

TRABALHO DE GRADUAÇÃO

**IMPLEMENTAÇÃO DE INTERAÇÃO HUMANO-ROBÔ  
POR MEIO DE VOZ NA PLATAFORMA NAO.**

Felipe de Moraes Dalosto

Brasília, Dezembro de 2019



**ENGENHARIA  
MECATRÔNICA**  
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA  
Faculdade de Tecnologia  
Curso de Graduação em Engenharia de Controle e Automação

## TRABALHO DE GRADUAÇÃO

# IMPLEMENTAÇÃO DE INTERAÇÃO HUMANO-ROBÔ POR MEIO DE VOZ NA PLATAFORMA NAO.

Felipe de Moraes Dalosto

*Relatório submetido como requisito parcial de obtenção  
de grau de Engenheiro de Controle e Automação*

Banca Examinadora

Profa. Mariana Costa Bernardes Matias, \_\_\_\_\_  
FGA/UnB  
*Orientadora*

Profa. Carla Cavalcante Koike, \_\_\_\_\_  
CIC/UnB  
*Examinadora*

Prof. Geovany Araújo Borges, \_\_\_\_\_  
ENE/UnB  
*Examinador*

Brasília, Dezembro de 2019

## FICHA CATALOGRÁFICA

FELIPE DE MORAIS DALOSTO

Implementação de interação humano-robô por meio de voz na plataforma NAO.

[Distrito Federal] 2019.

vii, 36p., 210 x 297 mm (FT/UnB, Engenheiro, Controle e Automação, 2019). Trabalho de Graduação – Universidade de Brasília. Faculdade de Tecnologia.

1. Robótica
2. Interação humano robô
3. Processamento de linguagem natural
4. Comandos por voz

I. Mecatrônica/FT/UnB

## REFERÊNCIA BIBLIOGRÁFICA

DALOSTO, FELIPE DE MORAIS, (2019). Implementação de interação humano-robô por meio de voz na plataforma NAO. Trabalho de Graduação em Engenharia de Controle e Automação, Publicação FT.TG-*nº0xx*, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 36p.

## CESSÃO DE DIREITOS

AUTOR: Felipe de Moraes Dalosto

TÍTULO DO TRABALHO DE GRADUAÇÃO: Implementação de interação humano-robô por meio de voz na plataforma NAO.

GRAU: Engenheiro ANO: 2019

É concedida à Universidade de Brasília permissão para reproduzir cópias deste Trabalho de Graduação e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desse Trabalho de Graduação pode ser reproduzida sem autorização por escrito do autor.

---

Felipe de Moraes Dalosto

Departamento de Eng. Elétrica (ENE) - FT

Universidade de Brasília (UnB)

CEP 70919-970, – Brasília – DF – Brasil.

## **Dedicatória**

*Dedico este trabalho aos amigos da equipe UnBeatables.*

*Felipe de Moraes Dalosto*

## Agradecimentos

*Primeiramente, gostaria de agradecer aos meus pais por todo apoio em minha jornada acadêmica. Sou grato também à Prof. Mariana pela dedicação aos projetos de extensão e pela oportunidade de ter conhecido a robótica. Por fim, agradeço os amigos que fiz durante o curso de Engenharia Mecatrônica. O companheirismo e o apoio emocional que recebi foram fundamentais concluir essa etapa acadêmica.*

*Felipe de Moraes Dalosto*

---

## RESUMO

Palavras Chave: Robótica, Processamento de linguagem natural, Interação humano-robô, Comandos de voz.

Este projeto busca desenvolver um sistema que permita a interação entre um usuário humano e um robô por meio da linguagem natural. Para desenvolvimento do sistema objeto deste projeto, foi escolhida a plataforma NAO, tendo em vista a sua alta aceitabilidade pelo usuário e facilidade na implementação de respostas multimodais. O sistema desenvolvido deverá ser capaz de capturar o áudio com as frases proferidas pelo usuário, realizar a transcrição e processar o conteúdo para identificar qual a intenção o usuário demonstrou. Cada intenção do usuário demanda uma resposta diferente, ativando a síntese de voz e movimentação do robô. As intenções programadas permitem diálogos simples, diálogos com contexto elaborado e locomoção do robô. A implementação utilizou o *framework* Robot Operating System (ROS) para gerenciamento dos processos, coordenando o fluxo de dados entre as aplicações distribuídas em diferentes computadores. O sistema implementado é avaliado conforme métricas padrões de aplicações em interação humano-robô.

---

## ABSTRACT

Keywords: Robotics, Natural language processing, Human–robot interaction, Voice user interface

This project seeks to develop a system that allows interaction between a user and a robot through natural language. The NAO platform was chosen due to its high user acceptability and ease implementation of multimodal responses. The system should be able to capture the audio with the phrases uttered by the user, transcribe and process the content to identify the intention the user has expressed. Each intention demands a different response, activating voice synthesis and robot movement. The programmed intentions allow simple dialogues, elaborated context dialogues and robot locomotion. The implementation used ROS framework for process management, coordinating the data flow between the API's and different computers. Finally, the implemented system is evaluated according to standard application metrics in human-robot interaction.

# SUMÁRIO

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	CONTEXTUALIZAÇÃO	1
1.1.1	INTERAÇÃO HUMANO - ROBÔ	2
1.1.2	ROBÔS HUMANOIDES	2
1.2	JUSTIFICATIVA E CONTRIBUIÇÕES	3
1.3	DEFINIÇÃO DO PROBLEMA	3
1.3.1	OBJETIVOS DO PROJETO	4
<b>2</b>	<b>Fundamentos</b>	<b>5</b>
2.1	COMUNICAÇÃO	5
2.1.1	PROPRIEDADES	5
2.1.2	DINÂMICA DE DIÁLOGOS	6
2.1.3	ATOS DE FALA	7
2.2	PROCESSAMENTO DE LINGUAGEM NATURAL	8
2.3	COMPREENSÃO DA LINGUAGEM NATURAL	8
2.3.1	ANÁLISE LÉXICA	8
2.3.2	ANÁLISE SINTÁTICA	9
2.3.3	ANÁLISE SEMÂNTICA	10
2.3.4	ANÁLISE PRAGMÁTICA	10
2.4	SÍNTESE DE FALA EM LINGUAGEM NATURAL	10
2.4.1	SISTEMA TEXTO-FALA	11
2.4.2	MULTIMODALIDADE	11
2.5	ARQUITETURAS INTERNAS	12
2.5.1	SISTEMAS BASEADOS EM ESTADOS FINITOS	12
2.5.2	SISTEMAS BASEADOS EM FORMULÁRIOS	13
2.5.3	SISTEMAS AVANÇADOS	13
2.6	AVALIAÇÃO DE APLICAÇÕES HRI	13
2.7	INTERAÇÃO POR DIÁLOGOS	14
<b>3</b>	<b>Ferramentas de Desenvolvimento</b>	<b>15</b>
3.1	PLATAFORMA NAO	15
3.1.1	NAOqi	16
3.2	GOOGLE SPEECH-TO-TEXT	19

3.3	DIALOGFLOW .....	20
3.4	ROS.....	21
3.5	REDES DE COMPUTADORES .....	21
3.5.1	TCP/IP .....	21
<b>4</b>	<b>Implementação .....</b>	<b>24</b>
4.1	DESCRIÇÃO DO SISTEMA .....	24
4.2	MÓDULOS DO SISTEMA .....	25
4.2.1	DIALOG FLOW CLIENT .....	25
4.2.2	SOQUETES UDP.....	25
4.3	FUNCIONALIDADES .....	26
4.3.1	MOVIMENTAÇÃO DE PARTES DO CORPO DO ROBÔ.....	26
4.3.2	MOVIMENTAÇÃO ESPACIAL.....	28
4.3.3	CONTROLE DE VOLUME .....	28
4.3.4	MUDAR DE POSTURA .....	28
<b>5</b>	<b>Resultados.....</b>	<b>30</b>
5.1	INTERAÇÕES COM USUÁRIO .....	30
5.1.1	LISTA DE FUNÇÕES.....	31
5.2	AVALIAÇÃO DO SISTEMA.....	31
5.2.1	TAXA DE ERRO DE PALAVRA (WORD ERROR RATE - WER) .....	31
5.2.2	TAXA DE ERRO DE FALA .....	32
5.2.3	COBERTURA DE VOCABULÁRIO .....	32
5.2.4	TEMPO PARA REALIZAÇÃO DA TAREFA.....	33
5.2.5	TAXA DE CORREÇÃO.....	33
<b>6</b>	<b>Conclusões.....</b>	<b>35</b>
6.1	DIÁLOGOS .....	35
6.2	IMPLEMENTAÇÃO E DESEMPENHO.....	35
6.3	TRABALHOS FUTUROS.....	36
6.3.1	EXPANSÃO DAS INTENÇÕES RECONHECIDAS .....	36
6.3.2	COMANDOS DE VOZ EM OUTROS ROBÔS .....	37
	<b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>	<b>38</b>
	<b>Anexos.....</b>	<b>40</b>
<b>I</b>	<b>Transcrição das interações .....</b>	<b>41</b>
I.I	INTERAÇÃO 1 .....	41
I.II	INTERAÇÃO 2 .....	43
I.III	INTERAÇÃO 3 .....	45



# LISTA DE FIGURAS

2.1	Etapas de compreensão da linguagem natural.....	9
2.2	Síntese de fala implementada para o robô NAO.....	11
3.1	Plataforma NAO.....	15
3.2	Funcionamento do <i>framework</i> NAOqi.....	16
3.3	Eixos de movimentação do NAO.....	18
4.1	Fluxo de dados entre as ferramentas do sistema implementado.....	24
4.2	Ângulos de movimentação do ombro e cotovelo.....	27
4.3	Ângulos de movimentação da cabeça.....	27
4.4	Posturas em pé, sentado e descansando.....	28

# LISTA DE TABELAS

5.1	Avaliação da taxa de erro de palavra .....	32
5.2	Avaliação da taxa de erro de fala. ....	32
5.3	Avaliação do tempo de realização da tarefa. ....	33
5.4	Avaliação da taxa de correção após erro. ....	34

# Capítulo 1

## Introdução

*Este capítulo contextualiza o leitor no domínio de máquinas autônomas, das quais se destacam os sistemas robóticos. Em seguida, é abordado o histórico desses agentes e sua atuação no cotidiano humano. Por fim, o objeto de implementação desse trabalho é definido e justificado.*

### 1.1 Contextualização

A robótica é a combinação sinérgica da mecânica, controle, computação e eletrônica. Os sistemas desenvolvidos buscam substituir o ser humano em tarefas. Os primeiros registros de especificações para autômatos programáveis concebidos no século I d.C. foram encontrados em obras na biblioteca de Alexandria, contudo a possibilidade de máquinas autônomas substituírem humanos já era abordada na mitologia grega e na China antiga. A criação de máquinas automatadas se intensificou na Europa durante o século XIX, produzindo milhares de autômatos-relógios e brinquedos inspirados em animais [1].

A substituição de humanos em tarefas requer precisão e capacidade de reagir a estímulos. Para interpretar o ambiente ao seu redor máquinas automatadas podem utilizar diversos tipos de sensores. Dessa forma, a robótica é comumente definida como o estudo da conexão inteligente entre percepção e ação [2]. Os robôs são capazes de sentir e atuar através de dispositivos mecânicos e eletrônicos controlados por computadores. São aplicações típicas a manipulação de objetos, locomoção, interação com humanos e com outros robôs.

A funcionalidade de sistemas robóticos depende da combinação sinérgica de diversos subsistemas. A capacidade de locomoção e manipulação de objetos é concedida aos sistemas robóticos pelos atuadores, responsáveis por movimentar os componentes mecânicos do robô. Sistemas deste tipo são estudados pela área de controle do movimento, envolvendo componentes como servomotores e *drivers*.

A percepção do ambiente é realizada pelo sistema de sensoriamento, que deve obter informações sobre o próprio robô e o mundo a sua volta. As funções desse sistema envolvem o condicionamento de sinais e processamento de dados. Por último, para comandar os robôs é necessário um sistema

de controle que avalie os resultados das ações dos robôs sobre o ambiente e planeje ações com base em seus objetivos.

### **1.1.1 Interação Humano - Robô**

A área de pesquisa em Interação Humano - Robô (HRI) é um campo multidisciplinar que se dedica a estudar a interação entre humanos e robôs. A interação com robôs é tópico da ficção científica e produção acadêmica antes mesmo dos robôs existirem. As interações humano-robô são projetadas para reproduzir aspectos da comunicação e das relações humanas, buscando garantir maior aceitabilidade do usuário [3].

Os avanços tecnológicos permitiram que a robótica criasse sistemas para a realização de diversas atividades em cooperação com seres humanos. No início, o desenvolvimento foi focado em robôs industriais, capazes de executar tarefas repetitivas e de baixa complexidade [4]. Contudo, a pesquisa e o desenvolvimento em robótica permitiram que robôs conquistassem o espaço [5], salas de aula [6] e até hospitais [7].

A cooperação entre humanos e robôs permite maior eficiência na realização de determinadas tarefas. Por exemplo, robôs podem ajudar cirurgiões na realização de procedimentos cada vez mais seguros e menos invasivos [8]. No entanto, a proximidade dos robôs no cotidiano humano aumenta os riscos gerados por essas interações. Com robôs e humanos compartilhando tarefas e áreas de trabalho é necessário desenvolver tecnologias que garantam a integridade dos seres humanos.

Assim, o objetivo da área de pesquisa HRI é definir modelos para as expectativas humanas sobre como um robô deve se comportar durante uma interação, guiando a concepção e o desenvolvimento de robôs mais efetivos [2]. É importante ressaltar que os robôs que interagem de forma tão próxima aos humanos estão submetidos a regras sociais e culturais, devendo apresentar comportamento aceitável e confortável para os usuários. Desta forma, é imprescindível que o robô construa modelos dinâmicos para suas interações. Além de reconhecer e manipular objetos, para interagir com humanos os robôs devem ser capazes de localizar pessoas e até mesmo interpretar suas emoções.

O controle dos robôs deve ser simples para o usuário, necessitando de pouco conhecimento prévio sobre robótica. A medida que os computadores se tornam capazes de processar a linguagem natural humana é possível transformar interfaces de controle em simples conversas com o usuário. A comunicação por meio de fala e gestos aproxima o controle dos robôs da forma humana de expressar desejos e ordens, evitando curvas de adaptação e tempo despendido em aprendizagem.

### **1.1.2 Robôs Humanoides**

A robótica móvel representa a subdivisão da robótica que estuda robôs capazes de se locomover em um ambiente, dentre os quais para essa aplicação se destacam os robôs humanoides. Um robô humanoide possui formato baseado na aparência do corpo humano, permitindo sua interação com ferramentas e ambientes projetados para seres humanos [9]. De forma geral, robôs humanoides são bípedes, possuem dois braços, tronco e cabeça. Alguns robôs podem ainda ter face com olhos

e boca, permitindo replicar a fala e as expressões faciais humanas. A semelhança com humanos permite maior aceitação dos usuários e engajamento nas interações.

Por outro lado, a empatia do observador humano só é proporcional à antropomorfia do robô até certo ponto. Quando robôs se comportam de forma demasiadamente parecida com seres humanos a resposta do usuário pode se tornar uma forte repulsa, comportamento relatado pela teoria do vale da estranheza [10]. A aparência humana também pode elevar as expectativas do usuário quanto ao comportamento do robô, causando frustração quando o robô não é capaz de executar as tarefas como um humano.

## 1.2 Justificativa e Contribuições

Com a evolução da robótica é necessário simplificar a interação do usuário com as funcionalidades de um robô. A comunicação entre usuário e máquina pode ocorrer de forma mais natural por meio da fala, afinal o usuário utiliza a linguagem que está familiarizado a se comunicar. A descrição de tarefas em linguagem natural permite comandar um robô sem a necessidade de aprender uma nova linguagem ou ferramenta, possibilitando que mais pessoas acessem serviços e informações.

Aplicações de processamento de linguagem natural para consultas à banco de dados, por exemplo, dispensam que o usuário conheça a estrutura do banco ou alguma linguagem específica. A utilização da linguagem natural permite abstração da implementação do sistema, tornando o software mais simples para o usuário.

Um robô controlado por linguagem natural pode ser visto como a integração de quatro domínios relacionados com a Engenharia Mecatrônica:

- Interação humano-robô
- Percepção e capacidade sensorial
- Capacidade de decisão
- Aprendizagem

Partindo dessas premissas, este trabalho tem como contribuição o desenvolvimento de um *framework* que permite diálogos estruturados entre humanos e robôs utilizando a plataforma NAO. Os diálogos poderão também movimentar o robô e configurar parâmetros para melhorar a interação.

## 1.3 Definição do problema

Ao decorrer da história os seres humanos desenvolveram um complexo e estruturado sistema de sinais conhecido como linguagem natural. A capacidade de compreender esse tipo de linguagem é inata aos seres humanos e necessária para a organização da vida em sociedade. A linguagem

natural reflete aspectos culturais e históricos, sendo a principal forma de interagir com outras pessoas, tomar conhecimento de algo ou expressar um sentimento [11].

O processamento da linguagem natural (PLN) permite que os seres humanos se comuniquem com computadores da forma mais intuitiva possível, utilizando a linguagem com a qual mais estão acostumados. Desta forma, eliminam-se as adaptações e o aprendizado intrínsecos a uma linguagem artificial. [12].

Entretanto, extrair significado da linguagem natural requer conhecimento específico da linguagem utilizada e do contexto no qual a conversa está inserida [13]. As aplicações modernas que utilizam o processamento de linguagem natural consistem basicamente em dois segmentos [14].

- **Aplicações baseadas em texto:** sistemas que procuram documentos ou trechos específicos em uma base de dados, como tradutores de línguas e editores de planilhas.
- **Aplicações baseadas em diálogos:** interfaces de linguagem natural para bancos de dados, *call center* automatizado, sistemas tutores e sistemas de comandos de voz embarcados em *smartphones*.

Este trabalho propõe uma aplicação baseada em diálogos, em que é necessário reconhecer a linguagem natural, dar significado e reagir conforme o contexto e o histórico da interação. A variação gramatical das palavras acrescenta complexidade à interpretação, uma vez que o usuário pode usar diferentes modos verbais e sinônimos para expressar a mesma intenção. A ambiguidade de expressões impõe que o sistema seja capaz de avaliar o contexto da conversa e os conhecimentos comuns para interpretar o significado de uma frase. O uso de pronomes pessoais e demonstrativos também aumenta a complexidade do PLN, afinal é necessário resolver as referências feitas durante o discurso.

### 1.3.1 Objetivos do projeto

O objetivo do projeto é permitir a interação entre usuários e o robô NAO por meio de diálogos. Na simulação de diálogo o robô deve ser capaz de saudar o usuário, responder perguntas comuns e reagir aos comandos. A capacidade de conversar com humanos permite interações mais atrativas, contribuindo com o trabalho educacional realizado utilizando a plataforma humanoide NAO.

O projeto integra ferramentas frequentemente utilizadas em sistemas de processamento de linguagem natural e robótica para permitir a criação de uma aplicação de interação com o usuário. As etapas de transcrição da fala e compreensão da linguagem natural são realizadas por APIs desenvolvidas pelo Google, já o gerenciamento dos processos do robô é simplificado pelo *framework* Robot Operating System (ROS).

# Capítulo 2

## Fundamentos

*Este capítulo explora os principais conceitos associados à comunicação entre seres humanos. Em seguida, é apresentado o campo do Processamento de Linguagem Natural e as etapas necessárias para sua implementação computacional. Por fim, são categorizadas as principais arquiteturas para implementar sistemas desse tipo.*

### 2.1 Comunicação

A comunicação pode ser definida como uma forma intencional de trocar informações por meio da produção e percepção de sinais em um sistema convencional. Este processo social primário permite criar e interpretar mensagens que informam, provocam ações e mudam a realidade da sociedade [11]. As conversas entre seres humanos seguem o princípio cooperativo, de forma que um diálogo não pode ser constituído por falas aleatórias. Para desenvolver um diálogo é necessário que os interlocutores pratiquem ações conjuntas com um objetivo comum.

A área conhecida como análise da conversação tem como objetivo estudar como ocorrem as interações durante uma conversação. Tomassello [15] descreve três tipos básicos de comunicação entre humanos: solicitar, informar e compartilhar. Já os sistemas de diálogos com robôs podem ser caracterizados por sua aplicação, arquitetura interna e modalidade.

#### 2.1.1 Propriedades

Para implementar a comunicação homem-máquina é necessário entender e sistematizar a comunicação em linguagem natural utilizada entre humanos. Nickerson [16] identificou as propriedades presentes nas conversações humanas:

- **Bi-direcionalidade:** o fluxo de informações ocorre nos dois sentidos.
- **Iniciativa mista:** cada participante da comunicação pode conduzir a conversa, fornecer informações ou realizar perguntas por vontade própria.

- **Percepção de posse da palavra:** também conhecido como controle ou *turn*, o momento de cada participante falar no diálogo é definido implicitamente.
- **Regras para troca de posse na palavra:** a passagem da posse ou controle da palavra entre os participantes do diálogo é marcado por momentos específicos ou convenções sociais.
- **Sensação de presença:** a percepção de cada participante de que o outro está presente na comunicação, entendendo seu desenrolar.
- **Comunicação não verbal:** a comunicação entre humanos faz uso de gestos, acenos e olhares para complementar a passagem de informações e sentimentos.
- **Intolerância ao silêncio:** momentos de silêncio não esperado durante conversas podem ser considerados desagradáveis, podendo comprometer a sensação de presença necessária para o engajamento em uma conversa.
- **Estrutura:** a estrutura dos diálogos é comum e marcada por protocolos. Iniciar com uma saudação, tratar um assunto ao longo da conversa e finalizar com uma despedida exemplifica um tipo de estrutura de diálogo com adesão aos protocolos sociais.
- **Banda larga:** alta capacidade para transferência de informações, mesmo que alguns diálogos não usufruam completamente.
- **Linguagem informal:** os diálogos tendem a ser mais informais que textos escritos, marcados por frases gramaticalmente incorretas ou incompletas.
- **Contexto situacional compartilhado:** o contexto compartilhado pelos participantes possui informações imprescindíveis para o correto entendimento da comunicação, de forma que muitos diálogos só possuem sentido dentro de um contexto.
- **Senso comum:** durante conversas é permitido fazer referências, direta ou indiretamente, a conhecimentos tidos como gerais a todas as pessoas.
- **Conhecimento específico compartilhado:** Além dos conhecimentos gerais do mundo, também podemos compartilhar conhecimentos específicos em diálogos com familiares, amigos ou entre classes profissionais.
- **Participantes de mesmo nível:** Nickerson sugere que as conversas são mais agradáveis quando envolvem participantes do mesmo nível intelectual, sem negar com isso a existência de outros tipos de conversas.

### 2.1.2 Dinâmica de diálogos

Conversas que envolvem apenas dois agentes são denominadas diálogos. A estrutura de um diálogo pode ser descrita através da posse da palavra e da sua transferência durante a comunicação.



### 2.1.2.1 Posse da palavra

Diálogos são estruturados com a premissa que somente uma pessoa deve falar a cada momento, ocasionando momentos de posse da palavra. A posse da palavra é caracterizada pela produção de uma fala, ou sequência de falas, sem interrupções por parte dos outros participantes. A busca da posse de palavra pode gerar atropelamentos, contudo estudos indicam que o total de sobreposição nos diálogos é inferior a 5% [17]. A transferência da posse da palavra entre participantes em um diálogo é chamada “tomada de palavra” e geralmente ocorrem em pontos mais susceptíveis a mudança. Nesses pontos, observam-se as regras:

- Durante a sua fala o orador atual pode selecionar o próximo orador.
- Se o orador não selecionar o próximo orador, então qualquer pessoa pode tomar a palavra.
- Se ninguém tomar a palavra o orador atual pode continuar falando.

### 2.1.3 Atos de fala

Entende-se por ato de fala à produção de um enunciado linguisticamente funcional em contexto de interação comunicativa. Os atos de fala podem simbolizar ações como: avisar, informar, prometer, pedir e ordenar. Segundo Austin [18], os atos de fala podem ser divididos em:

- **Ato locutório:** ato de proferir um enunciado, formar frase com um determinado sentido.
- **Ato ilocutório:** corresponde às intenções e ao contexto da conversa. A motivação do agente para elaborar a sentença.
- **Ato perlocutório:** corresponde aos efeitos que um dado ato ilocutório produz. Verbos como convencer, persuadir ou assustar ocorrem neste tipo de atos de fala.

A plurissignificação das palavras permite que diferentes atos locutórios simbolizem o mesmo ato ilocutório. Dessa forma, é possível utilizar diferentes enunciados para expressar um mesmo sentimento ou solicitação. Os atos ilocutórios, ou motivações, foram divididos por Searle nas seguintes classes [19]:

- **Assertivos:** o orador afirma algo (jurar, concluir).
- **Diretivos:** o orador tenta provocar uma ação no ouvinte (perguntas, ordens, pedidos).
- **Compromissivos:** o orador compromete-se a agir de alguma forma (prometer, combinar).
- **Expressivos:** o orador expressa seus sentimentos (agradecer, desculpar, saudar).
- **Declarativos:** o orador muda o estado do mundo com suas palavras (nomear algo, contratar alguém, aceitar casamento).

## 2.2 Processamento de Linguagem Natural

O processamento de linguagem natural (PLN) é uma área de pesquisa e aplicações dedicadas a explorar o entendimento e manipulação da linguagem natural por computadores. A área envolve conhecimentos de ciência da computação e informação, linguística, inteligência artificial e robótica. Os algoritmos podem utilizar a codificação direta de regras ou a aprendizagem automática para realizar as etapas necessárias para o processamento [20].

Dentre as aplicações modernas de PLN destacam-se:

- **Tradutores:** tradução automática de uma linguagem humana para outra.
- **Geração de linguagem natural:** conversão de bancos de dados para intenções semânticas em linguagem natural inteligível.
- **Análise de subjetividade:** interpretação de sentimentos do interlocutor por meio da linguagem escrita.
- **Sumarização automática:** resumo legível de um texto escrito.

O processamento da fala em linguagem natural utilizada neste projeto envolve o reconhecimento de fala, a compreensão da linguagem natural e a síntese de falas em linguagem natural.

## 2.3 Compreensão da Linguagem Natural

A compreensão da linguagem natural permite extrair informação e intenções de textos escritos em linguagem natural. Para converter o discurso de humanos em texto escrito são utilizados sistemas de reconhecimento de fala, capazes de identificar as palavras em ondas sonoras capturadas por microfones.

De forma semelhante aos processadores de linguagens artificiais como compiladores, as análises linguísticas são abordadas em três etapas: lexical, sintática e semântica. Existem ainda análises pragmáticas, em que é considerado o interlocutor que utilizou as expressões e o contexto de uso. Esses aspectos do discurso não estão explicitamente contidos nas palavras usadas no texto, mas carregam informações importantes sobre a real intenção de cada fala.

### 2.3.1 Análise Léxica

A análise léxica é a primeira etapa do processamento de linguagem e tem como objetivo identificar quais são as palavras ou expressões utilizadas em uma sentença. Este processo é auxiliado por delimitadores como pontuação e espaços em brancos, permitindo que cada palavra receba uma classificação [21].

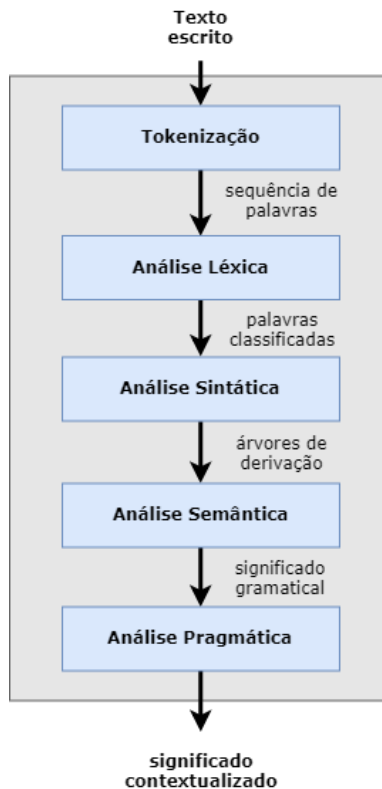


Figura 2.1: Etapas de compreensão da linguagem natural.

O reconhecimento das palavras é dificultado pelas flexões existentes na língua portuguesa, como conjugação de verbos e formação de plural. A taxa de erro na compreensão de palavra deve ser baixa, pois erros nesta etapa podem comprometer todos os passos subsequentes do processamento.

Existem ainda casos de ambiguidades linguística, em que palavras com a mesma grafia possuem classificações diferentes. A ambiguidade sintática ocorre quando a mesma palavra pode ser classificada em diferentes categorias sintáticas, enquanto a ambiguidade semântica ocorre quando o significado da palavra é diferente de acordo com o contexto [22].

### 2.3.2 Análise Sintática

Após identificar as palavras é necessário analisar a relação sintática entre elas para extrair o sentido do enunciado. Enquanto o analisador morfológico lida com a separação das palavras e suas classificações, o analisador sintático analisa os agrupamentos de palavras e as regras gramaticais da língua.

Também estudada por Rich [13], a análise sintática é definida como a extensão dos resultados da análise léxica para criar uma descrição estrutural da frase. Dessa forma, a lista de palavras é convertida em uma estrutura definida pela classe gramatical das palavras.

A análise sintática também é uma etapa indispensável para viabilizar o processamento semântico, pois as várias palavras são reduzidas em uma estrutura que simplifica a complexidade do processamento. A identificação das palavras com sua respectiva classe gramatical geralmente

ocorre com o auxílio de um glossário e regras gramaticais.

### **2.3.3 Análise Semântica**

Enquanto a sintaxe dedica-se a estudar o agrupamento das palavras para formar estruturas, a semântica está relacionada ao significado do agrupamento de palavras. As análises sintáticas e semânticas estão intrinsecamente relacionadas, uma vez que o agrupamento das palavras deve obedecer as regras gramaticais para fazer sentido.

A análise semântica, portanto, é realizada com base em regras sobre a posição e a relação das classes gramaticais. Por exemplo, a regra “se um verbo segue o sujeito, então o sujeito executou o verbo” é capaz de atribuir sentido a uma frase a partir de suas unidades léxicas. Tendo em mente que o real significado das palavras está conectado com conhecimentos sobre o mundo e emoções, a análise semântica é complexa e existem diferentes técnicas para realizá-la.

### **2.3.4 Análise Pragmática**

A última etapa da análise de linguagem natural se dedica a encontrar o real significado de um enunciado dentro de seu contexto. Harris [23] definiu a pragmática como o conjunto de aspectos do discurso que não estão presentes explicitamente nas palavras utilizadas no texto. Esta etapa é a mais importante para implementação de sistemas flexíveis.

A referenciação por meio de pronomes é um problema recorrente para a análise pragmática, pois é necessário resolver para qual dos sujeitos ou objetos presentes na conversa foi feita a citação. A utilização de pronomes permite também que a referência seja feita a um sujeito ou frase que ainda está por vir, dificultando ainda mais a análise.

## **2.4 Síntese de fala em linguagem natural**

A síntese de fala é o processo de reprodução artificial de uma linguagem natural. As falas são sintetizadas concatenando-se pedaços de fala gravadas e reproduzindo o som com o auxílio de um alto falante. Sistemas de síntese de fala inteligível permitem a implementação de aplicações para acessibilidade, atendimento automatizado, navegadores GPS e assistentes de voz.

Os sistemas se diferenciam pelo tamanho das palavras armazenadas no banco de dados. Um sistema baseado em sílabas e letras permite a reprodução de um grande espectro de palavras em detrimento da qualidade da síntese. O armazenamento de palavras ou frases inteiras possibilita saída de alta qualidade para espectro limitado de interações. Para analisar a qualidade de um sintetizador de fala é necessário considerar a similaridade com a fala humana e a sua capacidade de ser entendida pelos usuários.

### 2.4.1 Sistema texto-fala

A síntese de fala implementada para a interação humano robô neste trabalho utiliza os módulos de sistema texto-fala – Text-To-Speech (TTS) – disponíveis na biblioteca NAOqi [24]. Os subprocessos podem ser subdivididos em duas principais funcionalidades.

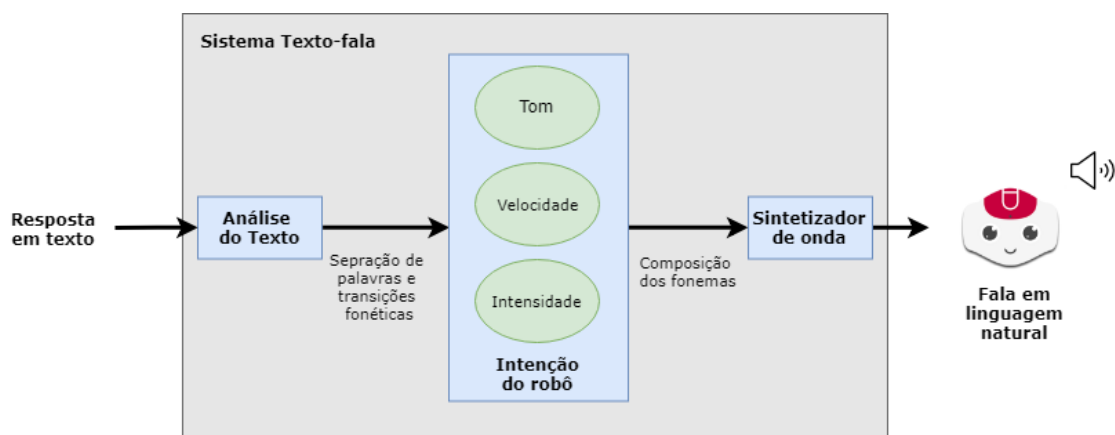


Figura 2.2: Síntese de fala implementada para o robô NAO.

Primeiramente, é preciso converter o texto bruto, com números e símbolos, no equivalente em palavras. Em seguida, são analisados os fonemas formados pelas transições entre as palavras. Os fonemas das palavras e das transições são concatenados e dão origem à representação linguística simbólica.

O sintetizador converte a informação simbólica em fala natural pela emissão de ondas acústicas [25]. Além da conversão de texto para linguagem falada, é função deste módulo transmitir informações paralinguísticas, como emoções ou intenções.

Desta forma, de acordo com o tipo de mensagem (informar, solicitar, saudar) devem ser utilizados os parâmetros de síntese (tom, intensidade, velocidade) adequados. A pesquisadora Cynthia Breazeal defende que mesmo em uma abordagem rudimentar os usuários reconhecem as emoções expressas na fala do robô [9].

### 2.4.2 Multimodalidade

A comunicação não verbal representa grande parte da comunicação interpessoal e é capaz de transmitir sentimentos, humor e opiniões. Robôs não possuem a capacidade de demonstrar sentimentos como os humanos, contudo estudos indicam que pequenas expressões em robôs podem ser interpretadas e reconhecidas por seres humanos. Pesquisas realizadas por Park, Moshkina e Arkin com a plataforma NAO demonstram também que humanos podem perceber emoções mesmo quando robôs não estão deliberadamente demonstrando sentimentos [26].

Além da modulação na sintetização de voz, neste trabalho será adotado o modelo multimodal de comunicação. As respostas do robô devem incluir gestos, movimento da cabeça, expressões faciais, orientação corporal adequada e direcionamento dos olhos. A utilização desses canais complementa

a comunicação, gerando maior engajamento e evitando ambiguidades.

As reações multimodais do robô devem expressar seu estado atual, podendo se mostrar confuso quando o algoritmo não identificar a intenção do usuário ou cansado quando sua bateria está baixa. As reações facilitam o entendimento de ações do robô, além de alinhar as expectativas das interações com as capacidades do robô sem comprometer a usabilidade dos sistemas.

A utilização de um padrão de gestos durante conversas permite criar traços de personalidade nos robôs que são reconhecidos pelos usuários e influenciam na comunicação. Um robô extrovertido, por exemplo, executa gestos mais expansivos, enquanto um robô introvertido evita o olhar do observador.

## 2.5 Arquiteturas internas

Para permitir o planejamento de ações e a tomada de decisões em um robô social são necessários modelos computacionais apropriados. A arquitetura interna deste tipo de sistema deve ser capaz de gerar planos orientados aos objetivos finais da interação, relacionando as expectativas do interlocutor com as intervenções do robô.

É necessário reavaliar os próximos estados do robô sempre que um objetivo é alcançado ou o interlocutor age de forma inesperada. Os modelos computacionais devem permitir que a adaptação seja rápida, ampliando a percepção de flexibilidade do sistema. As três principais categorias de sistema encontradas na literatura de robôs sociais foram descritas por Jurafsky e Martin [27] e são descritas a seguir.

### 2.5.1 Sistemas baseados em Estados Finitos

A arquitetura do sistema de diálogo pode ser baseada em uma máquina de estados finitos, de forma que todos os estados e transições são codificados antes da interação. Em sistemas deste tipo, a mudança de estado só ocorre quando o usuário fornece a informação pela qual o sistema espera naquele ponto, geralmente composta por expressões ou frases curtas.

O espectro de intenções suportada em cada momento da interação é limitado, permitindo melhor ajuste do reconhecimento de fala. Sistemas deste tipo restringem a liberdade do usuário, pois todas as respostas precisam ser mapeadas previamente na máquina de estados finitos. Respostas fora da ordem prevista ou com excesso de informações podem ocasionar erros no diálogo [28].

Nessa arquitetura o robô é responsável por guiar a conversa, solicitando as informações na ordem previamente programada e confirmando as operações para neutralizar eventuais erros de reconhecimento. Sistemas desse tipo são utilizados para automatizar centrais de atendimento, identificando o motivo da ligação e direcionando para o grupo de atendimento mais adequado. Perguntas comuns em sistemas desse tipo são “como posso te ajudar?”, “qual a data desejada?” e “você deseja (...) , isto está correto?”.

### 2.5.2 Sistemas Baseados em Formulários

Sistemas baseados em formulários fornecem interface em linguagem natural para um banco de dados. As perguntas feitas pelo agente estão associadas aos campos necessários para manipular o banco de dados. Por exemplo, o campo “horário” em sistemas de transporte pode ser preenchido com a resposta da pergunta “que horas você deseja partir?”.

O sistema deve ser capaz de gerir quais campos ainda não foram preenchidos e continuar a indagar o usuário até que possua os dados suficientes para consultar o banco de dados. Após encontrar a informação desejada, o robô pode informar ao usuário ou executar operações com os dados obtidos.

Sistemas deste tipo oferecem maior liberdade ao usuário quando comparados aos sistemas baseados em máquinas de estados finitos. Os sistemas são capazes de lidar com sobrecarga de informação e o usuário pode assumir a iniciativa do diálogo [28].

### 2.5.3 Sistemas Avançados

Aplicações modernas implementam algoritmos complexos para o planejamento, compreensão e gestão do diálogo, elaborando planos de várias etapas para concluir seu objetivo. Os sistemas são capazes de antecipar obstáculos e fazer inferências sobre o usuário, tornando os diálogos mais cooperativos e alcançando os objetivos de forma mais rápida.

Dentre as arquiteturas utilizadas ressaltam-se os modelos de crenças [29], os processos de decisão Markovianos [27] e os sistemas baseados no estado da informação [30]. Nesses casos, o comportamento do robô e as políticas de decisão podem ser alteradas de acordo com o contexto e o histórico do diálogo.

Arquiteturas de sistemas de diálogos podem ainda combinar múltiplas estratégias de diálogo no mesmo sistema, mudando de estratégia de acordo com as circunstâncias. Dessa forma é possível utilizar o melhor de cada uma das estratégias para tornar a interação homem-máquina o mais natural possível.

## 2.6 Avaliação de aplicações HRI

A avaliação de aplicações HRI com resultados verificáveis, confiáveis e reproduzíveis ainda é um grande desafio para a área. O uso de métodos e métricas padronizadas é defendido por alguns pesquisadores, mas não contempla totalmente as variações de cada interação. A mensuração dos efeitos da interação promove maior credibilidade e validade para a pesquisa científica, que deve ser reconhecida também nos campos da psicologia e das ciências sociais [31].

As interações devem ser avaliadas pela sua usabilidade, ou seja, o quanto atendem as expectativas e necessidades do usuário. Como boa parte da avaliação de usabilidade é subjetiva, Dybkjaer destacou métricas quantitativas para avaliar robôs sociais com base na sua capacidade de completar tarefas [32].

- **Taxa de erro de palavra (Word Error Rate - WER):** indica o quanto o sistema errou ao reconhecer as palavras.
- **Taxa de erro de fala:** indica quantos atos de fala tiveram pelo menos um erro de reconhecimento em relação a todos os atos de fala.
- **Cobertura de vocabulário:** porcentagem do discurso falado que é compreensível pelo sistema.
- **Tempo para realização da tarefa:** capacidade do sistema de gerar resposta suficientemente rápida para permitir a interação com humanos. Determinada pela quantidade de tomadas de fala ou pelo tempo necessário para completar determinada tarefa.
- **Taxa de correção:** quantos atos de fala foram utilizados para corrigir erros no diálogo.

## 2.7 Interação por diálogos

Para permitir que o usuário se comunique com a plataforma NAO por meio de diálogo é necessário respeitar as propriedades da comunicação em linguagem natural utilizada por seres humanos. Em especial, a intolerância ao silêncio requer que o tempo de processamento dos comandos seja curto e a iniciativa mista prevê que o robô seja capaz de reagir e também conduzir a conversa.

As métricas de avaliação HRI serão utilizadas para avaliar o sistema de diálogos, aferindo a capacidade de reconhecimento e recuperação do sistema. As etapas de transcrição da fala e compreensão da linguagem natural podem ser realizadas por ferramentas comerciais, como é descrito no capítulo a seguir.



## Capítulo 3

# Ferramentas de Desenvolvimento

*Para possibilitar a implementação do sistema foram utilizadas as bibliotecas fornecidas para o fabricante do robô NAO. Além disso, o sistema é gerenciado pelo framework ROS e a implementação utiliza API's disponíveis na Google Cloud.*

### 3.1 Plataforma NAO

A implementação deste projeto foi feita na plataforma NAO, produzida pela Aldebaran Robotics. Com cerca de 60 centímetros de altura e aparência cativante, o robô humanoide se popularizou pelo seu uso em diversas aplicações. Dentre essas destacam-se interações com crianças autistas [33], teleoperações [34] e a participação na categoria de plataforma padrão da Robocup [35].

O NAO possui processador Intel Atom de 1.6 GHz e 1GB de RAM, dois pares receptor/transmissor de sonares, giroscópio de dois eixos e acelerômetro de três eixos. Além disso, o NAO possui dois sensores infravermelhos, quatro sensores de pressão em cada pé e *encoders* em todas as juntas. O robô possui também microfone e duas câmeras com resolução 640x480 pixels.

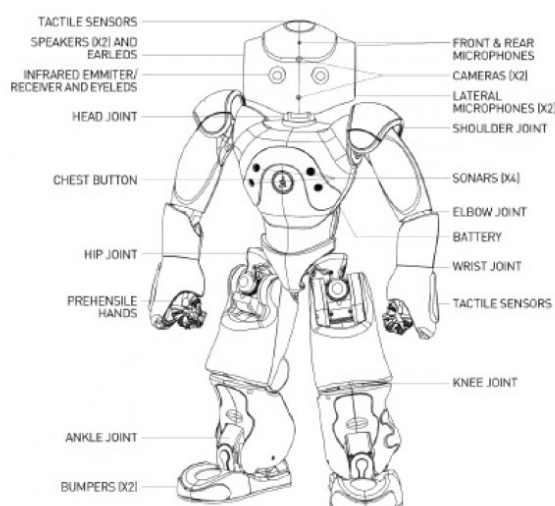


Figura 3.1: Plataforma NAO. [24].

O robô NAO conta com 25 graus de liberdade, sendo capaz de acessar com os braços qualquer ponto do espaço. O robô utiliza sistema operacional baseado em Linux na versão Gentoo.

### 3.1.1 NAOqi

Para comunicação com os motores, sensores e memória foi utilizado o *framework* NAOqi [24], desenvolvido pela fabricante do robô. Abaixo é ilustrado o funcionamento dos módulos do *framework* NAOqi.

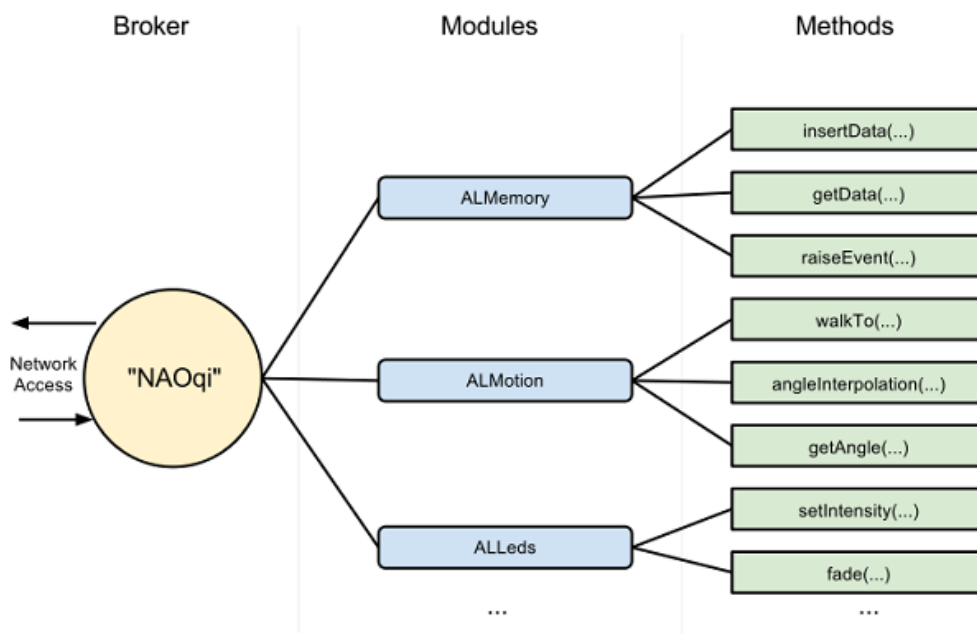


Figura 3.2: Funcionamento do *framework* NAOqi [24].

O *framework* acessa os diversos módulos por meio de um *Broker*, carregando as bibliotecas que contém os métodos para facilitar a programação da plataforma. Os principais módulos da NAOqi utilizados na aplicação são descritos a seguir.

#### 3.1.1.1 ALRobotPosture

O módulo ALRobotPosture permite que o robô se movimente para uma postura estável pré definida pelo fabricante. Para realizar o movimento o robô identifica a posição atual dos seus motores e calcula a trajetória da postura atual para a postura pré definida. O módulo utiliza os sensores inerciais e de pressão localizados nos pés do robô durante a movimentação para evitar quedas. Também é possível reduzir a velocidade de movimentação para garantir ainda mais a estabilidade.

A biblioteca é programada em 8 diferentes posturas, contudo foram priorizadas posturas pouco complexas que permitam o robô interagir em pé, sentado e com os joelhos dobrados. A intera-

ção foi programada para a postura em pé, dessa forma algumas expressões corporais podem ser prejudicadas se o robô estiver em outras posturas.

### 3.1.1.2 ALTextToSpeech

A sintetização das respostas enviadas pelo DialogFlow é realizada pela função ALTextToSpeech. A biblioteca está programada para sintetizar textos em português e permite configurar a entonação e a velocidade para customizar o discurso. Como a função não gera movimentação no robô, o módulo é utilizado principalmente durante as interações em que o robô não deve movimentar os seus braços ou pernas.

### 3.1.1.3 ALAnimatedSpeech

O módulo ALAnimatedSpeech permite que o robô se comunique de forma mais expressiva. Além de ativar a síntese de voz, a função sincroniza expressões corporais para reforçar a comunicação. Por ser uma plataforma educacional, o NAO possui mais de cem opções de linguagens corporais disponíveis para serem intercaladas com as falas.

Em sua configuração padrão, o módulo escolhe uma expressão corporal aleatoriamente para combinar com o tempo de reprodução da fala. Para sincronizar expressões corporais que reforcem a fala é necessário anotar o texto com *tags* que especifiquem o propósito da expressão corporal.

Por exemplo, para permitir uma apresentação animada foi utilizada a anotação a seguir:

```
^startTag(hello) Olá! Eu sou o NAO. ^stopTag(hello)
^startTag(you) Qual é o seu nome? ^stopTag(you) .
```

A *tag* "hello" indica que o NAO realizará aleatoriamente uma das opções de aceno disponibilizadas pela fabricante do robô. Após terminar o movimento, a *tag* "you" faz com que o robô realize uma das opções de animação em que aponte para o interlocutor. Além das *tags* "hello" e "you", foram utilizadas as *tags* descritas abaixo.

- **affirmative:** movimentos afirmativos com a cabeça e os braços do robô.
- **happy:** movimentos rápidos que demonstrem felicidade.
- **body language:** movimentos curtos e sem expressão de sentimento.
- **me:** robô referencia a si mesmo, apontando para o próprio corpo.

### 3.1.1.4 ALMotionProxy

O módulo ALMotionProxy facilita a movimentação espacial, disponibilizando funções que permitem o controle dos motores do robô. O módulo possui três principais métodos para controle dos atuadores:

- **Stiffness control:** controla o torque aplicado no atuador utilizando a escala de 0.0 a 1.0. A posição 0.0 indica que a junta está livre para ser movimentada manualmente, enquanto o outro extremo indica o máximo de torque.
- **Joint Control:** permite que a posição dos atuadores seja controlada diretamente. Cada junta pode ser controlada individualmente ou em paralelo com outras juntas, utilizando o feedback dos atuadores para atingir a posição desejada.
- **Locomotion Control:** ativa a marcha do robô para movimentação do corpo inteiro. O método utilizado para essa aplicação tem como parâmetros a distância ao longo dos eixos X e Y e um ângulo theta de rotação em Z. O ângulo theta é utilizado para compensar efeitos do desbalanceamento de peso do robô e permitir movimentos mais lineares.

Além de controlar os atuadores, o módulo implementa reflexos no robô que permitem desviar de obstáculos, evitar colisão entre os membros do robô, reduzir o gasto de bateria e diminuir o torque dos motores ao detectar eventuais quedas. Para permitir movimentos suaves, o módulo é executado em ciclos de 20ms (50Hz).

A movimentação espacial do robô utiliza o método *moveTo*, enquanto a movimentação de partes do robô utiliza o método *angleInterpolation*. Quando a intenção “andar para frente” é detectada, o método *moveTo* é executado tendo como parâmetro ( $x = 0.5$  e  $theta = 0.1$ ) para movimentar o robô meio metro para frente. Para a intenção “andar para direita” o método é executado com os parâmetros ( $y = -0.5$ ).

Os eixos de locomoção seguem a direção da ilustração abaixo.

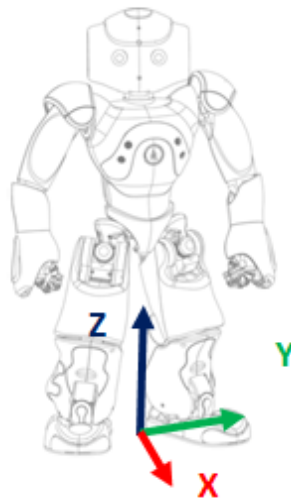


Figura 3.3: Eixos de movimentação do NAO.

O método *angleInterpolation* realiza a interpolação de múltiplas juntas e permite a movimentação de partes específicas do robô. Dessa forma, a movimentação da cabeça e dos braços é realizada executando o método com os ângulos desejados para os atuadores.

Para registrar os ângulos desejados foi utilizado o software Choregraphe, fornecido pela fabricante do robô. O torque dos atuadores é reduzido permitindo a movimentação manual dos membros do robô. É possível registrar a posição de todos os atuadores para formar uma linha do tempo com sequência de posturas.

A função depende de três principais parâmetros:

- **names:** nome dos atuadores que devem ser movimentados.
- **anglesList:** lista de ângulos da posição de destino.
- **timesList:** lista que determina em que instante cada posição deve ser assumida

### 3.1.1.5 ALMotALAutonomousLife

O módulo ALAutonomousLife é utilizado para tornar os movimentos do robô mais dinâmicos e parecidos com o dos humanos. Quando ativado, assume o controle dos LED's e dos atuadores enquanto o robô estiver em modo de espera. As três principais habilidades que o módulo confere ao robô são:

- **ALAutonomousBlinking:** gerencia os LED's dos olhos para causar a impressão de que o robô está piscando os olhos e reconhecendo os estímulos.
- **ALBackgroundMovement:** define pequenos movimentos corporais executados enquanto o robô está esperando por um novo comando.
- **ALBasicAwareness:** permite que o robô reaja ao ambiente e estabeleça contato visual com os usuários.

## 3.2 Google Speech-to-Text

A fala é a principal forma de comunicação utilizada por seres humanos e seu uso em aplicações facilita o controle de computadores. O primeiro algoritmo de reconhecimento de fala foi proposto por pesquisadores dos Laboratórios Bell em 1930, contudo o uso abrangente dessa tecnologia só foi possível em 1980 com os avanços na modelagem estatística da fala [36].

Atualmente o uso do reconhecimento de fala é impulsionado por algoritmos robustos desenvolvidos para aplicações comerciais. Buscando a difusão da tecnologia, desenvolvedores disponibilizam os algoritmos para uso em protótipos e pesquisa. Dentre os algoritmos disponíveis, foram analisados para a implementação o Google Speech-To-Text [37], o IBM Watson [38] e a Siri API [39].

O IBM Watson apresenta reconhecimento de alta precisão para 7 línguas, contudo sua compatibilidade com o *framework* ROS é limitada. A Siri API, desenvolvida pela Apple, oferece

reconhecimento compatível com microfones de celulares e *tablets*, no entanto seu uso gratuito é restrito a 30 requisições diárias. A ferramenta Google Text-To-Speech foi escolhida para esse projeto por possuir biblioteca de fácil integração com o ROS, permitindo a transcrição gratuita de até 60 minutos de áudio por dia.

A API disponibilizada gratuitamente pelo Google fornece a conversão de voz em texto a partir de redes neurais artificiais profundas. A transcrição é compatível com mais de 120 idiomas, podendo ser realizada a partir de arquivos de áudios ou processando *streaming*.

Para se conectar com a API é necessário gerar uma chave privada JSON, que deve ser exportada como variável de ambiente no terminal em que o código será executado. Após a sessão ser iniciada, a função Streaming Request é capaz de transcrever falas de até um minuto. A transcrição do áudio pode ser acompanhada pelo terminal, quando uma frase é finalizada o resultado é enviado ao Dialog Flow para reconhecimento do significado.

### 3.3 DialogFlow

Ferramentas de interpretação de linguagem natural (Natural Language Understanding - NLU) permitem extrair automaticamente o significado de textos. Dentre as ferramentas disponíveis para uso gratuito destacam-se a IBM Natural Language Understanding [40], e a SYSTRAN.io [41], no entanto foi escolhida a ferramenta Google Dialog Flow [42] por sua integração com o reconhecimento de voz.

O DialogFlow possui interface própria para construção de aplicações. Em 2016, a ferramenta foi adquirida pelo Google e integrada na infraestrutura Google Cloud. Os principais conceitos envolvidos na criação de uma aplicação são:

- **Agente:** representam as instâncias do módulo do *chatbot*.
- **Intenções:** mapeamento entre as sentenças do usuário e as ações que devem ser tomadas.
- **Entidades:** conceitos necessários para completar uma intenção, representam os parâmetros de entradas em linguagem natural.
- **Contextos:** a conexão entre as intenções e entidades gera contexto que aprofunda o nível do diálogo.

A criação de um agente no DialogFlow é feita pela definição de intenções e entidades. Para que o robô consiga interagir é necessário inserir as diferentes expressões que o usuário pode utilizar para expressar uma mesma intenção. As expressões inseridas na definição do *chatbot* são utilizadas para o treinamento do algoritmo de reconhecimento.

Sendo assim, a precisão depende da quantidade e da variedade de sentenças utilizadas no treinamento do chatbot. Por exemplo, para aumentar a robustez as expressões “Quantos anos você tem?” e “Qual a sua idade?”, devem utilizadas no treinamento para ativar a mesma intenção.

As sentenças transcritas pelo Google Speech-to-Text são enviadas para o DialogFlow e divididas em palavras. Cada oração é analisada em busca de uma semelhança com os exemplos, caso a sentença detectada na fala do usuário não corresponda a nenhuma opção disponível no agente, uma saída padrão informa a incompreensão e solicita que o usuário repita a expressão.

## 3.4 ROS

O ROS é um conjunto de bibliotecas *opensource* para facilitar o desenvolvimento de robôs, permitindo a integração de diferentes computadores e linguagens. As bibliotecas clientes possuem algoritmos de planejamento, percepção, mapeamento e ferramentas de simulação. Além do gerenciamento de processos programados em diferentes linguagens, possui também bibliotecas para realizar o controle de hardware em baixo nível.

Os três conceitos fundamentais do ROS são:

- **Nós:** elementos da arquitetura responsáveis por executar as diferentes tarefas do sistema.
- **Tópicos:** canais utilizados para troca de mensagens entre nós. Cada tópico pode ser lido ou receber publicações de diferentes nós ao mesmo tempo.
- **Mensagens:** definem as estruturas utilizadas para transmissão de dados. Além dos tipos padrão de dados o ROS permite a criação de estruturas para atender necessidades específicas de cada sistema.

Nesta aplicação o ROS é responsável por enviar para as API's Speech-to-Text e DialogFlow o áudio extraído do microfone do computador. Os resultados são publicados em tópicos e o ROS envia as informações para o robô NAO pela rede de computadores, permitindo a ativação da animação equivalente a intenção do usuário.

## 3.5 Redes de computadores

Para permitir a comunicação entre diferentes computadores é necessário que eles estejam conectados a uma rede. Os conceitos de transmissão de dados são abordados de acordo com as camadas dos protocolos Transmission Control Protocol / Internet Protocol (TCP/IP), com enfoque nos protocolos User Datagram Protocol (UDP) e Dynamic Host Configuration Protocol (DHCP) utilizados neste projeto.

### 3.5.1 TCP/IP

O TCP/IP é um conjunto de protocolos de comunicação entre computadores em rede organizados em camadas. Cada camada é responsável por uma das etapas de transmissão de dados

e fornece serviços para a camada superior. As camadas mais altas estão próximas do usuário, enquanto as camadas inferiores estão mais próximas do hardware.

A camada de aplicação gera requisições que são processadas e enviadas para protocolos da camada de transporte. A camada de transporte agrupa os dados em pacotes e encaminha os dados para a camada de Internet. Os dados são endereçados e transportados para a camada de rede responsável por enviar o pacote através da rede.

A padronização dos protocolos permite a conectividade entre sistemas não similares. Para permitir a comunicação em rede utilizada neste trabalho, foi necessário integrar a aplicação do diálogo com as camadas de transporte do computador e do robô. O protocolo UDP foi escolhido por oferecer maior velocidade na transmissão de mensagens e minimizar o atraso para o robô responder ao comando.

#### **3.5.1.1 User Datagram Protocol - UDP**

O protocolo UDP atua da camada de transporte da rede e permite o envio de dados encapsulados em pacotes. O serviço não cria uma conexão duradoura entre o cliente e o servidor, permitindo que um mesmo cliente envie pacotes para vários outros por meio dos serviços de broadcast e multicast. Não são necessárias mensagens para iniciar a comunicação ou negociar o tamanho de pacotes, permitindo um modelo simples e mais rápido que outros protocolos de rede.

Para estabelecer a comunicação os computadores devem definir um número de porta que não seja reservada para outros serviços. A porta indica o destino para a aplicação enviar os dados e onde o cliente deve aguardar pela chegada de mensagens. Além das portas de origem e destino, o cabeçalho de um pacote UDP deve conter o comprimento da mensagem para permitir a conferência de sua integridade.

Apesar de prover verificação da integridade dos pacotes recebidos pelo método de *checksum*, o protocolo não oferece garantia do recebimento da mensagem enviada. O uso do protocolo é recomendado para transmissão de dados pouco sensíveis, em que é preferível a perda de um pacote a esperar pela retransmissão.

As aplicações que se comunicam por este modelo estão sujeitas a instabilidade da rede em que os computadores estão conectados. O protocolo não deve ser utilizado em aplicações nas quais a checagem e correção de erros é necessária.

#### **3.5.1.2 Dynamic Host Configuration Protocol - DHCP**

O protocolo de configuração dinâmica de Host usa o modelo cliente-servidor com endereços IP's dinâmicos para conectar um computador a uma rede ou a outros dispositivo. O objetivo do protocolo é a simplificação da administração da rede, uma vez que um computador pode sozinho encontrar um endereço de IP válido.

Quando um computador entra na rede, o protocolo dispara um pacote para todas as máquinas com a requisição DHCP. Qualquer servidor pode responder a requisição, fornecendo uma confi-



guração para o computador solicitante. Ao se conectar com a configuração recebida pelo servidor DHCP, o computador tem acesso aos serviços disponíveis nos outros dispositivos conectados na rede.

O DHCP permite a criação de redes complexas sem a necessidade de alocação manual dos computadores, sendo utilizado para que computadores consigam acessar de forma simples a World Wide Web. Neste trabalho, o protocolo é utilizado para que o computador consiga enviar e receber dados das APIs Speech-To-Text e Dialog Flow armazenadas nos servidores do Google.

# Capítulo 4

## Implementação

*Neste capítulo são descritos os algoritmos implementados para permitir o controle do robô por comando de voz. São descritas também as funcionalidades disponíveis e como ativá-las.*

### 4.1 Descrição do sistema

O sistema implementado para possibilitar a interação por meio de diálogos com a plataforma NAO é composto por 5 principais etapas. As etapas 1 e 2 são responsáveis pela transcrição da fala do usuário, as etapas 3 e 4 efetuam a compreensão da linguagem natural e a etapa 5 comunica ao robô qual deve ser sua resposta.

A integração das ferramentas de desenvolvimento é ilustrada no diagrama abaixo.



Figura 4.1:

- **1:** A captura de ondas sonoras semelhantes com a fala humana inicia o sistema. O *framework* ROS é responsável por iniciar uma sessão no Google Cloud Plataform, integrando o microfone

do computador com a API do Google Speech-To-Text.

- **2:** O ROS envia o áudio capturado para a ferramenta Google Speech-To-Text e a transcrição parcial pode ser acompanhada pelo terminal.
- **3:** A interrupção da transcrição por mais de 2 segundos indica o fim da fala do usuário. A transcrição integral em formato de *string* é enviada para a ferramenta Google Dialog Flow
- **4:** O Dialog Flow retorna para o ROS uma estrutura de mensagem personalizada contendo a intenção identificada, as entidades complementares e a confiabilidade do reconhecimento.
- **5:** A intenção e as entidades são convertidas para o formato *string* e enviadas para o robô por conexão sem fio.

As etapas 1, 2, 3 e 4 são realizadas pelo módulo *intentionDetector.py*, a etapa 5 é executada com o auxílio de soquetes UDP implementados nos módulos *udpInterface.py* e *dialogNAO.py*.

## 4.2 Módulos do sistema

### 4.2.1 Dialog Flow Client

O script *intentionDetector.py* foi criado para conectar o áudio do computador com a API Dialog Flow. Inicialmente é necessário criar uma sessão para se comunicar com o agente. Para isso, é utilizado a biblioteca *flowbeta1*.

Em seguida, o áudio do microfone do computador é acessado para iniciar a identificação de palavras. Os áudios capturados são enviados em pacotes para o reconhecedor de fala e é retornada uma sequência de caracteres contendo a fala do usuário transcrita. A transcrição da fala é enviada ao servidor do Dialog Flow, que reconhece a intenção na fala do usuário e retorna a estrutura de dados com a resposta pré programada para a intenção.

### 4.2.2 Soquetes UDP

Para notificar o robô sobre o reconhecimento de uma intenção do usuário foram utilizados soquetes UDP. Optou-se pela implementação na linguagem Python pela simplicidade para criar e utilizar os soquetes com o auxílio da biblioteca *socket*. O soquete que envia a intenção para o robô é implementado no script *sendIntention.py*, enquanto o responsável por receber a mensagem é implementado no script *dialogNAO.py*

#### 4.2.2.1 udpInterface.py

O script *udpInterface.py* implementa um nó executado no computador que realiza a checagem recorrente do tópico `/intentions`, esperando que o nó conectado ao Dialog Flow publique as inten-

ções reconhecidas. Quando uma mensagem é postada no tópico, o script `udpIntercafe.py` captura a mensagem e a encaminha para o IP e a porta em que o robô está conectado.

Ao pressionar o botão localizado no peitoral do robô o número IP atribuído ao robô é informado em uma mensagem de áudio. O número IP deve ser usado como destino para a conexão UDP. A porta 5005 foi utilizada por ser padrão para esse tipo de implementação.

#### **4.2.2.2 dialogNAO.py**

Para que o robô execute uma resposta é necessário que o script `dialogNAO.py` identifique a chegada de uma nova mensagem no soquete UDP. Quando uma nova mensagem é detectada, seu conteúdo é interpretado e avaliado para planejar a resposta do robô.

O endereço utilizado para a conexão deve ser o próprio IP do robô, acompanhado pelo mesmo número de porta utilizado no módulo que envia as mensagens. As mensagens são capturadas com um buffer de 1024 bytes, seguindo a configuração padrão do protocolo

### **4.3 Funcionalidades**

#### **4.3.1 Movimentação de partes do corpo do robô**

Buscando garantir naturalidade e segurança para a interação, implementou-se o controle de voz para movimentação dos braços e cabeça do robô. A ativação desta funcionalidade pode ser feita em qualquer momento da interação, solicitando o tipo de movimento e especificando o membro desejado.

Os braços foram programados para se posicionarem em posturas pré definidas pelos ângulos dos motores do ombro e cotovelo. A movimentação dos braços pode ser acionada individualmente ou simultaneamente. Ao reconhecer o comando, o robô interrompe qualquer ação corporal que esteja realizando para posicionar os braços.

Para interpretar os comandos de movimentação de braços o algoritmo reconhece as intenções “levantar” e “abaixar” e busca pelas entidades “direita”, “esquerda” ou “ambos”. Para interpretar comandos de movimentação da cabeça o algoritmo reconhece as intenções "olhar" e "virar a cabeça para".

##### **4.3.1.1 Movimentação de ombros**

A movimentação lateral dos ombros do robô foi configurada para os perfis levantado ( $70^\circ$ ) e abaixado ( $10^\circ$ ). A funcionalidade é ativada por sentenças como “abaixar o braço esquerdo” e “levantar os braços”.

### 4.3.1.2 Movimentação dos cotovelos

Os cotovelos estão programados para as poses “esticado” ( $5^\circ$ ) e “dobrado” ( $80^\circ$ ). “Esticar o cotovelo direito” ou “dobrar os cotovelos” são exemplos da ativação da movimentação.

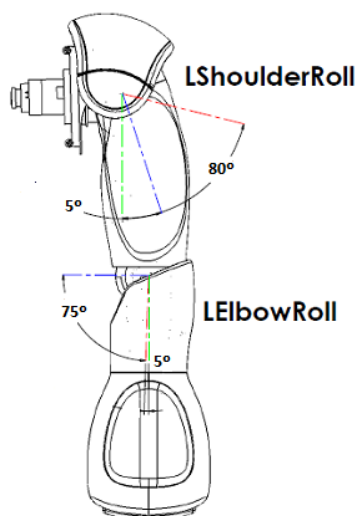


Figura 4.2: Ângulos de movimentação do ombro e cotovelo.

### 4.3.1.3 Movimentação da cabeça

O alinhamento dos olhos do robô com os do usuário é importante para garantir a naturalidade da interação. Para permitir o posicionamento da cabeça do robô por comando de voz são consideradas as posições “esquerda”, “frente” e “direita”.

São reconhecidas expressões como “virar a cabeça para a esquerda” ou “olhar para frente”.

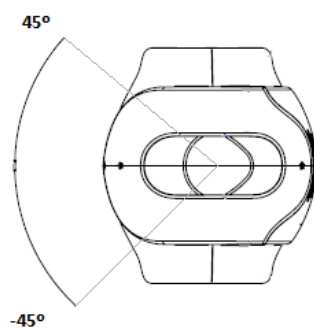


Figura 4.3: Ângulos de movimentação da cabeça.

### 4.3.2 Movimentação espacial

O posicionamento do usuário durante a interação pode variar e o robô deve ser capaz de se adequar. Para atender esse requisito utiliza-se da marcha implementada na biblioteca NAOqi para movimentá-lo em 4 direções. A intenção é formada por palavras como “andar” ou “caminhar” e as entidades “frente”, “trás”, “direita”, “esquerda”. Sendo assim, para movimentar o robô espacialmente são reconhecidas expressões como “caminhar para frente” e “andar para esquerda”.

### 4.3.3 Controle de volume

O robô NAO possui duas saídas de som localizadas em sua cabeça, permitindo a síntese de voz e reprodução de áudios em diversos formatos. Para garantir o conforto dos usuários é necessário adequar o volume dos alto falantes em diferentes ambientes. Sendo assim, implementou-se o controle de voz para permitir que a configuração do volume dos alto falantes seja feita sem o auxílio de um computador.

O controle implementado permite que o usuário varie o volume em passos de 10% ou utilize um dos três perfis programados (alto, baixo ou médio). A ativação do controle pode ser feita em qualquer momento da interação, sendo causada pelas intenções “aumentar” e “abaixar” ou pelo reconhecimento das entidades “baixo”, “médio” ou “alto”.

Sentenças como “NAO, aumentar o volume” ou “colocar o volume no médio” são reconhecidas para o controle do volume dos alto falantes do robô.

### 4.3.4 Mudar de postura

As interações foram programadas utilizando linguagem corporal, para a execução dos movimentos é necessário que o robô esteja em pé. Porém, outras posições corporais podem ser utilizadas para economizar bateria ou aumentar a estabilidade do robô.

A mudança de postura é reconhecida pelas intenções “ficar de pé”, “sentar” e “descansar”. A biblioteca NAOqi é acionada para alcançar a nova postura de forma estável, utilizando os sensores de posição e o acelerômetro do robô para corrigir o movimento das juntas.

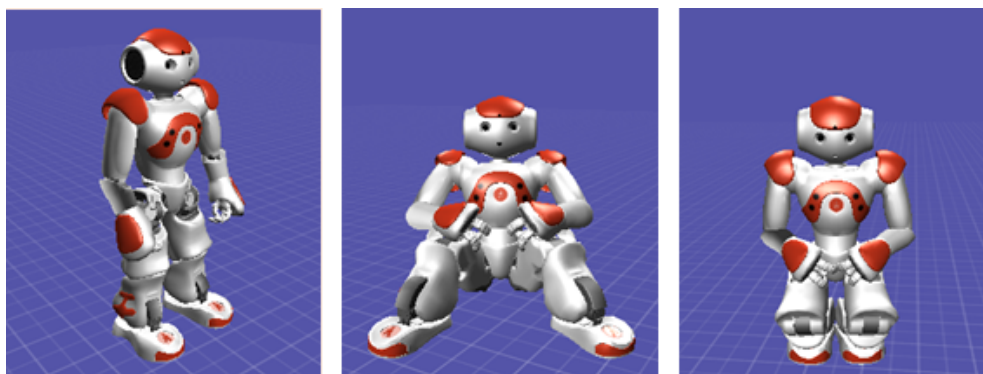


Figura 4.4: Posturas em pé, sentado e descansando.

Enquanto o robô estiver nas posturas sentado e descansando a expressão corporal será interrompida, permitindo apenas a síntese de fala das respostas e a movimentação de partes específicas do corpo.

# Capítulo 5

## Resultados

*Nesta sessão estão registrados os experimentos e os resultados obtidos durante o projeto, além disso é feita a avaliação do sistema com base nas métricas de avaliação humano robô propostas por Dybkjaer [32].*

### 5.1 Interações com usuário

O sistema de diálogos e comando de voz implementado na plataforma NAO foi testado com usuários que não tinham experiência com robótica. A descrição do funcionamento do sistema foi omitida do usuário, permitindo que a percepção sobre a usabilidade se baseasse apenas na interação.

Os experimentos aconteceram no Laboratório de Automação e Robótica da Universidade de Brasília (LARA - UnB). Foram selecionados três usuários com diferentes tons de voz e sotaques, buscando explorar a robustez da transcrição da fala. O robô foi posicionado no chão em frente ao usuário e conectado ao computador por conexão sem fio (Wi-Fi). O usuário foi informado que a captura de fala é realizada pelo computador e foi orientado a falar em sua direção.

Ao iniciar a interação, o usuário recebe uma pequena lista com os grupos de funcionalidades disponíveis no robô. As funcionalidades não são acompanhadas por instruções detalhadas ou exemplos, demandando que o usuário crie suas próprias frases para interagir com o robô. Após iniciar o experimento o usuário interage com o robô sem intervenções externas, explorando a capacidade do sistema ser facilmente entendido.

O robô indica que está pronto para receber comandos com seus LED's e alertas sonoros. O sistema não é capaz de lidar com atropelamentos de fala do usuário, uma vez que durante a execução dos movimentos de resposta o sistema de reconhecimento de fala é interrompido para evitar transcrições indesejadas.

O fim das interações é marcado pelo uso de todas as intenções programadas ou perda de engajamento do usuário, o que acontecer primeiro. As interações duraram de 3 a 4 minutos, variando de acordo com as funcionalidades escolhidas pelo usuário. O registro das interações foi feito por meio de vídeo com a fala do usuário e resultado da transcrição de texto, permitindo avaliar a



acurácia do reconhecimento. A transcrição das interações foi anexada ao final deste relatório, os vídeos obtidos estão disponíveis no CD anexo a este projeto.

### 5.1.1 Lista de funções

A lista de funções recebida pelos usuários facilita a interação e permite que ele explore sozinho a implementação. Os tópicos da lista devem ser explorados pelo usuário durante a interação do robô, permitindo a avaliação do desempenho do sistema.

- Saudações e despedidas.
- Movimentar os membros do robô.
- Locomoção do robô.
- Solicitar informações sobre o robô.
- Controlar o volume dos alto falantes.
- Mudar a postura do robô.
- Aprender sobre robótica.
- Ouvir uma piada.

## 5.2 Avaliação do sistema

### 5.2.1 Taxa de erro de palavra (Word Error Rate - WER)

Para avaliar a transcrição da fala do usuário é utilizado o critério de taxa de erro de palavra, largamente aplicado para avaliar a performance de transcritores e tradutores. A taxa de erro pode ser calculada por meio da expressão:

$$WER = \frac{(Substituições + Inserções + Omissões)}{Palavras}$$

Em que são considerados os casos:

- **Substituições:** quando uma palavra é transcrita como outra.
- **Inserções:** quando uma palavra não proferida pelo usuário é transcrita.
- **Omissões:** quando uma palavra da fala não é transcrita.
- **Palavras:** total de palavras proferidas na interação.

A WER foi avaliada para as três interações do robô, computando-se as substituições, inserções e omissões na transcrição do discurso. Na transcrição de resultados em anexos foram destacados os erros encontrados. Os resultados para as interações são mostrados na tabela abaixo:

	Substituições	Inserções	Omissões	Palavras	WER
Interação 1	1	0	1	85	2.35%
Interação 2	0	0	1	49	2.04%
Interação 3	1	0	1	52	3.84%

Tabela 5.1: Avaliação da taxa de erro de palavra

A taxa de erro WER reportada pelo Google durante entrevista em 2017 para o algoritmo de reconhecimento de fala é de 4.9%, superior ao encontrado nos experimentos. A diferença entre a WER nominal e a WER experimental pode estar associada ao uso de frases curtas e com poucos nomes próprios, cenários que facilitam a transcrição do texto.

### 5.2.2 Taxa de erro de fala

A taxa de erro de fala para este projeto demonstra a performance da implementação do agente no Dialog Flow. Para avaliação desse critério, registrou-se as intenções do usuário manualmente e comparou-se com as fornecidas pelo algoritmo de reconhecimento. A taxa de erro pode ser calculada por meio da expressão:

$$\text{Taxa de erro de fala} = \frac{\text{Atos com erro}}{\text{Total de atos}}$$

Sendo:

- **Total de atos de fala:** quantidade de frases que o usuário proferiu .
- **Atos de fala com erro:** quantidade de atos de fala com erro de reconhecimento.

Os resultados encontrados para as três interações experimentais estão registrados na tabela abaixo.

	Atos de fala com erro	Total de atos de fala	Taxa de erro de fala
Interação 1	2	22	9,1%
Interação 2	1	15	6,7%
Interação 3	2	16	12,5%

Tabela 5.2: Avaliação da taxa de erro de fala.

### 5.2.3 Cobertura de vocabulário

O treinamento dos algoritmos do Google Speech-To-Text utiliza a base de dados do Google para o treinamento das redes neurais de reconhecimento. A ferramenta é utilizada em outros serviços comerciais do Google e recebe melhorias contínuas de acurácia e extensão de vocabulário.

Durante a implementação e teste do projeto não foram identificadas palavras que o algoritmo não era capaz de transcrever. Mesmo expressões informais e nomes próprios foram reconhecidos pelo algoritmo de transcrição. Sendo assim, a cobertura de vocabulário foi suficiente para os requisitos do projeto e sua avaliação numérica desprezada neste projeto.

#### 5.2.4 Tempo para realização da tarefa

O tempo para realização de uma tarefa comandada por voz depende da quantidade de comandos necessários, do tempo para processamento dos comandos e do tempo necessário para executar a tarefa. As tarefas implementadas no robô possuem durações diferentes, uma vez que o tamanho da resposta dada pelo robô pode variar. Sendo assim, avaliou-se o tempo necessário para execução de uma resposta simples, uma movimentação e um diálogo com context, conforme indicado a seguir.

	Tempo de processamento	Tempo até o fim da tarefa
Resposta simples	548 ms	632 ms
Andar para frente	654 ms	5.806 ms
Contar uma piada	371 ms	8.426 ms

Tabela 5.3: Avaliação do tempo de realização da tarefa.

O tempo até o início das tarefas é satisfatório, não causando incômodo ao usuário durante a interação. Essa medida pode ser otimizada utilizando um roteador com maior velocidade de transmissão de mensagens. O tempo até o fim das tarefas não deve ultrapassar o limite de concentração do usuário. Como mesmo em um diálogo com contexto a duração da tarefa é curta, considera-se que o sistema atendeu os requisitos de tempo mesmo nas interações mais longas que envolvem a movimentação espacial do robô.

#### 5.2.5 Taxa de correção

A taxa de correção avalia a capacidade do agente continuar a conversa após uma detecção incorreta. Ao apresentar um comportamento incoerente com o comando do usuário ou contexto, o robô deve ser capaz de retornar ao estado anterior e corrigir seu planejamento.

Como o sistema implementado é baseado em estados finitos, a detecção incorreta de uma intenção leva o agente para um estado incorreto. Contudo, para ativar a transição correta o usuário só necessita proferir novamente a mesma sentença para que o estado desejado seja alcançado.

A avaliação da taxa de correção foi realizada com base na quantidade de tentativas que o usuário teve que realizar para conseguir ativar uma função após um erro de reconhecimento. A taxa de correção pode ser calculada pela expressão abaixo:

$$Taxa\ de\ correção = \frac{Frases\ de\ correção}{Quantidade\ de\ erros}$$

De forma que:

- **Quantidade de erros:** erros de detecção pelo Dialog Flow.
- **Total de frases de correção:** total de frases que representam a mesma intenção utilizadas após um erro de detecção.

A taxa de correção foi aferida para as três interações e os resultados estão registrados abaixo:

	Total de frases de correção	Quantidade de erros	Taxa de correção
Interação 1	2	1	2
Interação 2	2	2	1
Interação 3	3	2	1.5

Tabela 5.4: Avaliação da taxa de correção após erro.

A taxa de correção foi inferior a dois em todas as interações, demonstrando que é possível retomar o diálogo após um erro com, no máximo, duas orações. Caso os erros aconteçam durante momentos em que contexto é importante, a taxa de correção pode se elevar.

# Capítulo 6

## Conclusões

*O presente trabalho objetivou implementar interface para diálogo com o robô NAO utilizando comandos de voz. O sistema foi avaliado pela qualidade dos diálogos e pelo desempenho das ferramentas utilizadas. Ao final, são propostos trabalhos futuros para expandir o projeto.*

### 6.1 Diálogos

A interação humano robô por voz permitiu que o robô fosse controlado por usuários sem nenhum conhecimento prévio em robótica, eliminando a curva de aprendizagem de uma linguagem artificial. Além de obter respostas para perguntas simples, a implementação permitiu a locomoção espacial do robô e o controle de configurações como volume e rigidez dos motores.

Retomando as propriedades essenciais da comunicação apresentadas na introdução, conclui-se que a implementação atende os requisitos da linguagem natural. As propriedades de bidirecionalidade, iniciativa mista e percepção de posse da palavra foram atendidas, uma vez que as falas do usuário e do robô se intercalaram sem atropelamentos. A linguagem utilizada pelo robô é informal e as interações contam com o senso comum do usuário para fazer sentido. A sensação de presença é intensificada pelos LED's e movimentação da cabeça do robô, incentivando o usuário a continuar o diálogo.

### 6.2 Implementação e desempenho

O gerenciamento dos processos executados no computador pelo framework ROS possibilitou implementação simples e modular. A utilização dos tópicos para comunicação dos módulos permite que as funcionalidades sejam executadas separadamente para checagem e correção de erros. A execução do ROS não sobrecarregou o processamento do robô, uma vez que parte do processamento da aplicação é distribuída para o robô e para os servidores do Google.

A captura da fala do usuário pelo microfone do computador foi satisfatória, sendo capaz de capturar com velocidade, mesmo quando o usuário interage com o tom de voz baixo. Em alguns

momentos a síntese de voz do robô foi capturada pelo microfone do computador como sendo fala do usuário, sendo necessário interromper a captura durante a resposta do robô para evitar erros.

A transcrição de texto implementada pela API Google Speech-To-Text atendeu os requisitos de acurácia e velocidade de processamento. A transcrição considera os fonemas pronunciados e as outras palavras da oração, corrigindo palavras que não se encaixam no contexto. Apesar da alta acurácia da API, os erros foram mais frequentes em casos que o usuário profere apenas uma palavra.

O reconhecimento de intenções implementado pelo Dialog Flow é capaz de identificar a intenção a partir de diferentes falas, tons de voz e sotaques. Por acrescentar grande quantidade de dados ao treinamento da análise pragmática, a detecção acontece mesmo quando o usuário utiliza expressões não explicitamente previstas na programação do bot. A arquitetura utilizada para o planejamento e tomada de decisões é baseada em estados finitos, em que as transições são representadas pelas intenções proferidas pelo usuário e cada estado descreve uma possível resposta do robô.

O desempenho da rede implementada para a comunicação do computador com o robô foi suficiente para o diálogo acontecer naturalmente, com períodos de silêncio de menos de um segundo entre as falas do usuário e do robô. O protocolo UDP dispensa a autenticação da conexão entre o robô e o computador, reduzindo a complexidade de testes durante a implementação.

A plataforma NAO teve alta aceitabilidade pelos usuários, aumentando o engajamento nas interações. A capacidade de processamento do robô foi satisfatória para executar o algoritmo implementado na linguagem Python. Os dispositivos de rede do robô não apresentaram problemas de instabilidade ou baixa velocidade de transmissão.

A função ALProxy disponível na biblioteca NAOqi permitiu o controle dos atuadores e sensores por meio dos módulos disponíveis na API. Os módulos de postura e locomoção da NAOqi apresentaram estabilidade na realização dos movimentos, não ocasionando nenhuma queda do robô durante as interações. O módulo ALAnimatedSpeech sincroniza as falas com movimentações corporais do robô, dando maior naturalidade para interação pelo uso da comunicação não verbal.

## 6.3 Trabalhos Futuros

O comando de voz pode ser utilizado para acessar funcionalidades de diferentes robôs. Visando continuar a pesquisa em interação humano robô são propostas duas vertentes: a expansão de funcionalidades no NAO e o uso em outros robôs.

### 6.3.1 Expansão das intenções reconhecidas

O sistema implementado é capaz de reconhecer as intenções do usuário com alta acurácia, contudo é necessário definir as funcionalidades e respostas do robô antes da interação. Para abranger um espectro de interações ainda maior, é necessário expandir a lista de intenções programadas durante este projeto.

A expansão de diálogos é gerada pela inclusão de perguntas de respostas diretas e pela criação de trilhas para diálogos elaborados. A comunicação automatizada permite que o robô tenha um diálogo personalizado com cada usuário, adaptando os conteúdos de acordo com seu interesse. Nesse sentido, o projeto pode ser ampliado para explorar temáticas educacionais usando a metodologia de ensino adaptativo.

### **6.3.2 Comandos de voz em outros robôs**

O framework ROS é compatível com diversas plataformas robóticas, permitindo que o comando de voz implementado neste projeto seja adaptado para controlar outros robôs. Para receber as intenções identificadas um novo robô deve subscrever ao tópico */intentions*. O comando de voz pode ser utilizado, por exemplo, para comandar a posição de um robô manipulador ou solicitar a movimentação de robôs terrestres.

# REFERÊNCIAS BIBLIOGRÁFICAS

- [1] VILLAÇA, M. V. M.; SILVEIRA, J. L. Uma breve história do controle automático. *Revista Ilha Digital*, v. 4, p. 3–12, 2013.
- [2] SPONG, M. W. et al. *Robot modeling and control*. 1. ed. New York: Wiley, 2006. (1, v. 3). ISBN 9780471649908.
- [3] BARTNECK, C. et al. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, Springer, v. 1, n. 1, p. 71–81, 2009.
- [4] MASON, M. T.; JR, J. K. S. *Robot hands and the mechanics of manipulation*. The MIT Press, Cambridge, MA, 1985.
- [5] DIFTLER, M. A. et al. Robonaut 2-the first humanoid robot in space. *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, p. 2178–2183, 2011.
- [6] JR, A. F. J.; BARROS, F. d. A. Utilização de robôs de conversação como meio de aprendizagem para crianças e pré-adolescentes. *Anais do Workshop de Informática na Escola*, v. 1, n. 1, 2008.
- [7] FEIL-SEIFER, D.; MATARIC, M. J. Defining socially assistive robotics. *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, p. 465–468, 2005.
- [8] SUNG, G. T.; GILL, I. S. Robotic laparoscopic surgery: a comparison of the da vinci and zeus systems. *Urology*, Elsevier, v. 58, n. 6, p. 893–898, 2001.
- [9] BREAZEAL, C. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, Elsevier, v. 59, n. 1-2, p. 119–155, 2003.
- [10] SEYAMA, J.; NAGAYAMA, R. S. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, MIT Press, v. 16, n. 4, p. 337–351, 2007.
- [11] LEVINSON, S. C. Pragmatics. In: *International Encyclopedia of Social and Behavioral Sciences*. University of Nijmegen: Pergamon, 2001. v. 17, cap. 3, p. 11948–11954.
- [12] NETO, J. d. O.; TONIN, S. D.; PRIETCH, S. Processamento da linguagem natural e suas aplicações computacionais. *Escola Regional de Informática (ERIN)*, 2010.



- [13] RICH, E.; KNIGHT, K. Artificial intelligence. *McGraw-Hill, New*, 1991.
- [14] OLIVEIRA, F. A.; NAVAUX, P. O. A. Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa. *Rio Grande do Sul*, 2004.
- [15] TOMASELLO, M. Comunicação linguística e representação simbólica. *M. Tomasello, Origens culturais da aquisição do conhecimento humano*, p. 131–186, 2003.
- [16] NICKERSON, R. S. On conversational interaction with computers. *Proceedings of the ACM/SIGGRAPH workshop on User-oriented design of interactive graphics systems*, p. 101–113, 1976.
- [17] SACKS, H.; SCHEGLOFF, E.; JEFFERSON, G. A simple systematic for the organisation of turn taking in conversation. v. 50, p. 696–735, 12 1974.
- [18] AUSTIN, J. L. *How to do things with words*. Oxford: Oxford university press, 1975.
- [19] SEARLE, J. R. Indirect speech acts. In: *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press, 1979. v. 1, cap. 1, p. 30–57.
- [20] CHOWDHURY, G. G. Natural language processing. *Annual review of information science and technology*, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.
- [21] LEVINE, R. I. *Inteligência artificial e sistemas especialistas*. Nova York: McGraw-Hill, 1988.
- [22] KROVETZ, R.; CROFT, W. B. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 10, n. 2, p. 115–141, 1992.
- [23] HARRIS, R. Introduction to decision making. *Home page: <http://www.vanguard.edu/rharris/crebook5.htm>*. [Visited 14 October 2019], 1998.
- [24] Aldebaran Robotics. Naoqi framework. *Disponível em <http://doc.aldebaran.com/>, acessado em 15/08/2018*.
- [25] DUTOIT, T. *An introduction to text-to-speech synthesis*. Berlim: Springer Science & Business Media, 1997.
- [26] SENTIMENT apprehension in human-robot interaction with NAO, v. 4 de 5, (5, v. 4). The address of the publisher: Shen, Jie and Rudovic, Ognjen and Cheng, Shiyang and Pantic, Maja, 2015. 867–872 p. An optional note.
- [27] MARTIN, J. H.; JURAFSKY, D. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Nova Jersey: Pearson/Prentice Hall, 2009.
- [28] MCTEAR, M. F. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, ACM, v. 34, n. 1, p. 90–169, 2002.

- [29] BRATMAN, M. E.; ISRAEL, D. J.; POLLACK, M. E. Plans and resource-bounded practical reasoning. *Computational intelligence*, Wiley Online Library, v. 4, n. 3, p. 349–355, 1988.
- [30] LARSSON, S.; TRAUM, D. R. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural language engineering*, Cambridge University Press, v. 6, n. 3-4, p. 323–340, 2000.
- [31] BETHEL, C. L.; MURPHY, R. R. Review of human studies methods in hri and recommendations. *International Journal of Social Robotics*, Springer, v. 2, n. 4, p. 347–359, 2010.
- [32] DYBKJAER, L.; BERNSEN, N. O.; MINKER, W. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, Elsevier, v. 43, n. 1-2, p. 33–54, 2004.
- [33] SHAMSUDDIN, S. et al. Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. *Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium*, p. 188–193, 2012.
- [34] RODRIGUEZ, I. et al. Humanizing nao robot teleoperation using ros. *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, p. 179–186, 2014.
- [35] FORERO, L. L.; YÁNEZ, J. M.; SOLAR, J. Ruiz-del. Integration of the ros framework in soccer robotics: the nao case. *Robot Soccer World Cup*, p. 664–671, 2013.
- [36] JUANG, B.-H.; RABINER, L. R. Automatic speech recognition - a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, v. 1, p. 67, 2005.
- [37] Google Cloud Text-To-Speech. *Conversão de texto em voz com tecnologia de machine learning*. 2019. Disponível em: <<https://cloud.google.com/text-to-speech>>.
- [38] Watson Speech to Text. *Converta facilmente áudio e voz em texto escrito para uma compreensão rápida do conteúdo*. 2019. Disponível em: <<https://www.ibm.com/br-pt/cloud/watson-speech-to-text>>.
- [39] Apple. *Siri for Developers*. 2019. Disponível em: <<https://developer.apple.com/siri/>>.
- [40] Natural Language Understanding. *Natural language processing for advanced text analysis*. 2019. Disponível em: <<https://www.ibm.com/watson/services/natural-language-understanding/>>.
- [41] SYSTRAN Platform. *Welcome to SYSTRAN.io*. 2019. Disponível em: <<https://platform.systran.net/index>>.
- [42] Dialog Flow. *Build natural and rich conversational experiences*. 2019. Disponível em: <<https://dialogflow.com/>>.

# ANEXO I

## Transcrição das interações

### I.I Interação 1

Oi, quem é você?

>> Eu sou o NAO! Eu adoro conversar.

O que você sabe fazer?

>> Eu sei fazer malabarismo.

Você consegue andar para frente?

>> Tudo bem, vou andar para frente.

Levantar o braço direito.

>> Certo, vou levantar meu braço direito.

NAO dobre o seu braço esquerdo.

>> Certo, vou dobrar meu braço esquerdo.

O que você sabe ensinar?

>> Eu sei ensinar sobre robótica, sobre geografia e também matemática.

>> O que você quer aprender?

Eu quero aprender sobre robótica.

>> Eu adoro conversar sobre robôs! Você sabe como eu faço para me movimentar?

Não sei.

>> Eu utilizo motores elétricos. Tenho vários motores em meus braços e pernas.

>> Você sabe como eu consigo pensar tão rápido?

Como NAO?

>> Eu tenho um computador em minha cabeça, com ele eu consigo controlar meus motores.

>> Você sabe como eu faço para te ver?

Você usa uma câmera?

>> Isso mesmo, eu tenho uma câmera nos meus olhos. Com ela eu consigo enxergar tudo!

Quantos anos você tem?

>> Eu tenho menos de um ano, sou de última geração.

Olhar (substituição: oléo) para frente

>> Certo, vou olhar para frente.

Levantar o braço direito

>> Certo, vou levantar meu braço direito.

Dobrar o braço direito

>> Certo, vou dobrar meu braço direito.

Você sabe jogar futebol?

>> Eu adoro jogar futebol!

Qual música você pode tocar?

>> Sim, eu sei!

Me conta uma (omissão: uma) piada

>> Claro, o que cai em pé e corre deitado?

A chuva

>> Você já conhecia essa!

Você pode falar mais alto?

>> Tudo bem, vou falar mais alto.

De onde você veio?

>> Eu fui fabricado lá na França, mas agora eu moro na Universidade de Brasília.

Tchau

>> Até mais, adorei conhecer você.

## I.II Interação 2

Olá, tudo bem?

De onde você veio?

>> Eu fui fabricado lá na França, mas agora eu moro na Universidade de Brasília.

Você pode repetir?

>> Eu fui fabricado lá na França, mas agora eu moro na Universidade de Brasília.

Você gosta de futebol?

>> Eu adoro jogar futebol!

O que você (omissão: você) sabe fazer?

>> Eu sei fazer malabarismo.

Você consegue andar para frente?

>> Tudo bem, vou andar para frente.

Levante os braços.

>> Certo, vou levantar meus braços.

Sentar

>> Tudo bem, vou sentar.

Levantar

>> Tudo bem, vou me levantar.

O que você sabe fazer?

>> Eu sei tocar guitarra.

Você sabe lutar?

>> Sim, eu sei.

Você é uma criança?

>> Eu tenho menos de um ano, sou de última geração.

Conte uma piada

>> Claro, o que caí em pé e corre deitado?

Não sei

>> A chuva.

Até mais NAO

>> Até mais, adorei conhecer você.

### I.III Interação 3

Ei NAO

>> Olá, tudo bem com você?

O que você sabe fazer?

>> Eu sei fazer malabarismo.

Andar para direita

>> Tudo bem, vou andar para direita.

Quem fez (omissão: fez) você?

>> Eu fui fabricado lá na França, mas agora eu moro na Universidade de Brasília.

Você conhece uma música?

>> Sim, eu sei.

Você pode repetir?

>> Sim, eu sei.

Dobrar o braço direito.

>> Certo, vou dobrar o braço direito.

Olhar para esquerda.

>> Certo, vou olhar para esquerda.

Você gosta de futebol?

>> Eu adoro jogar futebol!

Quantos anos você (substituição: valsa) tem?

>> Eu tenho menos de um ano, sou de última geração.

Eu quero aprender.

>> Eu sei ensinar sobre robótica, sobre geografia e também matemática.

>> O que você quer aprender?

Quero aprender sobre robótica.

>> Eu adoro conversar sobre robôs! Você sabe como eu faço para me movimentar?

Você usa motores.

>> Isso mesmo, eu utilizo meu motores. Tenho vários motores em meus braços e pernas.

>> Você sabe como eu consigo pensar tão rápido?

Não sei.

>> Eu tenho um computador em minha cabeça, com ele eu consigo controlar meus motores.

>> Você sabe como eu faço para te ver?

Como?

>> Eu tenho uma câmera nos meus olhos. Com ela eu consigo enxergar tudo!

Tchau NAO.

>> Até mais, adorei conhecer você.