arXiv:2603.09307v1 [cs.SD] 10 Mar 2026

日语共情口语对话中的副语言情感感知验证时机检测

Zi Haur Pang、*Yahui Fu*、*Yuan Gao*、*Tatsuya Kawahara*

京都大学信息学研究科，日本京都{pang, fu, gao, kawahara}
@sap.ist.i.kyoto-u.ac.jp

摘要

情感验证是一种心理治疗沟通技巧，包括识别、理解并明确承认他人的感受和行为，从而加强治疗联盟并减少负性情绪。为最大限度地发挥验证所提供的情感支持，适时且适度地表达至关重要。本研究从语音角度探讨验证时机检测。通过结合副语言和情感信息，我们提出一种无需依赖文本语境的副语言与情感感知验证时机检测模型。具体而言，我们首先在不同的 HuBERT 骨干网络上进行持续自监督训练和微调，以获得：（i）具备副语言感知能力的自监督学习（SSL）编码器；（ii）多任务语音情绪分类编码器。随后，我们融合这两种编码器，并在下游验证时机检测任务上进一步微调整合后的模型。在 TUT 情感叙事语料库（TESC）上的实验评估比较了多种模型、融合机制和训练策略，结果表明，所提出的方法显著优于传统语音基线。研究结果显示，非语言语音线索与情感相关表征相结合时，能够提供足够的信息来判断何时应表达肯定，为更具共情能力的人机交互提供了一种以语音为先的途径。

索引词— 情感验证、共情、口语对话系统（SDS）

1. 引言

共情在人-机器人和人-智能体交互中至关重要：当系统表现出“理解”用户时，人们会透露更多信息、遵循指导并建立信任。在心理咨询和心理健康领域，共情虚拟智能体和机器人与更高的感知共情、信任度和满意度相关，但研究也提醒，时机不当或流于泛泛的表达会显得不真诚 [1, 2]。然而，常见的共情短语（例如，听到这个消息我很难过）往往听起来像套话，对那些压抑情绪或难以说出情绪的人尤其如此。在心理治疗中，一种更具体的技巧——**情感验证（肯定）**——能够发挥作用。肯定是指识别、理解并明确承认他人的感受和行为；早期研究将肯定划分为多个层次，并指出这一技巧与更好的治疗联盟及较低的负性情绪相关 [3]。后续研究报告称，在临床和青少年群体中，肯定能够预测痛苦程度的降低和参与度的提升 [4, 5]。

至关重要的是，日常生活中的肯定既通过言语，也通过副语言信号传达：语调、音高、响度、时机，以及简短的听者回应（反馈语）。语音科学研究表明，语音线索能够可靠地传递情绪，并影响人们对温暖感等特质的判断和可信度的判断，即使词汇内容保持不变 [6, 7]。由此带来一个实际问题：系统应当何时给予肯定？过度肯定可能显得不真诚；肯定不足则会错失提供支持的机会。会话研究表明，听者回应的时机受声学韵律模式和局部互动信号的引导。在日语会话中，简短的听者回应（相槌）出现频繁，且位于特定位置；低音高区域、停顿和边界前延长等韵律触发因素，可能引出反馈语，或预示话轮即将交出[8, 9]。这些发现表明，仅凭语音或许就能检测出合适的肯定时机，而无需依赖词汇语境。

在对话系统中，人们通过常识推理 [10], 情绪原因抽取 [11], 和角色设定条件化 [12, 13]来实现共情。然而，直接研究肯定的工作寥寥无几，而且大多基于文本。近期研究将肯定建模为一个流程，依次包括认可时机检测、情绪识别和认可回应生成，并在日语共情对话中取得了初步成果，但仅凭非语言语音线索能在多大程度上传递相关信息，仍是一个开放问题 [14, 15]。与此同时，语音研究中有大量关于话轮交替和附和语预测的工作，这些工作利用韵律和时机来预测听者行为，涵盖日语会话和多模态场景，为语音优先的认可时机检测奠定了良好基础 [16, 17]。

本文检验这样一个假设：通过联合建模日语情感对话中的副语言信息和情绪信息，可以仅凭语音检测出恰当的肯定时机。我们不使用词汇语境，单独考察时机决策（“现在是否应该肯定”），并提出一种双分支语音方法：(i) 在情感对话上对 HuBERT 编码器进行持续预训练，以突出副语言模式；(ii) 对第二个 HuBERT 编码器进行多任务微调，以学习情绪线索。随后，我们融合这两个编码器，并针对最终的认可时机检测器进行微调。我们使用 TUT情感叙事语料库（TESC）进行评估，全面比较该方法与强大的语音和语言基线，并分析融合方法和训练策略。

2. 提出的方法

2.1. 多任务情感识别

当前的自监督语音编码器（如 HuBERT 和 wav2vec 2.0）是在通用语音上预训练的，因此可能与我们的下游情感口语对话任务不匹配。为缩小这一领域差距，我们对预训练模型进行微调，并在多任务学习（MTL）框架下通过辅助任务丰富特征学习。如图 1 上半部分所示，给定输入波形 U，我们将主干网络的最后一层隐藏序列 $\mathbf{f} \in \mathbb{R}^{t \times d}$ 作为潜在表示，其中 t 表示话语长度，d 表示隐藏维度。我们对

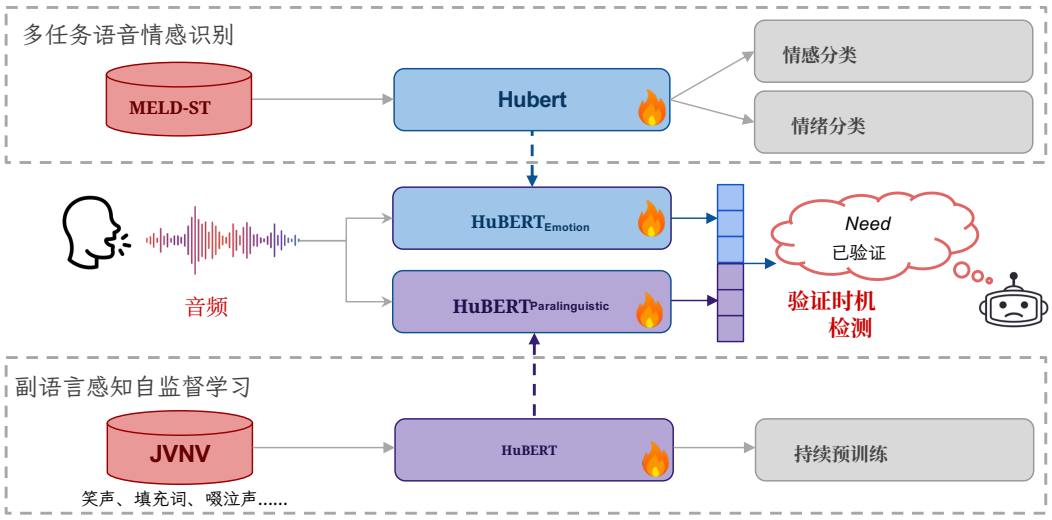


图 1。所提方法的架构 具备情绪感知能力的自监督学习 m 方法。该方法由以下部分组成：（1）多任务语音情感识别模型，（2）副语言感知自监督学习 tic- 型，以及（3）用于验证时机检测任务的特征融合。

时间维度进行均值池化（排除填充帧），以获得话语级向量

$$\bar{\mathbf{h}} = \frac{1}{t}\sum_{i=1}^{t} \mathbf{f}_i \in \mathbb{R}^d, \tag{1}$$

后接丢弃层（p=0.1）。两个线性头分别将 $\bar{\mathbf{h}}$映射为情绪和情感倾向逻辑值：

$$\mathbf{z}^e = \mathbf{W}^e\bar{\mathbf{h}} + \mathbf{b}^e, \quad \mathbf{p}^e = \mathrm{softmax}(\mathbf{z}^e), \tag{2}$$

$$\mathbf{z}^s = \mathbf{W}^s\bar{\mathbf{h}} + \mathbf{b}^s, \quad \mathbf{p}^s = \mathrm{softmax}(\mathbf{z}^s). \tag{3}$$

在这个多任务学习分支中，情绪被建模为细粒度的 7 类类别预测（愤怒、厌恶、恐惧、喜悦、中性、悲伤、惊讶），而情感倾向被建模为粗粒度的 3 类极性预测（负面、中性、正面）。同时学习这两项任务有助于获得互补的情感表征：情绪头捕捉离散的情感类别，而情感倾向头则提供辅助的极性信号，以正则化共享的话语表征。

为缓解情绪分类中的标签偏斜，我们采用类别加权交叉熵；情感倾向则使用标准交叉熵：

$$\mathcal{L}_{\text{emotion}} = -\sum_{c=1}^{C_e} w_c^{(e)}\, y_c^e \log p_c^e, \tag{4}$$

$$\mathcal{L}_{\text{sentiment}} = -\sum_{c=1}^{C_s} y_c^s \log p_c^s, \tag{5}$$

其中，C_e=7 和 C_s=3, y^e, y^s 是独热目标，$w_c^{(e)}$是在训练划分上估算的逆频率权重。

我们不固定任务权重，而是以端到端方式学习这些权重。具体而言，标量 logit 参数 $\tilde{\alpha}$ 与模型参数联合优化，并通过 Sigmoid 函数映射为 $\alpha = \sigma(\tilde{\alpha}) \in (0,1)$。总损失为

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{emotion}} + (1-\alpha)\,\mathcal{L}_{\text{sentiment}}, \tag{6}$$

该方法会在训练过程中调整各任务之间的权重，并免去手动调参。我们初始化 $\tilde{\alpha}$=0，使得在训练开始时 α=0.5。

2.2. 副语言感知自监督学习

受副语言学与验证之间关联的心理学证据启发，[7], 我们采用 HuBERT 式的掩码单元预测目标进行继续预训练，[18], 如图 1 下方所示。我们首先对 MFCC 帧运行 k-means，得到离散的伪目标单元，并将生成的单元 ID 与编码器帧率对齐。令K为簇的数量，$\{\mathbf{h}_t\}_{t=1}^{T}$ 为编码器输出的帧级隐藏状态。线性分类器将每个状态映射为单元 logits $\mathbf{z}_t \in \mathbb{R}^K$:

$$\mathbf{z}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b}, \qquad u_t \in \{1,\dots,K\}. \tag{7}$$

我们在特征轴上应用时间跨度掩码，通过采样跨度起点并掩码连续步数；只有被掩码的非填充位置会对损失产生贡献。记 $\mathcal{A}$ 为被掩码的非填充索引集合，

$$\mathcal{L}_{\text{SSL}} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} \mathrm{CE}(\mathbf{z}_t, u_t)\,. \tag{8}$$

特征速率长度通过主干网络的长度映射，根据原始注意力掩码计算得出，并据此生成特征速率掩码索引。填充步上的标签被设为从损失中排除。由此得到的编码器提供副语言感知表征，并将其作为下游融合中的一个分支。

2.3. 特征融合

为了检测表达肯定的正确时机，我们将第 2.1 节和第 2.2 节介绍的两个编码器进行拼接，如图 1 中部所示。每个编码器都会生成帧级隐藏状态序列；我们通过对未填充帧进行均值池化来获得话语嵌入，其中帧数由各编码器的特征提取器推导得出，以避免填充效应。

两种嵌入均通过线性层投影到共享空间（投影维度 256），随后接 GELU 与丢弃率 0.1。然后，我们将投影后的嵌入进行拼接，并应用最终线性层，生成用于“验证”或“不验证”的二分类 logits，并以交叉熵进行训练。

表 1。数据集的分布

数据集	#Validation	#Non-Validation
TESC	489	881
-训练	340	579
-*val*	74	157
-测试	75	145

3. 实验设置

3.1. 数据集

3.1.1. 预训练数据集

在多任务语音情感识别预训练中，我们使用 MELD-ST [19], MELD 的多语言扩展 [20]，它提供英语→日语（En–Ja）和英语→德语（En–De）配对。每组配对包含约 10,000 条带有情感标签的话语，这些话语摘自电视剧《老友记》，并由专业说话人以表演式语音呈现。在本研究中，我们仅使用日语部分，并重新划分语料库以消除说话人重叠，同时匹配划分规模和标签分布，得到训练/验证/测试比例为 $0.71/0.13/0.16$，说话人为 $199/36/46$，话语为 $6665/1623/1797$。

对于副语言感知自监督预训练，我们使用JVNV语料库 [21],，这是一个同时包含言语内容和非言语发声的日语情感语音数据集。JVNV包含约3.94小时由四名配音演员（两男两女）录制的表演式语音，涵盖六种情绪（愤怒、厌恶、恐惧、快乐、悲伤、惊讶）。每条话语只表达一种目标情绪。我们以8:2的比例进行说话人互斥划分，将其分为训练集和验证集。

3.1.2. 验证时机检测数据集

我们在TUT情感叙事语料库（TESC）[22]，上进行评估，这是一个日语双人多轮口语对话数据集，其中亲密朋友在八种普拉奇克情绪之一的提示下分享个人经历 [23]。TESC包含来自18个配对的247个场次（总计约9.2小时；平均每场次133.9秒）。标注人员利用音频及辅助的自动语音识别转录文本，将每个话语结束处标注为肯定 （应随后给出验证性回应）或非肯定。我们将标注数据按7:1.5:1.5的比例划分为训练/验证/测试集；结果见表1。

3.2. 实现细节

3.2.1. 多任务语音情绪识别

我们使用 Japanese HuBERT-Large[1] 作为主干网络，采用 AdamW [24] 进行微调，学习率为 1×10^{-5}，通过梯度累积实现有效批大小 16，并在验证集上采用早停策略（耐心值为 5）。我们每 500 步进行一次评估，并选择宏平均F1最高的检查点。最终模型的情绪识别非加权/加权准确率为 $30.43/46.13$，情感的非加权/加权准确率为 $52.06/54.81$。

3.2.2. 副语言学感知的自监督学习

对于持续预训练，我们使用 Japanese HuBERT-Large[1] with $K{=}100$ k均值梅尔频率倒谱系数单元（在 200k 帧上拟合）和跨度掩码（$p{=}0.065$，长度为 10）。我们以学习率训练 20 轮 1×10^{-5} with 批大小为 2，并进行 16 步梯度累积，每轮进行验证，并保存验证损失最低的检查点。

3.2.3. 认可时机检测

我们在单个 NVIDIA RTX 6000 Ada GPU 上训练所有模型，使用学习率 1×10^{-5} 在 TESC 上进行微调，批大小为 1，并采用 16 步梯度累积（最多 20 个轮次）、100 步预热和 0.01 的权重衰减。我们每 100 步进行一次验证，采用早停策略（耐心值为 5），根据 F1 选择最佳检查点，并将随机种子固定为 42。为缓解该有限规模数据集中肯定/非肯定样本的分布偏斜（340 对 579），我们通过下采样非验证话语，将训练划分平衡至 250 对 329。

3.3. 评估指标

在日常对话中，系统选择何时给予肯定，比它给予肯定的频率更重要。因此，我们将目标类精确率（即被预测为给予肯定 且确实值得肯定的轮次所占比例）视为主要指标。若模型将许多轮次标记为给予肯定（高召回率），就可能产生空洞或重复的回应，削弱人们感知到的共情；因此，如果高 F1 分数掩盖了较低的精确率，仅凭高 F1 分数可能会产生误导。为此，我们报告：**(i)验证精确率（V-Pre.）**，作为对话适切性的主要指标；**(ii)认可F1（V-F1）**，用于衡量目标类上的精确率-召回率权衡；**(iii) 非验证F1（NV-F1）**，用于衡量非目标类上的权衡；以及 **(iv)宏平均F1（M-F1）**，即对两个类别取平均，以确保多数非验证 类上的表现不会被忽视。

4. 结果与分析

4.1. 肯定时机检测结果

我们首先将本方法与公开可用的语音模型进行比较，即 Japanese HuBERT-Large[1] 和 xlsr-53-japanese[2]。表 2 中的结果显示，两个基线在此任务上的表现都较差，这可能是因为它们的预训练目标是一般性的声学规律，而非情感对话，从而造成了显著的领域失配。相比之下，我们结合了情绪感知线索和副语言线索的方法能够成功识别适合进行肯定的时刻，验证精确率达到 47.96，认可F1 达到 54.34，同时保持了较高的宏平均 F1，为 62.37。

除了语音编码器之外，由于我们的主要目标是检验在不理解文本语境的情况下，能否识别表达肯定的恰当时机，因此我们还将模型与公开可用的语言模型进行比较：传统语言模型（LM），例如 BERT[3] 和 ModernBERT[4],，以及大型语言模型（LLM），例如 Llama 3.1 8B Instruct[5] 和 GPT-4.1 Nano[6]。对于传统 LM，我们仅使用转写文本在自己的数据集上直接进行微调；对于 LLM，我们评估了零样本和三样本设置。结果表明，当前的 LLM（在零样本和

[1]https://huggingface.co/rinna/ 日语 HuBERT 大型模型
[2]https://huggingface.co/jonatasgrosman/ 日语 wav2vec 2.0 XLSR-53 大型模型[3] https://huggingface.co/tohoku-nlp/ 日语 BERT 大型模型[4] https://huggingface.co/sbintuitions/ 日语 ModernBERT 1.3 亿参数模型[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct[6]https://openai.com/index/gpt-4-1/

表 2。验证时机检测任务的结果 [%]。

	V-精确率.	V-F1	NV-F1	M-F1
语言模型				
BERT	39.96	41.62	66.76	54.19
ModernBERT	41.32	44.65	68.65	56.65
Llama 3.1 8B				
- 零样本	33.91	49.04	40.48	44.76
- 三样本	33.69	50.30	46.10	48.20
GPT 4.1 Nano				
- 零样本	39.97	49.43	58.97	54.20
- 三样本	39.44	46.55	61.93	54.24
语音模型				
HuBERT	38.51	51.12	49.77	50.45
Xlsr-53	37.04	47.62	52.17	48.89
本文方法	**47.96**	**54.34**	**70.41**	**62.37**

表 3。消融研究 [%]。

	V-精确率.	V-F1	NV-F1	M-F1
Para. HuBERT	**52.54**	46.27	**76.47**	61.37
情感 HuBERT	45.63	52.81	67.94	60.73
本文方法	47.96	**54.34**	70.41	**62.37**

三样本设置下）难以识别正确的肯定时机，而经过微调的传统 LM 表现更好，但仍逊于我们仅使用语音的方法。值得注意的是，尽管未使用文本语境，我们的方法仍优于所有基于语言模型的基线，证明了在无需单独进行转写的口语对话场景中，采用语音优先的验证时机方法是可行的。

4.2. 消融研究

为评估各组件的贡献，我们比较了：*Para.HuBERT*（在第 2.2 节中介绍的持续预训练 HuBERT，它能够学习副语言线索），以及情感 *HuBERT*（第 2.1 节介绍的独立情感编码器）。表 3 中的结果表明，情感编码器提升了时机检测性能（验证 F1 为 52.81，宏平均 F 为 60.73），这凸显了情感信息对于时机验证的重要性，与基于文本的发现一致 [14]。值得注意的是，尽管 MELD-ST 和 JVNV 是表演语料库，它们仍有助于在自然对话语料库 TESC 上进行建模。副语言分支*Para. HuBERT* 的验证精确率达到 52.54%，表明副语言线索能够支持验证时机检测，这与心理学证据一致 [7, 25]。

4.3. 不同融合策略的评估

接下来，我们评估不同融合策略对认可时机检测的影响。结果汇总于表 4。除了我们提出的简单拼接方法外，我们还测试了三种可学习的融合变体：**(i) 注意力**：通过两层多层感知机为每个编码器嵌入$\mathbf{h} \in \mathbb{R}^d$ 打分，并使用 $tanh$对各分支的分数进行归一化，再通过加权和得到融合向量（在 $\mathbb{R}^d$ 上进行分类）；**(ii) 门控**：将两个投影后的嵌入拼接起来，并通过 Sigmoid 门$\mathbf{g} \in (0,1)^d$生成逐元素混合结果$\mathbf{g} \odot \mathbf{h}_1 + (1-\mathbf{g}) \odot \mathbf{h}_2$（在 $\mathbb{R}^d$ 上进行分类）；**(iii) 多头注意力 (MHA)**：将两个嵌入视为长度为 2 的序列，并输入单层多头注意力层（4 个头），随后对输出进行池化并送入分类器（在 $\mathbb{R}^d$ 上）。与这些方法相比，**拼接**（我们提出的方法中的简单特征拼接方法）在验证精确率、认可F1 和宏平均 F1 上均表现最佳（47.96%、54.34%、62.37%）。我们将此归因于数据集规模较小且类别不平衡：容量更高的融合方法（注意力/门控/MHA）更容易过拟合，而拼接在低数据场景下是一种稳健且稳定的融合方法。

表 4. 不同融合策略的评估 [%]。

	V-精确率.	V-F1	NV-F1	M-F1
注意力	43.81	51.11	66.15	58.63
门控	44.00	50.29	67.17	58.73
MHA	46.67	50.91	**70.55**	60.73
拼接（我们的方法）	**47.96**	**54.34**	70.41	**62.37**

表5。不同训练策略的评估 [%]。Para.：副语言编码器；Emo.：情感编码器

Para.	Emo.	V-精确率.	V-F1	NV-F1	M-F1
冻结	冻结	0.00	0.00	79.45	39.72
微调	冻结	44.90	50.87	68.16	59.52
冻结	微调	**61.90**	27.08	**79.65**	53.37
LoRA	微调	48.53	46.15	74.07	60.11
微调	LoRA	48.61	47.62	73.72	60.67
微调	微调	47.96	**54.34**	70.41	**62.37**

4.4 不同训练策略的评估 s

我们评估了不同训练策略对验证时机表现的影响；结果汇总于表 5。我们比较了四种基础设置：(i) 冻结两个编码器；(ii) 微调其中一个编码器，同时冻结另一个编码器；(iii) 对其中一个编码器进行全面微调，同时使用 LoRA 对另一个编码器进行参数高效更新；以及 (iv) 微调两个编码器（默认设置）。

微调两个编码器取得了最佳的整体结果（V-F1 54.34 1 62.37, M-F1），这表明联合适配使两个分支能够针对目标标签分布及互补线索进行协同专门化。冻结两个编码器会导致目标类别表现崩溃（V-精确率/V-F1 均为零），这与冻结特征在专门化时机任务中的局限性相符。冻结副语言分支、仅微调情感编码器，虽然取得了最高精确率（V-精确率 61.90），但 V-F1 很低（27.08），这表明由于固定的副语言特征未能对齐，决策边界过于保守，召回率较低。对一个分支使用 LoRA（同时微调另一个分支）取得了有竞争力的宏平均分数（M-F1 60.11 / 60.67），但 V-F1 较低（46.15 / 47.62），说明与全面微调相比，其紧密协同适配的能力较弱。

5. 结论与未来工作

我们提出了一种副语言与情绪感知模型，以便在不依赖对话上下文的情况下，更好地捕捉表达肯定的正确时机。实验表明，同时纳入副语言线索与情感线索，使模型能够更准确地检测恰当的验证时机。所提方法在肯定类上的精确率和F1分别达到47.96%和54.34%，相较当前语音基线的绝对提升分别为10.92%和6.72%，相较当前文本基线的绝对提升分别为6.64%和9.69%。在未来工作中，我们将把该框架扩展至多模态和多语言场景，并将其部署到仿人机器人平台上，用于真实世界的人机交互。

6. 致谢

本研究得到JST NEXUS 251043539以及JST Moonshot研发计划JPMJPS2011的支持。

7. 参考文献

[1] Mate Szondy 和 Peter Fazekas，“对机器人的依恋与心理健康治疗效率”，《心理学前沿》，第15卷，第1347177页，2024年。[2] Iolanda Leite、Andr´e Pereira、Samuel Mascarenhas、Carlos Martinho、Rui Prada 和 Ana Paiva，“共情对人机关系的影响”，《国际人机研究杂志》，第71卷，第3期，第250–260页，2013年。[3] Marsha M. Linehan，“肯定与心理治疗”，载于《重新审视共情：心理治疗的新方向》，第353–392页。美国心理学会，1997年。[4] Kendra S Wasson Simpson、Anna Gallagher、Scott T Ronis、David AA Miller 和 Kate C Tilleczek，“青少年对否定和肯定在其心理健康治疗历程中所产生影响的感知”，《心理健康管理与政策及心理健康服务研究》，第49卷，第3期，第476–489页，2022年。[5] Amanda Carson-Wong、Christopher D Hughes 和 Shireen L Rizvi，“治疗师在个体辩证行为疗法治疗场次中采用认可策略对来访者情绪变化的影响”，《人格障碍：理论、研究与治疗》，第9卷，第2期，第165页，2018年。[6] Rainer Banse 和 Klaus R Scherer，“语音情绪表达的声学特征”，《人格与社会心理学杂志》，第70卷，第3期，第614页，1996年。[7] Joshua J Guyer、Pablo Bri˜nol、Thomas I Vaughan-Johnston、Leandre R Fabrigar、Lorena Moreno 和 Richard E Petty，“通过语音传达的副语言特征会影响自信评价和评价性判断”，《非言语行为杂志》，第45卷，第4期，第479–504页，2021年。[8] Guy Itzchakov 和 Harry T Reis，“倾听与感知到的回应性：揭示其意义并探讨关键研究方向”，《当代心理学观点》，第53卷，第101662页，2023年。[9] Michael W Kraus，“纯语音交流能提升共情准确度”，《美国心理学家》，第72卷，第7期，第644页，2017年。[10] Sahand Sabour、Chujie Zheng 和 Minlie Huang，“CEM：常识感知共情回复生成”，载于*AAAI*人工智能会议论文集，2022年，第36卷，第11229–11237页。[11] Jun Gao、Yuhan Liu、Haolin Deng、Wei Wang、Yu Cao、Ji- achen Du 和 Ruifeng Xu，“通过识别对话中的情绪原因改进共情回复生成”，载于*EMNLP 2021*，2021年，第807–819页。[12] Yahui Fu、Chenhui Chu 和 Tatsuya Kawahara，“StyEmp：通过多粒度前缀编码器和人格强化实现共情回复生成的风格化”，载于话语与对话特别兴趣组第*25*届年会论文集，2024年，第172–185页。

[13] Mingxiu Cai、Daling Wang、Shi Feng 和 Yifei Zhang，“Pecer：通过动态人格提取和上下文情感推理生成共情回复”，载于*ICASSP 2024—2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第10631–10635页。[14] Zi Haur Pang、Yahui Fu、Divesh Lala、Keiko OCHI、Koji IN- OUE 和 Tatsuya KAWAHARA，“基于情感故事语料库预测认可性回应”，载于日本人工智能学会第*37*届年会，2023年，第2O5OS2a03–2O5OS2a03页。[15] Zi Haur Pang、Yahui Fu、Divesh Lala、Keiko Ochi、Koji Inoue 和 Tatsuya Kawahara，“认可情绪状态：为共情对话生成认可性回应”，口语对话系统技术国际研讨会，2024年。[16] Nigel Ward 和 Wataru Tsukahara，“英语和日语中提示反馈回应的韵律特征”，《语用学杂志》，第32卷，第8期，第1177–1207页，2000年。[17] Agust´ín Gravano 和 Julia Hirschberg，“面向任务对话中的话轮交替提示”，《计算机语音与语言》，第25卷，第3期，第601–634页，2011年。[18]Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed，“HuBERT：通过隐单元掩码预测进行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。[19] Sirou Chen、Sakiko Yahata、Shuichiro Shimizu、Zhengdong Yang、Yihang Li、Chenhui Chu 和 Sadao Kurohashi，“MELD-ST：情感感知语音翻译数据集”，载于计算语言学协会 *ACL 2024* 年会论文集，2024年，第10118–10126页。[20] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea，“MELD：用于对话情感识别的多模态多方数据集”，载于计算语言学协会第*57*届年会论文集，2019年，第527–536页。[21] Detai Xin、Junfeng Jiang、Shinnosuke Takamichi、Yuki Saito、Akiko Aizawa 和 Hiroshi Saruwatari，“JVNV：包含言语内容和非言语表达的日语情感语音语料库”，*IEEE Access*，第12卷，第19752–19764页，2024年。[22]Hikaru Oishi、Mika Enomoto、Keiko Ochi 和 Yasunari Obuchi，“TUT情感故事语料库的设计与基础分析”，载于*O-COCOSDA*。IEEE，2021年，第43–48页。[23] Robert Plutchik，“情绪的本质：人类情绪有着深厚的进化根源，这一事实或可解释其复杂性，并为临床实践提供工具”，《美国科学家》，第89卷，第4期，第344–350页，2001年。[24] Ilya Loshchilov 和 Frank Hutter，“解耦权重衰减正则化”，载于国际学习表征会议，2019年。[25] Elina Weiste 和 Anssi Per¨akyl¨a，“心理治疗互动中的韵律与共情交流”，《心理治疗研究》，第24卷，第6期，第687–701页，2014年。