

Integração de um Modelo de Reconhecimento de Emoções ao Robô Humanoide NAO

Integration of an Emotion Recognition Model to the Humanoid Robot NAO

Joyce Maria do Carmo de Sá ¹  orcid.org/0000-0001-8224-1323

Ingyrd Vanessa de Sá Teles Pereira ^{1,2}  orcid.org/0000-0002-4561-4335

Alexandre Magno Andrade Maciel ¹  orcid.org/0000-0003-4348-9291

¹ Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

² Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil.

E-mail do autor principal: Joyce Maria do Carmo de Sá: jmcs@ecomp.poli.br.

Resumo

Este trabalho destaca a importância da robótica e conseqüentemente do robô NAO na educação e como o reconhecimento de emoções pode tornar a interação humano-robô mais natural. O objetivo é integrar um mecanismo de reconhecimento de emoções através da representação geral da fala ao robô humanoide NAO. O modelo proposto é constituído de 4 etapas, integrados sequencialmente. O experimento foi realizado com 84 amostras áudio gravadas por 4 pessoas divididos entre homens e mulheres. Os dados mostram que após a integração o NAO é capaz de verificar a emoção que o humano está sentindo durante a interação com uma acurácia de 60%, tornando a comunicação mais natural.

Palavras-Chave: Robótica; Emoções; IA; Machine Learning.

Abstract

This paper shows the importance of robotics and consequently of the NAO robot in education and how emotion recognition can make human-robot interaction easily and natural. The goal is to integrate a mechanism of emotion recognition in speech within NAO robot. The proposed model is composed by four steps, sequentially integrated. The experiment was performed with 84 audio samples recorded by four people, between men and women. Data shows that after integration NAO robot is capable of identifying the emotion felt by the human during the interaction.

Key-words: Robotics; Emotions; IA; Machine Learning.

1 INTRODUÇÃO

A robótica é uma área que tem tido enorme crescimento nos últimos anos, em virtude do amplo espectro de possíveis aplicações. Na educação, a robótica pode ser empregada em pelo menos três vertentes [1]: utilizando (i) o robô como uma ferramenta de programação, como (ii) foco de aprendizagem ou ainda utilizá-lo como (iii) um colaborador no aprendizado. Nesse último aspecto, o robô humanoide NAO tem uma predominância quando comparado com outros robôs, isso se deve a fatores como sua ampla disponibilidade, aparência atraente, preço acessível e robustez técnica [2].

O NAO é capaz de interagir com humanos por meio dos seus diversos sensores. No Brasil, algumas escolas já utilizam o robô na educação básica oferecendo situações diferenciadas de aprendizagem, pois permite a conexão entre teoria e prática. Além disso, o humanoide vem sendo utilizado em estudos a respeito do autismo, onde é considerado uma plataforma significativa para apoiar e iniciar interação em crianças com Transtorno do Espectro Autista (TEA) [3]. Embora todas suas funcionalidades, o NAO não tem em sua implementação um mecanismo de reconhecimento de emoções [4].

Grande parte da inteligência humana é a emocional, o que torna o homem capaz de identificar o estado em que as pessoas se encontram emocionalmente. O sistema de emoções é responsável pela avaliação e julgamento dos eventos para verificar o quanto esses eventos são benéficos para si. Na interação humano-robô (IHR) o objetivo é uma melhora na qualidade da comunicação, fornecendo apoio em compromissos sociais e comportamentais. Conhecer e compreender emoções humanas ajuda robôs a adaptar a comunicação em tempo real, melhorando e enriquecendo essa interação [5]. Ao inserir um sistema emocional em um robô, ele o ajudaria a ter mais flexibilidade em ambientes complexos e incertos, bem como ajudaria o robô a se comportar socialmente de forma aceitável e eficaz com as pessoas [6].

Há um grande desentendimento na literatura a respeito de uma definição para emoção, devido a sua abstração e às diferentes perspectivas em que pode ser analisada. Após um estudo em que aborda diversas definições diferentes, Kleinginna & Kleinginna propõem que emoção é um conjunto complexo de interações entre fatores subjetivos e objetivos, mediados por fatores neurais, sistemas que podem (i) dar origem a experiências afetivas, como sentimentos de excitação, prazer / desprazer; (ii) gerar processos

cognitivos como efeitos perceptivos emocionalmente relevantes, avaliações, processos de rotulagem; (iii) ativar ajustes fisiológicos generalizados às condições de excitação; e (iv) levar a um comportamento que é frequentemente, mas nem sempre, expressivo, direcionado ao objetivo e adaptação [7].

A voz humana é um dos meios básicos de comunicação, graças ao qual também se pode transmitir facilmente o estado emocional. No reconhecimento de emoções através da fala tanto é possível analisar o conteúdo do que se foi dito, como também a entonação em que se foi falado. Em Pereira I. é proposto um modelo de representação geral da fala, que é construído com um *Generative Adversarial Network* (GAN) e treinado de forma não supervisionada e depois incorporado em um modelo supervisionado, construindo assim o modelo semi-supervisionado de reconhecimento de emoções em voz. O uso da representação geral da fala, treinado de maneira não supervisionada, melhora o desempenho da aplicação e também constrói um modelo adaptativo para outros cenários, uma vez que a representação da fala não fica presa no cenário do conjunto de dados avaliado [8].

Nesse trabalho buscamos incorporar o modelo proposto em Pereira [8] ao robô humanoide NAO de forma a enriquecer e melhorar a interação com o ser humano, tornando-a mais natural. Para tal, foi desenvolvido um mecanismo que pode ser dividido em três partes: a primeira que é responsável pela manipulação dos sensores, para que o robô seja capaz de capturar o áudio externo; a segunda constitui a integração desse áudio com o modelo de detecção de emoções; a terceira é o tratamento da saída do modelo integrado que retorna a emoção identificada através do robô. Após esses passos, é realizada uma avaliação do modelo proposto, onde um grupo de pessoas interage com o robô em um ambiente não controlado, reproduzindo frases de maneira a representar cada emoção, e só então é verificado o quanto o robô foi assertivo nas previsões.

O restante do trabalho está organizado da seguinte forma: Na seção 2 é explorado o referencial teórico no qual será feito um overview a respeito do humanoide NAO, abordando suas funcionalidades e algumas de suas aplicações na educação. Além disso, serão abordados também nessa seção conceitos sobre o modelo de predição utilizado, emoções e detecção de emoções na fala. Na seção 3 é apresentado o modelo proposto para esse trabalho, detalhando cada parte do modelo e seu funcionamento geral. Na seção 4 discutiremos sobre a metodologia experimental

utilizada, e na seção 5 abordaremos os resultados obtidos na avaliação do modelo proposto. Para finalizar, na seção 6 serão discutidas as conclusões, abordando as dificuldades e limitações deste trabalho, além de propostas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

O NAO é um robô humanoide primeiramente desenvolvido pela empresa francesa Aldebaran Robotics. Em 2015, no entanto, ela foi adquirida pelo grupo japonês Softbank e renomeado como Softbank Robotics. O NAO é capaz de interagir com seres humanos por meio dos seus microfones, alto falantes, câmeras, sensores táteis, luzes (LED), mãos e movimentos corporais. Ainda possui reconhecimento de voz e de face, fala vários idiomas e tem um controle de voz ajustável. Ele é um robô de interação social utilizado como auxílio na educação e em várias outras vertentes. A programação do robô pode ser feita através do Choregraph, um software que facilita o desenvolvimento devido a sua fácil usabilidade e interface amigável e que também é desenvolvido pela Softbank. Além disso, é possível desenvolver aplicações para o robô utilizando as linguagens de programação Python e C++ [9].

A sua facilidade de manuseio permite que o NAO seja utilizado como ferramenta de ensino e aprendizagem para crianças, jovens e adultos. O robô já foi utilizado no ensino de inglês para crianças, onde elas aprendem enquanto ensinam ao NAO [10], o robô é também usado como tutor em atividades individuais ou em grupo [11]. Além dessas aplicações, o humanoide é utilizado para o ensino de ciência da computação, e nesse caso, os alunos aprendem enquanto desenvolvem aplicações para o hardware. Esses são apenas alguns exemplos de como o NAO já é utilizado na área educacional com diversos fins.

Todos os exemplos citados a respeito de aplicações do NAO exigem uma interação humano-robô e essa interação pode acontecer de forma mais natural quando o robô tem o contexto da situação. Uma vertente que pode auxiliar nesse caso é o reconhecimento de emoções.

No estado da arte há muitas divergências à respeito do que é emoção, e isso implica na sua representação. Em seu trabalho, Plutchik categoriza oito emoções básicas: alegria, confiança, medo, surpresa, tristeza, antecipação, raiva e desgosto [12]. Ele organiza essas emoções analogamente à uma roda de cores, onde como tal, há emoções primárias, e a combinação de duas ou mais emoções

dão origem a emoções mais complexas, podendo assim existir centenas de emoções.

Grande parte dos pesquisadores concordam que a emoção contém pelo menos duas características: excitação e valência. Tanto a valência quanto a excitação podem ser definidas como experiências subjetivas [13]. A valência ou valence é uma sensação subjetiva de prazer ou desagrado; excitação ou arousal é um estado subjetivo de sentir-se ativado ou desativado.

Na área de reconhecimento de emoções, o reconhecimento através da fala se faz pertinente, já que essa é forma básica de comunicação entre seres humanos. O primeiro estudo de reconhecimento de emoção na fala foi realizado em meados da década de 1980, utilizando propriedades estatísticas de certas características acústicas [14]. Em seu trabalho, Pereira I. propôs um modelo de reconhecimento de emoções através da fala que busca identificar e classificar emoções básicas. O modelo contém dois módulos: o primeiro é uma representação geral da fala composto por um autoencoder treinado por uma *Generative Adversarial Network* (GAN), enquanto o segundo é o modelo de classificação, responsável pela distribuição entre as classes em uma classificação de emoções, ou por predição dos valores de um modelo dimensional treinado de maneira supervisionada [8].

3 MODELO PROPOSTO

Este trabalho propõe a integração de um mecanismo de reconhecimento de emoções através da fala ao robô humanoide NAO, para tal, utilizamos a versão 5 do humanoide lançada em 2015. A Figura 1 ilustra a sequência dos passos para integração, que serão descritos de forma mais específica nas subseções posteriores.



Figura 1: Modelo de reconhecimento de emoções integrado ao NAO.

Fonte: Autor (2019).

3.1 Captura do áudio

O primeiro passo da integração é a captura do áudio, para isso utilizamos o módulo ALAudioRecorder presente na biblioteca Naoqi SDK disponibilizada para o desenvolvimento de aplicações para o NAO [4]. Esse módulo ativa os microfones do robô por um tempo determinado, após passado esse tempo o microfone é desligado e o áudio é salvo na memória interna do humanoide.

3.2 Pré-processamento do áudio

O procedimento de pré-processamento foi realizado exatamente como descrito em Pereira [8] de modo a manter uma padronização entre os áudios. O primeiro passo foi a conversão do áudio capturado para 16kHz, em seguida, cada faixa de áudio foi decomposta em partes de um segundo sem se sobrepor. Depois disso, o áudio bruto foi convertido em um espectrograma via Transformada de Fourier de Tempo Curto, onde os parâmetros utilizados foram uma janela de tamanho 1024 e um comprimento de 512.

3.3 Modelo de Reconhecimento de Emoções

O modelo de reconhecimento de emoções é carregado previamente em um micro serviço de modo que a resposta seja mais rápida. O áudio pré-processado é submetido ao modelo através de um POST-request. O modelo retorna para cada segmento do áudio as emoções e seus pesos correspondentes, e é elencada a emoção que contém o maior peso para cada segmento de áudio. Por exemplo, se o áudio contém três segundos, para cada segundo do áudio o modelo retorna todas as sete emoções com seus respectivos pesos, é eleita a emoção correspondente àquele segmento a que obteve o maior peso. Após definir a emoção correspondente para cada segmento de áudio, é observada qual emoção mais se repetiu ao longo de todo o áudio. Se a emoção que mais se repetiu foi raiva dentre os segmentos, por exemplo, então essa será a emoção identificada para todo o áudio. Em caso de empate, é escolhida a emoção que obteve o maior peso.

3.4 Saída do Modelo

Após a identificação da emoção, utilizamos o módulo ALTextToSpeech para que o robô fale a

emoção identificada. Esse módulo recebe um texto como argumento e esse texto é emitido pelo robô em forma de áudio representado pela voz do humanoide. Os microfones são ligados novamente para que a interação continue.

4 METODOLOGIA EXPERIMENTAL

Para realização dos experimentos, primeiro foram elencadas frases de teste para serem usadas na interação com o robô. Foram elaboradas 3 frases distintas para cada emoção, resultando 21 frases que são mostradas na Tabela 1. O texto das frases foi baseado nos datasets da EmoDB, já que essa foi a base utilizada para treinamento do modelo utilizado na integração. As interações foram realizadas em um laboratório, porém havia ruídos.

Tabela 1: Frases utilizadas no experimento de interação com o robô.

Emoção	Frase 1	Frase 2	Frase 3
Alegria	"This is delicious"	"You can stay at home"	"You are so funny ah ah"
Medo	"Get away from me"	"Can you keep a secret?"	"She is a crazy psycho"
Neutro	"Dinner is served"	"You are in the right track"	"Do wanna come in?"
Nojo	"The president is a moron"	"blah, blah, blah"	"You silly thing"
Raiva	"Open the Goddamm door!"	"You need to grow up"	"Pick up the phone!"
Tristeza	"I don't feel very good today"	"okay"	"You cracked me up"
Tédio	"You call that fan"	"You don't know how"	"You will see very soon"

Quatro pessoas, sendo dois homens e duas mulheres falaram as frases no idioma inglês de forma a representar cada uma das emoções enquanto interagem com o robô, resultando em 84 amostras de áudio com 3 segundos de duração em média um como demonstrado na Tabela 2.

Tabela 2: Quantidade de amostras geradas por emoção.

Emoção	# frases distintas	# pessoas que reproduziram	# de amostras geradas
Alegria	3	4	12
Medo	3	4	12
Neutro	3	4	12
Nojo	3	4	12
Raiva	3	4	12
Tristeza	3	4	12
Tédio	3	4	12

4.1 Experimento 1

No primeiro experimento testamos o modelo integrado treinado no dataset da EmoDB, uma base alemã, que contém cerca de 500 enunciados com declarações de 10 locutores variados entre homens e mulheres num ambiente interno e controlado [15]. Utilizamos esse modelo para testar a assertividade em situações adversas e verificar sua abrangência em outros idiomas, levando em consideração que o NAO é utilizado em todas as partes do mundo e dá suporte há mais de 20 línguas incluindo alemão, inglês e português (Brasil) [16].

4.2 Experimento 2

Não há nenhuma carga para figuras de cor. Uma vez que Revista de Engenharia e Pesquisa Aplicada é um diário eletrônico, as figuras coloridas são produzidas automaticamente como parte da publicação de periódicos. Por favor, assegure a escolha adequada das cores para a exposição inequívoca do que está sendo mostrado.

Nos rótulos dos eixos das figuras, use palavras em vez de símbolos. Por exemplo, escreva a quantidade "Pressão", ou "Pressão, P", e não apenas "P." No entanto, se não houver espaço suficiente no eixo para especificar a quantidade, escreva apenas o símbolo "P", mas defina na legenda da figura. Como, por exemplo, escreva "Intensidade (W/m^2)" ou "Intensidade, I (W/m^2)" (mas não apenas " W/m^2 ").

As etiquetas das figuras devem ser legíveis, aproximadamente de 8 a 10 pontos, quando reduzidas à largura da coluna do artigo.

5. RESULTADOS

Para a avaliação do modelo proposto, primeiro realizamos um experimento de interação com o robô usando o modelo integrado onde obtivemos como saída a matriz de confusão demonstrada na Tabela 3. Observando a matriz de confusão, podemos identificar que "raiva" foi a única emoção onde o classificador acertou em 100% dos casos, obtendo o maior grau de reconhecimento, ou seja, todas as vezes que uma emoção é identificada como "raiva" o modelo consegue classificá-la corretamente. Além disso, é possível observar que todas as emoções podem se confundir com a emoção "raiva", principalmente as emoções "alegria", "medo" e "nojo", isso pode demonstrar uma correlação dessas categorias. A emoção "raiva", por exemplo, assim como a "alegria" contém um alto arousal, ou seja, uma alta excitação, assim como as emoções "nojo", "medo" e "raiva" contém uma baixa valence, ou seja, são carregadas negativamente, característica comum em sentimentos negativos.

Tabela 3: Matriz de confusão do Experimento 1.

	Classe prevista						
	Alegria	Medo	Neutro	Nojo	Raiva	Tristeza	Tédio
Classe real	Alegria	5			1	6	
	Medo		3		3	6	
	Neutro	3		3	1	3	2
	Nojo				6	6	
	Raiva					12	
	Tristeza	3			2	1	3
	Tédio				5	5	

Para o primeiro experimento, é ilustrado na Tabela 4 as métricas de recall e precisão para cada uma das emoções individualmente. Essas são métricas comumente utilizadas no aprendizado de máquina, o recall é a proporção de exemplos que foram classificados como classe x, entre todos os exemplos que realmente têm classe x, ou seja, quanto parte da classe foi capturada, já a precisão é a proporção dos exemplos que realmente têm classe x entre todos aqueles que foram classificados como classe x [18].

Tabela 4: Métricas de avaliação para o Experimento 1.

Emoção	Recall	Precisão
Alegria	42%	45,5%
Medo	25%	100%
Neutro	25%	100%
Nojo	50%	33,3%
Raiva	100%	30,7%
Tristeza	25%	100%
Tédio	17%	28,7%

No geral, obtivemos uma acurácia de 40%, uma precisão média de 62,6% e um recall médio de 40%. É possível observar na Tabela 4 que as emoções "medo", "neutro" e "tristeza" obtiveram um baixo recall e uma alta precisão, isso pode significar que o classificador para esses casos consegue classificar poucos rótulos, mas esses estão corretos quando comparados com sua categoria real. Isso pode ser melhor observado na matriz de confusão da Tabela 3, todas as vezes que o classificador classificou uma emoção como "medo", ela realmente era "medo". Já para emoção "raiva" acontece o contrário, ela apresenta um alto recall e uma precisão baixa, indicando que o classificador para esse caso identifica várias emoções como "raiva", porém erra quando a emoção é comparada com sua classe real.

Além dessas métricas, observamos que quando o experimento foi gravado por mulheres a taxa de acerto foi de 23,8% em relação ao total, enquanto que os áudios gravados por homens obtiveram uma taxa de acerto de 16,6% também quando comparado com o total. A baixa acurácia do experimento pode ser atribuída ao fato das interações serem realizadas em inglês e as condições de gravação do experimento serem diferentes do que foi utilizado no treinamento de modelo.

Para o segundo experimento, as métricas de avaliação estão ilustradas na Tabela 5. Nessa tabela podemos observar que as emoções, "surpresa", "raiva" e "tristeza" apresentam precisão e recall relativamente alto, o que de fato é esperado, isso significa que o classificador é criterioso, e ainda assim classifica corretamente as emoções quando comparada com sua classe real. No geral para esse experimento obtivemos uma acurácia de 65%, o recall de 64% e uma precisão de 74%.

Ao comparar os resultados do experimento 1 com os do experimento 2 é evidente que o segundo apresenta um melhor desempenho, e isso já era esperado pois a base de dados no qual o modelo do experimento 2 foi treinado tem um contexto parecido com o cenário do conjunto de dados usado para testes.

Tabela 5: Métricas de avaliação para o Experimento 2.

Emoção	Recall	Precisão
Alegria	39%	92%
Medo	100%	58%
Neutro	62%	42%
Nojo	80%	67%
Raiva	69%	75%
Tristeza	67%	67%
Tédio	100%	50%

6. CONCLUSÕES

Nesse trabalho é proposta a integração de um mecanismo de reconhecimento de emoções através de um modelo que realiza a representação geral da fala ao robô humanoide NAO. Essa integração pode facilitar a interação entre ser humano e máquina, deixando-a mais natural. A principal contribuição desse trabalho foi a implementação desse mecanismo de reconhecimento de emoções no robô NAO, deixando-o ainda mais completo e eficaz na interação com seres humanos.

Foram realizados dois experimentos para verificar a assertividade do mecanismo de integração proposto, onde conseguimos uma acurácia de 40,48% para o modelo integrado e 65% para o modelo treinado em condições semelhantes ao conjunto de dados utilizado para teste. O fato de as interações serem realizadas em inglês e as condições de gravação do experimento serem diferentes pode ter influenciado nos resultados obtidos. Além desses resultados, observamos que quando o áudio é gravado por uma pessoa do sexo feminino, a taxa de acerto é maior que quando gravado por alguém do sexo masculino.

Uma dificuldade encontrada foi em relação ao processamento do modelo, devido à pouca capacidade computacional. Para trabalhos futuros pode-se abordar uma forma de aumentar os recursos computacionais. Outra proposta de trabalho futuro é a construção de um módulo próprio para o software Choregraph, utilizando o

mecanismo aqui proposto. Além disso, pode-se pensar no desenvolvimento de outros mecanismos de reconhecimento de emoções para integração ao NAO, tal como reconhecimento de emoções através da análise facial, com o intuito de deixá-lo mais completo e assertivo.

REFERÊNCIAS

- [1] MILLER, D. P.; NOURBAKHSI, I. R.; SIEGWART, R. **Robots for Education**, Springer Handbook of Robotics, 2007.
- [2] BELPAEME, T. *et al.* **Social robots for education: A review**. Science Robotics, vol. 3, Issue 21, 2018.
- [3] Shamsuddin, S. *et al.* **Initial Response in HRI- a Case Study on Evaluation of Child with Autism Spectrum Disorders Interacting with a Humanoid Robot NAO**. Procedia Engineering, vol 41, pp. 1448 – 1455, 2012.
- [4] Robotics, A. (2014a). **Aldebaran documentation**. Disponível em <http://doc.aldebaran.com>. Último acesso: 14/04/2020.
- [5] PICARD, R. W. (2000). **Affective computing**. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, 2000.
- [6] BREAZEL, C. **Function meets style: insights from emotion theory applied to HRI**. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(2):187–194, 2004.
- [7] KLEINGINNA, P. R.; KLEINGINNA, A. M. **A categorized list of emotion definitions, with suggestions for a consensual definition**. Motivation and Emotion, vol 5, pp. 345–379, 1981.
- [8] PEREIRA, I., *et al.* **Semi-supervised model for emotion recognition in speech**. In Artificial Neural Networks and Machine Learning – ICANN 2018, pages 791–800, 2018.
- [9] Robotics, S. (2018). **Nao the humanoid robot – softbank robotics emea**. Disponível em <https://www.softbankrobotics.com/emea/en/nao>. Último acesso: 14/04/2020.
- [10] TANAKA, F.; MATSUZOE, S. **Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning**. Journal of Human-Robot Interaction, vol 1, pp. 78-95, 2012.
- [11] SERHOLT, S. **Breakdowns in children’s interactions with a robotic tutor: A longitudinal study**. Computers in Human Behavior, vol. 81, pp. 250-264, 2018.
- [12] PLUTCHIK, R. **The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice**. American scientist, vol. 89(4), pp. 344–350, 2001.
- [13] RUSSELL, J.; KELLERMAN, H. **Emotion: theory, research and experience – The Measurement of Emotions**. Academic Press, 1989.
- [14] Kim, E. H. *et al.* **Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments**. In The 16th IEEE International Symposium on Robot and Human Interactive Communication, pages 689–694, 2007.
- [15] Robotics, S. (2014b). **Softbank robotics documentation what’s new in naoqi 2.5?** Disponível em <http://doc.aldebaran.com/2-5/news/index.html>. Último acesso: 14/04/2020.
- [16] BURKHARDT, F. *et al.* **A database of german emotional speech**. In 9th European Conference on Speech Communication and Technology (INTERSPEECH), 2005.
- [17] BARROS, P. *et al.* **The OMG-emotion behavior dataset**. In International Joint Conference on Neural Networks (IJCNN), pages 1–7, (2018)
- [18] BOUCKAERT, R. R. *et al.* **Weka manual for version 3-9-1**. University of Waikato, Hamilton, New Zealand, 2016.