ClusterVAD：基于预训练神经表征的无阈值无监督语音活动检测

Karl El Hajal[1,2],Mathew Magimai.-Doss[1]

[1]瑞士Idiap研究院[2]瑞士洛桑联邦理工学院

{karl.elhajal,mathew}@idiap.ch

arXiv:2611.33059v1 [eess.AS] 24 Nov 2026

摘要

语音活动检测（VAD）是语音处理流水线中的首要步骤之一，应用范围涵盖转录、说话人分离和语音语言模型。有监督检测器需要带标注数据，因此会沿袭其训练语料库的领域特点和标注惯例。现有的无监督检测器依赖人工设计的声学特征，无法从数据中获益，并且在远场、噪声或非典型语音场景下性能会下降。我们提出了ClusterVAD，这是一种基于冻结的语音基础模型表征构建的无监督VAD。该方法只需在一小组未转录语音上拟合一次，便可对表征进行聚类，并利用响度这一声学度量将每个簇标记为语音或非语音。在推理时，根据簇分配计算帧级语音概率，并针对每段录音拟合一个双状态隐马尔可夫模型，从而实现无阈值检测。在涵盖会议、远场和噪声录音、普通话及构音障碍语音的基准测试中，ClusterVAD在停顿级参考标注上的检测错误最低，在说话人轮次级参考标注上也具有竞争力。这些结果表明，语音基础模型本身能够捕捉语音与非语音之间的区别，无需训练检测器即可利用这一能力。

1 引言

语音活动检测（VAD）是语音和音频处理中的首要步骤之一。其目标是检测音频信号中的语音片段，并为语音语言模型等系统中的后续处理提供信息（Rabiner 和 Sambur，1975年；Sharma 等人，2022年；Landini 等人，2022年；Bredin，2023年；Kahn 等人，2020年；He 等人，2024年；Nguyen 等人，2023年）。因此，语音活动检测效果不佳会对语音技术的性能产生负面影响。例如，假阳性可能导致语音识别系统生成错误文本，而假阴性则可能导致信息丢失，进而影响后续处理。自电话通信兴起以来，语音活动检测一直是活跃的研究领域，至今仍是一个难题。语音活动检测方法大致可分为两类：（a）无监督语音活动检测和（b）有监督语音活动检测。

无监督方法通常是知识驱动的。最简单的方法是将能量、过零率、频谱形状或基频等特征与阈值进行比较（Rabiner 和 Sambur，1975年；Ramírez 等人，2004年；Ghosh 等人，2011），或者对一段录音的各帧拟合双高斯模型，并将均值较大的高斯分布作为语音模型（Bimbot 等人，2004年）。这类方法在干净环境下表现良好，但在噪声环境中可能失效。基于统计模型的检测器则分别对语音和噪声的分布进行建模，并应用似然比检验（Sohn 等人，1999年）。近期的方法通过片段级信噪比估计（Tan 等人，2020年）或零频滤波（Sarkar 等人，2022年）对声源和声道的知识进行编码。

有监督检测器则在带标签帧上训练分类器，涵盖深度信念网络和循环网络（Zhang andWu，2013年；Hughes andMierle，2013年）到卷积模型（Jia 等人，2021年；Silero Team，2024年）以及说话人分离中使用的多说话人分割网络（Bredinand Laurent，2021年；Plaquet and Bredin，2023年）。有监督模型会随着带标注数据的增加而改进；当测试数据与其训练数据匹配时，其性能优于无监督检测器，而在领域不匹配的情况下，无监督检测器更加可靠（Tan 等人，2020年）。与此同时，知识驱动的无监督检测器无法同样有效地利用更多数据。

本文旨在通过基于语音基础模型（SFM）的神经表征，调和语音活动检测（VAD）的这两种方向。这些表征通常通过自监督学习（SSL），利用大量未转录语音进行学习。wav2vec 2.0、Hu- 等模型生成的表征，

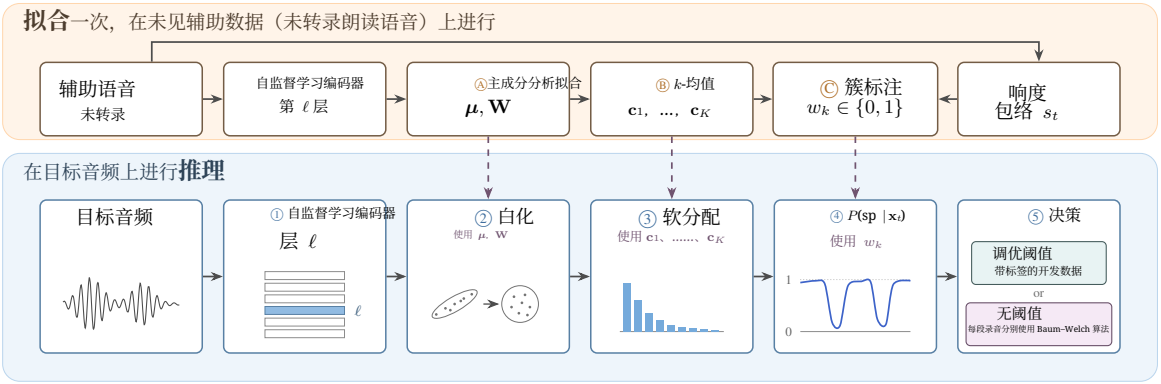


图 1：ClusterVAD 概览。**拟合**在无标签辅助数据上运行一次。**推理**在目标音频上运行。

BERT、WavLM 和 data2vec 2.0（Baevski 等人，2020年；Hsu 等人，2021年；Chen等人，2022年；Baevski 等人，2023年）在多种语音处理任务中，一贯被证明能比传统的知识驱动特征带来更好的系统性能。我们推测，原因之一是这些模型本质上应当能够学会区分语音与非语音：为了从数百或数千小时的异质语音数据中有效提取语音信息，语音基础模型必须在训练过程中隐式学会检测语音。除了以数据驱动的方式学习表征之外，这可能也是语音基础模型表示区别于知识驱动特征的因素之一；对于后者，我们主要应用关于语音的知识。换言之，语音与非语音的区分信息应当以一种下游任务能够轻易解读的形式嵌入所学表征中，无需大量监督或有监督训练。事实上，近期研究（van Niekerk et al.，2023年）已有效地将语音基础模型表示划分为响音、阻音和静音，以进行语音节律建模。后续研究将这种分割方法扩展到把构音障碍语音转换为典型语音，并发现，与传统语音活动检测相比，该方法在构音障碍语音中分离非语音更加可靠（ElHajalet al.，2025年）。

受这些工作的启发，我们提出了 ClusterVAD，这是一种基于冻结的语音基础模型表示的无监督语音活动检测方法（图1）。首先，在一小批辅助未转录语音上执行一次拟合阶段，对表征进行聚类，并使用响度这一声学度量将每个簇标记为语音或非语音。在推理时，ClusterVAD 根据簇分配计算帧级语音概率，并采用无阈值决策来检测语音片段。我们评估了 ClusterVAD 在基准测试中与有监督和无监督检测器进行比较，这些基准测试涵盖不同的领域、语言、录音条件和标注粒度。结果表明，ClusterVAD 在不同领域中表现稳健，并能以精细粒度检测停顿。代码将在论文发表后发布。

2 所提出的方法

ClusterVAD 将语音活动检测视为对使用语音基础模型提取的特征帧进行的无监督分类，该模型称为自监督学习编码器（见图 1）。它由拟合步骤（Ⓐ– Ⓒ）和推理步骤（①– ⑤）组成。拟合步骤在辅助未转录语音上运行一次，而推理步骤随后可在任意目标音频上运行。拟合过程构建一个带标签的码本：Ⓐ使用主成分分析对白化辅助数据的自监督学习帧，Ⓑ 使用 k-均值进行聚类，并且Ⓒ根据每个簇中帧的响度，将其标记为语音或非语音。在推理阶段，① 目标录音的自监督学习帧被 ② 白化并 ③软分配给各个簇；④ 再根据簇标签对分配结果进行边缘化，为每个帧得到一个语音概率，随后 ⑤ 通过决策将其转换为语音片段。

2.1 拟合

冻结的自监督学习编码器将波形映射为帧序列 $\mathbf{x}_1, \ldots\ldots, \mathbf{x}_N \in \mathbb{R}^D$，每 20 毫秒一帧，其中 N 为帧数，D 为特征维度。特征从单个中间层 ℓ中提取。辅助数据的帧构成码本数据 $\mathcal{D}$。

Ⓐ **PCA 白化。** 自监督学习层具有各向异性：少数幅值较大的维度对每帧的范数贡献不成比例地大（附录 B），这可能使基于距离的帧间比较偏向这些维度。

因此，ClusterVAD 会对每帧进行中心化，将其投影到码本数据的前 d 个主成分上，并将保留的成分缩放至单位方差，

$$\tilde{\mathbf{x}}_t = \mathbf{\Lambda}^{-1/2}\,\mathbf{V}^{\top}(\mathbf{x}_t - \boldsymbol{\mu}), \tag{1}$$

其中 $\boldsymbol{\mu}$ 是码本数据的均值帧，而 $\mathbf{V}$ 和 $\mathbf{\Lambda}$ 分别为其协方差矩阵的 d 个最大特征向量和特征值。

B◯ **k均值码本。** 对码本数据的白化帧进行 K均值聚类，可得到质心 $\mathbf{c}_1, \ldots, \mathbf{c}_K$。码本会为任意帧分配一个关于各簇的后验概率，

$$p(k \mid \mathbf{x}_t) = \frac{\exp\big(\cos(\tilde{\mathbf{x}}_t, \mathbf{c}_k)/T\big)}{\sum_{j=1}^{K} \exp\big(\cos(\tilde{\mathbf{x}}_t, \mathbf{c}_j)/T\big)}, \tag{2}$$

其中 $\cos(\cdot,\cdot)$ 表示余弦相似度，而温度参数 T 控制分配的软硬程度。

C◯ **簇标注。** 需要将各簇标注为语音或非语音。ClusterVAD 使用辅助音频的响度包络 $s_t \in [0,1]$ 对各簇进行标注。响度包络是一种衡量音节突出度的声学度量：伽马音调滤波器组的频带包络驱动调谐至音节 θ 节律（5 Hz）的阻尼振荡器，随后对能量最高的频带输出求和（Räsänen 等人，2018）。元音和浊辅音的响度通常较高，而非语音部分的响度较低。簇 的平均响度为

$$\bar{s}_k = \frac{\sum_{t \in \mathcal{D}} p(k \mid \mathbf{x}_t)\, s_t}{\sum_{t \in \mathcal{D}} p(k \mid \mathbf{x}_t)}. \tag{3}$$

这些 K 值 $\{\bar{s}_k\}$ 通过一维二均值聚类（采用两个中心的 Lloyd算法）被划分为低响度组和高响度组，中心初始化为最小值和最大值 $\bar{s}_k$，并且每个簇只计数一次。低组标记为非语音（$w_k = 1$），高组标记为语音（$w_k = 0$）。响度仅在辅助码本数据上计算；推理期间不需要计算响度。图2 展示了一个编码器的划分结果。

2.2 推理

在辅助数据上完成拟合后，便可对任意目标音频运行检测。

1◯– 2◯**特征提取和白化。** 同一个冻结的编码器和层 ℓ 将输入波形映射为帧 $\mathbf{x}_t$，并使用在 A◯ 中拟合的统计量，根据式 1 对每一帧进行白化。

3◯ **软分配。** ClusterVAD 计算每个白化帧与每个 K 质心之间的余弦相似度，并对这些相似度进行 Softmax 运算，为该帧给出其在各簇上的后验概率 $p(k \mid \mathbf{x}_t)$（式 2）。

4◯**语音概率。**一帧的语音概率是C◯中被标记为语音的簇所分配的后验概率之和，

$$P(\mathrm{sp} \mid \mathbf{x}_t) = \sum_{k=1}^{K} (1 - w_k)\, p(k \mid \mathbf{x}_t), \tag{4}$$

其中对簇标签的后验概率进行了边缘化。对于较长的录音，将其划分为重叠窗口进行处理，并使用汉宁窗权重对窗口级语音概率取平均。

5◯ **决策。** ClusterVAD 将 $P(\mathrm{sp} \mid \mathbf{x}_t)$ 以两种方式划分为语音片段和非语音片段。

调优阈值。 当 $P(\mathrm{sp} \mid \mathbf{x}_t)$ 超过阈值时，该帧被标记为语音；阈值和最短片段时长（即挂起时间）则根据有标注开发数据进行调优。

无阈值决策。ClusterVAD 也可以针对每段录音拟合决策，无需调优阈值。设 z_t 为 $P(\mathrm{sp} \mid \mathbf{x}_t)$ 的对数几率。双状态隐马尔可夫模型包含状态 $q_t \in \{\mathrm{ns}, \mathrm{sp}\}$，以及高斯发射 $\mathcal{N}(z_t; m_q, v_q)$，每个状态各有一个均值和一个方差，并配有 2×2 转移矩阵。发射概率根据 $\{z_t\}$ 的分布进行初始化，其中非语音高斯分布取较低百分位数，语音高斯分布取较高百分位数。随后，Baum–Welch 算法（Rabiner，1989年）根据该录音估计高斯参数和转移概率；自转移起到挂起时间的作用。在计算发射概率之前，将 z_t 截断到 $[m_{\mathrm{ns}}, m_{\mathrm{sp}}]$，以确保决策随得分单调变化。维特比解码会找出最可能的状态序列；每个状态中的连续帧构成语音片段和非语音片段。采用无阈值决策时，无需调优决策参数：码本及其簇标签来自未转录辅助数据，而决策则针对每段录音进行拟合（图3）。我们展示了采用这两种决策的结果。

3 实验设置

3.1 实现

自监督学习编码器。 我们从 11 个模型中选取冻结的基础版和大型编码器，并在开发数据上选择其层（附录A.1）。我们使用表现最好的三个模型

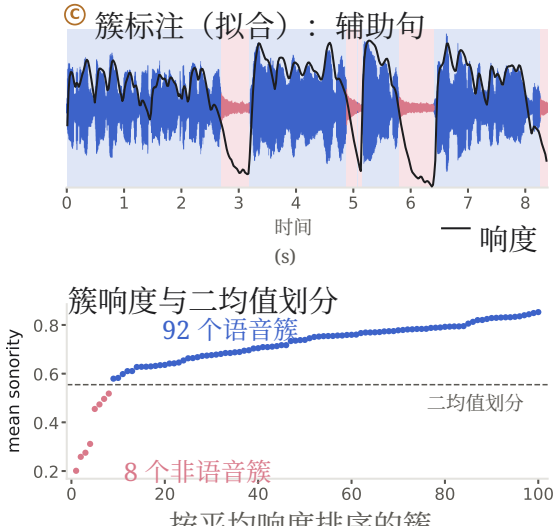
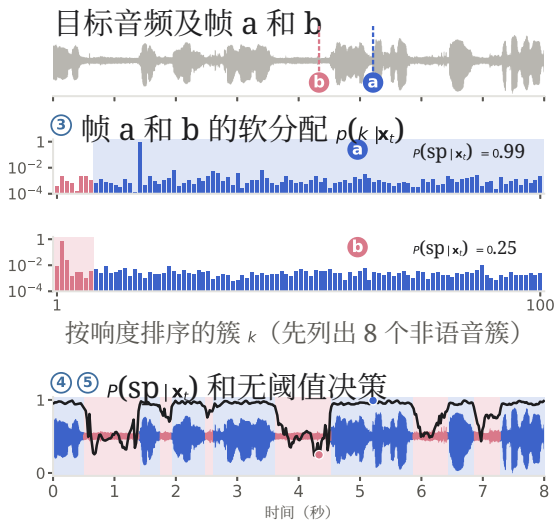


图2：OLIVE-J 的簇标注。上：一个辅助句及其响度，每帧按其所属簇的标签着色，语音或非语音。下：各簇的平均响度及二均值划分。



对一段 8 秒长的 AMI 会议录音进行推理。③：一个语音帧和一个非语音帧的软分配（对数尺度）。④–⑤：语音概率曲线，以及以颜色阴影标示的无阈值决策。

：HuBERT Large（Hsu 等人，2021年）第 10 层、data2vec 2.0 Base（Baevski 等人，2023年）Transformer 输入层（第 0 层），以及 OLIVE-J（El Hajal 和 Magimai-Doss，2026年）第 4 层。OLIVE-J 是 OLIVE 的仅掩码变体，与波形解码器联合训练；借助该解码器，我们还可以从帧和质心合成音频，以检查这些簇。[1]

码本。码本数据由从 LJSpeech 中随机选取的 1,000 条话语组成（Ito 和 Johnson，2017年），对应于由单个说话人朗读的 1.8 小时英语语音，而评测语料库中均不包含该说话人。未使用任何转录文本。我们保留 $d = 128$个主成分，使用 $K = 100$ 个簇（k均值）以及式 $T = 0.1$ 中的2。响度包络使用 20 个伽马音调滤波器组，频率范围为 50 至 7,500 Hz；每个频带采用一个频率为 5 Hz 的阻尼振荡器，并选取能量最高的 10 个频带（Räsänen 等人，2018）。图2展示了由此得到的簇标注。附录C对每项选择进行了消融。

推理。录音以 5 秒为窗口、2.5 秒为步长进行处理。随后，采用第2节介绍的两种决策之一，将语音概率转换为片段。调优阈值决策将语音概率与阈值进行比较，并移除短于最短时长的语音片段和非语音片段。无阈值决策则针对每段测试录音进行拟合：Baum–Welch 算法执行 10 次迭代，初始高斯分布取自 $\{z_t\}$ 的第 10 和第 90 个百分位数，方差为其四分之一，初始状态概率相等，自转移概率为 0.95。

3.2 基线方法

我们将 ClusterVAD 与无监督和有监督检测器进行比较。无监督检测器依赖信号处理：rVAD（Tan 等人，2020年），先进行两轮降噪，再根据基频和能量特征检测语音；de Jong 和 Wempe（2009）的停顿检测器，对强度轮廓进行阈值判定；WebRTC VAD（Google，2011），对子带能量应用高斯混合模型；以及基于帧能量阈值的检测器。有监督检测器是使用带标注语音训练的神经网络：pyannote 3.0 分段模型（Plaquet 和 Bredin，2023年），执行逐帧幂集多类分类；Silero VAD（SileroTeam，2024年），一种在大型多语言语料库上训练的卷积循环网络；MarbleNet（Jia 等人，2021年），一种由时间通道可分离卷积构成的逐帧多语言网络；以及 INA SSL-VAD（Pelloin 等人，2026年），一种使用冻结的 data2vec 2.0 Base 编码器的特征训练的分类器。后者是 ClusterVAD 的有监督对应方法：两者都使用冻结的自监督学习特征，但 INA SSL-VAD 使用带标注数据训练其分类器。附录E 介绍了各个基线的训练数据和配置。

[1]音频样本：karlhajal.github.io/ClusterVAD

数据集	语言	领域	噪声（信噪比）	标注	调优	测试文件	小时
AMI 头戴式麦克风混音	英语	会议、近讲	–	停顿级（强制对齐）、轮次级（人工）	开发集	16	9.1
AMI 远场	英语	会议，远场	–	停顿级（强制对齐），轮次级（人工）	开发集	16	9.1
AliMeeting	普通话	会议，远场	–	停顿级（强制对齐），轮次级（人工）	开发集	20	10.8
Buckeye	英语	访谈	–	停顿级（人工校正对齐）	开发集	68	10.6
Torgo	英语	构音障碍语音	–	停顿级（人工对齐）	留一说话者法	662	0.6
DIHARD III	英语、中文	11个领域	–	说话人分离标签	两倍	259	33.0
VoxConverse	英语	网络视频	–	说话人分离标签	开发集	232	43.5
AMI +MUSAN	英语	会议，近讲 20 至 0 dB		停顿级（强制对齐）	开发集	16 × 4	36.2
LibriVAD	英语	朗读语音	20 至 −5dB	停顿级（强制对齐）	验证集	14,148	30.2

表 1：评测基准。测试文件：测试集中的音频文件数量；小时：它们的总时长。

3.3 评测

数据集。我们在基准测试上评估语音活动检测的性能，这些基准在领域、语言、说话人、录音条件和噪声方面各不相同（表 1）。AMI（Carletta 等人，2006年）和 AliMeeting（Yu 等人，2022年）分别是英语和普通话会议语料库。AMI 的评估既包括近讲头戴式耳机混音，也包括远场麦克风录音；AliMeeting 则使用远场录音进行评估。两者均提供两种粒度不同的参考标注，详见下文。Buckeye（Pitt 等人，2005）包含带有人工校正词对齐的访谈录音。Torgo（Rudzicz 等人，2012年）涵盖非典型语音，其中包含由构音障碍说话人朗读提示语的录音，并附有人工作成的音素对齐。DIHARD III（Ryant 等人，2021年）和 VoxConverse（Chung 等人，2020年）是说话人分离基准，两者均使用各自的说话人分离标签进行评分：前者覆盖十一个领域，从有声读物和电话通话到临床访谈和餐厅对话；后者则包含多说话人网络视频。我们还通过在 AMI 中加入 MUSAN噪声（Snyder 等人，2015年），并将信噪比设为 20 至 0 dB，进一步评估对噪声的鲁棒性（Tan 等人，2020年；Dinkel 等人，2021年）；同时也在 LibriVAD（Stylianou 等，2025）上进行评估，该数据集将 LibriSpeech（Panayotov 等人，2015年）的朗读语音与九种噪声类型混合，信噪比为 20 至 −5 dB。

附录 F 详细介绍其划分和参考标注。

参考标注。 数据集中哪些内容算作非语音，取决于其标注惯例，而标注惯例又取决于预期用途。按照 Heldner 和 Edlund (2010)，我们将说话人轮次内的非语音区间称为停顿，将两位说话人之间的非语音区间称为间隙。为语音识别构建的参考标注通常会将语义连贯的片段保留为整体，并填补说话人轮次内部的停顿；而用于说话人分离的参考标注则更为精细：例如，DIHARD 会在时长达到或超过 200 毫秒的停顿处切分说话人片段（Ryant 等人，2021）。一种填补……的参考标注

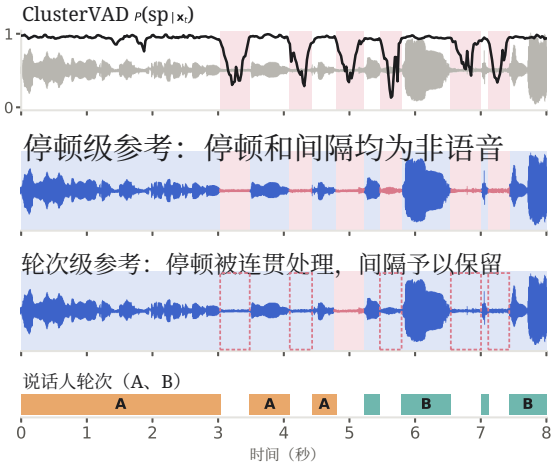


图 4：包含两位说话人的 AMI 片段的停顿级和轮次级参考标注。ClusterVAD（上方）的语音概率在停顿级参考标注中的每个非语音区间都会下降；轮次级参考标注将轮次内的停顿标记为语音，但会保留说话人之间的间隙。

停顿会惩罚那些检测出停顿的检测器：例如，在 AMI 上，人工计时词语几乎完全覆盖了通过强制对齐找到的 0.2–2 秒停顿（附录 G）。有监督检测器会学习其训练标签所采用的惯例，并能将其迁移到其他领域；而无监督检测器则不会继承任何惯例：它会按照特征的分辨率处理停顿；如果目标标注惯例有此要求，就必须将其片段跨越停顿合并。因此，我们区分两类参考标注（图 4）。停顿级参考标注保留每个长度不低于 0.2 秒的非语音区间（包括停顿和间隔）。这类标注通常来自对人工转录文本进行强制对齐。Kraljevski 等人(2015年) 表明，强制对齐标签的准确度与专家标注员的标注相当；此类标签广泛用于语音活动检测研究（Tan 等人，2020年；Bovbjerg 等人，2024年），也用于需要精确边界的说话人分离基准测试（Watanabe 等人，2020年；Horiguchi 等人，2025年）。对于 AMI 和 AliMeeting，我们采用停顿-

	AMI		AMI 远场		AliMeeting						
	停顿级	轮次级	停顿级	轮次级	停顿级	轮次级	Buckeye	Torgo	DIHARD III	Vox-对话	AMI + MUSAN	LibriVAD
全部语音（简单基线）	51.8	24.3	51.8	24.3	20.5	8.7	37.1	140.6	26.4	11.7	51.8	18.9
有监督												
pyannote 3.0	17.8±2.6	**6.2**±1.2	20.6±2.9	**7.4**±1.5	11.6±2.3	**3.8**±0.8	15.4±1.5	24.4±2.2	**7.9**±0.6	**5.0**±0.4	19.5±1.2	10.4±0.2
Silero	10.8±1.1	7.9±1.9	15.3±2.5	11.0±2.8	17.0±5.8	10.0±4.4	11.2±1.5	17.3±1.8	11.7±1.3	5.7±0.5	16.6±1.7	10.9±0.2
MarbleNet	14.0±1.1	8.9±2.0	21.7±2.9	12.7±2.7	20.8±5.0	9.3±2.9	13.6±1.5	18.1±1.7	18.7±1.7	5.6±0.5	20.6±1.8	12.5±0.2
INA SSL-VAD	10.6±1.0	8.5±1.9	16.7±1.8	10.9±2.1	13.6±1.9	5.4±1.1	9.2±1.4	26.9±2.5	13.0±1.3	5.6±0.5	18.3±1.8	13.6±0.3
无监督												
rVAD	16.8±1.5	11.0±1.7	19.7±2.6	14.4±2.9	14.5±2.5	9.0±2.7	17.4±2.2	23.8±1.9	13.9±0.9	8.1±0.8	24.5±1.6	12.0±0.2
de Jong & Wempe	22.0±4.4	16.3±3.1	40.4±12.6	23.3±6.0	15.6±3.7	6.8±1.9	11.7±2.3	35.4±4.5	16.3±1.7	9.9±1.0	29.9±2.4	16.0±0.2
WebRTC	16.9±2.4	11.9±2.7	19.6±2.3	14.5±3.0	14.1±2.9	6.5±1.7	14.7±2.2	24.2±2.7	17.4±1.5	9.3±0.9	31.0±2.5	13.7±0.1
能量	15.9±1.9	12.1±2.6	24.8±3.8	16.3±2.7	11.5±2.8	6.1±2.3	12.0±2.3	30.3±3.4	13.6±1.3	9.6±0.9	27.9±2.7	14.2±0.1
ClusterVAD，调优阈值												
OLIVE-J	**8.1**±0.8	7.5±1.8	11.0±1.7	9.6±2.3	10.8±2.4	5.9±1.9	**7.4**±1.3		8.6±0.8	5.3±0.5		8.8±0.3
data2vec 2.0	8.8±1.0	9.0±2.2	12.4±1.6	10.7±2.4	9.9±1.7	5.5±1.4	8.0±1.4		9.7±1.0	6.2±0.6		8.3±0.1
HuBERT-L	8.8±0.9	9.3±2.0	11.5±1.4	11.9±2.5	**9.6**±1.6	4.9±1.0	7.6±1.2		8.2±0.7	5.5±0.5		**6.5**±0.1
ClusterVAD，无阈值决策												
OLIVE-J	8.3±0.9	7.2±1.8	11.2±1.7	9.8±2.2	11.6±3.2	6.7±2.6	8.3±1.4		9.0±1.0	5.1±0.5		7.8±0.2
data2vec 2.0	8.7±1.0	8.5±2.2	12.9±1.6	11.4±2.5	11.7±2.8	6.6±2.2	7.6±1.6		10.2±1.0	6.2±0.6		10.5±0.2
HuBERT-L	9.5±1.2	8.4±2.3	**10.9**±1.3	9.5±2.1	9.6±1.5	4.7±1.0	8.7±1.5		10.5±2.0	5.3±0.5		7.3±0.1

表2：基准测试集上的检测错误率（DetER（%），越低越好）。评估得分以95%置信区间报告。粗体：最佳；下划线：次佳。

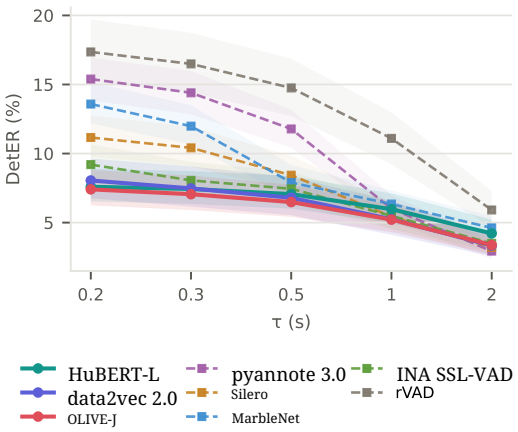


图5：在 Buckeye 数据集上，当参考中将短于 τ 的停顿标记为语音时的检测错误率（DetER，%）（$\tau = 0.2$ 秒：停顿级参考），并附有 95% 置信区间。系统根据各参考进行调优。

级参考标注，来自Horiguchi 等人(2025年）。轮次级参考标注遵循说话人轮次：它们来自人工标注，会跨越一个轮次内的大多数停顿，并保留间隔。

指标。 我们在帧级和片段级评估检测性能。在帧级上，我们报告检测错误率（DetER），即假正（将非语音检测为语音）和假阴性（将语音检-测为非语音）的总时长相对于参考语音时长的比例（Bredin，2017年）。此外，我们还使用受试者工作特征（ROC）曲线，该曲线将真正率与假正率绘制为应用于检-测器的帧级语音分数发生变化时的函数。该曲线下面积（曲线下面积，AUC）概括了这些分数将语音帧与非语音帧区分开的效果，且独-立于任何决策。在片段级上，我们使用事件 F1（Mesaros 等，2016年），即预测语音片段的 F1 分数；如果预测片段与参考片段的起始时间相差不超过 200 毫秒，且结束时间相差不超过 200 毫秒或参考时长的 20%，则认为二者匹配（Dinkel 等，2021年）。

调优协议。 由于标注规范各不相同，说话人分离工具包会针对每个数据集重新调优检测阈值和最短时长（Bredin 和 Laurent，2021年；Bredin，2023年）。我们同样针对每个基准测试的调优数据，对每个系统的决策进行调优（表1），并针对每种参考标注分别调优，以尽量降低检测错误率。调优阈值的搜索范围为 0.02 至 0.98，步长为 0.02；语音和非语音片段的最短时长则从 0、0.1 和 0.2 秒中选择；这三个值联合调优。对于具有离散设置的基线方法（例如 rVAD 阈值或 WebRTC 模式），相应设置也纳入同一搜索过程。最后，为匹配跨越停顿的参考标注，时长合并会将所有短于最短时长的非语音片段重新标记为语音 τ'；针对每种目标参考标注，最短时长从 0 至 5 秒之间的 11 个值中选取。此操作会在任一决策之后应用于每个系统。对于 ClusterVAD 的无阈值决策，τ' 是唯一需要调优的值。

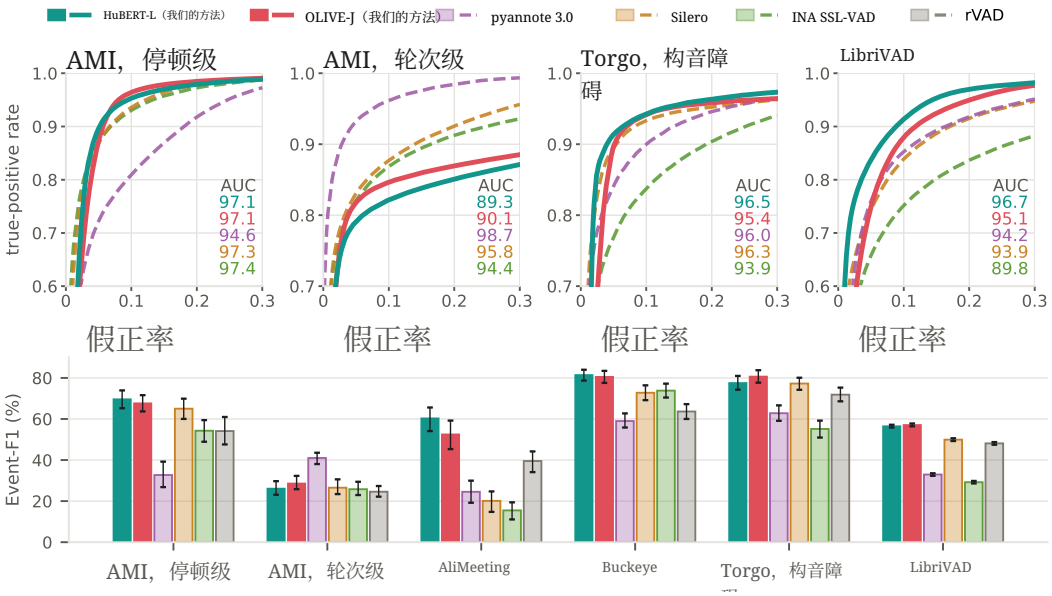


图6：上　方：系统表现有差异时的 ROC曲线和曲线下面积。下方：事件 F1及其 95%置信区间　　s.

4 结果与分析

4.1 检测错误与参考粒度

表 2 报告了十二个基准测试中的检测错误。ClusterVAD 模型在其中七个基准测试上取得了最低错误率，涵盖近讲和远场会议、普通话、访谈、构音障碍语音以及噪声语音。ClusterVAD 的语音概率粒度能够识别单个停顿，因此适用于停顿级参考标注。pyannote 在轮次级参考标注和两个说话人分离基准上表现最强；它使用这些语料库的数据进行训练，因此这些数据属于其适用领域，这也发挥了有监督方法的优势。ClusterVAD 在这些基准上仍表现良好：其错误率均为第二低，且处于 pyannote 的置信区间内。ClusterVAD 模型的表现也优于无监督基线。此外，采用 HuBERT-L 编码器和无阈值决策的 ClusterVAD 在所有基准测试上的表现均优于其有监督对应方法 INA SSL-VAD。采用与 INA SSL-VAD 相同的基础版编码器时，ClusterVAD 使用 OLIVE-J 在十一个基准测试上胜过它，使用 data2vec 2.0 则在九个基准测试上胜过它。

为单独考察参考标注粒度的影响，图5 展示了在参考标注中合并短于 τ 的停顿时，Buckeye 上的检测错误。当参考标注保留短停顿时，ClusterVAD 优于所有基线。随着 τ 增大，有监督检测器的错误率下降，神经系统的结果趋于一致，而 ClusterVAD 的变化最小：它能识别停顿，并通过时长合并使其适应更粗略的参考标注。

4.2 无阈值与调优决策

表 2还比较了 Cluster-VAD 的两种决策。在 OLIVE-J 上，无阈值决策的错误率低于调优阈值，或在每个基准中均处于其报告的置信区间内。对于其他编码器，两种决策主要在 Torgo 和 LibriVAD 上存在差异；这两个数据集中的短语音片段可能导致无阈值决策的拟合不够可靠。

4.3 ROC 分析与片段检测

图6 首先展示了帧级语音分数的 ROC曲线和曲线下面积，这些指标不依赖于决策。在停顿级 AMI和构音障碍语音上，ClusterVAD 的曲线下面积接近最佳有监督检测器的水平。曲线表明，在较低假正率下，有监督检测器能够达到更高的真正率；而在中等假正率下，ClusterVAD 的真正率更高。这些工作点上的差异有助于解释检测错误率的差异。在轮次级 AMI上，ClusterVAD 的分数区分语音与非语音的能力不如 pyannote，因为在参考标注为语音的停顿处，其分数会降低；时长合并无法由 ROC曲线体现，但在决策阶段弥补了这一差异，使 ClusterVAD 在检测错误率方面表现良好。在 LibriVAD 上，采用 HuBERT-L 的 ClusterVAD 具有最高的曲线下面积。

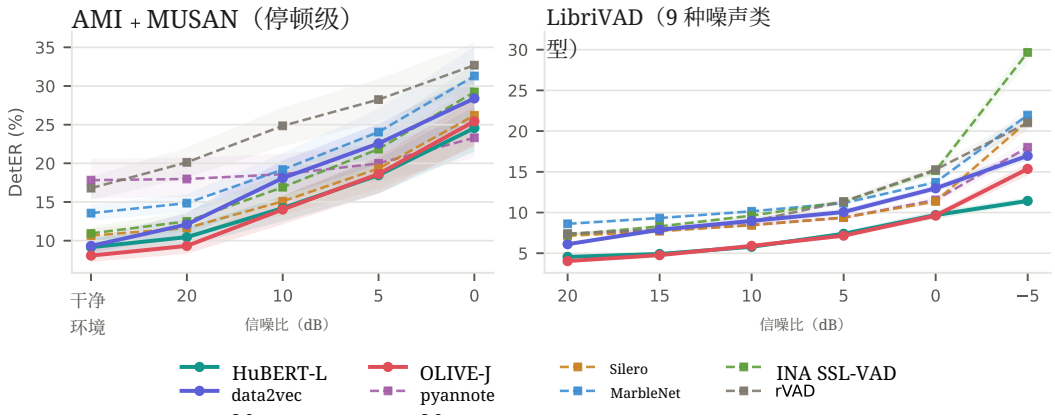


图 7：检测错误率（%）随信噪比变化的情况，以及 95%置信区间。每个系统均使用一个在所有信噪比上调优的决策；对于 AMI + MUSAN，ClusterVAD 使用调优阈值绘制，对于 LibriVAD，则使用无阈值决策绘制。

预测片段的事件 F1 也报告于图6中。除轮次级AMI外，ClusterVAD 在所展示的所有基准测试中均获得了最高的事件 F1，表明其片段边界与参考标注的边界最为接近。

4.4 对噪声的鲁棒性

图 7考察了随着信噪比降低，检测错误如何变化。在 LibriVAD 的所有信噪比下，以及 AMI 上除 0dB 之外的所有信噪比下，ClusterVAD 的错误率均最低 + MUSAN 中，pyan- note 的错误率略低，但仍处于置信区间之内。需要注意的是，对于 SSL 编码器而言，LibriVAD 属于其领域内数据，因为这些编码器是在有声读物语音上预训练的。

5 讨论与结论

在本研究中，我们证明了语音与非语音信息之间的区别会内在地编码在冻结的 SSL 表征中，并且可以通过聚类显式利用这一信息，从而在无需训练检测器的情况下，实现继承这些表征的鲁棒性和泛化能力的语音活动检测。事实上，尽管 PCA 白化、码本和簇标签仅在 1.8 小时的干净的朗读英语语音上拟合一次，且每个编码器所采用的层也只根据一个开发集进行选择，ClusterVAD 仍能跨领域泛化。我们将这一配置原样应用于会议、远场和噪声录音，以及普通话和构音障碍语音；与使用带标注语音训练的检测器相比，它取得了最低或次低的错误率。这凸显了这些表征的鲁棒性。层的选择也并不关键：每个编码器除少数几层外，其余各层的表现都接近最佳层（附录 A.1），这表明语音与非语音之间的区别贯穿整个编码器。最后，使用响度为簇添加标签——这是一种通过声学度量进行的自监督形式——就已经足够：尽管响度本身是一个较弱的检测器，并且在噪声条件下会迅速退化（附录 D），由其标注的簇仍能产生本身具有鲁棒性的语音概率。

在轮次级参考标注上，pyannote 仍然表现最佳。除了属于这些基准测试的领域之外，它还是一个说话人分离模型，能够利用说话人信息区分轮次中的停顿和说话人变化。相比之下，ClusterVAD 仅根据停顿时长进行停顿衔接。未来的工作将研究如何整合说话人信息，以实现说话人感知的停顿衔接。我们还展示了，在各项基准测试中，无阈值决策与调优阈值的效果相当，从而无需调优阈值。然而，在短语音片段上，其可靠性较低，这可能是因为可用于拟合模型的音频较少。未来的工作将探索如何提高其对这类输入的鲁棒性。

总体而言，ClusterVAD 将知识驱动的无监督语音活动检测跨领域泛化能力，与有监督语音活动检测利用数据的能力结合起来，充分利用从大规模未转录语音语料库中学得的冻结自监督学习表征。ClusterVAD 将这些表征、基于无标签辅助数据拟合的小型码本，以及作为声学度量的响度结合起来，实现了在不同领域均可与有监督检测器相媲美的无监督语音活动检测。

局限性

计算效率。ClusterVAD 继承了其自监督学习编码器的计算成本，该成本高于紧凑型检测器。不过，由于编码器只需运行到选定层（附录 H），因此成本仍然有限。例如，data2vec 2.0 变体仅使用卷积特征编码器。尽管如此，基于更轻量表征（例如蒸馏编码器）构建码本，仍有待未来研究。

输入时长。 由于无阈值判定是在每段录音上拟合的，因此在较短的输入上可靠性较低，因为可用于拟合的音频很少；对于某些编码器，其表现也不如调优阈值。

说话人感知暂停合并。 为语音识别构建的参考通常会合并同一说话人语音片段之间的停顿。说话人分离系统能够合并这类停顿，因为它们在检测时会整合说话人信息，而 ClusterVAD 仅根据片段时长进行合并。在流水线中整合说话人信息，可以让 ClusterVAD 调整检测方式，以适应此类参考。

语言。 该码本基于英语语音构建，基准测试仅涵盖英语和普通话。ClusterVAD 无需任何改动即可迁移到普通话，但其在其他语言上的表现仍有待评估。

伦理考量

本研究使用现有研究语料库，并遵循各自的许可协议，未收集任何新数据。语音活动检测器是语音处理流水线的重要组成部分，有助于实现并推动语音语言模型、语音识别器和语音合成系统等的发展。语音技术在许多领域都很有价值，但也可能遭到滥用，例如用于监控或语音克隆。提高人们对这些风险的认识，并推进水印技术和深度伪造检测，对于限制恶意使用至关重要。

致谢

本研究部分得到了瑞士国家科学基金会通过“病理性语音合成（PaSS）”项目提供的支持，资助协议编号为219726。

参考文献

Alexei Baevski、Arun Babu、Wei-Ning Hsu 和 Michael Auli。2023年。利用上下文化目标表征为视觉、语音和语言实现高效的自监督学习。载于 *ICML* 论文集。

Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。2022年。data2vec：用于语音、视觉和语言自监督学习的通用框架。载于 *ICML* 论文集。Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。wav2vec 2.0：语音表征自监督学习框架。载于 *NeurIPS* 论文集。

Frédéric Bimbot、Jean-François Bonastre、Corinne Fredouille、Guillaume Gravier、Ivan Magrin-Chagnolleau、Sylvain Meignier、Teva Merlin、Javier Ortega-García、Dijana Petrovska-Delacrétaz 和 Douglas A. Reynolds。2004年。文本无关说话人验证教程。*EURASIP* 应用信号处理期刊，2004年(4):430–451。

Holger Severin Bovbjerg、Jesper Jensen、Jan Østergaard 和 Zheng-Hua Tan。2024年。恶劣条件下鲁棒个性化语音活动检测的自监督预训练。载于 *ICASSP* 会议论文集。

Hervé Bredin。2017年。pyannote.metrics：用于说话人分离系统可复现评估、诊断和误差分析的工具包。载于 *Interspeech* 会议论文集。

Hervé Bredin。2023年。pyannote.audio 2.1 说话人分离流水线：原理、基准与方案。载于 *Interspeech* 会议论文集。

Hervé Bredin 和 Antoine Laurent。2021年。面向重叠感知重分段的端到端说话人分割。载于 *Interspeech* 会议论文集。

Jean Carletta、Simone Ashby、Sebastien Bourban、Mike Flynn、Mael Guillemot、Thomas Hain、Jaroslav Kadlec、Vasilis Karaiskos、Wessel Kraaij、Melissa Kronenthal等。2006年。AMI会议语料库：预发布公告。载于多模态交互机器学习（*MLMI*）会议论文集。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao等。2022年。WavLM：面向全栈语音处理的大规模自监督预训练。*IEEE*精选信号处理主题期刊，16(6)：1505–1518。

Joon Son Chung、Jaesung Huh、Arsha Nagrani、Triantafyllos Afouras 和 Andrew Zisserman。2020年。发现对话：真实环境中的说话人分离。载于 *Interspeech* 论文集。

Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。2021年。用于语音识别的无监督跨语言表征学习。收录于*Interspeech* 论文集。

Nivja H. de Jong 和Ton Wempe。2009。用于自动检测音节核并测量语速的 Praat 脚本。《行为研究方法》，41(2):385–390。

Heinrich Dinkel、Shuai Wang、Xuenan Xu、Mengyue Wu 和 Kai Yu。2021年。野外语音活动检测：一种使用教师-学生训练的数据驱动方法。*IEEE/ACM*《音频、语音与语言处理汇刊》，29：1542–1555。

David Doukhan、Christine Maertens、William Le Personnic、Ludovic Speroni 和 Reda Dehak。2024年。In-aGVAD：一个具有挑战性的法国电视和广播语料库，已针对语音活动检测和说话人性别分割进行标注。载于*LREC-COLING* 会议论文集。

Karl El Hajal、Enno Hermann、Sevada Hovsepyan 和 Mathew Magimai-Doss。2025年。无监督节奏与语音转换：改进构音障碍语音的自动语音识别。载于*Interspeech* 会议论文集。

Karl El Hajal 和 Mathew Magimai-Doss。2026年。OLIVE：用于语音自监督学习的视图增强潜在预测与波形重建。*arXiv* 预印本 *arXiv:2606.30356*。

Prasanta Kumar Ghosh、AndreasTsiartas 和 Shrikanth Narayanan。2011。利用长期信号变化性实现稳健的语音活动检测。*IEEE* 音频、语音与语言处理汇刊，19(3):600–613。

Google。2011年。WebRTC语音活动检测器。https://webrtc.org；Python接口：https://github.com/wiseman/py-webrtcvad。

Haorui He、Zengqiang Shang、Chaoren Wang、Xuyuan Li、Yicheng Gu、Hua Hua、Liwei Liu、Chen Yang、Jiaqi Li、Peiyang Shi等。2024年。Emilia：用于大规模语音生成的广泛、多语言且多样化的语音数据集。载于*IEEE SLT*会议论文集。Mattias Heldner和Jens Edlund。2010年。对话中的停顿、间隙与重叠。《语音学杂志》，38(4)：555–568。Shota Horiguchi、Naohiro Tawara、Takanori Ashihara、Atsushi Ando和Marc Delcroix。2025年。能否重新利用多说话人自动语音识别语料库进行说话人分离？载于*IEEE ASRU*会议论文集。Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov和Abdel-rahman Mohamed。2021年。HuBERT：通过隐藏单元掩码预测进行语音表征自监督学习。*IEEE/ACM*《音频、语音与语言处理汇刊》，29：3451–3460。Thad Hughes和Keir Mierle。2013年。用于语音活动检测的循环神经网络。载于*ICASSP*会议论文集。Keith Ito和Linda Johnson。2017年。LJ语音数据集。https://keithito.com/LJ-Speech-Dataset/。

Fei Jia、Somshubra Majumdar和Boris Ginsburg。2021年。MarbleNet：用于语音活动检测的深度一维时间通道可分离卷积神经网络。载于*ICASSP*会议论文集。

Jacob Kahn、Morgane Rivière、Weiyi Zheng、Evgeny Kharitonov、Qiantong Xu、Pierre-Emmanuel Mazaré、Julien Karadayi、Vitaliy Liptchinsky、Ronan Collobert、Christian Fuegen 等。2020年。Libri-Light：一个用于低监督或无监督自动语音识别的基准。发表于 *ICASSP* 论文集。

Herman Kamper、Aren Jansen 和 Sharon Goldwater。2017年。用于完全无监督大词汇量语音识别的分段式框架。*ComputerSpeech & Language*，46:154–174。

Ivan Kraljevski、Zheng-Hua Tan 和 Maria Paola Bissiri。2015年。强制对齐语音识别与人工方法在生成参考语音活动检测方面的比较。发表于 *Interspeech* 论文集。

Federico Landini、Ján Profant、Mireia Diez 和 Lukáš Burget。2022年。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）：标准任务上的理论、实现与分析。*Computer Speech & Language*，71：101254。

Annamaria Mesaros、Toni Heittola 和 Tuomas Vir-tanen。2016年。复音声音事件检测的指标。*Applied Sciences*，6(6)：162。

Tu Anh Nguyen、Eugene Kharitonov、Jade Copet、Yossi Adi、Wei-Ning Hsu、Ali Elkahky、Paden Tomasello、Robin Algayres、Benoît Sagot、Abdelrahman Mo- hamed 等。2023年。生成式口语对话语言建模。*Transactionsof the Association forComputational Linguistics*，11：250–266。Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015年。LibriSpeech：基于公版有声读物的自动语音识别语料库。载于*Proc.ICASSP*。Valentin Pelloin、Lina Bekkali、Reda Dehak 和 David Doukhan。2026年。数据选择对法语视听广播音频表征自监督学习的影响。载于*Proc.LREC*。Mark A. Pitt、Keith Johnson、Elizabeth Hume、Scott Kiesling 和 William Raymond。2005年。Buckeye会话语音语料库：标注规范及转录员可靠性测试。《语音通信》，45(1)：89–95。Alexis Plaquet 和 Hervé Bredin。2023年。用于神经说话人分离的幂集多类交叉熵损失。载于 *Proc.Interspeech*。Lawrence R. Rabiner。1989年。隐马尔可夫模型及其在语音识别中的若干应用教程。*Proceedings ofthe IEEE*，77(2)：257–286。

Lawrence R. Rabiner 和 Marvin R. Sambur。1975年。一种确定孤立话语端点的算法。*TheBell System Technical Journal*，54(2)：297–315。

Javier Ramírez、José C. Segura、Carmen Benítez，Á ngel de laTorre 和Antonio Rubio。2004年。使用长期语音信息的高效语音活动检测算法。《语音通信》，42(3–4)：271–287。

Okko Räsänen、Gabriel Doyle 和 Michael C. Frank。2018。将语音预语言阶段分割为类音节单元。《认知》，171：130–150。

Frank Rudzicz、Aravind Kumar Namasivayam 和 Talya Wolff。2012年。针对构音障碍说话人的声学与发音语音TORGO数据库。《语言资源与评测》，46(4)：523–541。

Neville Ryant、Prachi Singh、Venkat Krishnamohan、Rajat Varma、Kenneth Church、Christopher Cieri、Jun Du、Sriram Ganapathy 和 Mark Liberman。2021年。第三届DIHARD说话人分离挑战赛。载于*Interspeech*论文集。

Eklavya Sarkar、RaviShankar Prasad 和 Mathew Magimai-Doss。2022年。使用零频滤波对声源和系统信息建模的无监督语音活动检测。载于*Interspeech* 论文集。

Mayank Sharma、Sandeep Joshi、Tamojit Chatterjee 和 Raffay Hamid。2022年。面向电影和电视节目的现代语音活动检测方法综合实证综述。*Neurocomputing*，494：116–131。

Silero Team。2024年。Silero VAD：预训练的企业级语音活动检测器（VAD）、数字检测器和语言分类器。`https://github.com/snakers4/silero-vad`。

David Snyder、Guoguo Chen 和 Daniel Povey。2015年。MUSAN：音乐、语音和噪声语料库。*arXiv* 预印本 *arXiv:1510.08484*。

Jongseo Sohn、NamSoo Kim 和 WonyongSung。1999年。一种基于统计模型的语音活动检测方法。《*IEEE*信号处理快报》，6(1):1–3。

Ioannis Stylianou、Achintya kr. Sarkar、Nauman Dawalatabad、James Glass 和 Zheng-Hua Tan。2025年。LibriVAD：一个可扩展的开放数据集，包含用于语音活动检测的深度学习基准测试。*arXiv*预印本 *arXiv:2512.17281*。

Zheng-Hua Tan、Achintya kr. Sarkar 和 Najim Dehak。2020年。rVAD：一种基于片段的无监督稳健语音活动检测方法。*Computer Speech &Language*，59:1–21。

Benjamin vanNiekerk、Marc-André Carbonneau 和 Herman Kamper。2023年。用于语音转换的节奏建模。《*IEEE*信号处理快报》，30：1297–1301。

Shinji Watanabe、Michael Mandel、Jon Barker、Em- manuel Vincent、Ashish Arora、Xuankai Chang、San- jeev Khudanpur、VimalManohar、Daniel Povey、Desh Raj 等。2020年。CHiME-6挑战赛：应对未分段录音中的多说话人语音识别。载于 国际日常环境语音处理研讨会论文集（*CHiME*）。

Fan Yu、Shiliang Zhang、Yihui Fu、Lei Xie、Siqi Zheng、Zhihao Du、Weilong Huang、Pengcheng Guo、Zhijie Yan、Bin Ma 等。2022年。M2MeT：2022年ICASSP多通道多方会议转录挑战赛。载于*ICASSP*论文集。

Xiao-Lei Zhang 和 Ji Wu。2013年。基于深度置信网络的语音活动检测。*IEEE* 音频、语音与语言处理汇刊，21(4)：697–710。

A 编码器和层

A.1 编码器和层选择

我们在 11 个冻结编码器（基础版和大型）以及它们的每一层上运行 ClusterVAD：wav2vec 2.0（Baevski 等人，2020年）、XLSR-53（Conneau等人，2021年）、HuBERT（Hsu 等人，2021年）、WavLM（Chen等人，2022年）、data2vec（Baevski 等人，2022年）、data2vec 2.0（Baevski 等人，2023年）和 OLIVE-J（El Hajal andMagimai-Doss，2026年），并根据停顿级参考，在 AMI 开发数据上选择编码器和层。OLIVE-J 的第 4 层、data2vec 2.0 的第 0 层以及 HuBERT Large 的第 10 层取得了最低的三个错误率（表 3）。对于这些编码器，大多数层的表现都接近最佳层，并且每一层都优于在梅尔频率倒谱系数上运行相同流程的结果（图 8）。

编码器	Size	最佳层检测错误率	中位数
OLIVE-J	Base	12 层中的第 [illegible]	8.3	11.8
data2vec 2.0	Base	12 页中的第 [illegible]	9.0	10.2
HuBERT	大型	24 页中的第 1[illegible]	9.3	10.9
data2vec	Base	12 页中的第 9 页	10.1	12.6
data2vec	大型	24 个中的 14 个	10.5	13.9
WavLM	大型	24 个中的 7 个	10.9	13.0
WavLM 基础版+	Base	12 / 12	11.0	13.2
wav2vec 2.0	大型	7 / 24	11.9	13.8
HuBERT	Base	12 个中的 10 个	12.1	13.7
XLSR-53	大型	24 个中的 12 个	12.2	13.5
wav2vec 2.0	Base	12 个中的 9 个	16.7	21.5
39 个梅尔频率倒谱系数	–	–	30.2	–

表 3：各编码器最佳层上的 AMI开发集检测错误率（%）（停顿级参考），以及各层的中位数。

A.2 手工特征

为评估自监督学习表征的贡献，我们在 39 个梅尔频率倒谱系数上原样运行 ClusterVAD流程

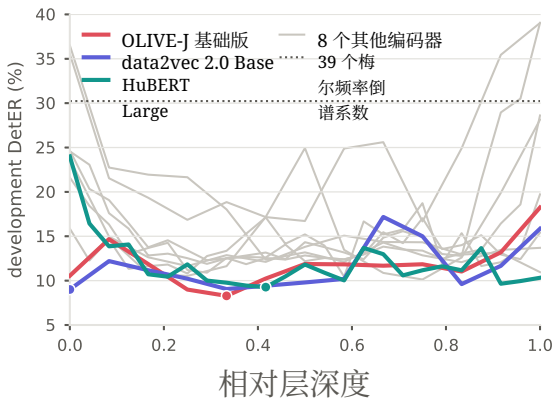


图 8：11 个编码器各层在 AMI开发集（停顿级参考）上的检测错误率（%）。

。结果如表4所示，凸显了表征对该方法的重要性。

	OLIVE-J		data2vec 2.0		HuBERT-L		MFCC	
	thr.	free	thr.	free	thr.	free	thr.	free
AMI，停顿级	7.6	7.7	7.9	7.8	8.2	8.6	29.6	13.9
AMI，轮次级	7.6	7.4	8.4	8.1	9.2	8.3	17.9	10.2
AMI 远场	9.7	9.8	11.1	11.4	10.4	9.9	32.1	18.6
AliMeeting	10.9	11.8	10.6	12.1	10.0	9.9	11.5	19.7
Buckeye	7.4	8.0	6.9	7.1	7.6	8.3	9.4	7.9
VoxConverse	3.2	3.2	3.7	3.6	3.3	3.3	5.9	6.1

表 4：使用自监督学习特征和 39 个梅尔频率倒谱系数的 Cluster-VAD 在开发集上的检测错误率（%），采用调优阈值（thr.）和无阈值决策（free）。

A.3 推理窗口与层组合

表 5 改变推理窗口，并通过平均或拼接来组合三个相邻层。较长的窗口表现与论文中的 5 秒窗口相当，较短的窗口表现略差；组合层并无帮助。

	OLIVE-J	data2vec 2.0	HuBERTLarge
推理窗口			
2 s	9.1	9.2	10.4
5 秒（论文）	8.3	9.0	9.3
10 s	8.2	9.0	8.9
20 s	8.1	9.0	8.7
层			
1 层（论文）	8.3	9.0	9.3
3 层，均值	8.3	9.7	9.9
3 层，拼接	9.8	9.5	9.5

表 5：不同推理窗口及三个相邻层组合在 AMI开发集（以停顿为参考）上的检测错误率（%）。

B PCA 白化

自监督学习特征具有各向异性：少数坐标占据了帧平方范数的很大比例，在 Large编码器中尤为明显（图 9），这可能使余弦相似度偏向这些坐标。PCA 白化，即拟合步骤 A◯，消除了这一影响并降低了开发集误差（表 6）。表 7 展示了保留的主成分所保留的方差比例。

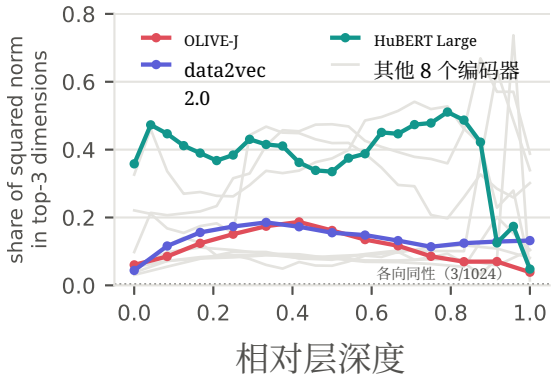


图 9：11 个编码器中，每层特征范数平方由最大的三个坐标所占的比例。

归一化	dev
none	11.8
PCA 白化，$d = 64$	8.6
PCA 白化，$d = 128$	**8.3**

表 6：OLIVE-J 在 AMI开发集上的检测错误率（%）（以停顿级别为参考），未使用和使用 PCA 白化。

编码器（层）	D	$d = 64$	$d = 128$
OLIVE-J（4）	768	74.9	85.9
data2vec 2.0（0）	768	70.9	85.5
HuBERT 大型（10）	1024	66.5	76.8
Base编码器，所有层	768	36–75	54–86
HuBERT Large，所有层	1024	46–89	62–96

表 7：d 个主成分保留的码本数据方差（%）（D：特征维度）。

C 码本消融

表8 逐一测试了 OLIVE-J 在 AMI开发集上的码本。检测错误对码本数据的来源基本不敏感，对其数量从100到

10,000条话语、簇数量在50到200之间，以及分配方式（温度在0.05到0.2之间，余弦或欧几里得距离）也基本不敏感。 K-均值和球形 k-均值是等价的。在不同数据子集和随机种子下，错误率也保持稳定。

因素	设置	停顿级	轮次级
码本数据（1k 条语句）	LJSpeech	**8.3**	**15.2**
	LibriSpeech	8.4	16.0
	AMI 训练集	10.8	15.6
	三者混合	8.9	15.6
LJSpeech 规模	100 条语音	9.3	16.8
	1k 条语音	**8.3**	**15.2**
	10k 条语音	9.1	16.8
簇 K	2	46.5	27.3
	10	11.1	16.4
	50	8.9	15.8
	100	**8.3**	**15.2**
	200	9.0	17.2
	500	9.6	17.9
聚类	k-均值	**8.3**	**15.2**
	球面 k-均值	8.7	16.5
分配	余弦，$T = 0.05$	8.5	**15.0**
	余弦，$T = 0.1$	**8.3**	15.2
	余弦，$T = 0.2$	8.6	15.6
	余弦，$T = 0.5$	9.9	16.8
	欧几里得，$T = 0.5$	8.5	15.3
	欧几里得，$T = 1$	9.6	16.6
随机种子	3 个子集 × 3随机种子（标准差）	8.5 ±0.4	16.2 ±0.6

表8：OLIVE-J的码本消融：AMI开发集上的检测错误率（%）。每个区块仅改变一个设置；粗体表示论文中的配置。

D 簇标注

表 9 比较了用于标注各簇的五种信号：响度、帧能量、SSL特征范数、WebRTC VAD决策以及节奏分组（van Niekerk 等人，2023年）。对于三个编码器，响度的开发集误差均为最低或并列最低，且无需依赖其他检测器。

标注信号	OLIVE-J	data2vec 2.0	HuBERT 大型
Urhythmic分组	20.2 (L5)	12.7 (L7)	12.3 (L6)
WebRTC VAD	9.8 (L3)	10.0 (L0)	9.3（L8）
率	11.8（L9）	9.7（L6）	14.1（L14）
帧能量	8.3 (L4)	9.0 (L0)	9.3 (L10)
响度	10.2 (L4)	10.6 (L3)	11.7 (L23)
特征范数			

表9：在 AMI开发集（停顿级参考）上，各标注信号在其最佳层（括号内）的检测错误率（%）。

将响度作为检测器。 直接将响度用作语音分数时，在干净环境语音上的表现明显劣于 Cluster-VAD，并且在噪声条件下性能下降得更快（表10）。

E 基线方法

每种基线方法均采用默认配置运行，并在每个基准上调整其决策（第 3.3 节）。对于无监督基线方法，搜索范围还包括离散设置。

有监督。 对于 pyannote segmentation 3.0，语音分数等于 1 减去没有说话人处于活动状态的概率；该模型使用 AISHELL、AliMeeting、AMI、AVA-AVD、DIHARD、Ego4D、MSDWild、REPERE 和 VoxConverse 训练。Silero VAD 的训练数据未公开。MarbleNet（帧级多语言 v2.0）使用约 3,900 小时的数据训练，其中包括 AMI、ICSI、LibriSpeech、Fisher 和 Multilingual LibriSpeech。对于 INA SSL-VAD，我们采用维特比解码前的帧概率；其编码器在法国电视和广播数据上预训练，分类器则在 InaGVAD 上训练（Doukhan 等人，2024年）。

无监督。 对于 rVAD，我们使用作者提供的 MATLAB 实现，并搜索其阈值（0.2、0.4 和 0.6）。对于de Jong 和 Wempe （2009），我们搜索最小停顿时长和静音级别（默认值为 0.3 秒，对应 -25 dB，以及 0.1 秒，对应 -20、-25 和-30dB）。对于 WebRTC VAD，我们搜索其四种模式（0 到 3）。对于帧能量检测器，我们搜索按录音和全局缩放的帧对数能量。

F 数据集

本附录详细介绍表1中的基准测试。

会议。 对于 AMI（Carletta 等人，2006年），我们使用标准说话人日志划分和 Landini 等人（2022年）基于词的轮次级参考，以及 Horiguchi 等人（2025年）提供的强制对齐的停顿级参考；远场版本使用麦克风阵列中的第一个麦克风。对于 AliMeeting（Yu 等人，2022年），我们使用Horiguchi 等人（2025年）的停顿级参考，并将原始标注作为轮次级参考。

噪声。 MUSAN 噪声（Snyder 等人，2015年）以 20、10、5 和 0 dB 信噪比添加到 AMI 头戴式麦克风混音中，并在干净音频和各个信噪比条件下汇总的开发集上调优一个决策。

	检测错误率（%，越低越好）				曲线下面积（%，越高越好）			
	AMI	+ MUSAN	LibriVAD		AMI	+ MUSAN	LibriVAD	
	干净环境	0 dB	20分贝	−5 分贝	干净环境	0 dB	20分贝	−5 分贝
响度	20.8±2.7	37.5±5.7	7.4±0.2	18.4±0.5	91.8±1.7	74.7±2.8	96.7±0.1	72.6±0.7
帧能量	15.8±1.7	39.7±6.1	8.2±0.2	19.3±0.4	90.6±2.0	69.2±2.1	93.3±0.1	63.2±0.4
ClusterVAD OLIVE-J	8.1±0.8							
ClusterVAD HuBERT-L	9.2±1.0							

表 10：在每个噪声基准测试中最干净和最嘈杂的条件下，直接用响度和帧能量作为语音评分，以及使用 ClusterVAD 时的检测错误率和曲线下面积（%）。

	AMI		AliMeeting	
	停顿	轮次	停顿	turn
语音占比（%）	65.9	80.4	83.0	92.0
每分钟语音片段数	22.2	5.6	20.2	7.1
片段中位时长（秒）	1.24	3.64	1.77	4.13
每分钟停顿数	14.0	2.0	11.8	2.7
每分钟间隔数	8.1	3.6	8.4	4.3
非语音中位时长 (s)	0.48	1.10	0.37	0.49

表11：AMI 和 AliMeeting 测试集的停顿级和轮次级参考标注统计。停顿用于分隔同一说话人的两个语音片段；间隙用于分隔两位说话人。

	参数	RTF		内存 (GB)	
编码器（层）	（M）	CPU	GPU	内存	GPU
OLIVE-J（0–4）	36.5	0.27	0.0015	4.2	1.7
data2vec 2.0（0）	8.1	0.19	0.0011	3.5	1.6
HuBERT-L（0–10）	139.1	0.56	0.0026	4.4	2.3

表 12：在 30 分钟 AMI 音频上，ClusterVAD 截至所选层的编码器参数、实时因子（RTF，即处理时间与音频时长之比；数值越低，速度越快）和峰值内存；测试使用一个 CPU 线程或一块 RTX 3090 GPU。

对于 LibriVAD（Stylianou 等人，2025年），我们使用 NonConcat 版本，并使用其中型测试集和小型验证集进行调优。

Buckeye。 我们根据人工校正的词对齐（Pitt 等人，2005年）对参与者的语音进行评分，并不对采访者的轮次进行评分；说话人划分采用 Kamper 等人（2017年）。

Torgo。 我们在具备人工音素对齐的三位构音障碍说话人上进行评估（Rudzicz 等人，2012年）。调优时采用留一说话人法：对每位说话人的评分，均使用在另外两位说话人上调优的决策。

说话人分离基准。 对于 DIHARD III（Ryant 等人，2021年），我们使用带有官方参考标注的评测集，并按领域分层后拆分为两折，每折使用在另一折上调优的决策进行评分。对于 VoxConverse（Chung 等人，2020年），我们在开发集上调优，并在测试集上测试。

置信区间。 所有置信区间均为基于测试文件计算的 95% 百分位自助法区间（1,000 次重抽样）。

G 参考标注

表11比较了 AMI 和 AliMeeting 的停顿级与轮次级参考标注。轮次级参考标注会连接一个轮次中的大多数停顿，并保留说话人之间约一半的间隔。因此，它们将更多音频标记为语音，且片段更少、更长。在 AMI 上，轮次级参考标注中的人工计时词语几乎完全覆盖了停顿级参考标注中 0.2–2 秒的停顿，因此，能够检测出这些停顿的检测器会受到惩罚。

H 计算成本

表12 报告了在提取过程于选定层停止时的编码器大小和推理成本。