音频、语音与语言处理汇刊》**，第 28 卷，第 1370–1384 页（2020） 19. Y. Luo、Z. Chen 和 T. Yoshioka：双路径循环神经网络：用于时域单通道语音分离的高效长序列建模，载于：ICASSP 2020，第 46–50 页（2020） 20. Y. Hao、J. Xu、P. Zhang 和 B. Xu：WASE：学习何时关注鸡尾酒会环境中的说话人提取，载于：ICASSP 2021，第 6104–6108 页（2021） 21. N. Zeghidour 和 D. Grangier：Wavesplit：通过说话人聚类实现端到端语音分离，**IEEE/ACM《音频、语音与语言处理汇刊》**，第 29 卷，第 2840–2849 页（2021）