边际混合：一种用于多说话人音频的鲁棒说话人验证方法

Jenthe Thienpondt, Nilesh Madhu, Kris Demuynck

IDLab，电子与信息系统系，根特大学 - imec，比利时 jenthe.thienpondt@ugent.be, nilesh.madhu@ugent.be, kris.demuynck@ugent.be

arXiv:2304.03515v1 [eess.AS] 7 Apr 2023

摘要

本文研究多个说话人重叠语音中的说话人验证任务。大多数说话人验证系统的设计都假设给定的音频片段中只有一位说话人。然而，在真实场景中，这一假设并非始终成立。本文表明，当前的说话人验证系统无法稳健地处理说话人明显重叠的音频。为解决这一问题，我们提出边际混合，这是一种简单的训练策略，可轻松应用于现有说话人验证流程，使生成的说话人嵌入能够稳健地应对多说话人音频。与其他方法不同，边际混合无需改动常规说话人验证架构，同时还能取得更好的效果。在基于 VoxCeleb1 构建的多说话人测试集上，与最先进的说话人验证基线系统相比，所提出的边界混合策略平均将等错误率降低了 44.4%。

索引词— 说话人验证，多说话人，mixup，ECAPA-TDNN，ResNet

1. 引言

说话人验证旨在回答两个话语是否源自同一个人的问题。近年来，得益于大量带标签的训练数据 [1, 2] 以及基于深度学习的说话人识别架构 [3, 4]，说话人验证系统取得了令人瞩目的成果。

大多数说话人识别流程都基于这样一个假设：训练和测试语句中只有一位说话人。尽管气泡噪声 [5]等增强技术常用于提升系统对背景说话人的鲁棒性，但我们认为，在多说话人音频场景中，这些系统无法区分不同说话人。

针对多说话人音频中的说话人验证，先前已有一些研究提出了若干方法。最直接的方法是在说话人验证之前进行说话人日志分析 [6]。这些系统在说话人轮次之间时间重叠有限时表现良好，但其设计并未考虑处理说话人之间存在显著干扰的混合音频。在 [7], 目标说话人提取 [8] 与说话人验证采用联合训练方案，以应对显著的说话人重叠。另一种方法见于 [9], , 其架构会在统计池化层之前，将目标说话人的嵌入注入隐藏的帧级特征中。然而，这些方法采用特定任务的架构，或要求对现有说话人验证流程进行大量改造。

为缓解这一问题，我们提出了 margin-mixup，这是一种基于mixup [10]的训练策略，可在多说话人场景下生成显著更稳健的说话人嵌入。Margin-mixup 将批次级的输入波形混合与改进版加性角度边际（AAM） [11]softmax 损失函数的使用相结合，该损失函数已用于当前最先进的说话人验证系统 [3, 12]。边距混合训练策略可以由现有的说话人验证模型采用，无需任何额外计算开销，也不需要对架构进行改动。

本文的组织结构如下：第 2 节概述当前最先进的说话人验证系统和我们的基线模型。第 3 节介绍所提出的边距混合训练策略。随后，第 4 节解释用于验证所提出的边界混合策略的实验设置，而第 5 节分析实验结果。最后，第 6 节给出一些总结性说明。

2. 基线系统

我们验证了三种最先进模型的多说话人验证性能，并将其作为所提出的边距混合训练策略的基线。我们的第一个基线系统是ECAPA-TDNN [3], , 该系统通过采用改进版的压缩-激励（SE） [14], 通道和上下文相关统计池化，以及聚合多层特征，改进了 x-vector [13] 架构。其次，我们采用 ECAPA-CNN-TDNN 模型[4], , 这是一种混合架构，通过在 ECAPA 模型中加入二维卷积干网络，增强其建模频率无关特征的能力。最后，我们考虑 fwSE-ResNet34 [12], , 这是一种基于ResNet 的架构，采用改进的频率维度 SE 层以及位置编码。

本研究得到弗兰德斯研究基金会（FWO）的支持，资助编号为 G081420N和S004923N。

3. 边际混合

当前最先进的说话人验证流程在单人语音音频上表现出色。尽管这些系统通常能够抵御类似嘈杂人声的背景噪声，但我们认为它们无法处理多位说话人造成的显著干扰。这主要是因为大多数说话人识别训练数据集侧重于单人语音音频，因而缺少适用于说话人重叠情况的损失函数。此外，使用基于间隔的损失函数（例如常用的 AAM-softmax）进一步加剧了这一问题。基于间隔的损失函数通过在训练期间对单个目标说话人施加间隔惩罚，提高说话人识别系统的判别性能。然而，这使得基于间隔的损失函数不适合对多个重叠说话人进行建模。

为缓解这一问题，我们提出了边际混合这一训练策略，使说话人嵌入空间能够对重叠说话人进行建模。正如我们所展示的，边际混合可以轻松应用于常规说话人验证模型，无需对说话人嵌入提取器架构做任何改动，因此，与针对特定任务的多说话人验证方法相比，它更加灵活。

3.1. 所提出的边际混合训练策略

边际混合基于 [10], 中提出的 mixup 训练策略。该策略在训练过程中通过对输入特征进行线性插值，从邻近分布中采样输入特征。结合目标标签的相应插值，模型将嵌入空间扩展到包含类间区域。

本文将插值限制在两个说话人之间。为了将 mixup 训练策略应用于说话人识别系统，按如下方式构造插值后的输入波形 $\hat{\mathbf{x}}$：

$$\hat{\mathbf{x}} = \lambda \frac{\mathbf{x}_a}{\|\mathbf{x}_a\|} + (1-\lambda)\frac{\mathbf{x}_b}{\|\mathbf{x}_b\|} \tag{1}$$

其中，λ 表示两个单说话人波形 $\mathbf{x}_a$ 和 $\mathbf{x}_b$ 之间的插值强度。加权插值前，输入波形会进行能量归一化。随后，相应的插值目标标签 $\hat{\mathbf{y}}$ 定义如下：

$$\hat{\mathbf{y}} = \lambda \mathbf{y}_a + (1-\lambda)\mathbf{y}_b \tag{2}$$

其中$\mathbf{y}_a$ 和$\mathbf{y}_b$ 是独热标签向量。我们注意到，mixup 训练最初的目标是使系统在未混合任务上获得更好的泛化能力，并降低标签被破坏的可能性 [10]。然而，我们在 mixup 训练中的主要目标是学习一个能够应对说话人重叠的嵌入空间。

3.2. 间隔惩罚混合

插值后的目标标签$\hat{\mathbf{y}}$ 给基于边际的损失函数带来了问题，因为这类函数旨在对单个目标标签施加边际惩罚。由于当前最先进的说话人验证系统都基于此类损失函数，因此我们在所提出的边距混合训练策略中使用了 AAM-softmax 的改进版本。

AAM-softmax 基于说话人嵌入 $\mathbf{e}_i$ （说话人标签为 i ）与对应类别中心 $\boldsymbol{W}_i$ （ $\boldsymbol{W} \in \mathbb{R}^{D\times N}$ ）之间的余弦距离。 D 和 N 分别表示嵌入维度和训练说话人的数量。此外，在说话人嵌入与目标类别之间的夹角上施加边距惩罚，以便在说话人类别周围形成更紧密的边界，从而提升说话人验证性能。

在边界混合训练期间，边际惩罚 m 被应用于混合输入语句的嵌入与$\mathbf{e}_{a,b}$ 之间的角度，该语句包含两位说话人 a 和 b 及其对应的类别中心 $\boldsymbol{W}_i$，并根据插值值 λ 进行加权：

$$\hat{\theta}_i = \begin{cases} \theta_i + \lambda m, & \text{if } i = a \\ \theta_i + (1-\lambda)m, & \text{if } i = b \\ \theta_i, & \text{else} \end{cases} \tag{3}$$

采用公式 3 中给出的边际惩罚后，边际混合损失函数 L 随后定义为：

$$L = \lambda \log \frac{e^{s\,cos(\hat{\theta}_a)}}{\sum_{j=1}^{N} e^{s\,cos(\hat{\theta}_j)}} + (1-\lambda)\log \frac{e^{s\,cos(\hat{\theta}_b)}}{\sum_{j=1}^{N} e^{s\,cos(\hat{\theta}_j)}} \tag{4}$$

^^其中， θ_a 和 θ_b 分别表示多说话人输入嵌入$\mathbf{e}_{a,b}$与 a 和 b 的类别中心之间施加边距惩罚后的角度。超参数 s 是一个缩放因子，用于优化训练期间的梯度流。

我们遵循 mixup [10] 原论文的做法，通过从参数 α 和 β 均为 0.2 的贝塔分布中采样，确定插值权重 λ。这样可以确保特征插值主要集中于其中一段话语，因为我们不希望影响常规单说话人测试话语上的说话人验证性能。

4. 实验设置

4.1. 多说话人测试数据集

为了分析针对多说话人音频场景提出的边距混合训练策略，我们基于原始 VoxCeleb1 [1] （Vox1-O）测试集构建了一个自定义测试集 Vox1-M。重叠说话人集合 Vox1-M 的构建方式是在 Vox1-O 语句中加入一条来自 VoxCeleb1 未使用训练分区的干扰语句。相应的信噪比（SNR）均匀采样。

在0-5dB之间。如有必要，我们会重复干扰说话人的话语，以使其与目标说话人的话语完全重叠。由于域内话语混合与低信噪比水平相结合，最终的多说话人测试集颇具挑战性。所有模型均使用VoxCeleb2 [2] 的训练分区进行训练，并将通过等错误率指标进行评估。

4.2. 训练配置

我们的ECAPA-TDNN 基线模型遵循 [3] 中定义的相同结构，隐藏特征维度为1024，并在帧级特征提取器的末端额外添加了一个核大小为 3、膨胀因子为 5 的 SE-Res2Block。fwSE-ResNet34和ECAPA-CNN-TDNN 模型对应于 [4] 中介绍的架构。有关这些模型的更多详细信息，请参阅随附的论文。

在训练过程中，我们随机裁剪 2 秒的音频，以防止过拟合。此外，我们基于 MUSAN 库 [5]（加性噪声、音乐和 babble）或 RIR 语料库 [15]（混响）应用一种随机增强。我们的输入特征由 80 维对数 Mel 滤波器组组成，窗口长度和步长分别为 25ms 和 10ms。随后，我们应用 SpecAugment [16]，随机遮蔽 0-10 个连续频率 bin 和 0-5 个时间帧。最后，我们沿时间维度对对数滤波器组能量进行均值归一化。AAM-softmax 损失函数的边际惩罚 m 和缩放因子 s 分别设为 0.2 和 30。

为了优化说话人嵌入提取器，我们使用 Adam [17] 优化器，权重衰减设为 2e-4。采用循环学习率（CLR） [18] 策略，并使用 *triangular2*衰减策略，学习率的最小值和最大值分别为 1e-8 和 1e-3。周期长度设为 130K 次迭代，批量大小为 128。

在初始训练阶段之后，我们进行大间隔微调（LM-FT） [19]，以促使说话人内嵌入更加紧凑，并增大说话人间的距离。间隔惩罚和裁剪长度分别提高至 0.5 和 5s。微调期间禁用所有数据增强。CLR 的最大学习率降至 1e-5，周期长度缩短至 60K。

训练完说话人嵌入提取器后，我们通过计算注册说话人嵌入与目标说话人嵌入之间的余弦距离，对测试试验进行评分。随后，我们采用 top-1000 自适应 s 归一化 [20, 21]，其中冒名者队列由 VoxCeleb2 中训练语句的平均说话人嵌入组成。

4.3. 边际混合配置

我们在初始训练阶段和微调训练阶段均采用第 3 节所述的边距混合训练策略，其中插值权重 λ 从贝塔分布中采样，α 和 β 均设为 0.2。

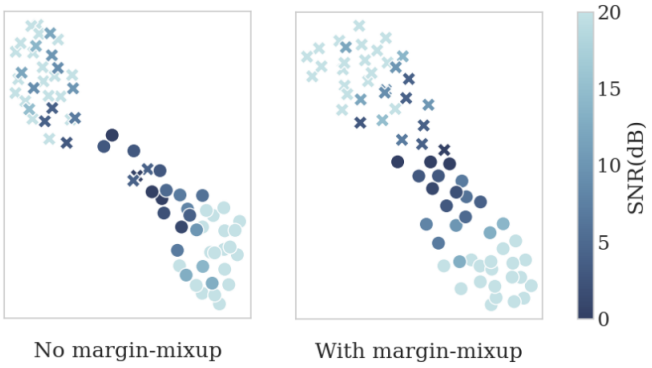

图1。来自两位说话人的话语的 UMAP降维 [22] 嵌入向量（以圆圈和叉号标记表示），与经过不同信噪比混合的样本一起展示，分别来自未使用（左）和使用（右）边际混合训练的模型。请注意，使用边际混合训练的模型中的嵌入空间对说话人之间的过渡状态有更好的表征能力。

系统	M-M	等错误率(%)	
		Vox1-O	Vox1-M
ECAPA-TDNN	-	0.76	22.31
	✓	**0.74**	**12.42**
ECAPA-CNN-TDNN	-	**0.66**	22.15
	✓	0.68	**12.13**
fwSE-ResNet34	-	**0.63**	21.11
	✓	0.66	**11.85**

表1。边际混合（M-M）策略在 Vox1-O 和多说话人 Vox1-M 测试集上的性能分析。

5. 结果

表 1 概述了将所提出的边界混合策略应用于第 2 节所述基线模型的影响。当像 Vox1-M 测试集那样在测试试验中混入另一名说话人时，我们注意到所有基线模型的性能都显著下降。这支持了我们的假设：大多数说话人验证设置所采用的单说话人假设，会导致嵌入无法对显著的说话人重叠进行建模。然而，尽管边界混合对单说话人测试集 Vox1-O 的性能影响可以忽略不计，但在 Vox1-M 上，等错误率相较于基线系统平均提升了 44.4%。这表明，所提出的边距混合训练策略能够显著帮助构建一个成功对说话人之间的过渡状态进行建模的嵌入空间，如图 1 所示，同时不会影响单说话人场景下的性能。

为了更好地理解边距混合训练策略各个组成部分的影响，

	方法	Vox1-O	Vox1-M
	fwSE-ResNet34	0.66	11.85
A	无混合边界	0.77	13.15
B	无 mixup 损失	0.74	17.23
C	仅输入 mixup	0.78	17.65

表 2。边际混合各组成部分的消融研究。

我们在表 2 中进行了消融研究。基线系统是使用边际混合训练的 fwSE-ResNet34 模型。请注意，该表中的所有消融实验仍采用输入 mixup，如第 3.1 节所述。在实验 *A* 中，我们训练模型时不混合 AAM-softmax 的边距，并通过在公式 3 中设置 $\lambda = 1$，仅对原始说话人施加边距惩罚。不混合边距会导致 Vox1-O 和 Vox1-M 上的性能均有所下降，这表明要在多说话人设置中充分发挥 AAM-softmax 损失函数的作用，恰当地混合边距惩罚至关重要。通过在公式 4 中设置 $\lambda = 1$，不应用 mixup 损失，因此也不再要求模型显式检测两个说话人，此时我们观察到 Vox1-M 上的性能大幅下降，相对等错误率增加了 45.4%。实验 *C* 更进一步，在公式 3 和 4 中均设置 $\lambda = 1$，结果两个测试集上的性能都进一步下降。值得注意的是，Vox1-M 上的性能仍优于表 1 中不使用边际混合的 fwSE-ResNet34 基线模型。我们推测，这是因为输入仍与另一位说话人的音频混合，可将其视为常规数据增强，使训练条件更接近 Vox1-M 的多说话人测试设置。

系统	Vox1-O	Vox1-M		
		0dB	2dB	0-5dB
基线	**0.63**	33.67	23.11	21.11
$\alpha,\ \beta = 0.1$	0.64	19.12	13.12	12.82
$\alpha,\ \beta = 0.2$	0.66	18.64	12.57	11.85
$\alpha,\ \beta = 0.4$	0.74	17.32	11.63	11.12
$\alpha,\ \beta = 0.8$	0.94	15.23	10.62	10.31
$\alpha,\ \beta = 1$	1.12	**14.84**	**9.44**	**8.93**

表3。对边界混合策略中用于采样插值权重 λ 的贝塔分布的不同 α 和 β 参数影响的分析。

随后，我们在表3中分析了第3.2节所述贝塔分布的 α 和 β 参数对插值强度 λ 的影响，该插值强度用于边界混合训练期间的采样。我们观察到，在不同信噪比水平下，标准单说话人测试集Vox1-O与多说话人测试集Vox1-M上的性能之间存在权衡。在 $\alpha,\ \beta = 1$时均匀采样概率为 λ 的情况下，一个

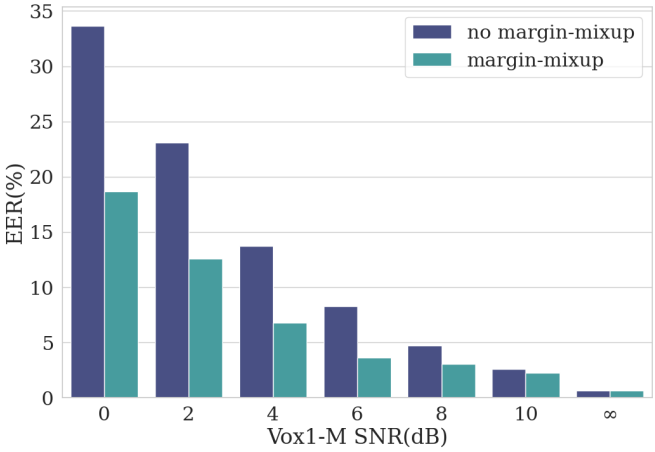

图2。柱状图展示了在不同信噪比下，边距混合训练策略对多说话人Vox1-M测试集上评估的 fwSE-ResNet34 模型的影响。

在*0dB* 条件下观察到显著的性能提升，而在常规Vox1-O测试集上则出现相应的性能下降。这表明，训练期间更激进的混合条件会在严重信噪比测试场景中产生更具鲁棒性的嵌入向量，但代价是常规说话人验证性能下降。如 $\alpha,\ \beta = 0.2$的情况所示，在多说话人场景中仍能获得更具鲁棒性的嵌入向量，同时Vox1-O上的性能没有出现显著下降。

最后，为了分析信噪比值对重叠说话人验证的影响，我们使用图2中的fwSE-ResNet34 基线模型，在固定信噪比值的情况下评估了Vox1-M测试集的多个版本。正如预期，信噪比值会严重影响说话人验证性能，信噪比与等错误率之间呈反向相关。我们注意到，在低信噪比值的多说话人设置中，边际混合的影响最大。我们推测，这主要是由于在较高信噪比场景下，基线模型越来越倾向于将混合说话人视为背景噪声。

6. 结论

本文介绍了边际混合（margin-mixup），这是一种使说话人嵌入在多说话人音频设置中更加稳健的训练策略。与其他方法不同，边界混合策略无需对说话人验证流程进行架构上的修改，即可适应多说话人场景，同时在这一具有挑战性的条件下取得显著的性能提升。在原始 VoxCeleb1 测试集的多说话人版本上，使用所提出的边界混合策略训练我们的基线模型，使等错误率相较于基线性能平均实现了44.4%的相对提升。未来，我们计划采用比余弦距离更合适的相似度度量来扩展边际混合，从而增强多说话人建模能力。

7. 参考文献

[1] Arsha Nagrani、Joon Son Chung和 Andrew Zisserman，《VoxCeleb：大规模说话人识别数据集》，载于 *Interspeech 2017* 论文集，2017 年，第 2616–2620 页。

[2] Joon Son Chung、Arsha Nagrani和Andrew Zisserman，“VoxCeleb2：深度说话人识别”，载于 *Interspeech2018*会议论文集，2018 年，第 1086–1090 页。

[3] Brecht Desplanques、Jenthe Thienpondt和Kris Demuynck，《ECAPA-TDNN：时延神经网络中用于说话人验证的强调通道注意力、传播与聚合》，发表于 *Interspeech2020*会议论文集，2020 年，第 3830–3834 页。

[4] Jenthe Thienpondt、BrechtDesplanques和Kris De- muynck，“在时延神经网络中整合频率平移不变性并在二维残差网络中引入频率位置信息以增强说话人验证”，载于*Interspeech2021*会议录，2021年，第2302–2306页。

[5] David Snyder、Guoguo Chen 和 Daniel Povey，《Musan：音乐、语音和噪声语料库》，*arXiv* 预印本 *arXiv:1510.08484*，2015年。

[6] David Snyder、Daniel Garcia-Romero、Gregory Sell、AlanMcCree、DanielPovey 和Sa njeevKhudanpur，《使用 x 向量进行多说话人对话的说话人识别》，发表于*ICASSP2019*，2019年，第5796–5800页。

[7] Wei Rao、Chenglin Xu、Eng Siong Chng 和 Haizhou Li，《用于多说话人说话人验证的目标说话人提取》，发表于*Interspeech 2019* 论文集，2019年，第1273–1277页。

[8] Marvin Borsdorf、Chenglin Xu、Haizhou Li 和 Tanja Schultz，《在目标说话人存在和缺席的情况下，对单说话人和双说话人语音进行通用说话人提取》，发表于*Interspeech 2021* 论文集，2021年，第1469–1473页。

[9]Ahmad Aloradi、Wolfgang Mack、Mohamed Elmin- shawi 和 E. A. P. Habets，“多说话人环境中利用时序特征融合进行说话人验证”，载于*EUSIPCO 2022*，2022 年。

[10] Hongyi Zhang、Moustapha Ciss´e、Yann N. Dauphin 和 David Lopez-Paz，“mixup：超越经验风险最小化”，载于*ICLR 2018*，会议论文集，2018 年。

[11] Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou，“ArcFace：用于深度人脸识别的加性角度间隔损失”，载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议，2019年，第4685–4694页。

[12] Jenthe Thienpondt、Brecht Desplanques 和 Kris De- muynck，“通过利用语言信息解决跨语言说话人验证中的分数偏移问题”，载于*ICASSP 2022*，2022 年，第 7187–7191 页。[13] David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur，“X向量：用于说话人识别的鲁棒深度神经网络嵌入”，载于*ICASSP 2018*，2018 年，第 5329–5333 页。[14]Jie Hu、Li Shen 和 Gang Sun，“挤压与激励网络”，载于*2018 IEEE/CVF Conferenceon Computer Vision and PatternRecognition*，2018 年，第 7132–7141 页。

[15] Tom Ko、Vijayaditya Peddinti、Daniel Povey、Michael L. Seltzer 和 Sanjeev Khudanpur，“面向鲁棒语音识别的混响语音数据增强研究”，载于*ICASSP 2017*，2017 年，第 5220–5224 页。

[16] Daniel S. Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D. Cubuk 和 Quoc V. Le，“SpecAugment：一种用于自动语音识别的简单数据增强方法”，载于 *Interspeech 2019* 论文集，2019 年，第 2613–2617 页。

[17] DiederikKingma 和 Jimmy Ba，《Adam：一种随机优化方法》，*ICLR* 论文集，2015 年。

[18] Leslie N. Smith，《用于训练神经网络的循环学习率》，载于*2017*年*IEEE*冬季计算机视觉应用会议，2017 年，第 464–472 页。

[19] Jenthe Thienpondt、Brecht Desplanques 和 Kris De- muynck，《idlab voxsrc-20参赛方案：大间隔微调以及基于质量感知分数校准的 DNN 说话人验证》，载于*ICASSP 2021*，2021 年，第 5814–5818 页。

[20] Pavel Matejka、Ondˇrej Novotn´y、Oldˇrich Plchot、Lukas ˇ Burget、Mireia Diez 和 Jan Cernock´y，《多语言说话人识别中的分数归一化分析》，载于*Interspeech 2017* 论文集，2017 年，第 1567–1571 页。

[21] Sandro Cumani、Pier Batzu、Daniele Colibro、Claudio Vair、Pietro Laface 和 Vasileios Vasilakakis，“面向实际应用的说话人识别方法比较”，载于*Proc.Interspeech 2011*，2011 年，第 2365–2368 页。

[22] Leland McInnes、John Healy、Nathaniel Saul 和 Lukas Großberger，“Umap：统一流形近似与投影”，*Journalof Open SourceSoftware*，第 3 卷，第 29 期，第 861 页，2018 年。