arXiv:2304.05922v1 [eess.AS] 12 Apr 2023

使用困难类别挖掘和类别间焦点损失的填充词检测

致远 赵宇成[1], 吴立军[1], *Chuanxin Tang*[1],殷达成[2],赵宇成[2],*Chong Luo* [1]

[1]微软亚洲研究院，北京，中国[2]中国科学技术大学，合肥，中国

摘要

“嗯”或“呃”这类填充词在自发语音中很常见。在录音中自动检测并去除这些词很有必要，因为它们会影响语音的流利度、自信度和专业性。先前的研究和我们的初步实验表明，填充词检测面临的最大挑战是，填充词很容易与“a”或“I”等其他难分类别混淆。本文提出了一种新颖的填充词检测方法，通过动态添加辅助类别并应用额外的类别间焦点损失，有效应对这一挑战。辅助类别通过挖掘难分类别，迫使模型显式地对易混淆词进行建模。此外，类别间焦点损失会自适应地调整“填充词”与“非填充词”类别之间的惩罚权重，以处理“非填充词”类别中遗留的其他易混淆词。我们的系统取得了最佳结果，在PodcastFillers数据集上的表现远超其他方法。

索引词— 填充词检测、关键词检测、语音不流畅、困难样本挖掘、焦点损失

1. 引言

没有接受过专业演讲训练的人往往会频繁使用“嗯”或“呃”等填充词 [1, 2, 3]。填充词会让讲话显得不够专业，降低说话者的可信度 [4, 3, 5]。在多媒体创作场景中，例如播客、语音会议、语音访谈、语音消息或视频演示，内容创作者迫切需要一种自动工具来检测并删除填充词，从而节省语音编辑的时间和精力。许多研究致力于填充词检测任务。

已有多种基于自动语音识别（ASR）的填充词检测方法得到研究 [6, 7, 8, 9]。在这些系统中，首先由自动语音识别模型识别语音，然后在转录文本中检测填充词。但大多数自动语音识别系统难以转录填充词，因为这些系统通常使用几乎不含填充词的口语文本语料库进行训练 [1]。此外，一些研究证明，即使在含有填充词的数据集上重新训练专用的自动语音识别系统，其填充词检测性能仍难以超过专门的填充词检测模型 [10], ，更不用说训练和使用自动语音识别模型所需的巨大计算成本。AVC-FillerNet [1]也利用自动语音识别模型，但它们使用自动语音识别来剔除含有已转录词语的片段，从而获取填充词候选片段。

已有研究提出了一些无需自动语音识别的非流利现象检测、社交信号检测或口吃检测方法 [11, 12, 13]。这些方法也能检测填充词或感叹词，后者是填充词的另一种称呼。除了填充词检测之外，它们还可以检测其他声音事件，例如笑声、长时间静默、重复或纠正等。这些方法需要同时检测多种事件，因此需要带有更全面标注的数据，这使得方法设计较为复杂。

最近，也提出了一些主要关注填充词的无需自动语音识别的方法 [2, 1]。Sagnik 等人提出了一种能够检测填充词并将其从语音中去除的系统 [2]。他们使用卷积循环神经网络（CRNN）进行填充词检测。然而，他们发现，该系统难以区分听起来类似“呃”或“嗯”的语音片段。VC-FillerNet [1] 是一种两阶段的无需自动语音识别的填充词检测方法。它首先利用语音活动检测（VAD）模块查找语音区域，然后使用分类模型预测帧级分类结果。

在传统方法中，所有非填充词都被混合在一起，形成一个“未知”类别或“其他”类别。基于先前工作 [2] 和我们的初步实验（第2.1节）的观察，我们发现，填充词很容易与发音类似于填充词的非填充词混淆。在本文中，我们提出了一种新颖的端到端、无需自动语音识别的填充词检测方法，该方法结合了动态辅助类别选择和额外的类别间焦点损失。作为一种困难类别挖掘策略，我们根据排名靠前的假阳性（易混淆）词添加一些辅助类别，以迫使网络对易混淆词进行显式建模。借助新的训练方案，我们只需一个训练循环即可自动选择辅助类别。此外，我们设计了一种类别间焦点损失，可以自适应地控制任意两个类别之间的惩罚权重。在我们的

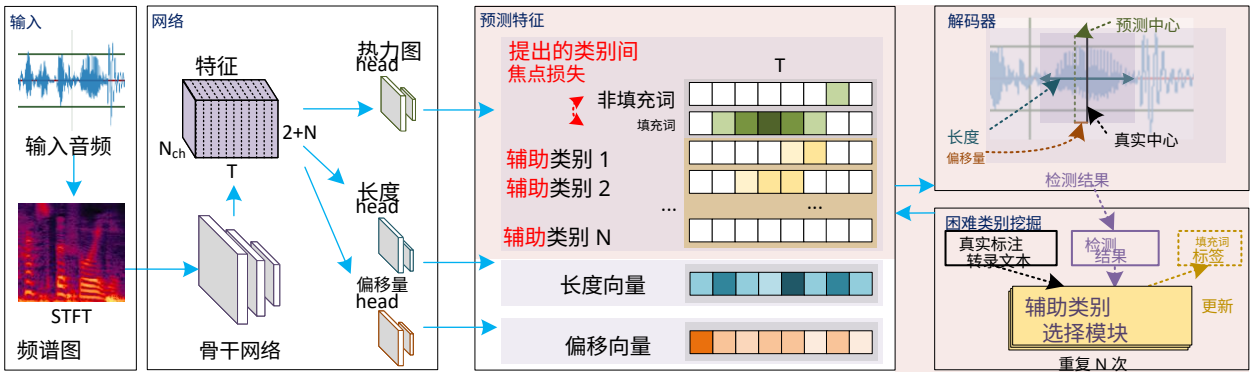


图 1。我们提出的方法概览。困难类别挖掘模块可选择辅助类别。类别间焦点损失应用于“非填充词”和“填充词”类别。

在设置中，我们对“非填充词”类别和“填充词”类别应用类别间焦点损失，以处理“非填充词”类别中除所选辅助类别之外的其他难分类别。我们的系统在 PodcastFillers数据集 [1] 上取得了当前最先进（SOTA）的结果。消融实验表明，我们设计的辅助类别和类别间损失是有效的。

2. 方法

2.1. 初步实验（问题的挑战）

填充词只包含一个音节。有许多单音节词或多音节词的部分，其发音相似。Sagnik 等人 [2]发现，他们的系统难以区分听起来类似于“uh”或“um”的语音片段，例如“a”、“the”或“umbrella”的开头部分。我们通过计算分类结果的混淆矩阵，开展了一项初步实验来研究这一现象。图 2 展示了我们在 PodcastFillers 验证集上的初步结果中，排名靠前的假阳性（FP）分析。本文将容易被误分类的词（类别）称为“易混淆词（类别）”。我们发现，易混淆词大多集中在某些特定词上。值得注意的是，在传统方法中（如上文所述），这些易混淆词会与其他词简单混合，组成一个“未知”或“其他”类别。因此，在这项初步研究中，我们采用焦点损失，通过困难样本挖掘来应对这一问题。尽管如此，根据图 2 所示的分析结果，仍有许多误分类未能消除。这促使我们设计更好的方法和算法来解决这一问题。

2.2. 我们的方法概述

简而言之，我们设计了一种新方案，在训练过程中动态选择辅助类别，并引入一种新的类别间.....

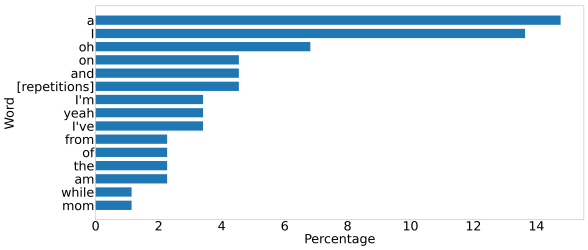


图 2. 我们在 PodcastFillers 验证数据集上的初步结果中，排名靠前的假阳性（FP）词语。

类别焦点损失，以应用于指定类别。由于我们的目标是在连续语音中检测特定的短语音信号，因此当关键词为填充词时，该任务可视为关键词检测（KWS）。因此，我们采用 AF-KWS [14] 来构建我们的骨干网络模型。如图1所示，我们的方法是一种端到端的无需自动语音识别的方法。它以原始语音作为输入，提取短时傅里叶变换（STFT）特征，然后使用残差网络（ResNet）[15]提取高级特征，接着采用三个可训练的预测头，分别对关键词热图、词长和偏移量进行回归。经过解码器模块后，它可以直接输出检测到的关键词位置。我们的训练方案可以自动选择易混淆词作为新类别，然后继续训练。下面我们将介绍所提出的两种机制的详细信息。

2.3. 辅助类别选择

由于“填充词”和“非填充词”在很大程度上分布不平衡（非填充词和容易混淆的难例数量较多），我们采用焦点损失 [16]来缓解训练难度。该损失能够动态增加对难例的惩罚。然而，这些非填充词只是简单地混在同一个“非填充词”类别中，这仍然使焦点损失难以学习易混淆（难）词的特征。

为了减少易混淆词与填充词之间的误分类，我们将一些易混淆词设为独立的辅助类别，促使模型显式学习这些易混淆词各自的特征。然而，对于每个数据集，在不同训练阶段检测到的易混淆词可能各不相同，因此我们需要在模型训练后挑选一些易混淆词作为新类别，再调整模型分类层并重新训练模型。显然，简单地重复这一过程非常复杂且耗时。因此，我们设计了一种训练方案，动态地逐个将易混淆词添加为辅助类别，使模型只需经过一次训练就能区分易混淆词。

如图 1 右侧所示，我们首先添加 h个分类头作为占位符。随着训练的进行，每经过 k 次迭代，我们就进行评估，分析当前训练阶段最容易混淆的词，并将其作为一个新类别添加到占位符中，以供后续训练；重复这一过程，直至所有占位符都被填满。

2.4. 类别间焦点损失

我们已将最容易混淆的词划分为不同类别。然而，由于模型骨干网络的容量有限，我们无法添加过多的辅助类别，而且“非填充词”类别中仍存在一些容易混淆的词。

焦点损失 [17, 16] 将训练重点放在少量难例上，防止大量易分类的负例在训练过程中压制检测器。它会一视同仁地处理所有类别中的难样本，但对于我们选出的辅助类别中的难样本，以及仍存在于未填充类别中的难样本，其区分难度并不相同。

我们设计了一种类别间焦点损失来解决该问题。它可以关注任意两个类别，允许控制正负样本损失的权重，并能自适应地调整惩罚。类别间焦点损失的形式如下：

$$L_{AB} = -\frac{1}{N}\sum_{t,c=c_A}\left(\hat{Y}_{t,c_B}\right)^{\gamma}\left(\mu_{AB}L_{pos} + \omega_{AB}L_{neg}\right) \quad (1)$$

$$L_{pos} = \left(1-\hat{Y}_{t,c_A}\right)^{\alpha}\log\left(\hat{Y}_{t,c_A}\right) \quad (2)$$

$$L_{neg} = \left(\hat{Y}_{t,c_A}\right)^{\alpha}\log\left(1-\hat{Y}_{t,c_A}\right), \quad (3)$$

其中 L_{AB} 关注模型将类别 A 误分类为 B的样本。N 表示样本数量，t表示时间维度上的点，$c = c_A$ 表示我们关注类别 A中的样本，L_{pos} 是类别 A 的正损失，而 L_{neg}是负损失，$(\hat{Y}_{t,c_B})^{\gamma}$是动态权重，其来源是类别 A 被错误分类为 B的概率，α 和 γ 是 L_{AB}的权重因子，μ 和 ω 则控制正损失与负损失之间的平衡。

整体训练损失由优化焦点损失 [17]组成

$$L_{main} = -\frac{1}{N}\begin{cases} \sum_{t,c}\left(1-\hat{Y}_{t,c}\right)^{\alpha}\log\left(\hat{Y}_{t,c}\right) \text{ if } Y_{t,c}=1, \\ \sum_{t,c}\left(1-Y_{t,c}\right)^{\beta}\left(\hat{Y}_{t,c}\right)^{\alpha}\times \\ \sum_{t,c}\log\left(1-\hat{Y}_{t,c}\right) \end{cases} \text{otherwise} \quad (4)$$

以及“非填充词”类别与“填充词”类别之间的一对类别间焦点损失：

$$L = L_{main} + L_{FN} + L_{NF}, \quad (5)$$

其中，L_{FN} 表示模型将“填充词”误分类为“非填充词”时产生的损失，由 L_{AB}定义；L_{NF} 的定义方式相同。

2.5. 实现细节

我们训练模型时将初始学习率设为 5e-3。在第 300、350 和 400 个训练轮次，学习率将降至前一学习率的十分之一。总共训练 450 个轮次。对于焦点损失中的因子，我们设为 $\alpha = 2$、$\beta = 4$。在数据预处理、数据增强、特征提取和模型骨干网络设置方面，我们遵循 AF-KWS [14] 方法中的设置。为动态添加辅助头，我们从第 120 个训练轮次开始添加辅助头，此后每隔 k =10 个训练轮次添加一个辅助头。

3. 实验

3.1. 实验设置

我们在 PodcastFillers 数据集 [1] 上评估所提出的方法。我们将系统与三种表现强劲的基线方法进行比较：两阶段方法 VC-FillerNet [1], 、基于神经网络的方法 Filler-CRNN [2], 以及基于强制对齐器的方法 Gentle [18]。除 VC-FillerNet 外，Ge 等人 [1] 还提出了一种基于自动语音识别的方法 AVC-FillerNet，该方法利用了他们工作中经训练的自动语音识别模型所产生的结果。由于其他基线方法和我们的方法均无需自动语音识别，因此我们不纳入这种方法。

我们计算“填充词”类别的基于事件的指标（精确率、召回率、F1）[19] ，并遵循 [1] 采用 200 毫秒的容差水平。由于我们的方法能够直接从连续语音中检测填充词，无需将语音切分成短片段进行分类，因此本文不采用基于片段的指标。

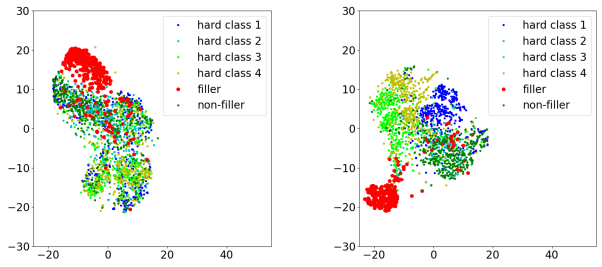
3.2. 结果与比较

结果如表 1 所示。我们可以看到，所提出的方法显著优于所有其他系统。当辅助类别数量为 $h = 8$ 时，模型获得的辅助类别为：“and”、“a”、“[重复]”，

方法	精确率	召回率	F1
Gentle*	77.0	64.9	70.4
Filler-CRNN*	37.5	78.3	50.7
VC-FillerNet	66.0	76.9	71.0
Ours	80.9	81.0	81.0

表 1。在 PodcastFillers数据集上的性能比较。带 * 的方法表示由 [1]重新实现的 PodcastFillers数据集版本。

“[breath]”、“the”、“I”、“[agree]”和“oh”。显然，它们的发音与填充词相似，很容易被混淆。



(a) 我们的方法不包含辅助类别和类别间焦点损失 (b) 我们的方法包含辅助类别和类别间焦点损失

图3。使用 t-SNE 投影对我们的模型学习到的潜在特征嵌入进行可视化。

图3展示了我们使用 t-SNE [20] 对模型学习到的潜在特征嵌入进行的可视化。困难类别1-4由我们的方法从非填充词类别中选出。我们提出的方法对填充词和非填充词表现出良好的区分能力。移除辅助类别和类别间焦点损失后，“填充词”类别的嵌入分布更加分散，并且与“非填充词”类别和困难类别有更多重叠。这表明我们的方法能够更好地建模困难类别，从而降低区分填充词与非填充词的难度。

3.3. 消融研究

为验证我们提出的辅助类别和类别间焦点损失的效果，我们将类别间焦点损失替换为 AF-KWS 中使用的优化焦点损失，作为“Ours w/o inter-category focal loss”版本，并移除辅助类别，作为“Ours w/o auxiliary categories”版本。如表2所示，辅助类别和类别间焦点损失能够有效提升性能。

为了研究辅助类别数量的影响，我们设置了不同数量的头。如表3所示，当辅助类别数量为 $h = 8$ 时，模型取得了最佳性能。

方法	精确率	召回率	F1
Ours	80.9	81.0	81.0
不使用类别间焦点损失	78.3	77.7	78.0
不含辅助类别	80.2	70.4	74.9

表2。消融结果。

辅助类别 h	精确率	召回率	F1
4	84.0	76.0	79.8
8	80.9	81.0	81.0
12	81.0	80.9	80.9

表3。所提出的方法在不同辅助类别数量下PodcastFillers数据集上的填充词检测结果。

μ_{FN}	ω_{FN}	μ_{NF}	ω_{NF}	精确率	召回率	F1
2	2	0.5	0.5	80.9	81.0	81.0
0.5	0.5	2	2	84.8	73.7	78.9
2	0.5	0.5	0.5	80.4	78.2	79.3
0.5	2	0.5	0.5	80.0	79.0	79.5

表 4。采用不同类别间焦点损失系数时，所提出的方法在PodcastFillers数据集上的填充词检测结果。

为了考察模型是否应更多关注将填充词误分类为非填充词的样本，或将非填充词误分类为填充词的样本，以及模型是否应更多关注正样本或负样本，我们调整了类别间焦点损失中的四个权重因子。结果如表 4 所示，在 $\mu_{FN} = 2$, $\omega_{FN} = 2$, $\mu_{NF} = 0.5$, $\omega_{NF} = 0.5$ 时取得了最佳结果，这表明模型应更多关注将填充词误分类为非填充词的错误，而正样本和负样本上的错误可以保持相同的权重。

4. 结论

本文验证了将所有其他词混为一个类别并不能促进填充词检测。我们设计了一种填充词检测方法，在训练过程中动态添加辅助类别，从而在网络中显式建模最容易混淆的词。对于仍留在“非填充词”类别中的其他易混淆词，我们提出了类别间焦点损失，可自适应地调整任意两个类别之间的惩罚权重。消融实验表明，辅助类别和类别间损失均能发挥作用。我们的系统在PodcastFillers数据集上取得了81.0的F1分数，大幅优于其他方法。

5. 参考文献

[1]Ge Zhu、Juan-Pablo Caceres 和 Justin Salamon，“填充词检测与分类：一个数据集与基准”，发表于*Interspeech*。Citeseer，2022。

[2]Sagnik Das、Nisha Gandhi、Tejas Naik 和 Roy Shilkrot，“通过语音增强提升表观公开演讲流利度”，载于*ICASSP 2019-2019 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第 6890–6894 页。

[3] Kathryn Womack、Wilson McCoy、Cecilia Ovesdotter Alm、Cara Calvelli、Jeff B Pelz、Pengcheng Shi 和 Anne Haake，“非流利现象作为认知加工的命题外层面指标”，载于计算语言学中意义命题外层面研讨会论文集，2012年，第 1–9 页。

[4] Emily Duvall、Aimee Robbins、Thomas Graham 和 Scott Divett，“探讨填充词及其影响”，*Schwa. Language & Linguistics*，第 11 卷，第 35–49 页，2014。

[5] Paul Lennon，“探究英语作为外语中的流利度：一种定量方法”，*Language Learning*，第40卷，第3期，第387–417页，1990年。

[6] Peter A Heeman、Rebecca Lunsford、Andy McMillin 和 J Scott Yaruss，“ 利用临床医师标注改进口吃语音的自动语音识别”，载于*Interspeech*，2016 年，第 2651–2655 页。

[7] James Ferguson、Greg Durrett 和 Dan Klein，《基于半马尔可夫模型和韵律特征的非流利现象检测》，载于*2015*年北美计算语言学协会分会会议论文集：人类语言技术，2015年，第257–262页。

[8] Hany Hassan、Lee Schwartz、DilekHakkani-T¨ur 和Gokhan Tur，“会话语音翻译的分段与非流利现象去除”，载于 国际语音通信协会第十五届年会，2014 年。

[9]Koharu Horii、Meiko Fukuda、Kengo Ohta、Ryota Nishimura、AtsunoriOgawa和Norihide Kitaoka，《使用不流利度标注的端到端自发语音识别》，*Interspeech 2022* 论文集，第 4108–4112 页，2022 年。

[10] Paria Jamshid Lou 和 Mark Johnson，“端到端语音识别与非流利现象去除”，*arXiv* 预印本 *arXiv:2009.10298*，2020。

[11] Hirofumi Inaguma、Koji Inoue、Masato Mimura 和 Tatsuya Kawahara，“使用双向 LSTM-CTC 在自发对话中进行社交信号检测。”载于 *Inter-speech*，2017年，第 1691–1695 页。

[12] Utkarsh Mehrotra、Sparsh Garg、Gurugubelli Krishna 和 Anil Kumar Vuppala，“利用结合声学与韵律特征的前语言自动音节划分检测语音中的多种不流利现象”，载于*2021*年亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）。IEEE，2021年，第761–768页。

[13]Tedd Kourkounakis、Amirhossein Hajavi 和 Ali Etemad，《FluentNet：使用深度学习端到端检测口吃语音非流利现象》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 2986–2999 页，2021年。

[14] Zhiyuan Zhao、Chuanxin Tang、Chengdong Yao 和 Chong Luo，“用于连续语音关键词检出的无锚框检测器”，载于*Interspeech*。Citeseer，2022。

[15] Bin Xiao、Haiping Wu 和 Yichen Wei，“用于人体姿态估计和跟踪的简单基线”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第466–481页。

[16] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Doll´ar，“用于密集目标检测的焦点损失”，载于*IEEE*国际计算机视觉会议论文集，2017年，第2980–2988页。

[17] Xingyi Zhou、Dequan Wang 和 Philipp Kr¨ahenb¨uhl，“将物体视为点”，*arXiv*预印本 *arXiv:1904.07850*，2019年。

[18] Robert M Ochshorn 和 Max Hawkins，《Gentle：基于 Kaldi 构建的稳健而宽松的强制对齐器》，2017年。

[19] Annamaria Mesaros、Toni Heittola 和 Tuomas Virta-nen，“多声部声音事件检测的指标”，*Applied Sciences*，第 6 卷，第 6 期，第 162 页，2016 年。

[20] Laurens van der Maaten 和 Geoffrey Hinton，“使用 t-SNE 可视化数据。”，机器学习研究杂志，第9卷，第11期，2008年。