人类反馈驱动的动态语音情感识别

Ilya Fedorov[1], *Dmitry Korobchenko*[2]

[1]NVIDIA, Switzerland

[2]NVIDIA, UK

ilyaf@nvidia.com, dkorobchenko@nvidia.com

Abstract

本工作旨在探索动态语音情感识别的新领域。与传统方法不同，我们假设每个音频轨道都与一系列在不同时刻活跃的情感序列相关联。该研究特别关注情感化 3D 虚拟角色的动画制作。我们提出了一种多阶段方法，包括训练经典语音情感识别模型、合成生成情感序列以及基于人类反馈进一步改进模型。此外，我们引入了基于狄利克雷分布的情感混合建模的新方法。这些模型是根据从 3D 面部动画数据集中提取的真实情感进行评估的。我们将我们的模型与滑动窗口方法进行了比较。实验结果表明，基于狄利克雷的方法在情感混合建模方面有效。结合人类反馈进一步提高了模型质量，同时提供了一种简化的标注程序。

Index Terms: 语音情感识别，角色动画

1. 介绍

语音情感识别（SER）通过分析语音表达来确定情绪状态，在促进人机交互方面发挥着重要作用。传统上，SER 为语音录音分配静态的情感标签，这有助于各种领域的应用，从客户服务到医疗保健和教育。虽然这种静态方法很有用，但它往往无法捕捉随着时间变化的情绪表达的流动性，从而限制了它在需要详细情感理解的情景中的适用性。

来自大型语言模型（LLMs）近期进展的一项应用是创建智能会话代理。这些代理以 3D 数字人类的形式出现时，可以作为视频游戏中的高级非玩家角色（NPC），提供动态且响应迅速的互动。此外，它们还可以在教育或治疗环境中充当虚拟助手，提供个性化的支持。这一技术飞跃不仅丰富了各个领域的用户体验，还为人工智能系统中情感智能的整合设定了新的标准。

我们探讨了这种数字人物的真实感，使用了一个预训练的 NVIDIA Audio2Face 网络 [1] 来为 3D 角色制作动画。该神经网络通过处理两个关键输入：人类语音的音频及其对应的情绪状态序列，生成面部动画，包括嘴唇、眉毛、眼睛和脸颊的动作。在此基础上，我们引入了动态语音情感识别（DSER），这是一种新的方法，可以从音频中预测随时间变化的情绪序列。DSER 超越了传统的 SER 方法，为 3D 角色的动画提供了情绪智能的新维度，达到了前所未有的真实感水平。

DSER 模型的开发和训练面临重大挑战，原因是数据标注过程非常复杂。对于经典的监督学习方法，需要手动为每个训练音频中的许多时间戳分配情感标签。这是一个细致的过程，要求注释者仔细地标记每个时刻的情感。此外，情感的模糊性进一步使这一过程变得复杂。

鉴于这些数据标注挑战，我们选择使用人类反馈来微调我们的模型。这种方法显著推动了 LLMs 领域的发展，在地面真相标签模糊的情况下提供了重大优势，使得识别所需行为比明确展示它更容易。

在这项研究中，我们提出了一种用于训练动态语音情感识别系统的多步骤管道。该过程包括经典 SER 模型的训练、基于滑动窗口方法为 DSER 生成合成序列到序列数据以及使用直接偏好优化算法 [2] 进行的人类反馈微调。此外，我们通过狄利克雷分布建模引入了一种新的形式化 SER 问题的方法。我们的模型在一个从面部动画数据集中通过优化过程提取的情感序列的小数据集上进行了评估。我们将我们的方法与滑动窗口方法进行

了比较。结果显示，所提出的方法在预测质量上显著优于启发式方法。

2. 相关工作

我们对动态语音情感识别（DSER）的探索采用了 Wav2Vec2.0 架构 [3]，相较于传统的神经网络模型如 CNN 和 LSTM，在预测情绪 [4] 方面展示了显著的进步。基于 Transformer 的架构，包括 Wav2Vec2.0 和 HuBERT[5]，之前已被用于情感识别 [6, 7]。

我们积极利用了 NVIDIA Audio2Face[1] 神经网络来可视化我们的结果并收集人类偏好的数据集。

我们利用直接偏好优化（DPO）[2]，这是一种推进了从人类反馈中进行强化学习（RLHF）[8, 9] 概念的技术。这种方法使得模型可以直接基于人类的反馈进行训练，而无需训练一个单独的奖励模型。

3. 提出的

我们的方法是多阶段的。在第一阶段，我们训练一个神经网络来预测整个音频轨道的情感表示。在第二阶段，我们训练一个序列到序列模型，该模型能够在一次正向运行中预测一系列情感，并利用第一阶段训练的模型生成的合成数据。在最终的第三阶段，我们进一步使用人类反馈对第二阶段的模型进行微调。这些阶段如图 1 所示。

在深入描述这些阶段的细节之前，让我们首先考虑我们对情绪识别任务形式化的做法。为了可视化我们的实验结果，我们使用神经网络根据语音音频和一序列情绪生成三维面部动画。这个序列由数值向量表示，范围从 0 到 1，其中每个向量对应特定时间点的一种连续的情绪描述。这些向量中的每个标量代表特定情绪活动的程度，如愤怒或喜悦。在这种情况下，我们的兴趣在于创建一种混合情绪的概率模型，而不是边缘类别的概率。与传统分类任务假设对象和不同类别之间存在直接联系不同，我们旨在将一个对象与反映混合情绪的分类分布关联起来。形式上，我们的目标是开发一个估计不是类别边际概率，

$$p(\text{emotion} \mid \text{audio}),\ \text{emotion} \in \{\text{anger, sadness, ..., joy}\},$$

而是情绪混合联合概率的概率模型，

$$p(\text{anger} = p_1, \text{sad} = p_2, \ldots, \text{joy} = p_n \mid \text{audio}),$$
$$p_i \in (0,1), \quad \sum_{i=1}^{N} p_i = 1.$$

我们提出使用狄利克雷分布来建模情感混合，这是一种由 N 个正数值参数化的 N 维连续分布，$\alpha = \{\alpha_i > 0\}_{i=1}^{N}$，其定义域为单形：

$$\text{Dir}(x \mid \alpha) \propto \prod_{i=1}^{N} x_i^{\alpha_i - 1}, \quad x \in \mathcal{C}, \tag{1}$$

$$\mathcal{C} = \{(x_1, x_2, \ldots, x_N) \in \mathbb{R}^N \mid x_i > 0, \sum_{i=1}^{N} x_i = 1\}. \tag{2}$$

此公式表明来自狄利克雷分布的样本表示一个分类分布。因此，我们旨在建模的情感混合实际上是来自特定狄利克雷分布的样本。为了定义此分布，我们使用参数为 θ 的神经网络 f 预测其参数：

$$\alpha = f_\theta(\text{audio}) \tag{3}$$

$$\text{emotion} \sim \text{Dir}(\text{emotion} \mid \alpha) \tag{4}$$

模型 f_θ 是通过最大化训练数据集 $\{(\text{audio}_i, \text{emotion}_i)\}_{i=1}^{K}$ 上的似然性进行训练的：

$$\theta^* = \arg\max_\theta \frac{1}{K} \sum_{i=1}^{K} \log \text{Dir}(\text{emotion}_i \mid f_\theta(\text{audio}_i)). \tag{5}$$

经此方式训练的模型能够估计任何情绪混合的概率。为了推断特定预测，我们使用狄利克雷分布的期望值：

$$\text{emotion}_i = \frac{\alpha_i}{\sum_{i=1}^{N} \alpha_i}. \tag{6}$$

3.1. 建模奇异情感表示

在第一阶段，我们训练模型来预测整个音频轨道的一种单一情感混合。我们使用经典的监督

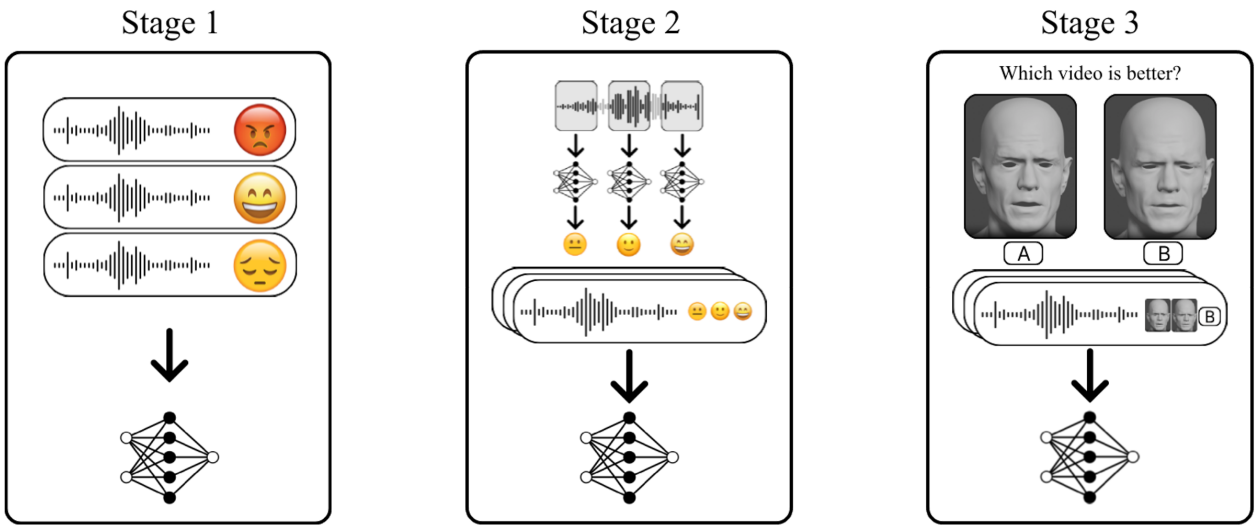

图 1: 提出方法的三个阶段。在第一阶段，我们训练模型 A 来预测整个音轨的一种情感。在第二阶段，我们使用模型 A 通过滑动窗口方法生成顺序数据，并在此数据上训练模型 B。在最终阶段，我们通过结合人工反馈进一步改进模型 B。

方法和目标函数（5）解决这个问题。我们利用了一个由（音频，情感）对组成的数据库，其中情感以离散的一热编码标签表示。我们使用了一组六个情感类别：愤怒、厌恶、恐惧、快乐、中性和悲伤。这种选择基于公共领域中数据集的可用性。

3.2. 情感建模序列

在动态语音情感识别中，我们的目标是从为整个音频轨道识别单一的情感向量转变为预测带有相应时间戳的一系列情绪。实现这一转变的明显方法是通过滑动窗口技术。该技术涉及将一个在第一阶段训练好的模型应用于音频的小型且均匀间隔的部分，从而构建情感时间线。然而，这种方法遇到了两个主要挑战：由于模型仅处理短波形段而受限的情境理解能力；以及窗口固有的固定网格假设，限制了情绪的时间戳位置，并需要搜索过程来确定最优的窗口大小和步幅。

为了克服这些限制，我们建议训练一个能够在一个前向传递中直接预测整个情感序列的模型。这使得模型可以观察音频的完整上下文，并精确决定何时改变情感。由于缺乏具有时间特定情感标注的数据集，我们通过滑动窗口方法创建了合成数据以实现这一目的。我们将第一阶段训练的模型应用于生成此类样本，然后训练了一个新模型来重现这种行为。

3.3. 纳入人类反馈

在最终阶段，我们通过学习人类反馈范式进一步微调了第二阶段训练的模型。为此，我们为每个音频轨道生成了多条情感序列，将这些序列输入面部动画神经网络，并渲染出展示数字人物表现各种情感序列的视频。然后，我们通过对这些视频对进行标注来选择更优的选项。

数据集创建后，我们使用直接偏好优化（DPO）算法对第二阶段的序列到序列模型进行了微调。该方法利用以下损失进行优化：

$$\mathcal{L} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\mathrm{Dir}_\theta(y_w\,|\,x)}{\mathrm{Dir}_{\mathrm{ref}}(y_w\,|\,x)} - \beta\log\frac{\mathrm{Dir}_\theta(y_l\,|\,x)}{\mathrm{Dir}_{\mathrm{ref}}(y_l\,|\,x)}\right)\right]. \tag{7}$$

在此公式中，期望是在三元组 (y_w, y_l, x) 上进行的——即优选情绪、不优选的情绪和相应的音频。$\mathrm{Dir}_\theta(y_w\,|\,x)$ 是要训练的模型，其参数为 θ，而 $\mathrm{Dir}_{\mathrm{ref}}(y\,|\,x)$ 则是如 DPO 论文中详细解释的参考模型。我们使用第二阶段的模型冻结拷贝作为参考模型。β 参数控制可训练模型的行为与参考模型之间的差异程度。损失（7）为序列中的每种情绪

表 1: 第一阶段训练数据集的详情。

	愤怒	厌恶	恐惧	喜悦	中性	悲伤	
CREMA-D[10]	1271	1271	1271	1271	1087	1271	7442
RAVDESS[11]	192	192	192	192	96	192	1056
锦 XL[12]	240	-	-	480	240	240	1200
私有	1404	1404	1404	1404	1404	1404	8424
	3107	2867	2867	3347	2827	3107	18122

计算，然后取平均值。此公式清楚地说明了为什么开发了带有 Dirichlet 分布的方法：如果我们仅仅有一个边缘情绪的概率模型，而不是它们的混合物，我们将无法表达公式中使用的联合概率。

4. 实验

4.1. 数据集和评估

阶段 1。为了增加数据的多样性和规模，我们的训练数据集是从几个公开的数据源和一个私有数据集中编译而成的，我们不披露该私有数据集是因为商业利益。最终数据集的组成详见表 1。它包括 18122 个样本，对应 15 小时的音频，每个音频剪辑的平均时长为 3 秒。所有录音均为英文，并且性别比例大约为 1:1。

为了更好地评估模型准确性，我们采用了跨数据集验证。我们独立收集了一个包含 516 个样本的测试数据集，每个情绪有 86 个样本。音频记录来自 7 名说话者，每个人用英语录制了一半的样本，另一半则使用其母语进行录制。我们将标准分类准确率作为评估指标。

阶段 2。为了生成随时间变化的情感数据，我们要求一个情感动态的数据集。第一阶段使用的标准数据集不适合，因为这些音频样本主要具有静态情绪特征。相反，我们使用了 VoxMovies[13, 14] 集合，该集合包含电影和电视节目中的对话音频轨道。我们使用了一个大小为 1.4 秒、步长为 1 秒的滑动窗口来预测数据集中每个音频轨道的情感序列，共得到了 676 个序列到序列样本。

评估这样一个模型构成了一个显著的挑战。一种方法是使用保留样本的标准方法，并将预测与预先生成的数据进行比较。然而，在这种情况下，度量值仅会指示模型复制滑动窗口方法的程度。

如前所述，我们有一个预训练的面部动画神经网络，能够根据输入音频和情绪序列生成 3D 数字角色。此外，我们还有一个包含 3D 面部动画和相应音轨的小数据集。为了创建用于评估我们的序列到序列模型的真实数据集，我们通过优化 NVIDIA Audio2Face 网络（A2F）相对于输入情绪的权重来从该数据集中提取情绪序列。具体来说，我们将 A2F 的权重冻结，并找到了使真实动画与 A2F 预测之间的均方误差（MSE）最小的情绪序列。形式上，对于动画中的每一帧，我们解决了以下优化问题：

$$\text{emotion}^* = \arg\min_{\text{emotion}} \text{MSE}(\text{A2F}(\text{audio}, \text{emotion}), \text{GT}). \tag{8}$$

使用这种方法，我们直接从真实面部运动中提取了 40 个情绪序列。虽然这些数据不足以训练一个 DSER 模型，但足以评估其质量。由于商业利益，我们不公开此数据集。

作为度量标准，采用了平均绝对误差（MAE）。由于驱动的神经网络使用了不同的情感空间，特别是零情感向量对应于中性情绪，因此 Kullback-Leibler 散度的应用受到了限制。

第三阶段 To gather the human feedback, we utilized the manually filtered MELD [15] dataset. This dataset represents speech recordings from a comedy show. Because of that it contains a lot of samples with background laughter, biasing the prediction towards the 喜悦 emotion. We removed those samples, resulting in a total of 400 samples. For each audio track included, we generated 5 distinct sequences of emotions. Randomization involved applying the model from the 1st stage to the samples using the sliding window approach with varying window sizes and strides. These parameters were randomly selected from a range between 0.25 sec to 1.25 sec, under the condition that similar parameters could not repeat, to increase the diversity of generated samples. Subsequently, we rendered videos for each sequence of emotions. For the same audio, the annotator was shown 2 videos. The goal was to choose the one where the

表 2: 单情感预测模型的评估。

损失	骨干结构	准确性
Cross-Entropy	Large	**0.84**
Dirichlet Likelihood	Large	0.81
Cross-Entropy	Base	**0.8**
Dirichlet Likelihood	Base	0.79

表 3: 顺序模型的评估。

方法和损失函数	骨干结构	平均绝对误差
Seq2Seq + DPO	Large	**0.195**
Seq2Seq + Dirichlet	Large	0.200
Seq2Seq + Cross-Entropy	Large	0.201
Sliding Window + Dirichlet	Large	0.206
Sliding Window + Cross-Entropy	Large	0.207
Seq2Seq + DPO	Base	**0.203**
Seq2Seq + Dirichlet	Base	0.206
Seq2Seq + Cross-Entropy	Base	0.208
Sliding Window + Dirichlet	Base	0.228
Sliding Window + Cross-Entropy	Base	0.231

sequence of emotions was more preferable. In the result, we obtained 1500 samples of side-by-side comparisons of emotional sequences. We do not disclose this dataset due to commercial interests.

4.2. 模型架构。

在我们所有的实验中，我们都使用了 Wav2Vec2.0 模型。该模型处理音频波形以生成特征序列。在第 1 阶段，我们通过时间上的平均池化这些特征，为每个音频片段创建一个单向量。在第 2 和 3 阶段，我们应用基于滑动窗口的平滑处理，其中 kernel=5 和 stride=1。之后，我们应用全连接层将特征维度降低到 6，以匹配目标情感的数量。然后通过映射 $g(x) = x^2 + 10^{-6}$ 将这些输出转换为适合狄利克雷分布参数的正值。

4.3. 训练详情

我们试验了通过 Hugging Face[16] 平台提供的两个预训练 Wav2Vec2.0 检查点：*facebook/wav2vec2-base*（94M 参数）和 *facebook/wav2vec2-large-lv60*（315M 参数）。此后，我们将分别将这些模型称为 Base 和 Large。

我们进行了大约 200 次运行以找到第一阶段模型的最佳超参数，因为此模型随后用于生成合成数据。验证是在我们收集的数据集的保留部分上进行的。总体而言，该模型证明了对超参数变化具有较强的鲁棒性。在最佳检查点中，我们采用了以下设置：100 个纪元，Adam 优化器的学习率从 5×10^{-5} 衰减到 10^{-5}，批量大小为 16，权重衰减为 10^{-4}。此外，我们还使用了诸如音频位移、裁剪和噪声注入等增强技术。对于后续阶段，我们使用相同的参数。在所有实验中，我们将输入音频样本重采样到 16 kHz。

4.4. 结果与讨论

表 2 和表 3 分别展示了单情感模型和序列到序列模型的评估结果。虽然传统的交叉熵损失在预测整个音频的情感时在准确性方面显示出略好的结果，但狄利克雷方法对于顺序混合建模的结果更好。加入人类反馈进一步提升了结果，达到了最先进的质量。

我们考察了损失函数（7）中正则化参数 β 的影响，注意到较高的 β 值使模型更紧密地与参考模型对齐，这符合 DPO 理论的假设。这种相关性通过实证观察得到了证实。此外，过低的 β 值导致模型丧失预训练特征，从而降低性能指标。表 4 表示了该参数对大型检查点的影响。值 $\beta = 0.5$ 显示了从人类反馈中学习和利用参考模型知识之间的最优平衡。报告的数值代表了训练过程中达到的最佳指标分数。

尽管基于人类反馈的方法为注释和扩展数据集提供了更简单的程序，但在实际操作中，有时很难比较所提出的感情序列以进行选择，这降低了标签的效率。这可以被视为该方法的局限性。

我们的方法在一个配备了 NVIDIA RTX 3090 GPU 的节点上进行了训练和评估。每个阶段的训

表 4: β 参数的影响。

β	平均绝对误差
0.01	0.215
0.1	0.197
0.5	**0.195**
10.0	0.200

练需要 1-2 小时。在 FP16 格式下使用 NVIDIA TensorRT 加速的最大检查点推理，处理 6 ms 的输入音频大约需要 1 sec，从而实现模型的实时应用。

5. 结论

在这项研究中，我们介绍了动态语音情感识别（DSER），这是一种适用于 3D 角色动画的新型历时语音情感识别方法。我们的方法利用合成数据生成并结合人类反馈微调，旨在克服传统情感识别技术中的挑战。结果显示了我们的方法在准确建模情绪混合序列方面的效率，并强调了人类反馈在提升模型性能方面的重要性。我们看到进一步的研究方向包括扩展数据集和对情绪混合建模进行更详细的控制。

6. References

[1] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, jul 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073658

[2] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://arxiv.org/abs/2305.18290

[3] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219966759

[4] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451 – 3460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[6] L. Pepino, P. E. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *ArXiv*, vol. abs/2104.03502, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233181984

[7] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *ArXiv*, vol. abs/2111.02735, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:242757022

[8] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[9] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, "Learning to summarize from human feedback," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[10] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[11] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, p. e0196391, May 2018. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0196391

[12] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," 09 2018, pp. 2768–2772.

[13] A. Brown, J. Huh, A. Nagrani, J. S. Chung, and A. Zisserman, "Playing a part: Speaker verification at the movies," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021*, 2020.

[14] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6