利用标签自适应混合学习用于语音情感识别的鲁棒自注意力特征

Lei Kang†、*Lichao Zhang*‡、*Dazhi Jiang*†

†中国汕头大学计算机系
‡中国空军工程大学航空工程学院{lkang, dzjiang}@stu.edu.cn,
lichao.zhang@outlook.com

arXiv:2305.06273v1 [cs.CL] 7 May 2023

摘要

语音情感识别（SER）旨在机器与人类进行自然语言交互的场景中识别人类情感。由于人类情感具有模糊性，这被认为是一项具有挑战性的问题。尽管近年来语音情感识别取得了进展，但最先进的模型仍难以达到令人满意的性能。我们提出了一种基于自注意力的方法，将标签自适应混合增强（mixup）与中心损失相结合。通过在混合增强过程中调整标签概率，并使中心损失适配混合增强训练方案，我们提出的方法取得了优于现有最先进方法的性能。

索引词— 语音情感识别、自注意力特征、混合增强、中心损失

1. 引言

情感语音识别（SER）是人机交互领域最重要的研究课题之一。情感语音识别旨在将输入的语音信号分类为相应的情感类别。然而，由于人类情感具有内在的复杂性、模糊性和高度的个体差异性，这是一项极具挑战性的问题。如何有效提取情感特征是解决情感语音识别问题的关键。

近年来，基于深度神经网络（DNN）的方法主导了情感语音识别领域。尤其是卷积神经网络（CNN）在计算机视觉领域取得成功后，研究人员通常会将语音信号转换为手工设计的频谱图特征作为输入，以便利用 CNN 模型 [1, 2, 3, 4, 5]。不过，得益于循环神经网络（RNN）的发展，原始语音波形也可以直接用作输入 [6]。然而，对于较长的语音信号，基于 RNN 的模型始终难以解决梯度消失问题。

自注意力机制在语音处理领域引起了广泛关注 [7, 8]。近年来，优秀的自监督模型相继涌现，其中，wav2vec2.0 [9] 和 HuBERT [10] 是最受欢迎、性能最出色的模型。除此之外，已有许多经过预训练的 wav2vec2.0 和 HuBERT 模型可供使用，它们在语音领域已具备适用于通用任务的良好权重分布。我们以 HuBERT 为基线架构，并通过一些必要的修改使其适用于情感语音识别。

为了进一步提升情感语音识别模型的泛化能力，数据增强技术被广泛使用，其中，mixup 策略通过混合成对的训练数据及其标签 [11]，被证明是一种简单而有效的方法。Dai 等 [12]提出了一种同时采用中心损失和识别损失作为学习目标的情感语音识别方法。中心损失将同一类别中的特征拉近其类别中心，而识别损失则分离不同情感类别的特征。然而，mixup 与中心损失的结合使用尚未见报道，这是因为 mixup 生成的是带有概率的混合标签，而中心损失需要类别索引。我们提出了一种同时使用 mixup 和中心损失的有效方法，通过学习鲁棒的情感特征，旨在提升情感语音识别任务的性能。

我们的主要贡献有三点：首先，我们改进了基于 HuBERT 的自注意力模型，以更有效地提取情感特征，并通过全面的消融研究加以验证。其次，我们提出了一种标签自适应 mixup 方法，显著提升了情感语音识别性能。第三，据我们所知，这是首次将中心损失与 mixup 结合用于情感语音识别。我们提出的方法在 IEMOCAP 语音数据集上取得了优于先进方法的性能，采用留一会话（LOSO）方式时，WA 为 75.37%，UA 为 76.04%。我们的代码可在 https:// github.com/leitro/LabelAdaptiveMixup-SER 获取。

2. 语音情感识别

在本节中，我们提出了图 1 所示的情感语音识别模型，该模型由 3 个主要部分组成：标签自适应 Mixup 模块、情感特征提取器和投影模块。设 $\{\mathcal{X}, \mathcal{Y}\}$ 为一个情感语音数据集，其中包含

本研究部分得到了国家自然科学基金项目（项目编号：62206163 和 62006245）、汕头大学项目（项目编号：140/09421059）以及汕头大学数字人文与新文科研究孵化项目的资助。

语音信号 $x \in \mathcal{X}$ 及其对应的独热编码情感类别 $y \in \mathcal{Y}$。E 指的是愤怒、快乐、悲伤和中性这几种情感类别。

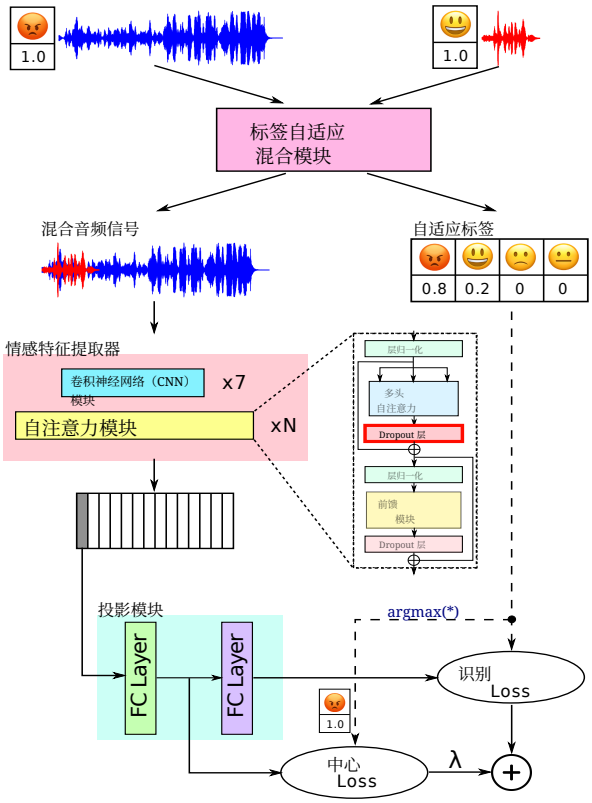

图 1。我们提出的情感语音识别模型示意图。

2.1. 标签自适应混合增强

Mixup [13]是一种广受欢迎、与数据无关的数据增强技术，通过对样本及其标签进行成对凸组合来训练神经网络。给定随机训练样本对（x_i，y_i）和（x_j，y_j），我们可以按照传统的 mixup 策略获得一对合成样本（x_{ij}，y_{ij}），具体如下：

$$x_{ij} = \lambda x_i + (1-\lambda)x_j \quad (1)$$

$$y_{ij} = \lambda y_i + (1-\lambda)y_j \quad (2)$$

其中，$\lambda \sim \mathcal{B}(\alpha,\alpha) \in [0,1]$ 和 $\mathcal{B}$ 表示参数为 $\alpha \in (0,\infty)$ 的贝塔分布。因此，Mixup 是一种直接的方法，通过在特征空间中进行线性插值来扩充训练数据。

语音数据的长度会随着其文本内容而变化，但其标签是情感类别，且概率为 1。因此，像公式 2 所示，将标签视为与语音片段相同并不准确。我们提出标签自适应 mixup 方法来取代它，具体如下：

$$y_{ij} = \left(\frac{\lambda l_i}{\lambda l_i + (1-\lambda)l_j}\right) y_i + \left(\frac{(1-\lambda)l_j}{\lambda l_i + (1-\lambda)l_j}\right) y_j \quad (3)$$

其中，y_{ij} 是情感类别列表，$[z_1, z_2, ..., z_{|E|}]$的总和为 1；l_i 是第 i 个样本的长度。简单来说，我们将 λ 设为常数 0。5。因此，情感类别的概率仅取决于输入语音数据对的长度。

2.2. 情感特征提取

情感特征提取器和投影模块共同构成了有效情感特征提取的处理流程。我们选择最新发布的隐藏单元 BERT（HuBERT） [10]作为情感特征提取器的基线模型。HuBERT 共有三种架构，分别为 HuBERT-Base、HuBERT-Large 和 HuBERT-XLarge。我们选择 HuBERT-Large 作为基线模型，该模型在来自 Libri-Light 数据集 [14]的 60,000 小时无标注音频数据上进行了预训练。HuBERT-Large 模型由卷积部分和 Transformer 部分组成。我们保持卷积部分不变，重点调整后者，以适应情感语音识别（SER）任务。Transformer 部分由 24 个自注意力模块组成，如图 1 中虚线矩形所示。我们减少了自注意力模块的数量，并修改了多头自注意力模块与前馈模块之间的 dropout 概率，如红色矩形所示。我们将在第 3.3 节中进一步讨论这些修改。

我们将语音数据 $x \in \mathcal{X}$ 输入情感特征提取器，从而生成高层次的情感特征表示F_e。F_e 是一个特征向量序列，其长度会根据输入语音信号的长度而变化。我们没有采用平均池化 [15] 将特征向量序列聚合为固定大小的表示，而是直接选取第一个特征向量 F_e^0作为整个序列的情感特征表示，这得益于自注意力模块在长距离特征探索与提取方面的强大能力。我们将在第3.3节中将其与平均池化方法进行比较。随后，如图1底部所示，投影模块由两个全连接层堆叠而成，分别记为 f_0 和 f_1，对应第一层（绿色）和第二层（紫色）。

2.3. 学习目标

2.3.1. 识别损失

我们采用对数 Softmax 库尔贝克-莱布勒散度损失作为识别损失，引导情感语音识别模型进行情绪分类，具体如下：

$$\mathcal{L}_r = \sum_{k=1}^{|E|} z_k \log\left(\frac{z_k}{\hat{z}_k}\right) \quad (4)$$

其中，z_k 是 y_{ij}中第 k类情感的真实概率，$\hat{z}_k$ 是 E 中第 k类情感的预测概率。$\hat{z}_k \in \hat{y}_{ij}$是对输出特征 $f_1(f_0(F_e^0))$应用 Softmax 后得到的。

2.3.2. 中心损失

中心损失最初是为人脸识别而提出并应用的 [16]。它会按小批次更新训练数据的特征中心，并尝试减小特征空间中的类内差异。Dai 等人 [12]将中心损失用于说明其能够为情感语音识别任务学习更有效特征。为了在训练过程中配合 mixup 策略，我们对中心损失公式作如下修改：

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\| f_0(F_e^0) - \mu_{argmax(y_{ij})}\|_2^2 \tag{5}$$

其中，N 是一个小批次中的训练样本数量，$\mu_{argmax(y_{ij})}$ 是情感类别 $argmax\ (y_{ij})$ 的特征质心。y_{ij} 是使用 mixup 方法得到的情感类别 E 概率列表，仅选择概率最高的情感类别作为中心损失的真实标签。通过这种方式，我们不仅解决了 mixup 与中心损失无法协同工作的问题，还通过引入混合噪声学习到了鲁棒的情感特征。因此，模型使用如下联合损失进行训练：

$$\mathcal{L} = \mathcal{L}_r + \lambda\mathcal{L}_c \tag{6}$$

其中，λ 是用于平衡两种损失的权衡超参数。

3. 实验

3.1. 数据集与评估指标

我们采用 IEMOCAP [17]数据集评估所提出的方法。该数据集包含约 12 小时的多模态数据，包括语音、转录文本和面部录制数据。本文仅关注其中的语音数据。语音数据包含 5 个会话，每个会话都涉及两名不同说话人之间的对话。为了使我们的结果能够与当前最先进的方法 [2, 3, 18], 进行比较，我们将“excited”合并到“happy”类别中，并使用“angry”“happy”“sad”和“neutral”四个类别的语音数据，最终从 5 个会话和 10 名说话人中得到共 5531 条声学话语。

我们采用广泛使用的留一会话法（LOSO）五折交叉验证来报告最终结果。因此，每一折中，使用 4 个会话中的 8 名说话人进行训练，另外 1 个会话中的 2 名说话人用于测试。我们选用加权准确率（WA）和非加权准确率（UA）作为评估指标。

3.2. 实现细节

在优化过程中，模型使用 Adam 算法进行训练，并采用动态学习率方案（每个 epoch 将学习率降低至原来的 1.25 倍，直至第 20 个 epoch），该方案同时适用于识别损失和中心损失。学习率的初始值为 $1e$-4 和 $1e$-3，分别对应识别损失和中心损失。所有实验均在 NVIDIA RTX3090 上完成。模型使用 PyTorch 1.12 实现，更多细节请参阅我们的代码。

3.3. 基线模型

我们尝试探索 HuBERT-Large 模型在情感语音识别任务中的最佳用法。本节中的所有实验均在前 4 个会话中前 8 名说话人的语音数据上进行，训练恰好 5 个 epoch，并在第 5 个会话中剩余的 2 名说话人上评估，以报告 WA 和 UA 结果。这样既能确保实验采用与说话人无关的设置，也能避免寻找最佳 epoch，从而有效开展实验。

首先，HuBERT-Large 模型规模庞大，包含 24 个自注意力模块，因此我们想了解情感语音识别性能与自注意力模块数量之间的关系。从图 2 可以看出，使用 22 个自注意力模块时性能最佳。我们还发现，增加层数并不总能提升性能；选择 12 层也能在性能与效率之间取得良好平衡。不过，本文的目标是充分发挥所提方法的最佳性能，因此最终选择 22 层。

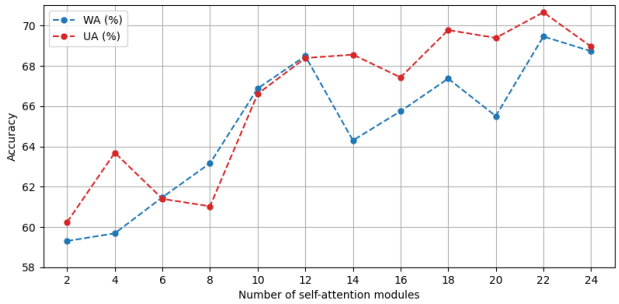

图2。根据所使用的自注意力模块数量绘制的消融实验曲线。

表1。多头自注意力模块与前馈模块之间的投影 dropout 层的 dropout 概率。

Dropout 概率	加权准确率 (WA, %)	非加权准确率 (%)
0	69.46	70.66
0.1	69.94	70.49
0.2	69.46	70.41
0.3	70.59	70.61
0.4	**70.99**	**72.83**
0.5	61.97	67.35

其次，将视角聚焦于图 1 虚线矩形所示的自注意力模块，多头自注意力机制提取序列语音特征之间的上下文信息，而前馈模块则试图获取高层情感特征。因此，中间的投影 dropout 层发挥着关键作用，需要对其进行调整，以防止模型针对特定任务发生过拟合。根据表 1，我们选择 0.4 作为投影 dropout 的取值。

在情感特征提取器的每个自注意力模块中添加一层。

3.4. 消融实验

基于上一节的结果，我们找到了最适合作为情感特征提取器的 HuBERT-Large 模型架构。本节将进一步讨论特征降维方法、mixup 方法以及中心损失的使用。在实验中，我们仍使用前 4 个 session 的数据训练模型，并在剩余的 session 上进行评估，报告加权准确率和非加权准确率结果。不过，我们随机抽取 10% 的训练数据作为验证集，并采用 10 个 epoch 的早停策略来确定最佳模型权重。随后，通过在测试数据上评估最佳模型，得到加权准确率和非加权准确率结果。

如图 1 所示，情感特征 F_e，即情感特征提取器的输出，是一个长度可变的向量序列，需要将其汇总为固定维度的向量，以供投影模块使用。本文比较了两种简单的方法：采用自适应平均池化进行下采样，即 $Avg(F_e)$；或者直接选取 F_e 中的第一个向量，即 F_e^0。根据表 2 前两行所示的实验结果，后一种方法取得了更好的性能。这是因为在训练过程中，相关的情感特征已经聚合到这一个向量中，因此，相较于人工设计的池化方法，该方法具有更强的鲁棒性和可靠性。

表 2。对我们提出的方法进行消融研究，从左到右分别为特征降维方法、Mixup 方法和中心损失方法。

特征降维		混合增强		中心损失	加权准确率 (%)	非加权准确率 (%)
$Avg(F_e)$	F_e^0	卷积。	自适应。	λ		
✓	–	–	–	0	70.91	71.80
–	✓	–	–	0	70.99	72.83
–	✓	✓	–	0	70.83	74.06
–	✓	–	✓	0	73.97	75.03
–	✓	–	✓	0.0005	74.54	76.20
–	✓	–	✓	0.001	74.21	75.99
–	✓	–	✓	**0.002**	**74.86**	**76.31**

为了评估我们提出的标签自适应 mixup 方法的有效性，我们采用传统 mixup [13] 方法进行比较。由于 mixup 被认为是数据增强技术之一，因此在接下来的实验中，我们还将一些常见的数据增强技术与 mixup 结合使用，例如高斯噪声、削波失真、增益、增益过渡、极性反转、双曲正切失真、时间遮蔽、时间拉伸和音高移位。通过随机组合这些常见的数据增强技术并使用传统 mixup 方法，情感语音识别模型的加权准确率和非加权准确率分别达到 70.83% 和 74.06%，如表 2 第 3 行所示。与传统 mixup 策略相比，我们提出的标签自适应 mixup 方法将加权准确率提升了约 3%，将非加权准确率提升了约 1%，如图所示。

见表 2 第 4 行。之所以取得如此显著的提升，是因为所提出的方法根据语音片段的不同长度重新平衡了情感类别的权重。在常见情况下，尤其是在 IEMOCAP 数据集中，同一种情感类别可能出现在简短的感叹词中，也可能出现在较长的独白中，因此传统的 mixup 会将感叹词和独白同等对待，从而引入强噪声。

此外，我们尝试在训练阶段加入中心损失。如第 2.3.2 节所述，λ 是一个用于权衡中心损失与识别损失的超参数。从表 2 的第 5 行至第 7 行，我们展示了不同 λ 对性能的影响。当 λ =0.002 时，性能最佳。

3.5. 与先进方法的比较

最后，根据 IEMOCAP 第 5 个 session 中最后 2 位说话人的数据评估结果（仅包含一个折），我们确定了适用于情感语音识别的最佳神经网络架构和超参数。因此，我们按照留一说话人（LOSO）的方式进行完整的 5 折交叉验证，并报告平均加权准确率（WA）和非加权准确率（UA），如表 3 所示，其性能优于现有先进方法。

表 3。通过留一会话交叉验证（LOSO）五折交叉验证与先进方法进行比较。

方法	Year	加权准确率 (%)	非加权准确率(%)
人类表现 [4]	2017	69.00	70.00
TDNN-LSTM-attn等 [6]	2018	70.10	60.70
长短期记忆网络 等人 [19]	2019	56.99	53.07
IS09分类等 [7]	2019	64.33	64.79
CNN-GRU-SeqCap 等人 [20]	2019	72.73	59.71
HGFM等人 [21]	2020	66.60	70.50
ACNN等人 [5]	2020	67.28	67.94
ASR-SER等人 [22]	2020	68.60	69.70
轻量级模型等 [1]	2020	70.39	71.72
SSL与CMKT融合等 [23]	2021	61.16	62.50
音频$_{25}$，$_{250}$+BERT等 [2]	2021	69.44	70.90
选择性多任务学习等 [24]	2022	56.87	59.47
MFCC+频谱图+W2E等 [18]	2022	69.80	71.05
CNN-SeqCap 等人 [3]	2022	70.54	56.94
提出	**2023**	**75.37**	**76.04**

4. 结论

本文提出了一种基于自注意力的情感语音识别方法，并对其架构和超参数进行了深入修改与评估。此外，我们提出了一种简单有效的标签自适应 mixup 方法，大幅提升了性能。据我们所知，我们首次将 mixup 与中心损失结合起来训练情感语音识别模型，从而促使模型学习更鲁健的特征。与当前最先进的方法相比，我们提出的方法在 IEMO-CAP 语音数据集上取得了更优的性能。

5. 参考文献

[1] Ying Zhong、Ying Hu、Hao Huang 和 Wushour Silamu，《基于可分离卷积的轻量级模型用于语音情感识别》，载于 *INTERSPEECH*，2020 年，第 11 卷，第 3331–3335 页。[2] Wen Wu、Chao Zhang 和 Philip C Woodland，《通过融合时间同步与时间异步表征进行情感识别》，载于 *ICASSP*。IEEE，2021 年，第 6269–6273 页。[3] Xixin Wu、Shoukang Hu、Zhiyong Wu、Xunying Liu 和 Helen Meng，《用于语音情感识别的神经架构搜索》，载于*ICASSP*。IEEE，2022 年，第 6902–6906 页。[4] Vladimir Chernykh 和 Pavel Prikhodko，《利用循环神经网络进行语音情感识别》，*arXiv* 预印本 *arXiv:1701.08071*，2017 年。[5] Mingke Xu、Fan Zhang 和 Samee U Khan，《通过注意力头融合提高语音情感识别准确率》，载于 *2020* 年第 *10* 届年度计算与通信研讨会及会议（*CCWC*）。IEEE，2020 年，第 1058–1064 页。[6] Mousmita Sarma、Pegah Ghahremani、Daniel Povey、Nagendra Kumar Goel、Kandarpa Kumar Sarma 和 Najim Dehak，《使用深度神经网络从原始语音信号中识别情感》，载于 *INTERSPEECH*，2018 年，第 3097–3101 页。[7] Lorenzo Tarantino、Philip N Garner、Alexandros Lazaridis 等，《用于语音情感识别的自注意力》，载于 *INTERSPEECH*，2019 年，第 2578–2582 页。[8] Jiaxing Liu、Zhilei Liu、Longbiao Wang、Lili Guo 和 Jianwu Dang，《结合局部与全局感知深度表征学习的语音情感识别》，载于 *ICASSP*。IEEE，2020 年，第 7174–7178 页。[9] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli，《wav2vec 2.0：语音表征自监督学习框架》，载于 神经信息处理系统进展。2020 年，第 33 卷，第 12449–12460 页，Curran Associates, Inc.。[10]Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed，《Hubert：通过对隐藏单元进行掩码预测实现自监督语音表征学习》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021 年。[11] Wenjing Zhu 和 Xiang Li，《基于多尺度特征表征的全局感知融合语音情感识别》，载于 *ICASSP*，2022 年，第 6437–6441 页。[12] Dongyang Dai、Zhiyong Wu、Runnan Li、Xixin Wu、Jia Jia 和 Helen Meng，《使用中心损失从频谱图中学习用于语音情感识别的判别性特征》，载于 *ICASSP*，2019 年，第 7405–7409 页。[13] Hongyi Zhang、Moustapha Cisse、Yann N. Dauphin 和 David Lopez-Paz，《mixup：超越经验风险最小化》，载于 *ICLR*，2018 年。[14] J. Kahn、M. Rivi`ere、W. Zheng、E. Kharitonov、Q. Xu、P.E. Mazar´e、J. Karadayi、V. Liptchinsky、R. Collobert、C. Fuegen、T. Likhomanenko、G. Synnaeve、A. Joulin、A. Mohamed 和 E. Dupoux，《Libri-light：有限或无监督条件下的自动语音识别基准》，载于*ICASSP*，2020 年，第 7669–7673 页。[15]Edmilson Morais、Ron Hoory、Weizhong Zhu、Itai Gat、Matheus Damasceno 和 Hagai Aronowitz，“使用自监督特征进行语音情感识别”，载于*ICASSP*。IEEE，2022 年，第 6922–6926 页。[16] Yandong Wen、Kaipeng Zhang、Zhifeng Li 和 Yu Qiao，“一种用于深度人脸识别的判别式特征学习方法”，载于*ECCV*，Bastian Leibe、Jiri Matas、Nicu Sebe 和 Max Welling 编，Cham，2016 年，第 499–515 页，Springer International Publishing。[17]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N. Chang、Sungbok Lee 和 Shrikanth S. Narayanan，“IEMOCAP：交互式情感双人动作捕捉数据库”，语言资源与评估，第 42 卷，第 4 期，第 335–359 页，2008 年 12 月。[18] Heqing Zou、Yuke Si、Chen Chen、Deepu Rajan 和 Eng Siong Chng，“基于协同注意力的多层级声学信息语音情感识别”，载于*ICASSP*。IEEE，2022 年，第 7367–7371 页。[19] Jack Parry、Dimitri Palaz、Georgia Clarke、Pauline Lecomte、Rebecca Mead、Michael Berger 和 Gregor Hofer，“跨语料库语音情感识别的深度学习架构分析”，载于 *INTERSPEECH*，2019 年，第 1656–1660 页。[20] Xixin Wu、Songxiang Liu、Yuewen Cao、Xu Li、Jianwei Yu、Dongyang Dai、Xi Ma、Shoukang Hu、Zhiyong Wu、Xunying Liu 等，“使用胶囊网络进行语音情感识别”，载于 *ICASSP*。IEEE，2019 年，第 6695–6699 页。[21] Yunfeng Xu、Hua Xu 和 Jiyun Zou，“HGFM：用于声学情感识别的分层粒度特征模型”，载于*ICASSP*。IEEE，2020 年，第 6499–6503 页。[22] Han Feng、Sei Ueno 和 Tatsuya Kawahara，“结合声学到词语 ASR 模型的端到端语音情感识别”，载于*INTERSPEECH*，2020 年，第 501–505 页。[23] Sheng Zhang、Min Chen、Jincai Chen、Yuan-Fang Li、Yiling Wu、Minglei Li 和 Chuanbo Zhu，“结合跨模态知识迁移与半监督学习进行语音情感识别”，知识型系统，第 229 卷，文章编号 107340，2021 年。[24] Heran Zhang、Masato Mimura、Tatsuya Kawahara 和 Kenkichi Ishizuka，“利用不同风格语料库进行语音情感识别的选择性多任务学习”，载于*ICASSP*。IEEE，2022 年，第 7707–7711 页。