在DareFightingICE中的深度强化学习盲目AI

摘要--本文在IEEE CoG 2022年的DareFightingICE竞赛中，介绍了一个在DareFightingICE平台上使用声音作为输入的深度强化学习AI。在这项工作中，只使用声音作为输入的人工智能被称为盲人工智能。虽然最先进的人工智能主要依靠其环境提供的视觉或结构化观察，但只从声音中学习玩游戏仍然是新事物，因此具有挑战性。我们提出了不同的方法来处理音频数据，并将近似策略优化算法用于我们的盲目人工智能。我们还建议使用我们的盲目人工智能来评估提交给比赛的声音设计，并为这项任务定义三个指标。实验结果表明，不仅我们的盲目人工智能有效，而且所提出的三个衡量标准也有效。

索引词--声音，盲目人工智能，深度强化学习，近似策略优化，格斗游戏，FightingICE，DareFightingICE

I. 简介

长期以来，视频游戏中的声音一直是一个重要因素[1]-[3]。在游戏中，有不同类型的声音效果，从用户界面声音、环境声音、背景音乐等等。这些不同类型的声音效果提高了人类玩家的兴致。人类玩家利用游戏中的这些声音输入来完成各种任务，例如，寻找物品或敌人的位置，并通过其特定的声音来识别物体。视频游戏中的声音可以以不同的方式帮助人类玩家，但我们在本文中讨论的问题是。人工智能（AI）玩家可以从视频游戏的声音中学习吗？

最近的研究表明，人工智能播放器可以使用声音作为输入来检测物体的位置和方向[4]。一个将声音作为输入与其他信息一起使用的人工智能播放器被证明比那些不使用的人表现更好[5]。然而，使用声音作为输入的人工智能的研究仍然处于不成熟的阶段，以前的研究要么集中在只从音频线索学习玩简化的游戏，要么在他们的人工智能中使用视觉和其他输入，如游戏数据与音频数据。我们的研究引入了一个只使用声音作为输入的人工智能，来玩一个名为 "DareFightingICE "的格斗游戏。

DareFightingICE[6]是FightingICE[7]的增强版，FightingICE是一款格斗游戏，自2013年以来一直被用作格斗游戏AI竞赛系列的平台。DareFightingICE是2022年IEEE游戏大会的一个正式比赛。DareFightingICE有一个 "仅有声音"的选项，只允许AI选手接收音频数据作为输入。

这项工作的贡献如下：第一，创造了第一个格斗游戏人工智能，这是比赛中的官方样本人工智能，只使用声音作为输入（盲目的人工智能）；第二，我们使用人工智能来评估一个给定的声音设计在游戏中事件表现能力方面的有效性；第三，我们的工作为盲目的人工智能研究打开了一扇新的大门。

II. 相关的工作

A. 声音的AI技术

音频信号处理是人工智能的最重要领域之一。近年来，由于深度学习变得越来越普遍，它被应用于音频处理，因此在语音识别[8], [9]和文本转语音[10], [11]等应用中获得了成功。在深度学习中应用音频处理的一种常见方式是将音频数据转换为图像，然后像处理其他图像一样处理它们。这可以通过生成频谱图来实现，频谱图是二维图像，代表着沿两轴的时间和频率的频谱序列，颜色代表频率成分的强度。频谱图可以通过对音频信号应用短时傅里叶变换（STFT）得到，信号的STFT可以用快速傅里叶变换（FFT）计算。也可以对频谱图应用Mel-frequency scale，以形成Mel-spectrogram，这种方法更适合人类的感知，在[8]和[9]中得到了应用。在我们的工作中，作为我们的盲目人工智能的音频编码器，我们比较了三种类型的转换，它们使用2层卷积神经网络（CNN）、FFT的组合和2层全连接网络

(FCN)，以及Mel-spectrogram和一个2层CNN的组合，分别。

B. 基于声音的游戏AI

有一些现有的研究集中在使用声音玩游戏的人工智能。Gaina和Stephenson[4]扩展了通用视频游戏人工智能框架以支持声音，并训练人工智能只用声音作为输入进行游戏。Hegde等人[5]扩展了标准的VizDoom框架[12]，向AI提供游戏中的声音，并在一系列越来越复杂的场景中训练AI，以测试对声音的感知能力。这些研究的结果显示了对从声音中学习玩游戏的AI的研究潜力。然而，在这些研究中，AI只接受了一种声音设计的训练，并没有被用来评估游戏中声音设计的有效性。因此，据我们所知，我们的工作是第一次用不同的声音设计来训练人工智能，并评估其有效性。

C. 近似的政策优化

近端策略优化（PPO）[13]算法是recent年流行的强化学习方法。PPO在以前的策略梯度方法的基础上提供了一个可靠的信任区域策略优化，并优于传统的方法，如Q-learning。其策略损失函数如下。

$$L_t^{CLIP}(\theta) = \hat{E}_t[\min(\rho_t(\theta)\hat{A}_t, clip(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \quad (1)$$

$$\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{老}}(a_t|s_t)} \quad (2)$$

$$\hat{A}_t = \sigma_t + (\gamma\lambda)\sigma_{t+1} + \ldots + (\gamma\lambda)^{T-t+1}\sigma_{T-1} \quad (3)$$

$$\sigma_t = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t) \quad (4)$$

在上述方程中，s_t 和 a_t 是状态和行动。在时间步数t，分别为$\pi_\theta(a_t|s_t)$和$\pi_{\theta_{老}}(a_t|s_t)$为分别为当前政策和先前政策的$s_t$，$t$的概率。$V_\theta(s_t)$是状态$s_t$的价值函数，$E$、γ、λ分别是剪裁、广义优势估计（GAE）和折扣系数。在格斗游戏人工智能竞赛中，PPO曾被用于一些研究中[14]-[16]并取得了令人瞩目的成功，特别是可以从2021年的冠军和亚军都使用PPO作为他们的人工智能的一部分看出。[1].此外，PPO在音频处理任务中表现突出，例如在多语言环境中基于音频的导航[17]和语义视听导航，其中物体的声音与它们的语义一致[18]。在这项工作中，我们使用PPO来训练我们的盲人AI。

[1] https://www.ice.ci.ritsumei.ac.jp/ ftgaic/index-R.html

D. 格斗游戏AI

许多最先进的算法在格斗游戏AI竞赛中测试了他们的性能，包括深度强化学习（DRL）。Kim等人[14]利用深度强化学习与自我游戏和蒙特卡洛树搜索（MCTS）创建了一个格斗游戏人工智能代理，他们后来提出了一个强化学习代理，通过重用神经网络表征对环境进行轻微调整[16]。2020年，Tang等人[19]提出了一种将滚动地平线进化算法与对手模型相结合的方法，并赢得了当年的比赛。2021年，Liang等人[15]扩展了上述工作，提出了具有基于Elo的选择机制的PPO，其中强大的历史AI在训练中被更频繁地选择。然而，在2021年之前，所有这些AI都使用由比赛提供的框架数据。从2022年开始，使用名目繁多的游戏平台的DareFightingICE比赛[6]已经启动，参赛AI只需要根据音频数据进行比赛，而无法获得帧数据信息。因此，我们的工作是第一次努力训练人工智能玩格斗游戏，只使用声音作为输入。

III. 方法论

A. 预处理

DareFightingICE提供的原始音频数据的形式是向量$s \in [-1, 1]^n$，包含n个归一化的音频样本。原始音频数据的大小为800，每个两个通道（左和右），但每个通道用零填充，这样它就有1024个样本，以便在Java中进行FFT[6]。然而，在我们的工作中，我们选择只使用每个通道的800个原始样本。激励我们的是Hedge等人[5]，我们提出并比较了三种编码器来在将音频数据送入深度神经网络之前，下文将对其进行处理。图1显示了以下架构这些编码器的所有参数都是基于以前的工作[5]，但根据经验为我们的工作进行了调整。

*1) 1D-CNN。*音频数据通过每8 个 th，并送入两个一维卷积层进行降采样。降采样有助于降低计算的复杂性。最后。得到一个32x5的音频特征向量。

*2) FFT。*使用FFT将输入的音频数据s转换到频域，并将FFT数据与幅度的自然对数$sFFT$ = 纵向比较。$logFFT(s) \in R^{n/2}$。然后对所得数据进行下采样最后，一个一维的 "一 "字，被送入一个两层的FCN。得到256的音频特征向量。

*3) Mel-spectrogram。*输入的数据用短期傅里叶变换（STFT）转换为频域频谱图，STFT是一个以给定跳数移动的窗口信号的傅里叶变换序列。然后用Mel scale处理这些频率。对于超参数的设置，我们选择了10毫秒的跳数，25毫秒的窗口大小和80个梅尔频率成分。然后，频谱数据被送入一个由两个二维卷积层组成的网络。最后，得到一个32x40x1的音频特征向量。

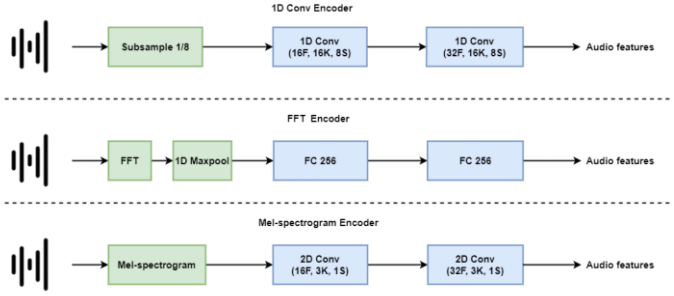

图1.音频编码器。1D-CNN（顶部），FFT（中间）和Mel-spectrogram（底部）。

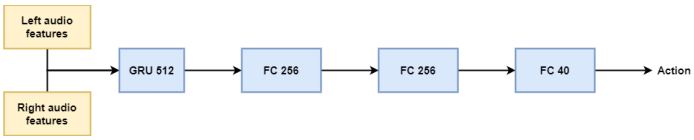

图2.我们的盲目人工智能架构。

表一
PPO的超参数设置。

超参数	价值
亚当步长	$3 * 10^{-4}$
优化代用的历时数	10
小批量生产的规模	64
折扣 (γ)	0.99
GAE参数（λ）。	0.95
GRU隐藏单位	512

B. AI设计

DRL指的是越来越多的强大算法，这些算法使用深度神经网络在具有高维状态和行动的环境中学习。在我们的工作中，正如前面在II-C的结尾所说，我们使用PPO，其架构和奖励在下文中描述。

*1) 网络结构。*我们的模型由一个选定的上一节给出的音频编码器，一个门控递归单元[20]，以及一个全连接的三层网络来产生动作概率。全连接层网络由三层组成。该网络的输入是音频编码器的输出。有两个隐藏层，每层有256个节点。输出层包含40个节点代表40项行动[2].我们遵循PPO超参数从[15]中设置，如表I所示。我们的人工智能的结构在图2中描述。

*2) 奖励的定义。*按照以前工作中的配方[21]，我们将奖励函数定义如下。

$$奖励_t = 奖励_t^{offense} + 奖励_t^{防卫} \quad (5)$$

$$奖励_t^{进攻} = HPopp_t - HPopp_{t+1} \quad (6)$$

$$奖励_t^{防御} = 自身HP_{t+1} - 自身HP_t \quad (7)$$

其中t和$t+1$分别代表当前帧的步长和后续步长；HP表示感兴趣的角色的命中率，如果该角色（*自己*）收到来自对手（*对手*）的伤害，HP将减少，在本作品中以及比赛中，在一个游戏回合开始时初始化为400值。

C. 竞争指标

在这里，我们提出了三个指标来评估一个给定的声音设计和/或一个给定的音频编码器的有效性：*面积*，$赢得_{ratio}$，和$avgHP_{diff}$。它们的定义是，一个声音设计或音频编码器越有效，其三个指标的值就越高。

*1) 学习曲线。*为了得到这个指标，对于一个给定的声音设计，我们首先训练三个盲目的AI，每个都使用不同的音频编码器。使用的对手AI是前面比赛中的MCTS样本AI的弱化版，即Kan等人[6]中讨论的MctsAi65。每次训练持续900个游戏回合。我们将一个历时定义为15个连续回合的间隔，并在每个历时中计算出以下数据的平均值$HP_r^{self} - HP_r^{opp}$在每轮结束时r超过最最近的15轮。因此，由60个点构建了一条学习曲线。然后我们对感兴趣的学习曲线进行多项式回归，计算从起点到终点的面积，如公式（8）所示。

$$面积 = \int_{x1}^{x60} poly(x)dx。 \quad (8)$$

其中x_1和x_{60}分别是学习曲线的起点和终点。

*2) 与对手AI的战斗性能：*我们通过让每个训练有素的盲人AI与上述对手AI战斗90个回合来评估其战斗性能。在90个回合中获胜的次数之比[3]式（9），以及训练有素的人工智能与其对手在回合结束时的平均HP差，式（10），然后是计算出来的。

$$胜率 = \frac{获胜轮次}{总回合数} \quad (9)$$

$$平均值_{扩散} = \frac{所有r的HP_t^{self} - HP_t^{opp}之和}{总回合数} \quad (10)$$

IV. 结果和讨论

图3和图4显示了我们的盲目人工智能使用三种编码器的学习曲线，分别用于2022年DareFightingICE比赛中的声音设计和2021年格斗游戏人工智能比赛中的FightingICE声音设计。表二总结了他们的*面积*的值。从这个表中可以看出，声音DareFightingICE的设计给出了一个更高的*面积*值比前两个音频编码器的FightingICE要好。在

[2] 使用中的行动包括2个投掷入地，12个攻击入地，3个技能入地，7个移动入地，2个守卫入地，12个攻击入空，2个技能入空行动。

[3] 在游戏中，当对手的HP达到零时，拥有非零HP的一方，或在达到60秒的回合长度限制时，拥有较高HP的一方为回合赢家。

图3.我们在DareFightingICE声音设计中使用不同编码器的盲目人工智能的学习曲线。

图4.我们在FightingICE声音设计中使用不同编码器的盲目人工智能的学习曲线。

表二
从我们的盲人AI的学习曲线来看，在与对手AI战斗时，有不同的声音设计和不同的音频编码器。

声音设计	编码器	*地区*
敢打敢拼ICE	1D-CNN	-30.11
敢打敢拼ICE	FFT	**173.94**
敢打敢拼ICE	梅尔谱图	-120.71
FightingICE v4.5	1D-CNN	-54.41
FightingICE v4.5	FFT	77.03
FightingICE v4.5	梅尔谱图	79.3

表三
我们的盲目人工智能在不同的声音设计和不同的音频编码器下的表现，以及在对抗
OPPONENT AI.

声音设计	编码器	*胜率*	$avgHP_{diff}$
敢打敢拼ICE	1D-CNN	0.29	-41.49
敢打敢拼ICE	FFT	**0.7**	**42.9**
敢打敢拼ICE	梅尔谱图	0.51	8.81
FightingICE v4.5	1D-CNN	0.28	-61.10
FightingICE v4.5	FFT	0.54	15.87
FightingICE v4.5	梅尔谱图	0.46	6.64

图5.在DareFightingICE的声音设计中，对手AI用火球攻击时的音频数据的FFT结果。

此外，当FFT编码器用于该声音设计时，可以达到最高的*面积*。

表三显示了我们的盲目人工智能在声音设计和音频编码器的每个组合中的战斗性能。

图6.在FightingICE声音设计中，对手AI用火球攻击时的音频数据的FFT结果。

DareFightingICE的声音设计在每个编码器的两个性能指标上都优于FightingICE的声音设计。这是预料之中的，因为DareFightingICE的声音设计是其前身的增强版，并以视障玩家为目标，尽管由于其作为2022年比赛的样本和基线声音设计的性质，还有改进的空间。

现在，我们讨论人工智能的行为[4].特别是，我们专注于当声音设计是DareFightingICE和FightingICE的时候，AI行为的差异，两者都使用FFT编码器。在DareFightingICE的声音设计中，只要不攻击，盲人AI就会倾向于自我防卫，但在其他声音设计中，AI不会这样做。此外，在FightingICE的声音设计中，盲人AI不能避免对手的火球技能，因为在技能发射时没有声音提示。相反，因为在DareFightingICE的声音设计中，当对手释放火球技能时，会播放声音提示，所以盲人AI似乎能够识别声音提示，并尽可能地避开该技能。图5和图6分别显示了在DareFightingICE声音设计和FightingICE声音设计中，对手AI通过火球动作攻击时，所产生的FFT音频数据。当我们的盲人AI试图避开火球攻击时的游戏画面序列如图7所示。

上述结果证实，所提出的三个指标可以用来评估声音设计。在2022年的比赛中，使用FFT编码器的盲法人工智能将从头开始重新训练，以评估比赛的声音设计轨道中的每个参赛声音设计。此外，在这项工作中训练的版本是公开可用的[5]作为官方的盲目人工智能样本，并将在比赛的人工智能轨道中作为基线人工智能使用。

V. 结论

在本文中，我们在DareFightingICE平台上引入了一个只使用声音作为输入的盲人AI。我们还评估了该人工智能在与对手人工智能作战时使用不同音频编码器的性能，而对手人工智能的性能在以前的工作中已经为视障玩家做了调整。我们的盲人AI能够在FFT的情况下击败对手的AI。

4
六种声音设计和音频编码器组合的战斗视频样本，其中P1是盲人AI，P2是对手AI，可在 ailable https://tinyurl.com/BlindAICoG2022。
5https://tinyurl.com/DareFightingICE/SampleAI/BlindAI

编码器或 Mel-spectrogram编码器。还发现，FFT编码器是最好的。

为了评价DareFightingICE竞赛中的声音设计，我们提出了三个衡量标准。第一个指标是获得学习曲线的面积。第二个和第三个指标是与上述对手AI战斗时的胜率和平均HP差。我们的体验结果表明，DareFightingICE的声音设计比FightingICE的声音设计更有效。这证实了所提出的三个指标可以用来评价比赛中的参赛声音设计。

在未来，我们计划改进我们的盲目人工智能，使其更好地从声音观察中理解游戏状态，并在研究中使用人工智能来程序化地产生有效的声音设计。