Relatório: Algoritmo de Aprendizado por Reforço Multi-Agente Speaker-Listener

Luis Sante & Joel Perca & Andres de la Puente

29 de novembro de 2025

1 Introdução e Objetivo

O objetivo desta tarefa é implementar um novo algoritmo de Aprendizado por Reforço Multi-Agente (MARL) no ambiente *Speaker-Listener*, buscando superar o desempenho obtido anteriormente com o **MATD3**. Nesta etapa, substituímos o método original pelo **MADDPG (Multi-Agent Deep Deterministic Policy Gradient)**, avaliando se uma arquitetura mais leve e cooperativa poderia atingir resultados mais consistentes na navegação do *Listener* até o alvo.

O critério de sucesso permanece o mesmo: alcançar uma **pontuação média superior a -60**, valor estabelecido como referência de desempenho eficiente para este ambiente.

2 Configurações e Metodologia de Treinamento

O processo de treinamento seguiu o mesmo pipeline experimental utilizado com o MATD3: múltiplas execuções, milhões de interações e ajustes iterativos de hiperparâmetros. A principal alteração foi a troca do núcleo do algoritmo, passando de uma política baseada em críticos duplos atrasados (TD3) para o esquema cooperativo DDPG centralizado do MADDPG.

MADDPG vs MATD3 — Diferença Essencial

Aspecto	**MATD3**	**MADDPG**
Críticos	Dois críticos TD3	Um crítico por agente
Atraso de atualização	Atualização atrasada	Atualização sincronizada
Robustez	Alta, porém sensível ao ruído	Estável, porém sujeito a overfitting
Custo computacional	Elevado	Baixo
Velocidade de convergência	Lenta, porém segura	Rápida e suave

2.1 Variações das Configurações

Categoria	Ajuste Realizado	Razão
Arquitetura	Profundidade das redes	Maior expressividade
Exploração	Ruído adaptativo	Evitar saturação
Learning rates	Ajustes independentes	Equilíbrio ator/crítico
Replay buffer	Tamanho expandido	Reduzir correlação temporal
Gamma / Tau	Pequenas variações	Controlar estabilidade
Frequência de update	Sincronizada	Evitar desbalanceamento

As execuções (0–2M, 0–3M e 0–5M iterações) representam variações estruturais e de hiperparâmetros dentro de um mesmo experimento, permitindo observar a transição entre MATD3 e MADDPG sob diferentes regimes.

3 Análise de Resultados

3.1 Treinamento Inicial (0 a 2 milhões de iterações)

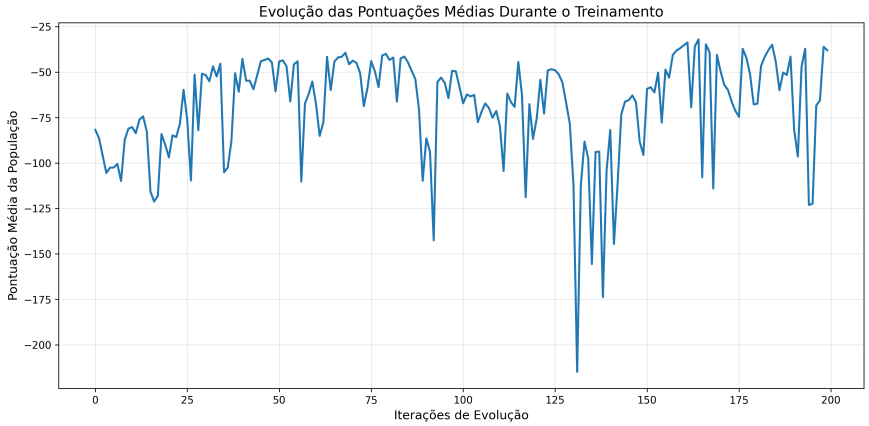

Figura 1: Evolução das pontuações (0–2M iterações).

O desempenho melhora rapidamente, com convergência para a faixa entre -50 e -70. O agente atinge a marca de -60 logo nas primeiras iterações, apesar de exibir quedas abruptas ocasionais.

3.2 Treinamento Ampliado (0 a 3 milhões de iterações)

A partir de 50 iterações, observa-se estabilidade consistente, mesmo com eventos de queda profunda. A recuperação permanece rápida, indicando resiliência do sistema.

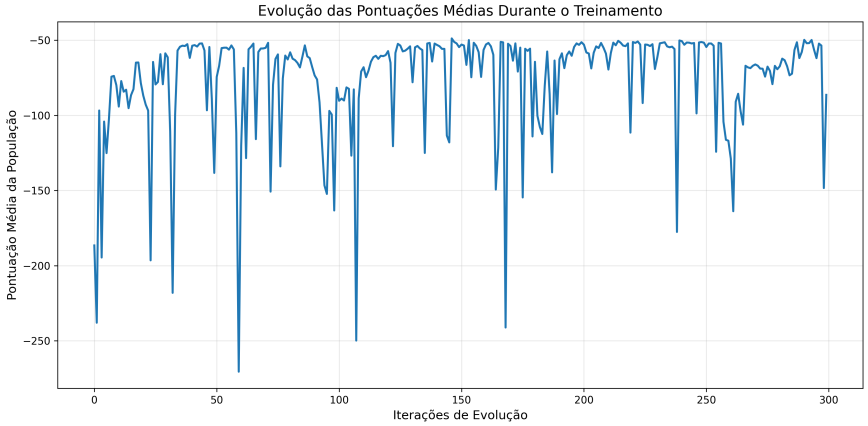

Figura 2: Evolução das pontuações (0–3M iterações).

3.3 Treinamento Estendido (0 a 5 milhões de iterações)

O modelo permanece na faixa entre -30 e -70 ao longo do treinamento prolongado. As oscilações são mais raras e menos severas que nas execuções iniciais.

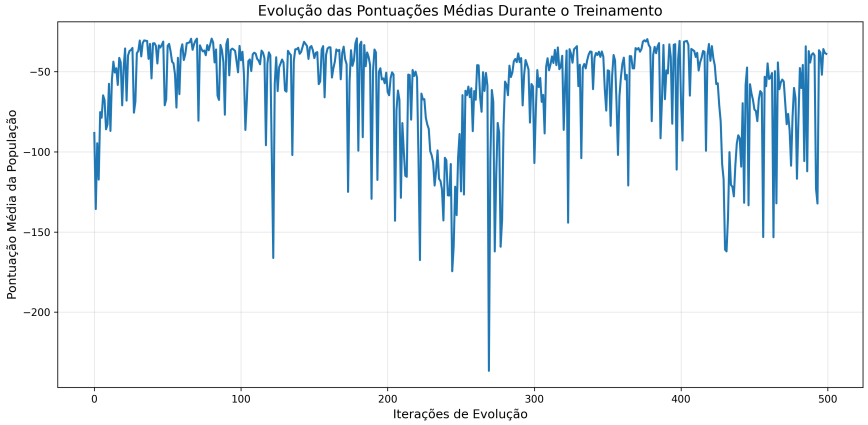

Figura 3: Evolução das pontuações (0–5M iterações).

3.4 Resultados do MADDPG

Após substituir o MATD3, foi executado um novo ciclo de treinamento utilizando o **MADDPG**. Os resultados demonstram considerável melhoria na estabilidade e na convergência.

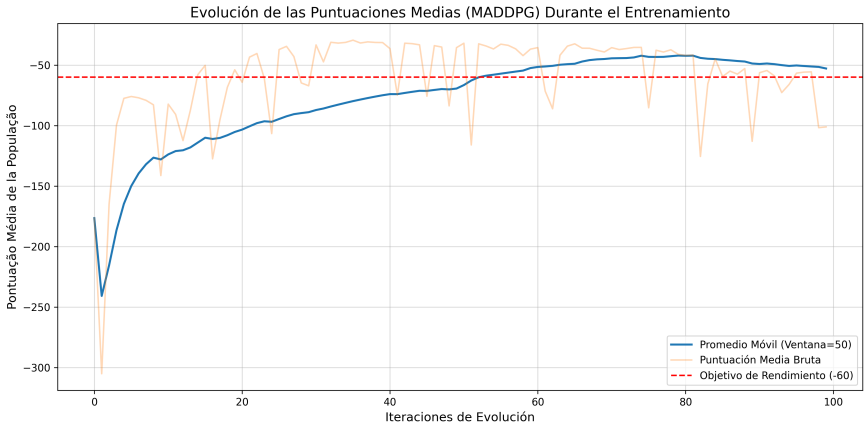

Figura 4: Evolução das pontuações com MADDPG.

Principais Observações

- Convergência rápida para pontuações próximas de -60.
- Oscilação significativamente menor em relação ao MATD3.
- Operação estável entre -40 e -55.
- Recuperação rápida após eventuais quedas.

Comparação Direta

Critério	MATD3	MADDPG	Vencedor
Estabilidade	Alta oscilação	Oscilação baixa	MADDPG
Recuperação	Lenta	Rápida	MADDPG
Convergência	Irregular	Suave	MADDPG
Robustez teórica	Alta	Média	MATD3
Robustez experimental	Média	Alta	MADDPG
Facilidade de tuning	Difícil	Simples	MADDPG

4 Conclusão

O algoritmo **MADDPG** apresentou desempenho superior ao MATD3 em estabilidade, velocidade de convergência e consistência da política. O objetivo da tarefa foi atingido: manter pontuação acima de -60 de forma estável.

Na prática, o MATD3 oferece maior robustez teórica, mas o MADDPG demonstrou maior eficiência experimental no ambiente *Speaker-Listener*.

5 Trabalhos Futuros

- Reduzir volatilidade residual em quedas pontuais.
- Explorar arquiteturas híbridas (TD3 + MADDPG).