The Newsbridge - Telecom SudParis

VoxCeleb 说话人识别挑战赛 2022 系统描述 n

Yannis Tevissen[1,2]*, Jérôme Boudy*[1]*, Frédéric Petitpont*[2]

[1] SAMOVAR、Telecom SudParis、巴黎综合理工学院、法国[2] Newsbridge、法国yannis.tevissen@telecom-sudparis.eu

摘要

本文介绍了我们团队在 VoxCeleb 说话人识别挑战赛 2022（VoxSRC 2022）说话人分离赛道中使用的系统。我们的解决方案围绕一种全新的语音活动检测算法组合而设计，该组合充分利用了多个系统的优势。我们提出了一种新颖的多流方法，其决策协议基于分类器熵。

我们将这种方法称为多流语音活动检测，并将其与标准基线说话人分离嵌入、聚类和重新分段结合使用。通过这项工作，我们成功证明，只要采用强大的基线方法，并仅对语音活动检测进行改进，就能取得接近最先进水平的结果。

索引词：说话人分离、语音活动检测

1. 引言

说话人分离的任务，相当于回答在给定音频场景中“谁在何时说话？”这一问题。说话人日志是几乎所有现代语音处理应用中的关键环节，尤其是那些后续会使用语音识别算法的应用。

如 [1] 和 [2] 所示，说话人日志方法多种多样，但其原理往往相同。首先，从音频文件中分离出同质语音片段，然后对这些片段进行聚类，以确定每个片段由哪位说话人说出。在过去十年中，说话人日志的性能和鲁棒性得到了显著提升，部分原因在于神经网络的发展，如今它们能够处理大量数据，尤其是新的语音嵌入特征 [3], [4]。

按照这一算法模式，几乎所有说话人日志方法的前端分析都包括确定媒体中何时有人声。这一必要步骤称为语音活动检测（VAD）。

语音活动检测方法也多种多样[5]–[8]，但在说话人日志预处理过程中，选择合适的算法往往被忽视。

VoxCeleb说话人识别挑战赛（VoxSRC）的说话人分离赛道 4 专注于分析脱口秀节目、讨论、辩论和访谈等在线内容。在这项挑战赛中，我们决定研究语音活动检测系统对说话人分离结果的影响。

2. 系统描述

为了提升说话人分离的性能，我们选择从预处理入手，尤其关注语音活动检测。我们实现了一种多流方法，其决策协议基于分类器熵。

获得语音活动检测结果后，我们使用了标准的 VBx 系统。该系统包含一个标准的 x向量提取器，后接贝叶斯隐马尔可夫模型聚类。随后对说话人日志结果重新分段，以准确处理重叠区域。

我们的工作从测试多种语音活动检测算法开始，其中包括 pyannote 2.0 和 GPVAD 系统。

2.1. 语音活动检测

我们的工作首先测试了多种语音活动检测算法，其中包括 pyannote 2.0 和 GPVAD 系统。

2.1.1. Pyannote 2.0 语音活动检测

该方法完全基于神经网络，因为它采用了 [7] 中所述的 PyanNet架构。梅尔频率倒谱系数用作二元分类器的输入特征。该方法能够适应领域不匹配，关于循环神经网络的详细介绍见 [7]。本实验使用了开源的 pyannote 模型。该模型已在 AMI数据集 [9] 上预训练。

2.1.2. GPVAD

为实现更稳健的语音活动检测，[8] 的作者提出了一种神经网络方法，该方法结合了卷积层和循环层，并采用弱监督训练方案。在 Audioset数据集上训练后，该方法能够检测 517 种声音类别，其中包括语音。随后，大幅缩减了数量庞大的标签，得到一个语音/非语音二元分类器。

2.1.3. 多流语音活动检测

我们发现，pyannote 2.0 和 GPVAD 的语音活动检测结果具有互补性：前者似乎能产生更稳定的结果，而后者似乎更适用于嘈杂的音频。因此，我们决定实现一种方案，以充分利用这两种语音活动检测系统的优势。

两个语音活动检测系统并行运行。应用了在 VoxConverse 开发集上优化的阈值，然后执行决策协议。

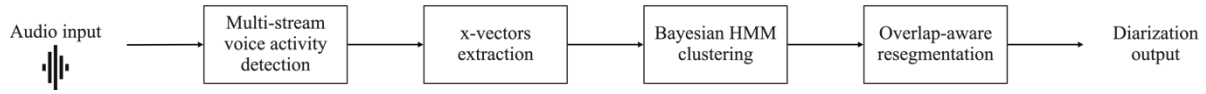


图 1：系统描述

表 1：系统使用的阈值。所有阈值均在 *VoxConverse* 开发集上进行了优化

方法	起始	结束	最短时长 on	Min. 关闭时长
Pyannote 语音活动检测	0.767	0.713	-	-
GPVAD	0.010	0.010	-	-
多流 VAD	-	-	0.182	0.501
重新分段	0.537	0.724	0.410	0.563

我们引入了一种基于熵的决策协议，用于动态选择语音活动检测分类器。对于每个分类器，都会在滑动的 250毫秒窗口上计算局部熵。对于每个窗口，按照以下公式 (1) 计算每个二元分类器的熵$h_{k,\ i}$ 。

$$h_{k,i} = -P(Speech|o_{k,i})\ log_2\ P(Speech|o_{k,i}) \quad (1)$$
$$-P(Non\ Speech|o_{k,i})\ log_2\ P(Non\ Speech|o_{k,i})$$

其中，i 是时间索引，k 是观测值的分类器索引，观测值记作 $o_{k,\ i}$。

随后，我们为每个窗口选择熵最低的分类器，从而得到多流语音活动检测结果，如图 3 所示。

两个语音活动检测器在 Nvidia T4 GPU 和 Intel Cascade Lake CPU 上并行运行，单个文件的最长推理时间为 14 秒。

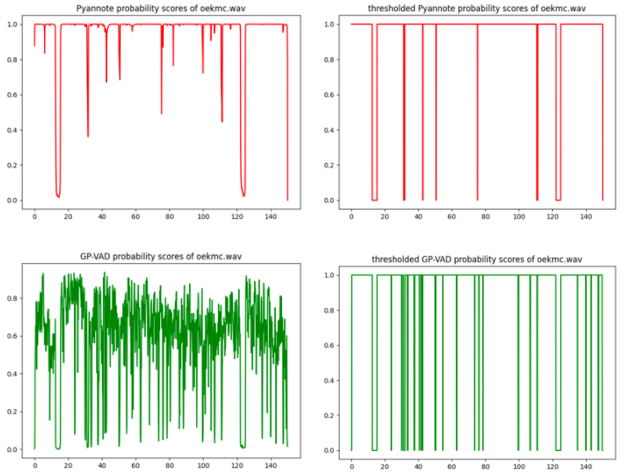


图 2： *pyannote 2.0*（红色）和 *GP-VAD*（绿色）的原始及阈值化语音活动检测分数

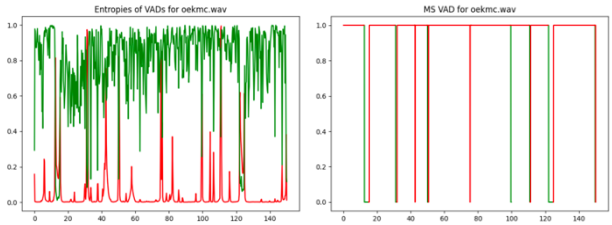


图 3： 两个语音活动检测分类器的局部熵（左）以及多流语音活动检测输出（左）

2.2. 语音嵌入

X向量[4] 提取基于 64 个梅尔滤波器组，这些滤波器组经过 ResNet101 [10] 网络处理，生成 256 维语音嵌入。该嵌入器使用 VoxCeleb 1 和 2 [11]进行训练。如[12], 所示，聚类前还会应用概率线性判别分析（PLDA）来降低维度。

2.3. 聚类

我们采用了 [13] 中介绍的贝叶斯隐马尔可夫模型聚类方法，每个说话人对应一个状态，并以凝聚层次聚类作为初始化。

2.4. 重新分段

此前的所有组件都无法处理重叠语音片段，因此我们加入了神经网络重新分段[14] 来恢复这些区域。这使我们将漏检语音方面的错误减少了一半。

3. 结果与分析

我们在 VoxConverse 0.3 测试集上评估了所提出的系统，并按照与挑战赛评测规则一致的 250 毫秒容差区间报告详细结果。

同时报告说话人日志错误率（DER）和杰卡德错误率（JER）[15]，以及漏检语音（MS）、误报（FA）和说话人混淆（SC）的百分比。

作为对比，我们还报告了使用 pyannote 2.0 神经网络说话人日志处理流程获得的结果。

表 2： *VoxConverse* 上的说话人分离结果
测试集

方法	MS	FA	SC	DER	JER
pyannote 说话人分离	2.29	1.31	4.36	7.96	43.38
pyannote 语音活动检测 + VBx	3.09	0.79	**2.78**	6.66	**29.61**
含重新分段	**1.55**	**1.65**	2.98	**6.18**	29.73
GP-VAD + VBx	3.78	2.30	3.68	9.76	31.82
带重新分段	**1.55**	**1.65**	3.5&	6.88	30.44
多流语音活动检测 + VBx	3.05	1.41	3.09	7.54	30.22
带重新分段	**1.55**	**1.65**	3.04	6.39	29.80

对于 VoxSRC 第4赛道，我们报告了三个最佳系统的结果。

表3：在挑战平台上获得的说话人分离结果

方法	DER	JER
挑战赛基线	19.60	41.43
Pyannote 2.0 语音活动检测 + VBx	8.30	31.14
Pyannote 2.0 语音活动检测 + VBx + 重新分段	7.32	30.12
多流语音活动检测 + VBx + 重新分段	**6.62**	**29.01**

虽然我们的所有系统都取得了相近的得分，但在挑战赛中表现最佳的方案并非在 VoxConverse 测试集上表现最佳。这可以用两个语料库之间轻微的领域差异来解释。

使用所提出的系统，我们在挑战赛排行榜上取得了 6.62% 的成绩，与最佳方案仅相差 2%。这表明，仅改进语音活动检测就能显著提升说话人分离得分。此外，我们的 JER 排名为 5th，这可能表明错误在各说话人之间分布得相对均衡。还需要进一步研究，以确定在何种条件下（噪声、说话轮次时长等）我们的多流方法优于仅使用一种语音活动检测的标准说话人分离方法。

3.1. 可复现研究

我们的所有实验、系统和结果均可复现，因为它们主要依赖开源算法。得益于相关作者，语音活动检测系统[1,2] 和我们的基线说话人分离[3]均可直接使用。

4. 结论

今年的挑战赛中，我们选择专注于语音活动检测。我们的工作显著提升了说话人分离得分，并证明改进说话人分离预处理是提升系统稳健性的直接途径。我们希望进一步改进该系统，提升其在各种数据集上的性能。为此，我们需要明确多流方法在哪些场景下最为适用。

5. 致谢

本研究获得了 Newsbridge 公司的全额资助（https://newsbridge.io/）。我们还要感谢 Hervé Bredin 与我们进行了愉快的讨论，并就我们的工作和系统提出了深刻见解。

6. 参考文献

[1] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, 和 O. Vinyals, 《说话人分离：近期研究综述》，*IEEE* 音频、语音与语言处理汇刊，第 20 卷，第 2 期，第 356–370 页，2012 年。

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, 和 S. Narayanan, 《说话人分离综述：深度学习的最新进展》，2021 年。[3] E. Variani, X. Lei, E. McDermott, I. L. Moreno, 和 J. Gonzalez-Dominguez, 《用于小型文本相关说话人验证的深度神经网络》，*ICASSP*，*IEEE* 国际声学、语音与信号处理会议论文集，第 4052–4056 页，2014 年。[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, 和 S. Khudanpur, 《X向量：用于说话人识别的鲁棒深度神经网络嵌入》，*ICASSP*，*IEEE* 国际声学、语音与信号处理会议论文集，第 2018 年 4 月卷，第 5329–5333 页，2018 年。[5] D. K. Freeman, G. Cosier, C. B. Southcott, 和 I. Boyd, 《用于泛欧洲数字蜂窝移动电话服务的语音活动检测器》，*ICASSP*，*IEEE* 国际声学、语音与信号处理会议论文集，第 1 卷，第 369–372 页，1989 年。[6] M. Lavechin, M. P. Gill, R. Bousbib, H. Bredin, 和 L. P. Garcia-Perera, 《端到端域对抗语音活动检测》，国际语音通信协会年会论文集 *INTERSPEECH*，第 2020 3689 卷，2020 3685 -Octob，第 – 页。[7] H. Bredin 等，《Pyannote.Audio：用于说话人分离的神经网络构建模块》，*ICASSP*，*IEEE* 国际声学、语音与信号处理会议论文集，第 2020 年 5 月卷，第 7124–7128 页，2020 年。[8] H. Dinkel, Y. Chen, M. Wu, 和 K. Yu, 《通过弱监督声音事件检测实现真实环境中的语音活动检测》，国际语音通信协会年会论文集 *INTERSPEECH*，2020 年。[9] I. McCowan 等，《AMI会议语料库》，2005 年。[10] K. He, X. Zhang, S. Ren, 和 J. Sun, 《用于图像识别的深度残差学习》，2016 年。[11] A. Nagraniy, J. S. Chungy, 和 A. Zisserman, 《VoxCeleb：大规模说话人识别数据集》，国际语音通信协会年会论文集 *INTERSPEECH*，第 2017 年 8 月卷，第 2616–2620 页，2017 年。[12] F. Landini 等，《VoxConverse挑战赛中的BUT说话人分离系统分析》，2020 年。[13] F. Landini, J. Profant, M. Diez, 和 L. Burget, 《说话人分离中x向量序列的贝叶斯隐马尔可夫模型聚类（VBx）：理论、实现及标准任务分析》，计算机语音与语言，第 71 卷，第 1–39 页，2021 年。[14] L. Bullock, H. Bredin, 和 L. P. Garcia-Perera, 《重叠感知说话人分离：使用神经网络进行重叠语音检测的重新分段》，*ICASSP*，*IEEE* 国际声学、语音与信号处理会议论文集，第 2020 年 5 月卷，第 7114–7118 页，2020 年。[15] O. Galibert, 《重叠语音情况下说话人分离和自动语音识别的评估方法》，国际语音通信协会年会论文集 *INTERSPEECH*，第 1131–1134 页，2013 年 8 月。

[1] https://github.com/pyannote/pyannote-audio [2] https://github.com/RicherMans/GPV

[3] https://github.com/BUTSpeechFIT/VBx