跨语言语音情感识别：韵律模式与自监督学习

GENG Tianqi HAN Zhichen FENG Hui 袁家宏
Korin Richmond 李远超*

摘要 本研究考察了中文（CN）、德语（DE）和英语（EN）情感语音的韵律模式，重点关注声学特征。通过方差分析和混合效应建模，我们揭示了跨语言及特定情绪的韵律模式。对语音情感识别（SER）性能的比较分析表明，在跨语言任务中，人类的表现优于自监督学习（SSL）模型；与此同时，SSL模型通过高效的微调过程，在跨语言SER方面取得了显著提升，达到可与母语使用者相媲美的性能。本研究还分析了模型与人类在语音情感分段识别（SED）方面的表现，凸显了SSL模型在迁移学习场景中的有效性。这些发现强调了普遍韵律特征（如强度、基频）与语言特异性因素之间的相互作用。此外，SSL模型的出色表现表明，这些声学特征在情感处理中的潜在价值，并意味着通过改进训练技术，跨语言情感识别有望取得进展。

关键词： 韵律模式；语音情绪；自监督模型；语音情感识别；语音情感说话人分离

arXiv:2611.29040v1 [eess.AS] 21 Nov 2026

跨语言的语音情感识别:韵律模式与自监督学习

耿天琪 韩智宸 冯卉 袁家宏 科林·里士满 李远超*

摘要：本研究旨在探讨汉语、英语和德语中语音情感表达的韵律模式及其跨语言差异。通过使用单因素方差分析和混合效应模型，本研究系统分析了情绪对相关声学参数的影响，并在控制性别和语言差异的情况下，揭示了跨语言情境下情绪的特定韵律表达。研究结果显示，不同语言的情感韵律表达存在一定共性，同时也展现出由语言类型和文化背景等因素导致的差异。具体而言，本文所用的汉语和德语语料在表达不同情绪时，均普遍依赖与强度相关的特征，而英语语料则不然。在基频方面，研究发现开心情感的基频普遍最高，愤怒次之，悲伤情绪的基频最低。然而，混合效应模型的结果显示，情绪对韵律表达的影响在不同文化和语言环境下存在差异，表明语言和文化背景在情感韵律的解读中扮演着重要角色。

在此基础上，本研究还对跨语言语音情感识别的性能进行了比较分析，重点探讨人类被试与自监督学习（Self-Supervised Learning，SSL）模型在跨语言情感识别任务中的表现差异。实验结果表明，人类在跨语言情感识别任务中的表现通常优于 SSL 模型。通过参数高效微调（Parameter-EfficientFine-Tuning）策略，SSL 模型在语音情感识别任务中的表现显著提升。在跨语言任务中，模型的识别能力逐渐接近母语使用者的水平。这一发现证明了 SSL 模型在跨语言语音情感识别中的潜力。

此外，研究还开展了语音情感分段识别（Speech Emotion Diarization，语音情感分段识别）任务，重点分析语音情感片段的精细化识别。研究表明，SSL 模型在识别情感片段时，能够通过针对某种语言的训练掌握该语言特有的情感模式；无论是在该语言还是多语言背景下，模型对情感特征的敏感度和识别准确度都明显提高。对人类被试切分出的最具代表性的语音情感片段进行进一步的韵律特征分析，结果表明，强度作为情感感知觉的普遍线索在不同语言中都具有重要性。同时，研究还发现，基频相关特征在三种语言的情绪感知中展现出跨语言的共性。

本研究强调了人类感知与自监督学习模型在语音情感识别任务中的异同。尽管自监督学习模型在情感识别中取得了一定进展，但人类感知往往涉及更多语境和语义信息，而这些信息目前尚未被自监督学习模型完全模拟。未来的研究可以进一步探讨如何将更多语境和语义信息整合到自监督学习模型中，以提高其情感识别的精度。同时，也需要更多样化的数据集，并控制发音人、文化背景和语言类型等变量，从而进一步优化和细化本研究的结论。

关键词：韵律模式；语音情感；自监督模型；语音情感识别；语音情感分段识别

*通讯作者（通讯作者）：yl809@cam.ac.uk。

1 引言

1.1 韵律与情绪语音

情绪通常通过离散类别（如“六大基本情绪”（Ekman 等，1969））或连续维度来分类，例如环形模型（Russell，1980）、效价—唤醒度—支配度框架（Mehrabian，1996）和FEELTRACE量表（Cowie 等，2000）。说话时，我们会运用语音韵律和非语言发声，并结合具体的词，不仅传达语言信息，也传达诸如情绪之类的非语言信息。

语音情绪一直受到研究者的广泛关注，相关研究考察了情绪的产生与感知（Hancil，2009；Mac 等，2012）。就产生而言，语音韵律被认为在传达情绪方面发挥着核心作用。基频、语速和强度等韵律特征与情绪状态密切相关（Banse 与 Scherer，1996；Williams 与 Stevens，1972）。既有研究还表明，普通话中的情绪感知是多线索的，而不同情绪可能依赖不同的感知线索。具体而言，平均时长和平均强度被发现对识别悲伤和愤怒发挥着重要作用（Zhang 等，2020）。此外，情绪的韵律表达可以独立于言语内容而运作（Kitayama 与 Ishii，2002），并且在不同文化中依赖相似的线索（Bolinger，1978）。

在感知方面，实验研究表明，人们能够从声音表达中识别出种类繁多的积极和消极情绪（Cowen 等，2019）。副语言特征使听者能够识别不同口语中的具体情绪（Pell、Monetta 等，2009；Scherer 等，2001）。有趣的是，尽管人们观察到了情感语音感知方面的跨语言差异（Elfenbein 和 Ambady，2003；Erickson，2010；Pell、Monetta 等，2009），听者不仅能够准确识别母语中的情绪，也能识别其他语言中的情绪，且识别准确率高于随机水平（Pell、Paulmann 等，2009；Thompson 和 Balkwill，2006），这表明语音情绪识别具有普遍性（Jeon 等，2013）。

1.2 语音情绪识别

Wav2vec 2.0（W2V2）（Baevski 等，2020）和 WavLM（Chen 等，2022）等强大的预训练语音模型及其多语言变体已得到开发，自监督学习（SSL）也在持续发展。

这些模型在多种下游语音任务中展现出卓越的性能，包括语音情感识别（SER）（Li et al., 2022）。为进一步提升自监督学习模型在不同语言和数据集上的语音情感识别适用性，研究者采用参数高效微调（PEFT），在增强模型性能的同时，减少对整个模型进行微调的必要性（Feng & Narayanan, 2023; Lashkarashvili et al., 2024）。

然而，语言和文化差异仍是跨语言语音情感识别面临的一大障碍（Shoukat et al., 2023）。提升跨语言语音情感识别性能通常需要大量目标语言训练数据，而对于情感语音数据集有限的语言而言，这可能颇具挑战（Ahn et al., 2021; Neumann, 2018）。

以往研究主要依赖目标语言数据来改进语音情感识别模型；然而，从单语设置转换到跨语言设置时，研究观察到性能显著下降（Latif et al., 2018；Neumann, 2018）。此外，研究人员还探索了对抗神经网络，以实现在非受控环境中的跨语言适应（Cai et al., 2021；Latif et al., 2019）。最近，Upadhyay et al. (2024) 建立了一种层锚定方法

以促进情绪迁移，同时考虑语音模型的层级结构和任务特定性。

1.3 研究问题

尽管自监督学习模型在跨语言语音情感识别方面已展现出潜力（Singh & Gupta, 2023），但仍存在两个关键空白：a）对声学韵律特征如何介导跨语言情感交流的理解不足；b）缺乏对机器与人类感知模式的系统比较。本研究通过考察不同情绪如何借助基频、时长和强度等声学特征进行表达，旨在识别情感韵律中普遍存在的模式和语言特定模式。此外，除传统的话语级语音情感识别外，本研究还将考察细粒度情感感知（即片段级语音情感识别）和微调技术，以比较自监督学习模型与人类表现。

为此，我们提出以下研究问题：

1）普通话、英语和德语中不同情绪各自具有哪些韵律特征？2）人类听者如何利用这些韵律特征识别情绪，尤其是在处理非母语语音时？3）基于自监督学习的模型能否达到与人类相当的情感语音识别性能？4）在跨语言场景中，如何更有效地微调用于语音情感识别的自监督学习模型？5）基于自监督学习的模型能否以类似人类行为的方式识别情感显著片段?

2 方法

Jeon 等（2013）发现，尽管跨语言问题对人类的影响较小，但支持向量机模型在单语语境中的表现更好。Werner 和 Petrenko（2019）的进一步研究表明，人类在语音情感识别（SER）中的跨语言能力通常很强。尽管取得了这些进展，人类与模型之间的比较研究仍然不足，使得这一领域尚未得到充分研究。

2.1 数据集和模型

使用了四个公开的情感语料库：

- ESD：十名以普通话为母语的中文（CN）说话人，包括五名男性和五名女性，在一个普通话情感语料库中用五种不同的情绪类别录制了话语（Zhou 等，2022）。

- PAVOQUE：一个德语（DE）情感语料库（Steiner 等，2013），由一名职业男演员录制，包含五种情绪类别。在该数据集中，带有中性情绪的语音占比超过一半。

- IEMOCAP：一个英语（EN）情绪语料库（Upadhyay 等，2024），包含九类情绪，数据来自五对男女说话者参与的脚本对话和即兴情感对话。

- ZED：一个专为 SED任务创建的英语情感数据语料库（Wang 等，2023），包含对语音数据在话语和子话语（片段）层面的人工标注。该语料库包含从 YouTube 视频转换而来的单声道 16kHz WAV 文件，涵盖喜剧节目、受害者访谈、体育解说等场景中的男性和女性语音。

我们采用了一个在英语情感语音上训练的 WavLM-large 模型[1]，以及三个分别在中文[2]，、德语[3] 和英语[4] 上预训练的 W2V2-base 模型（Wang 等，2023）。我们利用不同的模型和数据集开展了以下活动，以实现特定目标。

2.2 受试者

我们招募了两组受试者参加测试。第一组由六名普通话母语者组成（两男四女），他们从未学习过德语。第二组由六名德语母语者组成（一男五女），他们从未学习过中文。两组受试者的英语水平都很高，且均无听力或情绪理解方面的障碍。

2.3 模型训练

2.3.1 自监督学习模型的逐层分析

为探究单语和跨语言语音情感识别性能，我们使用了三个数据集：ESD（中文）、PAVOQUE（德语）和 IEMOCAP（英语）。分析的情绪包括愤怒、快乐、中性和悲伤。模型方面，我们采用了 W2V2-CN、-DE 和 -EN，分别代表在中文、德语和英语语音数据上进行预训练的自监督学习模型。

自监督学习模型会在不同层中编码语音信息，既往研究表明，中间层的表征通常能在语音情绪识别任务中取得更优性能（Li et al., 2023）。为确定单语和跨语言语音情感识别的最佳层，我们进行了逐层分析。在此方法中，自监督学习模型作为特征提取器使用，所有参数均被冻结，并以非加权准确率（UA）[5] 作为评估指标。

分析包含三种实验设置。在单语设置中，模型在与其预训练语言相同的语言数据上进行微调和测试。例如，W2V2-CN 使用中文数据（ESD）进行微调和测试。在跨语言设置中，模型使用其预训练语言的数据进行微调，但在另一种语言的数据上进行测试。例如，W2V2-CN 使用中文数据（ESD）进行微调，并在德语数据（PAVOQUE）或英语数据（IEMOCAP）上进行测试。最后，在迁移学习设置中，模型使用与其预训练语言不同的语言数据进行微调和测试。例如，W2V2-CN 使用德语（PAVOQUE）或英语（IEMOCAP）数据进行微调和测试。

2.3.2 用于跨语言语音情感识别的自监督学习模型的参数高效微调

在逐层分析之后，我们使用多种参数高效微调方法进一步优化了效果最佳的层。具体而言，我们采用了低秩适配器（LoRA）（Hu 等人，2021年）、

[1] https://huggingface.co/speechbrain/emotion-diarization-wavlm-large

[2] https://huggingface.co/TencentGameMate/chinese-wav2vec2-base

[3] https://huggingface.co/facebook/wav2vec2-base-de-voxpopuli-v2

[4] https://huggingface.co/facebook/wav2vec2-base-960h [5] 也称为总体准确率，计算预测正确的实例所占比例，而不考虑类别分布。

瓶颈适配器（BA）（Houlsby 等，2019）和加权门控（WG）技术。此外，我们还采用了两阶段微调过程（Lashkarashvili et al., 2024）：模型首先在源语言上进行微调，待初始阶段收敛后，再在目标语言上进行微调。

2.4 实验流程

2.4.1 韵律分析

为考察每种特定情绪的韵律模式，所有音频材料均由 xSegmenter（Xiong，2019）自动切分，并在 Praat 中进行人工校正。通过 Praat 脚本提取平均 f_0（单位为赫兹）、以及 f_0 最小值、f_0 最大值、时长和平均强度。所有与 f_0相关的测量值均采用半音转换进行归一化处理（Shi 等，2009），而非f_0 参数则保留其原始量表。语速通过词数除以句子时长（毫秒）计算，而 Int_SD 则计算为平均强度的标准差。

为考察 SED 任务中选定片段与未选定片段在声学特征方面的差异，所有声学参数均以未归一化形式进行分析。鉴于该分析聚焦于组内对比，归一化处理没有必要，也不会提高比较结果的有效性。

我们进行了单因素方差分析，以考察情绪对声学特征的影响。为进一步控制不同语言和性别间的个体差异，我们建立了如下线性混合效应模型：

$$\text{Parameter} \sim \text{emotion} + (1 \mid \text{gender}) + (1 \mid \text{language}). \tag{1}$$

在该模型中，情绪被视为固定效应，而语言和性别被视为随机效应，以控制它们可能对参数造成的影响。我们将“中性”情绪类别设为参照组，从而在考虑语言和性别所致变异的同时，评估情绪对参数的总体影响，确保结果的稳健性。

2.4.2 语音情绪识别任务

本任务使用三个数据集——ESD、PAVOQUE 和 IEMOCAP，以及所有选定的模型，研究语音情感识别。研究涵盖四种情绪：愤怒、快乐、中性和悲伤。

在模型训练方面，每种情绪选取 400 条话语，每个数据集共计 1600 条话语。验证集和测试集各由随机选取的 200 条话语组成。模型训练采用五折交叉验证，以提高可靠性。三种语言使用了等量数据，以尽量减少训练数据规模差异的影响；同时，每个数据集的四种情绪类别均保持平衡，确保每种情绪的样本数量相同。

在逐层分析中，采用了一个将 768 维投影到 4 维的分类头。训练参数包括学习率 1e-4、epsilon 值 1e-8 和权重衰减 1e-5。模型训练了 100 个训练轮数，批次大小为 32。使用的损失准则为交叉熵。如果验证损失连续 10 个训练轮数未改善，则停止训练。在参数高效微调配置中，分类头的设置与逐层分析中使用的设置相同。对于 LoRA 模块，我们将注意力头数量设为 8，将缩放因子（alpha）设为 16，并采用 0.1 的丢弃率。在 BA 模块中，缩减因子配置为 16。模型训练了 100

个训练轮数，批次大小为 16。逐层分析中的损失和停止准则得以保留。

在人类感知实验中，我们从每种情绪类别中选取了 12 个句子，最终共呈现 144 条话语作为刺激材料（12 个句子 * 4 种情绪 * 3 个数据集）。参与者通过 web-MUSHRA 界面（Schoeffler 等，2018）听取这些语音样本，并识别其中表达的情绪。与模型性能评估一致，采用 UA 作为评估指标。

2.4.3 SED任务

在此任务中，我们使用了 ESD、PAVOQUE 和 ZED 数据集及模型。考察的情绪包括愤怒、快乐和悲伤；由于中性情绪缺乏可感知的情绪变化，不适合进行分段，因此将其排除。为探究细粒度的语音情绪表达，我们进行了语音情感分段识别（SED）：参与者首先像在语音情绪识别任务中那样听取语音样本并标注情绪，随后剪辑语音，并选出最能清晰表达该情绪的片段。我们采用情绪分段错误率（EDER）作为评估指标（Wang 等，2023）。该指标用于计算说话人分离结果的错误率，并计入漏检情绪（ME）、虚警（FA）、重叠（OL）和混淆（CF）。分数越低，表示性能越好。

$$\mathrm{EDER} = \frac{\mathrm{ME} + \mathrm{FA} + \mathrm{OL} + \mathrm{CF}}{\text{Utterance Duration}}. \tag{2}$$

鉴于语音片段切分需要投入大量精力，模型仅使用 ZED 数据集中的英语话语执行语音情感检测任务：每种情绪随机选取 8 条话语，共计 24 条话语（8 *3 种情绪）。与此同时，人类参与者使用全部三种语言完成该任务，共评估 72 条话语（24*3 种语言）；其中，额外的中文和德语话语分别取自 ESD 和 PAVOQUE 数据集。

3 结果与讨论

3.1 韵律模式

3.1.1 基频

我们的语料库取自用于模型训练的开放数据集，因此不可避免地存在一些数据一致性方面的差异。考虑到这一局限，后续关于声学特征的分析与研究将主要聚焦于语言内比较。

1) f_0 **平均值**从图1可以清楚地看出，在中文数据集中，快乐情绪的平均基频（f_0）最高，中性情绪最低。在英语和德语中，愤怒情绪的平均值（ f_0 ）最高，悲伤情绪最低。相比之下，中文的整体平均值（ f_0 ）高于英语，甚至高于德语。德语和英语中的平均值（ f_0 ）呈现出更相似的情绪影响模式，从高到低依次为愤怒、快乐、中性和悲伤。英语和德语在发音模式上呈现出的相似性，可能归因于两者共有的类型学特征。相比之下，中文作为一种声调语言，说话人可能需要付出额外努力，将情绪表达编码到汉字固有的声调结构中。

表1所示的方差分析结果表明，情绪对中文和德语中的 f_0平均值产生了显著影响。然而，在英语中未观察到情绪对 f_0 平均值的显著影响。

混合效应模型的结果显示，截距显著（$\beta = 20.14$，标准误 $= 4.91$，t $= 4.10$，p < 0.001）。与中性基线相比，愤怒（$\beta = 4.81$，标准误 $= 0.55$，t $= 8.69$，p < 0.001）和快乐（$\beta = 4.55$，标准误 $= 0.55$，t $= 8.22$，p < 0.001）均显著提高了 f_0平均值。然而，悲伤与中性情绪相比没有显著差异（$\beta = 0.39$，标准误 $= 0.55$，t $= 0.70$，p$= 0.485$）。

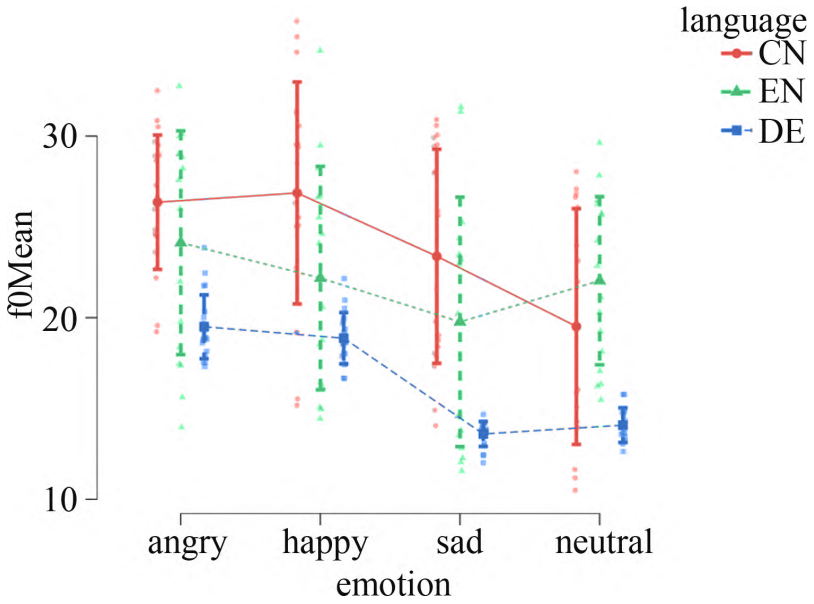

图1： f_0 中文、德语和英语中不同情绪下的平均值。

随机效应分析显示，语言（$\tau_{00} = 2.53$）和性别（$\tau_{00} = 46.16$）所导致的方差，其组内相关系数为 0.86，表明语言和性别之间的差异可以解释 f_0 平均值总方差的 86%。残差方差 σ^2 为 7.91。边际R^2 为 0.082，而条件 R^2 为 0.872，反映出随机效应对模型解释力的显著贡献。

表1： 情绪对 f_0 平均值影响的跨语言方差分析结果。

语言	平方和	df	统计量	p值
CN	601	3	6.37	<0.01
DE	554	3	111	<0.01
EN	153	3	1.41	0.249

为进一步探究这些语言中情感语音的具体模式，我们分析基频最小值（f_0 Min）和最大值（f_0 Max）。

2) f_0 最小值 从图2可以看出，在中文和德语数据集中，f_0 最小值在快乐情绪中最高，其次是愤怒、悲伤，最后是中性，数值最低。相比之下，英语数据集中的 f_0 最小值依次为愤怒、快乐、中性和悲伤。同样，德语数据集中的 f_0 最小值排序依次为愤怒、快乐、悲伤和中性。总体而言，与英语和德语数据集相比，中文数据集的最低音高值始终较高。

如表2所示，我们进行了单因素方差分析，以探究情感表达对三种不同语言中 f_0 最小值的影响。结果表明，情绪对中文和德语中的 f_0 最小值有显著影响。然而，在英语中未观察到显著影响，但结果接近显著。

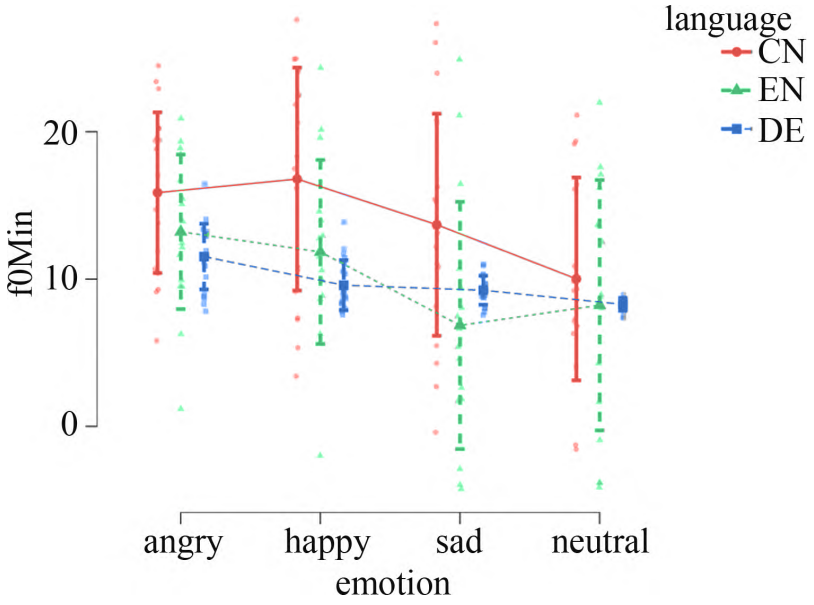

图2： f_0 中文、德语和英语中不同情绪的最小值。

表2： 跨语言情绪对 f_0 最小值影响的方差分析结果。

语言	平方和	统计量	p值
CN	475	3.31	<0.05
DE	103	14.5	<0.001
EN	428	2.72	0.0521

混合效应模型的结果显示，截距显著（$\beta = 10.09$，SE $= 4.03$，t $= 2.50$，p $= 0.013$）。与中性基线相比，愤怒（$\beta = 4.64$，SE $= 0.96$，t $= 4.83$，p < 0.001）和快乐（$\beta = 4.15$，SE $= 0.96$，t $= 4.32$，p < 0.001）均显著提高了 f_0最小值。然而，悲伤与中性相比没有显著差异（$\beta = 1.23$，SE $= 0.96$，t $= 1.28$，p$= 0.203$）。随机效应分析显示，语言（$\tau_{00} = 4.82$）和性别（$\tau_{00} = 28.30$）均存在可归因的变异，组内相关系数（ICC）为0.58，表明 f_0最小值中有58%的方差可归因于不同语言和性别之间的差异。

3） f_0 **最大值**对跨语言的 f_0 最大值进行分析（如图3所示），揭示了情感表达中鲜明的跨语言模式。在中文中，快乐情绪的 f_0 最大值最高，其次是愤怒、悲伤和中性。在英语中，愤怒的 f_0最大值最高，其次是中性、快乐和悲伤。相比之下，德语呈现出不同的排序：快乐最高，其次是愤怒、中性和悲伤。

表3总结了单因素方差分析的结果，该分析考察了情绪对f_0 Max的影响。在中文和德语中均观察到情绪对 f_0 Max的显著影响，表明情绪类别会影响这两种语言中的 f_0 Max。然而，在英语中未发现显著影响，这表明情绪变化不会显著影响该语言中的f_0 Max。这些发现反映出，与英语相比，中文和德语等声调语言和重音语言受到的影响更强。

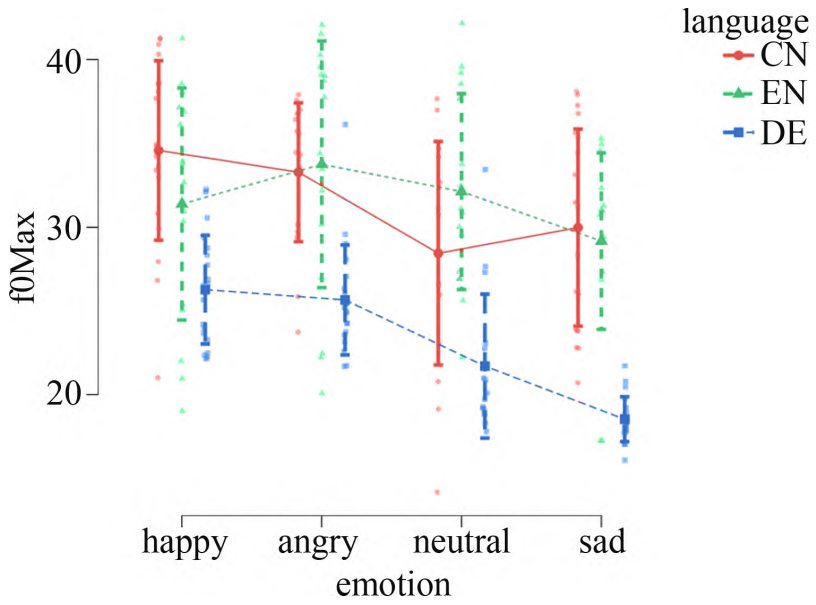

图3： f_0 中文、德语和英语中各情绪的最大值。

表3： 跨语言情绪对 f_0 最大值影响的方差分析结果。

语言	平方和	统计量	p值
CN	445	4.87	<0.05
DE	779	26.1	<0.001
EN	174	1.41	0.248

f_0 Max混合效应模型的结果显示截距显著（$\beta = 28.69$，SE $= 4.19$，t $= 6.85$，p < 0.001）。与中性基线相比，愤怒（$\beta = 3.46$，SE =0.83，t $= 4.17$，p < 0.001）和快乐（$\beta = 3.74$，SE $= 0.83$，t $= 4.51$，p < 0.001）均显著提高了 f_0 Max。然而，悲伤与中性相比没有显著差异（β =-1.57，SE $= 0.83$，t = -1.90，p $= 0.059$），显示出 f_0 Max有降低趋势，但未达到统计显著性。

随机效应分析表明，语言（$\tau_{00} = 7.81$）和性别（$\tau_{00} = 29.10$）均会造成方差，组内相关系数为 0.68，这表明语言和性别之间的差异可以解释 f_0 Max 总方差的 68%。残差方差为 17.73。边际解释力为 0.088，而条件解释力为 0.704，反映出随机效应解释了模型解释力的很大一部分。

3.1.2 强度

Int_SD 是通过计算每个句子中各音节平均强度值的标准差得出的，反映了句子强度的变化程度。Int_SD 值越高，表示强度波动越大；值越低，则表明强度越稳定。如图 4 所示，在中文数据集中，愤怒情绪下的强度波动范围最大，其次是快乐、中性，悲伤情绪下的波动范围最小。在英语数据集中，中性情绪的波动范围最大，其次是快乐；愤怒和悲伤的波动范围相近。在德语数据集中，快乐情绪下的波动范围最大，中性情绪下最小，而愤怒和悲伤呈现出相似的强度波动模式。

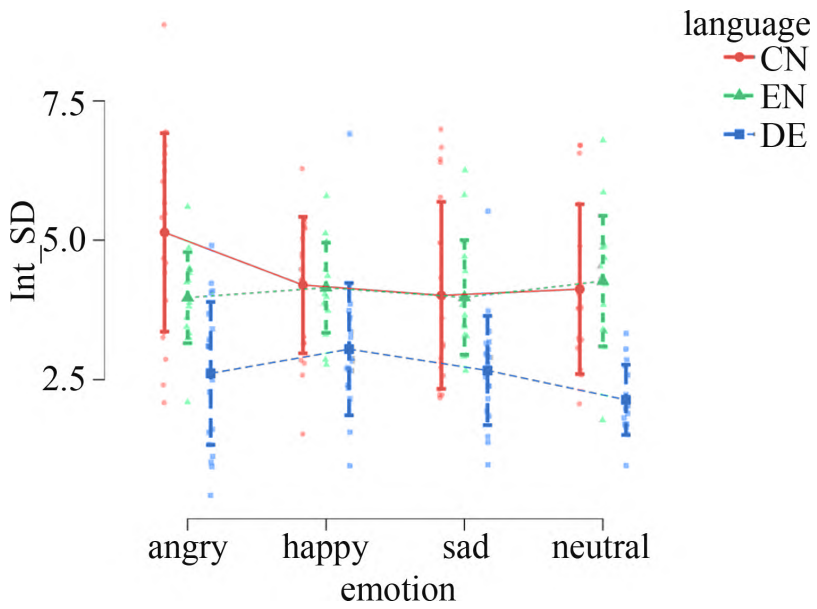

图 4： 中文、德语和英语中不同情绪下的 Int_SD。

由于我们比较的是强度标准差，因此将这一指标视为可在三种语言之间进行比较是合理的。总体而言，德语的强度波动程度最小，而中文和英语在快乐、悲伤和中性情绪下呈现出相似的模式。可以得出结论：尽管中文和英语在类型学上各不相同，但说话人在表达情绪时倾向于采用相似的强度策略，愤怒情绪是一个明显的例外。表达愤怒时，中文说话人的强度变化程度高于英语说话人。相比之下，德语与另外两种语言之间观察到的显著差异可能由多种因素造成。首先，德语数据集主要由一位男性说话人录制，他可能性格较为内向；也有可能是该数据集本身呈现出更稳定的强度表现。此外，德语说话人的文化背景也可能影响这些结果。这些可能性值得在未来研究中进一步探讨，例如纳入更多样化的刺激材料和说话人。

表 4： 跨语言情绪对 Int_SD 影响的方差分析结果。

语言	平方和	统计量	p值
CN	16	2.17	0.0986
DE	7.37	2.15	0.101
EN	1	0.357	0.785

然而，表 4 所示的方差分析检验结果表明，在三种语言中，情绪对 Int_SD 的影响均未达到统计显著性。在中文数据集中，情绪呈现边缘效应，显示出趋于显著的趋势。同样，在德语数据集中，情绪的影响也接近显著。然而，在英语数据集中，未观察到情绪的显著影响。这些结果表明，在所分析的语言中，情绪并未显著调节强度波动的变异性。此外，由于方差分析未发现显著效应，因此未继续进行混合效应模型分析，因为关于情绪影响的主要假设未得到支持。

3.1.3 语速

考虑到不同语言的类型，本研究中语速（spRate）的计算方式因语言而异：对于中文，语速通过将句子总时长除以音节数得出；对于英语和德语，则通过将总时长除以词数计算。spRate 越大，表示语速越慢；数值越小，则表示语速越快。

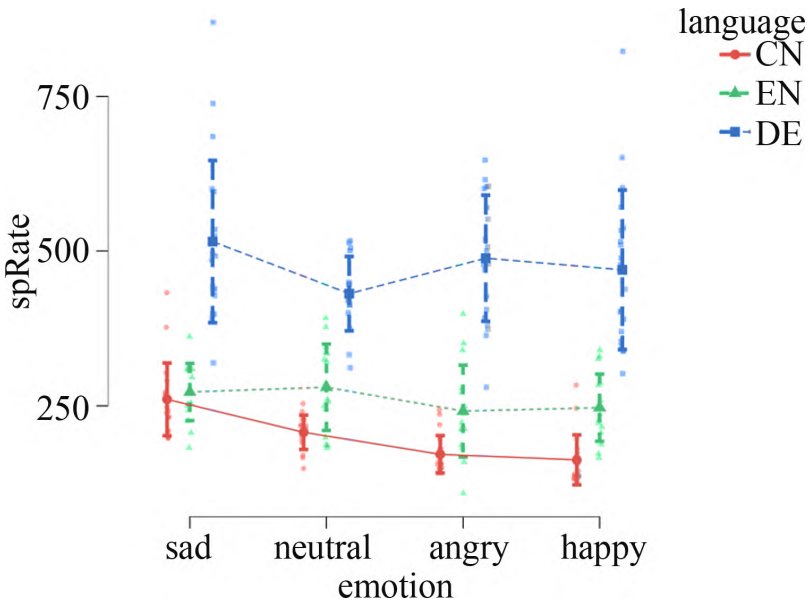

图5： 不同情绪下的中文、德语和英语语速。

如图5所示，在中文数据集中，平均语速从最慢到最快依次为悲伤、中性、愤怒和快乐。在英语数据集中，中性的平均语速最慢，其次是悲伤，而快乐和愤怒状态下的语速以相近的方式更快。在德语数据集中，悲伤状态下的语速最慢，其次是愤怒和快乐，中性状态下的语速最快。

表5： 跨语言情绪对语速影响的方差分析结果。

语言	平方和	统计量	p值
CN	117841	22.6	<0.001
DE	66438	1.79	0.158
EN	17014	1.47	0.232

方差分析结果（表5）显示，中文数据集中情绪对语速有显著影响，这表明情绪状态在影响中文语速方面起着关键作用。然而，在德语或英语数据集中，未观察到情绪对语速的显著影响。这些发现表明，情绪变化对中文语速有明显影响，而对德语和英语语速的影响似乎可以忽略不计。

我们采用混合效应模型分析情绪对语速的影响。截距显著（β = 307.69，SE = 86.86，t = 3.54，p < 0.001）。在情绪预测变量中，只有悲伤对语速表现出显著的正向影响（β = 45.00，SE = 15.66，t = 2.87，p = 0.004），这表明悲伤情绪下语速有所加快。其他情绪预测变量——愤怒（β =-5.13，SE = 15.66，t = -0.33，p = 0.743）和快乐（β= -13.14，SE = 15.67，t = -0.84，p = 0.402）——对语速均无显著影响。对于随机效应，语言的方差分量较大（τ_{00} = 22,105.07），而

性别的方差分量相对较小（$\tau_{00} = 84.33$），残差方差为 6332.15。组内相关系数为 0.78，表明语速的变异很大一部分归因于不同语言和性别之间的差异。边际 R^2 为 0.018，显示固定效应的解释力有限；而条件 R^2 为 0.782，反映随机效应对模型贡献显著。总之，尽管语言和性别能够显著解释语速的变异，情绪状态的影响有限，只有悲伤情绪表现出显著影响。

3.2 语音情绪识别

3.2.1 模型性能

图6展示了逐层分析的结果。在单语设置中，中文和德语模型在各自的源语言上均表现出色，这符合预期，因为它们是在这些语言上进行预训练的。然而，在跨语言设置中，两个模型的准确率都显著下降。尽管语言差异导致一定程度的性能下降是意料之中的，但降幅之大超出了最初的预期，尤其是考虑到情绪声学具有共同特性（Jeon等，2013）。

一种可能的解释是，自监督学习模型不仅能捕捉低层次的声学特征，还会将其转化为高层次的语言信息，例如词汇身份和含义（Pasad等，2021年）。这种转化会造成不同语言之间的语言差异，从而加剧准确率的下降。然而，在迁移学习设置下，模型能够达到与单语场景相当的性能，这凸显了在采用有效知识迁移技术时，自监督学习模型在跨语言语音情绪识别任务中的适应能力。观察到的轮廓变化可能受到自监督学习模型训练目标的影响，尤其是对比式掩码片段预测。

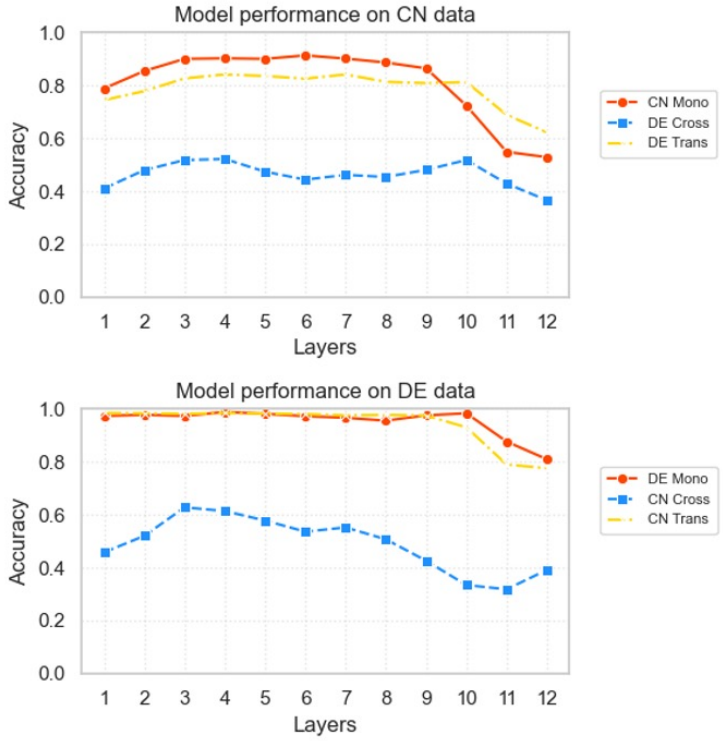

图6： 中文和德语模型在单语、跨语言和迁移学习设置下的逐层分析。

3.2.2 模型与人类

为探究基于自监督学习的模型能否达到与人类评估者相当的语音情感识别性能，我们通过比较模型与人类受试者的非加权准确率进行了比较分析。

1) 单语 如表6和表7所示，在总体非加权准确率方面，两个模型的表现均优于各自的母语人类受试者。进一步考察表8所示的中文和德语单语模型，以及中文和德文母语者在所有情绪类别上的混淆矩阵可以发现，与德语单语模型相比，德文母语者更容易将中性德语语音误分类为悲伤。相比之下，与中文单语模型相比，中文母语者在识别快乐和中性情绪时的精确率较低。这些发现表明，在获得充足训练数据的情况下，自监督学习模型能够在语音情绪识别任务中取得较强的单语性能。

表6： 采用不同参数高效微调策略进行单语、跨语言和迁移学习时的模型性能。

模型	设置	源语言	目标语言	参数高效微调策略			UA%
				LoRa	BA+WG	2-stg	
W2V2-CN	Mono	CN	CN				91.4
				✓			87.0
				✓	✓		93.9
	跨语言	CN	DE				62.8
				✓			65.3
				✓	✓		70.7
	翻译	DE	DE				98.5
				✓			98.8
				✓	✓		98.8
				✓	✓	✓	98.9
	翻译	EN	EN				65.5
				✓			66.3
				✓	✓		67.7
W2V2-DE	Mono	DE	DE				98.9
				✓			97.8
				✓	✓		97.9
	跨语言	DE	CN				52.2
				✓			58.5
				✓	✓		56.0
	翻译	CN	CN				84.2
				✓			83.6
				✓	✓		87.5
				✓	✓	✓	85.8
	翻译	EN	EN				62.4
				✓			65.0
				✓	✓		66.0

2) 跨语言 如表6和表7所示，在总体准确率方面，人类和模型在跨语言条件下的表现均有所下降，其中跨语言模型受到的影响明显大于人类，这表明人类能够更有效地处理跨语言场景。关于所有情绪类别上的表现，如表8所示，德语跨语言模型难以识别中文数据中的中性和悲伤情绪，导致召回率较低。此外，与人类参与者相比，该模型在两种语言中更容易混淆愤怒和快乐。另一方面，中文跨语言模型在解读德语语音时与中文母语者表现出较强的一致性，两个群体都经常将快乐误分类为中性。

表7： 所有语言中的人类表现（非加权准确率）（人类参与者听取与模型相同的语音样本，并对情绪进行分类）。

	CN	DE	EN
中文参与者	79.5	73.3	63.5
德语参与者	82.6	91.7	73.6

表8： 中文和德语说话者及模型在单语（第一列）、跨语言（第二列）和迁移学习（第三列）设置下的混淆矩阵。迁移学习设置下没有人类参与者（愤怒；快乐；中性；悲伤）。

	A	H	N	S
A	0.71	0.24	0.06	0.00
H	0.00	0.82	0.18	0.00
N	0.00	0.10	0.86	0.04
S	0.00	0.01	0.18	0.81

(a) 人类：中文（单语）

	A	H	N	S
A	0.81	0.13	0.06	0.01
H	0.21	0.38	0.42	0.00
N	0.01	0.01	0.93	0.04
S	0.01	0.00	0.15	0.83

(e) 人类：中文交叉（在德语上）

	A	H	N	S
A	0.54	0.22	0.13	0.11
H	0.14	0.61	0.19	0.06
N	0.04	0.28	0.58	0.10
S	0.03	0.01	0.15	0.81

(k) 人类：中文（英语上的第二语言）

	A	H	N	S
A	0.95	0.05	0.00	0.00
H	0.02	0.97	0.02	0.00
N	0.00	0.00	1.00	0.00
S	0.00	0.00	0.03	0.97

(b) 模型：中文单语（在中文上）

	A	H	N	S
A	0.95	0.02	0.03	0.00
H	0.08	0.18	0.51	0.22
N	0.00	0.00	0.95	0.05
S	0.00	0.00	0.00	1.00

(f) 模型：中文交叉（在德语上）

	A	H	N	S
A	0.70	0.13	0.10	0.07
H	0.08	0.57	0.35	0.00
N	0.07	0.18	0.75	0.00
S	0.00	0.05	0.02	0.93

(l) 模型：中文翻译（英语上的第二语言）

	A	H	N	S
A	0.97	0.03	0.00	0.00
H	0.00	0.94	0.05	0.00
N	0.00	0.00	0.83	0.17
S	0.00	0.00	0.04	0.96

(c) 人类：德语（单语）

	A	H	N	S
A	0.88	0.07	0.06	0.00
H	0.08	0.75	0.15	0.01
N	0.03	0.03	0.76	0.18
S	0.00	0.03	0.06	0.92

(g) 人类：德语交叉（基于中文）

	A	H	N	S
A	0.78	0.03	0.04	0.15
H	0.06	0.69	0.22	0.03
N	0.15	0.13	0.60	0.13
S	0.01	0.01	0.10	0.88

(m) 人类：德语（以英语为第二语言）

	A	H	N	S
A	0.98	0.02	0.00	0.00
H	0.02	0.98	0.00	0.00
N	0.00	0.00	0.93	0.07
S	0.00	0.00	0.00	1.00

(d) 模型：德语单语（在德语上）

	A	H	N	S
A	0.52	0.48	0.00	0.00
H	0.17	0.65	0.17	0.02
N	0.00	0.38	0.38	0.23
S	0.05	0.28	0.30	0.37

(h) 模型：DE交叉（在中文上）

	A	H	N	S
A	0.70	0.12	0.17	0.02
H	0.00	0.65	0.33	0.02
N	0.08	0.30	0.58	0.03
S	0.00	0.00	0.25	0.75

(n) 模型：德语翻译（基于英语）

此外，对人类的精确率、召回率和 F1 分数进行了双侧 Welch t 检验。结果表明，中文受试者与德语受试者在德语数据中对快乐的召回率存在显著差异 [$t(10) = -7.511, p < 0.001$], 此外，在中性情绪的精确率方面也存在显著差异 [

t (10) = -5.614, p < 0.001]。此外，与中文数据相比，中文受试者在德语数据中对快乐的召回率更低 [t (10) = -5.137, p < 0.001], ，这表明两个说话者群体之间存在语言和副语言知识差距。具体而言，在中文、德语和英语数据中，对悲伤的召回率存在显著差异 [t (10) = -2.708, p $= 0.022$] ，在中性情绪的精确率方面也存在显著差异 [t (10) = -7.511, p < 0.001]。中文受试者难以在德语数据中感知快乐，这显著影响了中性情绪的精确率，凸显了语言差异和副语言差异对跨语言悲伤感知的影响。

3) 迁移学习 迁移学习设置与人类习得第二语言的过程相似，可用于比较自监督学习模型与使用英语数据的人类受试者。如表6所示，采用迁移学习的自监督学习模型达到单语水平的表现，并且在中文和德语数据上优于人类受试者。在英语数据上，德语受试者的准确率高于中文受试者和所有受测模型。此外，两阶段微调并未显著提升表现。这些结果表明，迁移学习有助于自监督学习模型适应新语言，但模型表现会因目标语言数据集的具体特征而异。

分析各情绪类别的表现时，中文受试者仅在识别快乐情绪时优于模型，而中文迁移学习模型在其他三种情绪类别上的表现均优于人类。对于德语受试者，与德语迁移学习模型相比，人类在预测快乐和中性情绪方面表现更好。此外，在单语场景中有效的参数高效微调策略，在跨语言或多语言场景中未必能带来收益。表7进一步表明，尽管中文和德语受试者都是第二语言学习者，但英语情绪识别比中文和德语更具挑战性。这种难度的增加可能是因为仅使用了IEMOCAP数据集中的即兴话语；这些话语体现了更自然、更贴近现实生活的情绪，使语音情绪识别任务更具挑战性。

3.3 语音情绪分段

为探究模型能否以类似人类行为的方式识别情感显著片段，我们将模型性能与人类受试者的表现进行了比较。此外，为阐明人类听者如何利用韵律特征识别情绪，尤其是在处理非母语语音时，我们对参与者选出的片段的声学特征进行了详细分析。

3.3.1 模型与人类

表9所示结果表明，总体而言，两个受试者组的表现均优于模型，其中德语组的情绪分段错误率（EDER）最低。模型在识别快乐方面表现最佳，在识别悲伤方面表现最差。在人类受试者组中，中文受试者对愤怒的感知稍准确一些，而德语受试者在识别悲伤方面表现更好。这一模式与表8所示的语音情感识别结果一致：中文受试者识别悲伤的阈值较高，因此召回率较高，但精确率较低。相比之下，德语受试者的精确率较高，但召回率较低。中文受试者对悲伤的敏感度差异，使其在SED任务中对悲伤的假阴性数量更多。

3.3.2 选定片段的声学特征

完成 SED任务后，我们对人类参与者选定的语音片段进行了更细致的考察。研究的主要声学变量包括 intMax、intMin、intMean、 f_0 Max、 f_0 Min 和 f_0 Mean，以及这些片段的时长

表9： WavLM与人类在ZED数据上的EDER（%）比较。

	WavLM	中文参与者	德语参与者
愤怒	36.6	25.8	27.5
快乐	27.5	31.8	28.7
Sad	50.3	38.6	28.3
平均值	38.2	32.1	28.2

。音高相关变量采用半音音阶进行归一化，强度相关变量则采用最小-最大法进行归一化。选定片段的时长按其长度占整个句子长度的比例计算。

1) 选定片段的时长 在图7中，dur_per 表示选定片段的时长占整个句子时长的比例。

对于中文和德语听者而言，德语刺激材料始终比中文或英语刺激材料需要更高的加工难度，这一点体现在选定片段时长所占比例更大。

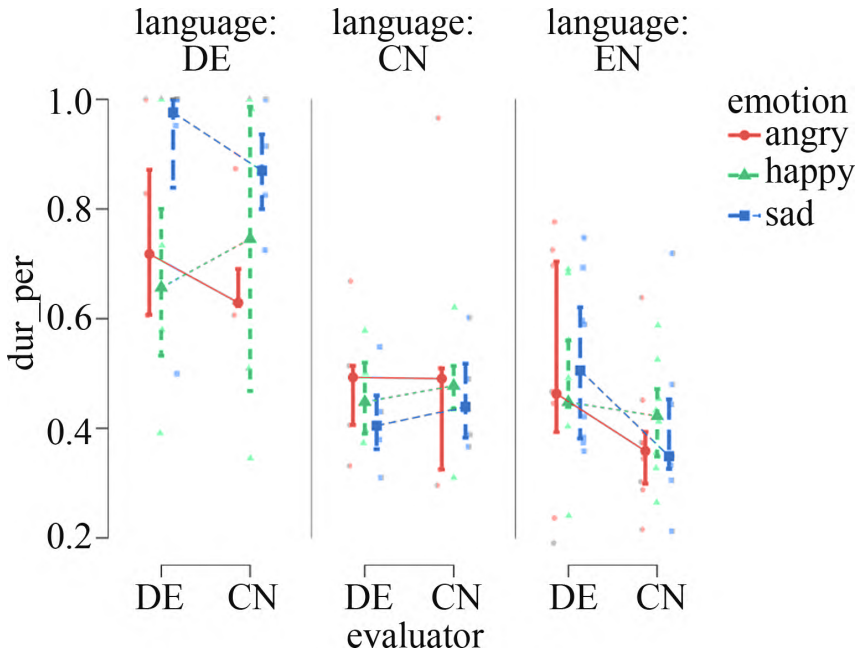

图7： 不同数据集下受试者对各类情绪的选定片段时长

德国参与者感知英语（在类型学上与德语本身相似）时的反应时间模式，与他们聆听母语时观察到的模式相似，呈现出悲伤 > 愤怒 > 快乐的顺序。不过，他们对英语的总体反应时间短于对德语刺激材料的反应时间。这种差异可能源于录音中德语使用者的发音特点。相比之下，聆听与德语差异显著的声调语言中文时，他们表现出明显不同的反应时间模式。

就单种情绪而言，在德语刺激材料中，中国参与者在标记为悲伤的片段上花费的时间最长，其次是快乐，愤怒最短。类似地，德国参与者在悲伤上花费的时间最长，其次是愤怒，快乐最短。

对于中文刺激材料，中国和德国参与者表现出相似的时间模式：在愤怒上花费的时间最长，其次是快乐，悲伤最短。不过，中国参与者处理中文刺激材料时，倾向于选择相对较长的语音片段。

对于英语刺激材料，德国参与者在悲伤情绪上花费的时间最多，其次是愤怒，快乐最少。相比之下，中国参与者在快乐情绪上花费的时间最多，其次是愤怒，悲伤最少。值得注意的是，在处理英语刺激材料时，与其他语言相比，中国参与者通常选取较短的语音片段。

以语言、情绪和评估者为固定因素进行了方差分析检验。分析显示，语言的主效应显著 [$F(2, 81) = 28.05, p < 0.001$]。然而，情绪和评估者均未表现出显著的主效应（两者的 $p > 0.40$）。尽管描述性图表显示，某些情绪类别在不同评估者之间存在数值差异，但这些差异在统计上并不显著。

2) 中文 在感知中文情绪时，来自两个国家的参与者所选取的语音片段在强度相关特征上呈现相似模式，如图8所示：在最大_强度组和平均_强度组中，愤怒的强度最高，其次是快乐，悲伤最低。而在未选取的片段中，中国参与者和德国参与者的最小强度与平均强度也呈现相似模式，这表明这些参数在他们的情绪感知中发挥着相似作用。

就基频而言，选定片段的最大值 f_0 在不同情绪之间存在明显差异；在两组均选取的片段中，快乐情绪的最大值平均值 f_0 显著高于未选定片段。关于最小值 f_0，中文参与者似乎更倾向于将较低值解读为悲伤的表现。

如表10所示，方差分析检验结果表明，情绪对若干声学测量指标产生了显著影响。对于强度相关特征，情绪显著影响 min_int 和 mean_int，但对 max_int 的影响未达到显著水平。对于基频相关参数，情绪对 max_f0 产生了显著影响，但对 min_f0 或 mean_f0 的影响不显著。

相比之下，评估者和选择状态对任何声学测量指标均未表现出显著影响，只有选择状态对 min_int 有轻微影响。总体而言，结果表明，情绪是影响声学变化的主要因素，而评估者的语言背景或选择状态差异的影响微乎其微。

3) 德语 如图9所示，在德语数据集中，中国参与者和德国参与者选定的语音片段在强度方面都表现出高度一致性。这表明，对于强度在德语语音情绪识别中的作用，大家看法一致：愤怒情绪下的强度最高，其次是快乐，最后是悲伤。然而，未选定片段之间的差异，尤其是在悲伤情绪下的差异，表明德语母语受试者感知母语情感线索时，并非只依赖强度。相比之下，中国参与者倾向于选取整个强度较低的片段，尤其是在悲伤情绪下。

从音高相关特征来看，两组参与者选定的片段也呈现出相似之处。具体而言，愤怒情绪的平均音高最高，其次是快乐，最后是悲伤。值得注意的是，德语悲伤情绪的音高动态范围较大（ f_0 Max较高， f_0 Min较低），这可能表明悲伤情绪的表现形式更为复杂。这种模式既包含低能量、延长的片段，也包含偶尔出现的高音高情绪峰值。选定片段中的这些特征表明，对于中国参与者和德国参与者而言，对德语悲伤情绪的感知可能与音高范围扩展有关。

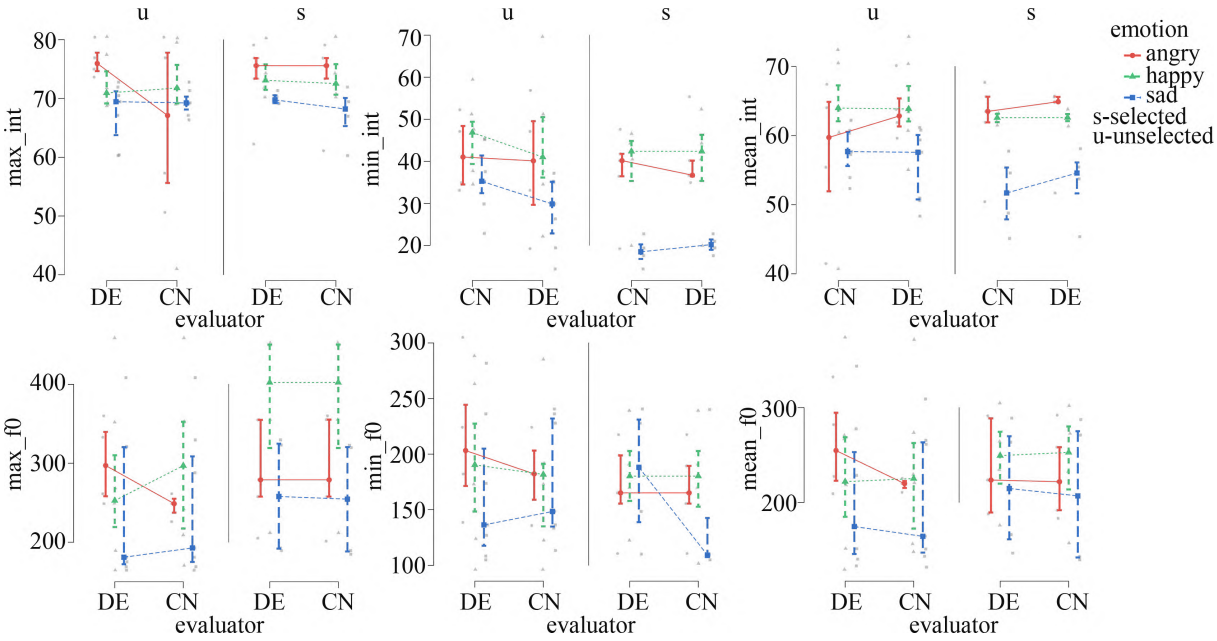

图8： 中文条件下按受试者、选择状态和情绪划分的声学参数比较 e.

表10： 中文条件下情绪、评估者和选择状态对声学参数影响的方差分析结果。

	案例	平方和	自由度	均方	F	p
max_int	情绪	200.71	2	100.36	1.929	0.156
	评估者	93.44	1	93.44	1.796	0.186
	选择状态	47.23	1	47.23	0.908	0.345
min_int	情绪	3018.48	2	1509.24	13.691	< .001
	评估者	11.85	1	11.85	0.108	0.744
	选择状态	669.36	1	669.36	6.072	0.017
mean_int	情绪	810.924	2	405.462	9.542	< .001
	评估者	34.317	1	34.317	0.808	0.373
	选择状态	19.158	1	19.158	0.451	0.505
max_f0	情绪	56890.2	2	28445.1	3.49	0.038
	评估者	602.7	1	602.7	0.074	0.787
	选择状态	17514.8	1	17514.8	2.149	0.149
min_f0	情绪	2934.48	2	1467.24	0.413	0.664
	评估者	2683.12	1	2683.12	0.755	0.389
	选择状态	2949.28	1	2949.28	0.829	0.367
mean_f0	情绪	10400.7	2	5200.4	1.098	0.342
	评估者	1583.2	1	1583.2	0.334	0.566
	选择状态	149.3	1	149.3	0.032	0.86

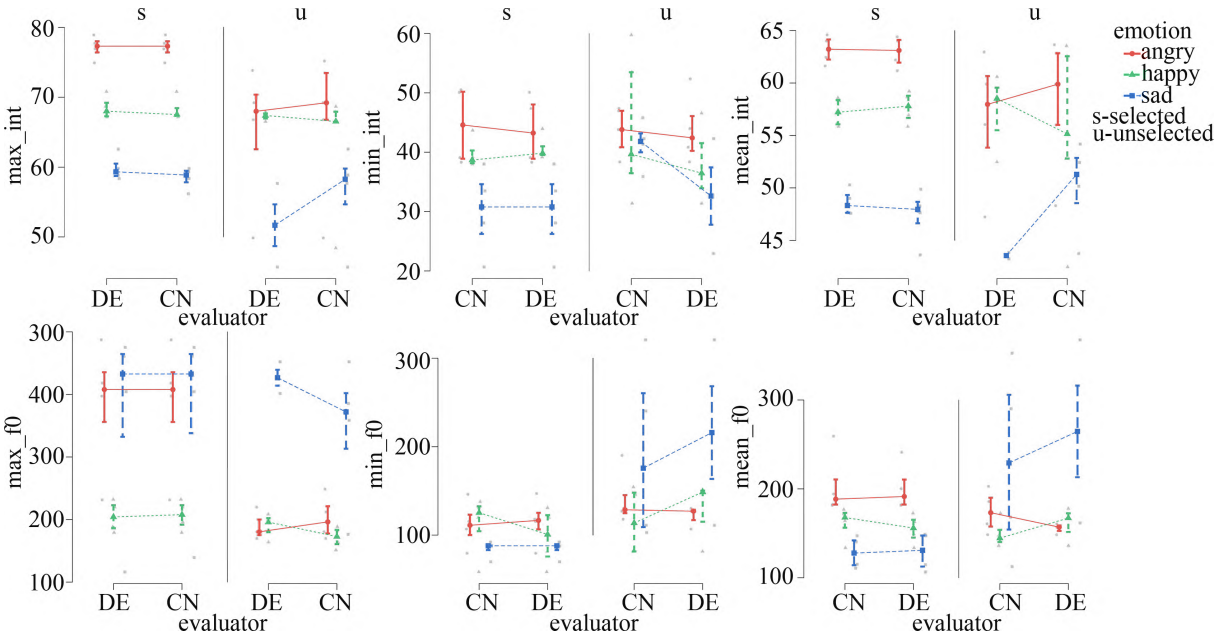

图9： 按受试者、选择状态和情绪划分的德语声学参数比较 n.

表11： 德语中情绪、评估者和选择状态对声学参数影响的方差分析结果。

	案例	平方和	自由度	均方	F	p
max_int	情绪	1664.305	2	832.152	22.164	< .001
	评估者	0.38	1	0.38	0.01	0.92
	选择状态	463.957	1	463.957	12.357	0.001
min_int	情绪	799.55	2	399.78	8.178	0.001
	评估者	62.06	1	62.06	1.269	0.268
	选择状态	67.93	1	67.93	1.389	0.246
mean_int	情绪	1237.393	2	618.697	29.319	< .001
	评估者	8.472	1	8.472	0.401	0.53
	选择状态	80.663	1	80.663	3.822	0.059
max_f0	情绪	230800	2	115400	13.065	< .001
	评估者	1935	1	1935	0.219	0.643
	选择状态	44421	1	44421	5.029	0.032
min_f0	情绪	7097.535	2	3548.767	1.431	0.254
	评估者	1.929	1	1.929	7.779×10^{-4}	0.978
	选择状态	28195.373	1	28195.373	11.369	0.002
mean_f0	情绪	8500.2	2	4250.1	1.889	0.168
	评估者	78.5	1	78.5	0.035	0.853
	选择状态	7099.69	1	7099.69	3.155	0.085

根据表11所示的方差分析结果，发现情绪对三个强度测量指标均具有显著影响。情绪显著影响最大_强度、最小_强度和平均_强度。选择状态也显著影响最大_强度，但其对最小_强度和平均_强度的影响未达到显著水平。在任何强度参数上均未观察到评估者的影响。

对于与 f_0相关的测量指标，情绪对最大_基频具有显著影响，但对最小_基频或平均_基频没有显著影响。选择状态显著影响最小_基频，同时对最大_基频也表现出较小的影响。与强度测量指标一样，评估者的语言背景对任何 f_0 参数均无显著影响。总体而言，这些结果表明，情绪是声学变异的主要来源，而选择状态仅对某些参数产生选择性影响，评估者背景的作用则微乎其微。

4) 英语 在识别英语情绪时，中国参与者和德国参与者所选择的片段在不同情绪类别之间均表现出明显的最大强度差异。愤怒和快乐等情绪的最大强度较高，而悲伤情绪的最大强度则显著较低。就最小强度而言，德国受试者倾向于选择强度较低的片段来代表悲伤情绪，这表明，与中国参与者相比，他们更多地依靠较低强度来表示悲伤。

从基频来看，两组参与者整体呈现出一致的模式：快乐的选定片段具有最高的 f_0 ，其次是愤怒，悲伤的则最低。然而，德国参与者更倾向于选择 f_0 值较高或明显较低的片段，作为情绪表达的代表。

英语数据集的方差分析结果汇总于表12。对于与强度相关的测量指标，情绪仅对最小_int有显著影响，而对最大值和平均强度的影响均不显著。选择状态显著影响最大强度，但不影响最小或平均强度。评估者对任何强度参数均无显著影响。

对于与 f_0相关的参数，在多项测量指标中均发现了显著的情绪效应。情绪显著影响最大 f_0、最小 f_0和平均 f_0。选择状态也显著影响最大 f_0和平均 f_0，但不影响最小f_0。与强度测量指标一样，评估者对所有 f_0 参数的影响均不显著。

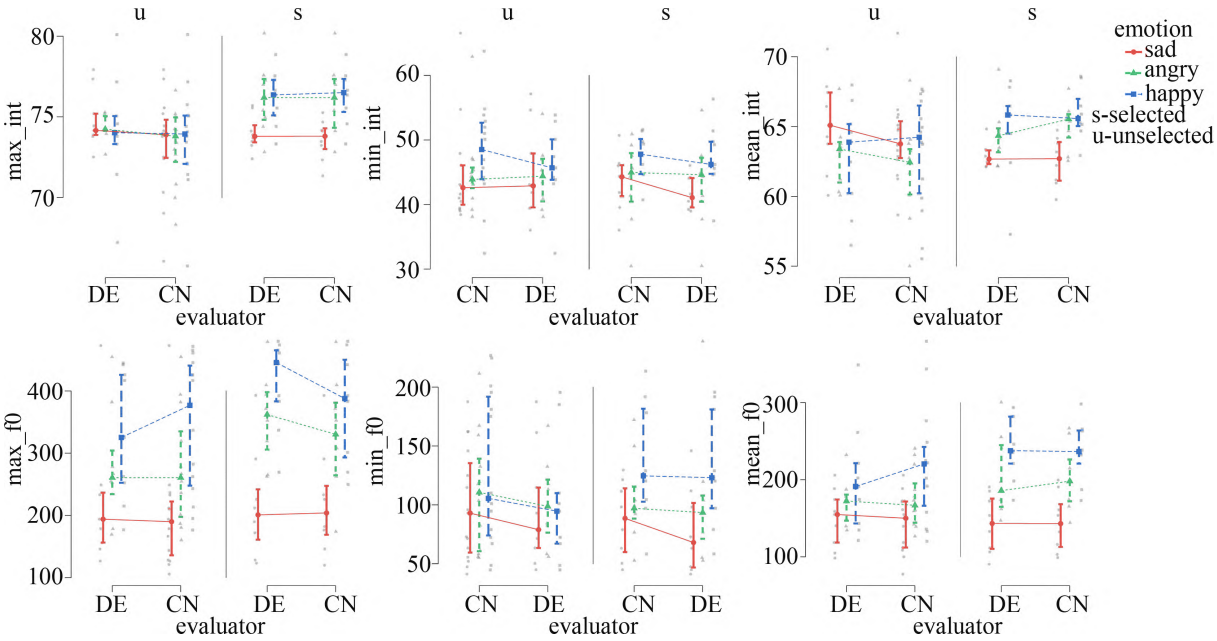

图10 ： 英语语境下按受试者、选择状态和情绪划分的声学参数比较 h.

表12： 英语语境下情绪、评估者和选择状态对声学参数影响的方差分析结果。

	案例	平方和	自由度	均方	F	p
max_int	情绪	27.666	2	13.833	2.359	0.1
	评估者	8.848	1	8.848	1.509	0.222
	选择状态	59.304	1	59.304	10.113	0.002
min_int	情绪	268.149	2	134.074	3.382	0.038
	评估者	13.582	1	13.582	0.343	0.56
	选择状态	25.951	1	25.951	0.655	0.42
mean_int	情绪	8.22	2	4.11	0.473	0.624
	评估者	0.446	1	0.446	0.051	0.821
	选择状态	24.067	1	24.067	2.772	0.099
max_f0	情绪	438467.6	2	219234	24.723	< .001
	评估者	4255.4	1	4255.41	0.48	0.49
	选择状态	51632.7	1	51632.8	5.823	0.018
min_f0	情绪	22787.251	2	11393.6	4.484	0.014
	评估者	2413.825	1	2413.83	0.95	0.332
	选择状态	4.545	1	4.545	0.002	0.966
mean_f0	情绪	119437.7	2	59718.8	25.134	< .001
	评估者	117.4	1	117.4	0.049	0.825
	选择状态	10132.8	1	10132.8	4.265	0.041

4 结论

为探究跨语言情感表达的韵律模式，并比较人类与自监督学习模型在跨语言语音情感识别和语音情感检测任务中的表现，本研究首先考察了三种语言（即中文、德语和英语）表达情绪时采用的韵律模式，重点关注基频和强度。我们使用方差分析和混合效应模型，在控制性别和语言的个体差异后，考察了情绪对这些参数的影响。研究结果揭示了细致的跨语言及情绪特异性模式，凸显了语言、文化因素与情感韵律之间复杂的相互作用。

随后，我们考察了人类与自监督学习模型在跨语言语音情感识别中的表现。通过比较单语和跨语言场景，以及迁移学习设置，结果表明，在跨语言语音情绪识别任务中，人类的表现始终优于自监督学习模型。然而，采用参数高效微调策略并进行逐层分析后，自监督学习模型取得了显著进步，在目标语言中达到了接近母语者的表现。这些发现凸显了自监督学习模型通过先进微调技术适应跨语言语音情感识别的潜力。

该研究还探讨了SED，重点关注对情感显著语音片段的细粒度识别。对所选语料库的韵律分析揭示了人类和自监督学习模型在检测片段级情绪时所采用的不同策略。人类擅长跨语言准确识别情绪片段，而自监督学习模型则展现出适应并捕捉片段级情绪的能力，在迁移学习场景下尤其能够达到出色的精确率。这些发现凸显了人类感知与机器学习在理解和处理片段级情感线索方面的互补优势。

对SED任务中选定语音片段的韵律模式进行进一步分析，可得出以下结论：首先，中文和德语参与者在不同语言中都表现出对强度相关特征的共同依赖，凸显了强度作为情绪感知线索的普遍重要性（Juslin & Laukka, 2001; Xiong et al., 2008）。与此同时，未选定片段中观察到的显著差异表明，仅凭本研究考察的声学参数不足以完全捕捉母语者所感知的细微情感线索，这凸显了其他声学或语言因素可能发挥的作用。此外，音高相关特征呈现出跨语言共性。例如，在所有语言中，f_0 对于快乐情绪始终最高，其次是愤怒，悲伤最低。然而，对音高特征的解读在不同文化和语言语境中有所不同。这些发现表明，情绪感知同时受到普遍声学特征和语言特异性因素的影响。未来研究应纳入更多样化的数据集，并控制说话人变异、文化背景和语言类型学等变量，以进一步完善这些结论。

参考文献

Ahn, Y., Lee, S. J., & Shin, J. W.（2021年）。基于少样本学习和领域自适应的跨语料库语音情感识别。IEEE Signal Processing Letters, 28, 1190–1194。

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M.（2020年）。wav2vec 2.0：语音表征自监督学习框架。神经信息处理系统进展, 33, 12449–12460。

Banse, R., & Scherer, K. R.（1996年）。情绪的声音表达中的声学特征。《Journal of Personality and Social Psychology》，70(3)，614。

Bolinger, D. L.（1978年）。跨语言语调。载于《人类语言的普遍性》（第471–524页）。斯坦福大学出版社。Cai, X., Wu, Z., Zhong, K., Su, B., Dai, D., & Meng, H.（2021年）。使用领域对抗神经网络进行无监督跨语言语音情感识别。第12届国际中文口语处理研讨会（ISCSLP），1–5。Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., & Xiao, X.（2022年）。Wavlm：用于全栈语音处理的大规模自监督预训练。IEEE信号处理精选主题期刊，16（6），1505–1518。Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D.（2019年）。映射24种情绪：人类简短发声所传达的情绪。《美国心理学家》，74（6），698–712。Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M.（2000年）。“FEEL-TRACE”：一种实时记录感知到的情绪的工具。ISCA语音与情感教程及研究研讨会（ITRW）。Ekman, P., Sorenson, E. R., & Friesen, W. V.（1969年）。情绪面部表情中的跨文化元素。《科学》，164(3875)，86–88。Elfenbein, H. A., & Ambady, N.（2003年）。情绪识别中的普遍性与文化差异。《Current Directions in Psychological Science》，12（5），159–164。Erickson, D.（2010年）。日本、韩国和美国听众对韩国说话者回忆过去情绪事件的感知：一些声学线索。2010年语音韵律——第五届国际会议。Feng, T., & Narayanan, S.（2023年）。PEFT-SER：探讨使用参数高效迁移学习方法和预训练语音模型进行语音情感识别。第11届国际情感计算与智能交互会议（ACII），1–8。Hancil, S.（2009年）。情感语音中的韵律作用（第97卷）。Peter Lang。Houlsby, N., Giurgiu, A., J astrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S.（2019年）。面向自然语言处理的参数高效迁移学习。国际机器学习会议，2790–2799。Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W.（2021年）。LoRA：大语言模型的低秩适配。国际学习表征会议。Jeon, J. H., Le, D., Xia, R., & Liu, Y.（2013年）。语音跨语言情感识别初步研究：自动分类与人类感知的比较。Interspeech 2013，2837–2840。Juslin, P. N., & Laukka, P.（2001年）。预期情绪强度对情绪声音表达中的线索利用和解码准确率的影响。《Emotion》，1（4），381。Kitayama, S., & Ishii, K.（2002年）。词语与声音：两种语言中对情感话语的自发注意。《认知与情感》，16(1)，29–59。Lashkarashvili, N., Wu, W., Sun, G., & Woodland, P. C.（2024年）。用于语音情感识别和领域自适应的参数高效微调。2024年IEEE国际声学、语音与信号处理会议（ICASSP），10986–10990。Latif, S., Qadir, J., & Bilal, M.（2019年）。用于跨语言语音情感识别的无监督对抗领域适应。第8届国际情感计算与智能交互会议（ACII），732–737。Latif, S., Qayyum, A., Usman, M., & Qadir, J.（2018年）。跨语言语音情感识别：乌尔都语与西方语言的比较。国际信息技术前沿会议（FIT），88–93。Li, Y., Bell, P., & Lai, C.（2022年）。在语音情感识别联合训练中融合自动语音识别输出。2022年IEEE国际声学、语音与信号处理会议（ICASSP），7362–7366。

Li, Y., Mohamied, Y., Bell, P., & Lai, C. (2023). 自监督语音模型探索：情感语料库研究。IEEE spoken language technology workshop (SLT), 868–875. Mac, D.-K., Aubergé, V., Castelli, E., & Rilliard, A. (2012). 局部与整体韵律线索：声调对法国人跨感知越南语态度韵律的影响。Speech Prosody. Mehrabian, A. (1996年). 愉悦—唤醒—支配：描述和测量个体气质差异的通用框架。Current Psychology, 14(4), 261–292. Neumann, M. (2018). 英语和法语的跨语言与多语言语音情感识别。2018 IEEE国际声学、语音与信号处理会议 (ICASSP), 5769–5773. Pasad, A., Chou, J.-C., & Livescu, K. (2021年). 自监督语音表征模型的逐层分析。2021 IEEE自动语音识别与理解研讨会 (ASRU), 914–921. Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). 外语中的情绪识别。Journal of Nonverbal Behavior, 33(2), 107–120. Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). 声音表达的情绪识别因素：四种语言的比较。Journal of Phonetics, 37(4), 417–435. Russell, J. A. (1980). 情感环形模型。Journal of Personality and Social Psychology, 39 (6), 1161. Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). 从声音表达推断的情绪在不同语言和文化中具有相关性。Journal of Cross-Cultural Psychology, 32(1), 76–92. Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., & Herre, J. (2018). webMUSHRA——用于基于网页的听音测试的综合框架。Journal of Open Research Software, 6(1), 187. Shoukat, M., Usama, M., Ali, H. S., & Latif, S. (2023). 打破壁垒：多语言基础模型能否弥合跨语言语音情感识别的差距？2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), 1–9. Singh, A., & Gupta, A. (2023). 解码情绪：用于语音情感识别的语音模型多语言综合研究（编号：arXiv: 2308.08713）。arXiv. https://doi.org/10.48550/arXiv.2308.08713. Steiner, I., Schröder, M., & Klepp, A. (2013). PAVOQUE语料库作为表现性语音分析与合成资源。Proceedings Phonetik & Phonologie, 9. Thompson, W. F., & Balkwill, L.-L. (2006). 五种语言中的语音韵律解码。Semiotica, 2006 (158). https://doi.org/10.1515/SEM.2006.017. Upadhyay, S. G., Busso, C., & Lee, C.-C. (2024). 提升跨语言语音情感识别的层锚定策略。Proceedings Interspeech 2024, 4693–4697. Wang, Y., Ravanelli, M., & Yacoubi, A. (2023). 语音情感分段识别：情绪何时出现？2023 IEEE自动语音识别与理解研讨会 (ASRU), 1–7. Werner, S., & Petrenko, G. K. (2019). 语音情感识别：人类与机器。Discourse, 5 (5), 136–152. Williams, C. E., & Stevens, K. N. (1972). 情绪与语音：若干声学相关因素。The Journal of the Acoustical Society of America, 52(4B), 1238–1250. Xiong, W., Cui, D., Meng, F., & Cai, L. (2008). 情感语音韵律特征的分析与转换。第八届中国语音学学术会议暨纪念吴宗济先生百年诞辰音系学前沿问题国际研讨会论文集， 5, 971–975. Xiong, Z. (2019). xSegmenter：音素自动分段与标注工具。Journal of Chinese Phonetics, 11, 27–34.

Zhang, G., Wang, Y., & Huang, J. (2020年). 普通话中声学线索被掩蔽时对情感语音的感知。《中国语音学报》, 13, 14–23.

Zhou, K., Sisman, B., Liu, R., & Li, H.（2022）。情感语音转换：理论、数据库与 ESD。Speech Communication，137，1–18。

作者信息

GENG Tianqi

卡内基梅隆大学迪特里希人文与社会科学学院语言、文化与应用语言学系硕士生。她的研究兴趣包括语音韵律、实验音系学和应用语言学。

电子邮箱：tianqig@andrew.cmu.edu

HAN Zhichen

硕士，毕业于爱丁堡大学哲学、心理学与语言科学学院。研究兴趣包括语音韵律和语音情感识别。

电子邮箱：z.han@shisu.edu.cn

FENG Hui

博士，天津大学外国语学院副教授。研究兴趣包括语音韵律和实验音系学。

电子邮箱：fenghui@tju.edu.cn

袁家宏

博士，中国科学技术大学人文与社会科学学院教授。目前的研究重点包括通过口语早期检测和预测阿尔茨海默病，运用机器学习方法研究汉语声调的起源与演变，以及从声调识别的角度探索模型-大脑对齐。

邮箱：jiahongyuan@ustc.edu.cn

Korin Richmond

博士，爱丁堡大学语音技术研究中心副教授。研究兴趣包括人类语言与语音技术、发音-声学关系以及语音合成。他曾担任多个国际及英国语音会议的组织委员会成员。

邮箱：korin@cstr.ed.ac.uk

李远超

博士，中国科学技术大学语言科学跨学科研究中心访问研究员。他的研究聚焦于利用以语音为中心的多模态人工智能检测心理、认知和健康状态。他的研究成果曾获多个机构颁发的奖项，包括 IEEE信号处理学会、ACM计算机-人机交互特别兴趣小组、艾伦·图灵研究所和施密特科学。他曾在多个知名语音会议上共同组织多项活动。他是本文的通讯作者。

电子邮箱：yl809@cam.ac.uk

附录

本研究所用代码现已发布于：
`https://github.com/zhan7721/Crosslingual SER`