
**UNIVERSIDADE ESTADUAL DO CEARÁ**
**CENTRO DE CIÊNCIAS E TECNOLOGIA**
**MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO**

**LEINYLSON FONTINELE PEREIRA**

# DESENVOLVIMENTO E AVALIAÇÃO DE DESEMPENHO DO MECANISMO DE RECONHECIMENTO AUTOMÁTICO DE VOZ DE UM SISTEMA TUTOR INTELIGENTE

**FORTALEZA - CEARÁ**
**2015**

LEINYLSON FONTINELE PEREIRA

# DESENVOLVIMENTO E AVALIAÇÃO DE DESEMPENHO DO MECANISMO DE RECONHECIMENTO AUTOMÁTICO DE VOZ DE UM SISTEMA TUTOR INTELIGENTE

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Centro de Ciências e Tecnologia (CCT) da Universidade Estadual do Ceará (UECE) como requisito parcial para obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação.

Orientador: Prof. Dr. Jorge Luiz de C. e Silva.

FORTALEZA - CEARÁ
2015

Dados Internacionais de Catalogação na Publicação

Universidade Estadual do Ceará

Sistema de Bibliotecas

Pereira, Leinylson Fontinele.
Desenvolvimento e Avaliação de Desempenho do Mecanismo de Reconhecimento Automático de Voz de um Sistema Tutor Inteligente [recurso eletrônico] / Leinylson Fontinele Pereira. - 2015.
1 CD-ROM: il.; 4 ¾ pol.

CD-ROM contendo o arquivo no formato PDF do trabalho acadêmico com 89 folhas, acondicionado em caixa de DVD Slim (19 x 14 cm x 7 mm).

Dissertação (mestrado acadêmico) - Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Acadêmico em Ciência da Computação, Fortaleza, 2015.
Área de concentração: Ciência da Computação.
Orientação: Prof. Dr. Jorge Luiz de Castro e Silva.

1. DTW. 2. FFT. 3. HMM. 4. Interação Humano Computador. 5. Reconhecimento Automático da Fala e Síntese de Voz. I. Título.

LEINYLSON FONTINELE PEREIRA

# DESENVOLVIMENTO E AVALIAÇÃO DE DESEMPENHO DO MECANISMO DE RECONHECIMENTO AUTOMÁTICO DE VOZ DE UM SISTEMA TUTOR INTELIGENTE

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Centro de Ciências e Tecnologia (CCT) da Universidade Estadual do Ceará (UECE) como requisito parcial para obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação.

Aprovada em: 28 de agosto de 2015.

BANCA EXAMINADORA

Prof. Dr. Jorge Luiz de Castro e Silva (Orientador)
Universidade Estadual do Ceará - UECE

Prof. Dr. Thelmo Pontes de Araújo
Universidade Estadual do Ceará - UECE

Prof. Dr. Francisco Heron de Carvalho Júnior (Membro externo)
Universidade Federal do Ceará - UFC

## AGRADECIMENTOS

Inicialmente agradeço a Deus que iluminou o meu caminho durante esta trajetória e à minha família, por todo o apoio dado, para conseguir superar mais esta etapa do meu percurso acadêmico.

Agradeço aos meus pais, Raimundo e Júlia, que me apoiaram em decisões e momentos difíceis, mas que principalmente comemoraram conquistas e felicidades da vida que aconteceram no decorrer desse período, e também a minha esposa Alessandra, que sempre me esperou, quando estava longe, vi que sem você as coisas seriam bem mais difíceis. Obrigado pelo amor e carinho.

Gostaria também de agradecer, aos meus colegas de classe e amigos do LADESC que me apoiaram, quando necessitado elucidar alguma dúvida, em especial ao meu mais recente amigo Athânio, companheiro de discussões de ideias e críticas profissionais e pessoais.

Gostaria ainda de agradecer ao Prof. Dr. Jorge Luiz, pela disponibilidade, confiança e orientação, ao o Prof. Dr. Thelmo Pontes pelos esclarecimentos inicias na área de processamento de sinais durante sua disciplina. Agradeço também ao Prof. Dr. Alberto Adade Filho, que ajudou-me na obtenção do acesso à base de dados cognitiva.

A todos os professores do curso e coordenação do MACC, que deram o máximo de si, tornando-se assim, importantes no desenvolvimento de minha vida acadêmica e pela aquisição da bolsa de auxílio concedida pela CAPES, que foi de muita valia.

Por fim, um muito obrigado a todas as pessoas que por diferentes razões, contribuíram para a realização desta dissertação e demais atividades pertinentes ao curso.

“Engenharia: onde os nobres semi-hábeis
trabalhadores executam a visão
daqueles que imaginam e sonham.
Olá, Ooompa-Loompas da ciência”

(Sheldon Cooper)

## RESUMO

Este trabalho apresenta um Sistema Tutor Inteligente (STI) para acessibilidade de pessoas deficientes, capaz de realizar tanto o reconhecimento automático da fala quanto a síntese de voz. Por meio de um diálogo interativo e através do reconhecimento de palavras isoladas e conectadas (frases), o STI retorna de forma audiovisual o significado e ilustrações associadas à palavra reconhecida, auxiliando desta forma, o usuário na compreensão de novas palavras e conceitos. Esta Interação Humano-Computador (IHC) motiva o usuário no processo de ensino-aprendizagem através do reconhecimento e da geração automática da fala. O trabalho avalia o STI, mais especificamente seu mecanismo de reconhecimento de voz, através de dois fatores fundamentais que regem sua usabilidade: a acurácia (taxa de acerto) e o tempo de processamento. O STI realiza a extração de características representativas dos sinais da voz através da Transformada Rápida de Fourier e da Transformada de Fourier de Tempo Curto. O STI também classifica os sinais de voz por meio do cálculo do erro médio, desvio padrão e covariância, e também através dos métodos *Dynamic Time Warping* (DTW) e *Hidden Markov Model* (HMM). O STI apresentou resultados satisfatórios quanto ao seu mecanismo de reconhecimento de voz, destacando-se a classificação do sinal de voz através da computação do erro médio do sinal.



**Palavras-chave**: DTW. FFT. HMM. Interação Humano-Computador. Reconhecimento Automático da Fala. Síntese de Voz. STFT.

**ABSTRACT**

This work presents an Intelligent Tutor System (ITS) for accessibility of disabled persons, capable of performing both the automatic recognition of speech and speech synthesis. Through an interactive dialogue, and through the recognition of isolated and connected words (phrases), the ITS returns audiovisual forms the meaning and graphics associated with the recognized word, thus assisting the user in understanding the concepts and new words. This Human-Computer Interaction (HCI) motivates the user in the teaching-learning process through recognition and automatic generation of speech. The study evaluates the ITS, specifically its voice recognition engine through two fundamental factors governing its usability: accuracy (hit rate) and the processing time. The ITS performs extraction of characteristics representative of speech signals through Fast Fourier Transform and Short Time Fourier Transform. The ITS also ranks the voice signals via the calculated average error, standard deviation and covariance, and also by the methods Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). The ITS showed satisfactory results regarding its voice recognition engine, especially the classification of the speech signal by the average signal error computing.

**Keywords:** Automatic Speech Recognition. DTW. FFT. HMM. Human-Computer Interaction. Speech Synthesis. STFT.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ACRÔNIMOS

| | |
|---|---|
| **AS** | Ator Sintético |
| **CAI** | *Computer Assisted Instruction* (Instrução Auxiliada por Computador) |
| **DFT** | *Discret Fourier Transform* (Transformada Discreta de Fourier) |
| **DTW** | *Dynamic Time Warping* (Alinhamento Dinâmico no Tempo) |
| **FFT** | *Fast Fourier Transform* (Transformada Rápida de Fourier) |
| **HMM** | *Hidden Markov Model* (Modelo Oculto de Markov) |
| **Hz** | Hertz |
| **IA** | Inteligência Artificial |
| **ICAI** | *Intelligent Computer Assisted Instruction* (Instrução Auxiliada por Computador Inteligente) |
| **IHC** | Interação Humano-Computador |
| **IHCM** | Interação Humano-Computador Multimodal |
| **MATLAB** | *MATrix LABoratory* |
| **MFCC** | *Mel-Frequency Cepstral Coefficient* |
| **PCM** | *Pulse-Code Modulation* |
| **PDS** | Processamento Digital de Sinais |
| **RI** | Recuperação de Informação |
| **SAPI** | *Speech Application Programming Interface* |
| **STFT** | *Short-Time Fourier Transform* (Transformada de Fourier de Tempo Curto) |
| **STI** | Sistema Tutor Inteligente |
| **TTS** | *Text-to-Speech* (Texto-para-Fala) |
| **UECE** | Universidade Estadual do Ceará |
| **VAD** | *Voice Activity Detection* (Detecção de Atividade de Voz) |

**SUMÁRIO**

# 1 INTRODUÇÃO

Neste capítulo são apresentados os principais fatores que levaram ao desenvolvimento deste trabalho, assim como os objetivos almejados. A estrutura do restante do trabalho é apresentada ao final.

## 1.1 MOTIVAÇÃO

A linguagem é o meio de comunicação mais importante para o homem e o ato de falar é o modo mais natural de comunicação entre as pessoas. A fala provê uma forma de diálogo entre pessoas que contribui para o entendimento da produção, percepção, processamento, aprendizagem e uso da linguagem. A interação oral entre interlocutores humanos e entre humanos e máquinas está incorporada nas condições da comunicação, que compreendem a codificação e decodificação do significado, bem como a mera transmissão de mensagens através de um canal acústico.

O interesse pela interação humano-computador através da fala tem aumentado consideravelmente, dando origem a uma demanda muito grande por sistemas capazes de reconhecer o que foi pronunciado, bem como produzir artificialmente a fala a partir do texto. Essa demanda gerou a necessidade do desenvolvimento de interfaces humano-computador mais amigáveis e simples de usar a partir da comunicação oral, permitindo assim o uso de computadores e outros aparelhos eletrônicos por um número maior de pessoas.

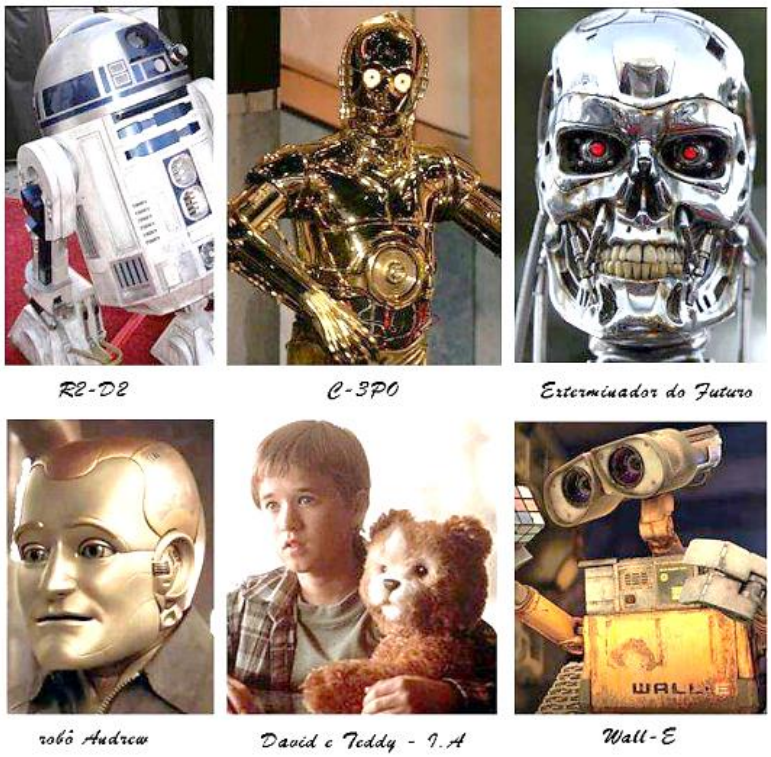
O reconhecimento de voz ou de fala refere-se à habilidade que uma máquina ou programa possui para interpretar o que foi pronunciado, ou ainda, compreender e executar comandos falados. Existem algumas dificuldades que devem ser superadas para tornar os sistemas de reconhecimento de voz aptos à compreensão de um discurso em qualquer contexto, falado naturalmente por qualquer indivíduo, em qualquer ambiente e em qualquer dialeto (SILVA, 2009). Em geral, deve ser possível em tais sistemas o funcionamento em condições de ruído de fundo e adaptação a vários tipos de locutores.

Existem atualmente, várias empresas que comercializam sistemas de reconhecimento de voz, contudo, nenhum desses sistemas possuem capacidade de entender corretamente 100% das palavras pronunciadas.

Alguns sistemas se destacam pelo alto padrão de reconhecimento como o Dragon® (BAKER, 1975) , outros pela popularidade alcançada nos últimos anos, como o *Google Now*[1], a *Siri*[2] da *Apple* e a *Cortana*[3] da *Microsoft*.

O reconhecimento automático da voz por máquinas tem inspirado grandes produções da ficção científica, tal como o robô R2D2 (Figura 1) de George Lucas no clássico filme 'Guerra nas Estrelas' (BRESOLIN, 2003).

**Figura 1 - Famosas Máquinas Falantes do Cinema**




Fonte: Santos (2013).

Um dos campos mais promissores na pesquisa de Inteligência Artificial (IA), é a modelagem de agentes inteligentes credíveis, aumentando o realismo através do comportamento, cognição (raciocínio), interação, percepção e ação. Esses agentes, conhecidos como Atores Sintéticos (AS), afetam emocionalmente o usuário, aumentando sua motivação e engajamento por meio de respostas rápidas e consistentes (LINO, TEDESCO e ROUSY, 2006), além de uma fácil compreensão.

[1] Trata-se de um assistente pessoal inteligente com uma interface de linguagem natural que responde perguntas e faz recomendações, porém não dispõe de personalidade.
[2] É um aplicativo no estilo assistente pessoal para *iOS*, recentemente disponibilizado em Português.
[3] Assistente pessoal que realiza chamadas, previsão do tempo, alarmes, dentre outros.

Um dos aspectos mais difíceis no desenvolvimento das pesquisas na área de reconhecimento de voz pelo computador é a sua interdisciplinaridade[4] natural, pois seu desenvolvimento requer o conhecimento e perícia de um largo espectro de disciplinas (RABINER e JUANG, 1993). Algumas áreas do conhecimento interagem mais com reconhecimento de voz, destacando-se (BRESOLIN, 2003):

(i) **Processamento de Sinais -** corresponde ao processo de extração de informações relevantes do sinal da fala, através da aquisição de dados, análise espectral[5] e vários tipos de pré-processamento e pós-processamento;

(ii) **Padrões de Reconhecimento -** formam um conjunto de métodos e algoritmos usados para agrupar dados e criar um ou mais padrões de um conjunto de dados, podendo ser posteriormente comparados com um dado sinal para efeito de reconhecimento;

(iii) **Análise de Padrões -** constitui o conjunto de procedimentos para estimação de parâmetros de modelos estatísticos;

(iv) **Linguística -** define as relações entre os sons (fonologia), palavras de uma linguagem (sintaxe), significado das palavras faladas (semântica) e o sentido derivado do significado (pragmático);

(v) **Fisiologia -** compreende a concepção dos mecanismos do sistema nervoso central do ser humano incluindo a produção e percepção da fala;

(vi) **Ciência da Computação -** estuda algoritmos eficientes para implementação de métodos usados em um sistema de reconhecimento de voz.

Durante o desenvolvimento da dissertação, implementou-se um Sistema Tutor Inteligente (STI) que facilitou a execução dos testes e a análise de desempenho das técnicas empregadas na solução dos problemas presentes no reconhecimento da fala. Sua finalidade vai desde o pré-processamento do sinal até o reconhecimento da palavra pronunciada, para uma posterior execução da tarefa correspondente.

---

[4] Entende-se por interdisciplinaridade, como uma colaboração entre diversas disciplinas, que leva a interações, isto é, uma certa reciprocidade, de forma que haja, em suma, enriquecimento mútuo.

[5] Quando um som é decomposto em seus componentes simples, estamos realizando uma análise espectral.

Uma das motivações para o estudo e desenvolvimento do STI é proporcionar, através de múltiplas percepções, acessibilidade a informação (imagens, sons, textos) para pessoas deficientes ou com limitações.

## 1.2 OBJETIVO GERAL

O objetivo geral da dissertação é construir um Sistema Tutor Inteligente que realize o reconhecimento automático da fala e síntese de voz e avaliar o desempenho das técnicas utilizadas conforme os níveis de acurácia e tempo necessário para o treinamento e classificação dos sinais.

### 1.2.1 Objetivos Específicos

Os objetivos específicos desta dissertação são:

I. Implementar um STI que faça o reconhecimento e síntese de voz.

II. Construir uma base de locuções das palavras a serem utilizadas nos testes.

III. Demonstrar a usabilidade do STI através de casos de uso.

IV. Medir e analisar a acurácia dos classificadores implementados.

V. Medir e comparar o tempo necessário para treinar e classificar os sinais.

## 1.3 ORGANIZAÇÃO DO TRABALHO

O conteúdo deste trabalho está organizado em 6 capítulos, incluindo a presente introdução. Cada capítulo é sucintamente apresentado a seguir:

**Capítulo 2 - Fundamentação Teórica:** Apresenta o aparato necessário para a compreensão do tema abordado nesta pesquisa, discorrendo sobre os conceitos de STI. Os tipos de sistemas de reconhecimento de voz que envolvem o Processamento Digital de Sinais (PDS) da fala, assim como o processo de síntese de voz também são apresentados. A transformada de Fourier como técnica de extração dos parâmetros representativos do sinal de voz e a classificação dos sinais por meio do cálculo do erro médio, desvio padrão, covariância e das técnicas de DTW e HMM, também tem suas características exibidas.

**Capítulo 3 - Trabalhos Relacionados:** Apresenta trabalhos relacionados aos STI's, na qual são expostos alguns sistemas das mais variadas áreas que utilizam de alguma forma o reconhecimento da fala ou a síntese de voz como mecanismo de automatização de tarefas a partir de várias abordagens, destacando suas principais características. Também é feito um levantamento de trabalhos voltados especificamente ao processamento de sinais da fala, utilizando diferentes métodos.

**Capítulo 4 - Metodologia e Cenário do STI:** Descreve em detalhes a composição da arquitetura do modelo de STI desenvolvido, utilizado para o problema de reconhecimento de palavras apresentado neste trabalho. Apresenta também o fluxo de funcionamento do STI, bem como a confecção da base de dados e do ator sintético.

**Capítulo 5 - Análise dos Resultados:** Relata o estudo analítico realizado para validação e avaliação da pesquisa, apresentando casos de usos do funcionamento da arquitetura visando demonstrar a usabilidade e facilitar seu entendimento. Em seguida, é realizada uma análise de desempenho do sistema desenvolvido como subsídio para esta pesquisa, a fim de verificar estatisticamente os resultados por meio de testes, corroborando para uma análise comparativa dos métodos utilizados no treinamento e classificação dos sinais. Os respectivos resultados são empregados para se chegar a conclusões acerca da viabilidade dos classificadores implementados.

**Capítulo 6 - Considerações Finais:** Relaciona as principais contribuições e conclusões deste trabalho, bem como suas limitações e os próximos passos a serem alcançados em trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

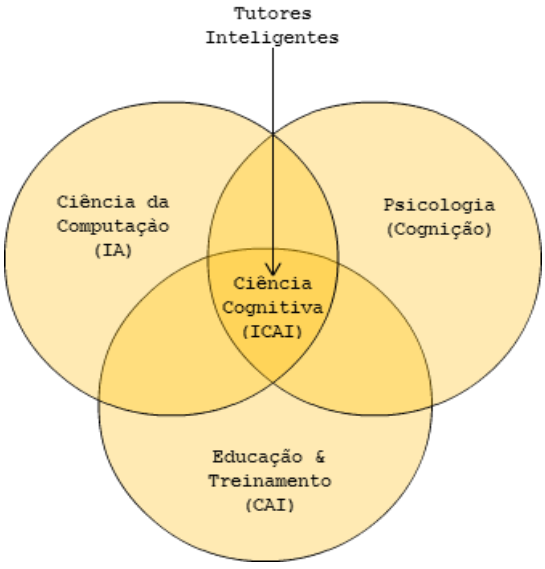
Este capítulo fornece um referencial teórico, facilitando o entendimento do STI e do estudo realizado para validação e avaliação do mecanismo de reconhecimento de voz. Expondo inicialmente uma visão geral de STI's, também são apresentados os tipos de sistemas de reconhecimento, dando uma ênfase maior ao problema de reconhecimento de palavras, por ser esse o objeto de estudo, promovendo uma melhor contextualização da linha de pesquisa de processamento de sinais da fala.

## 2.1 SISTEMAS TUTORES INTELIGENTES

De acordo com Fischetti e Gisolfi (1990), um STI seria como um livro ativo, no qual o leitor possui uma maior interação na transmissão do conhecimento em nível mais apropriado de entendimento. Nesse contexto, julgam compreensível a utilidade que os STI's podem vir a propiciar nos diversos campos do conhecimento humano.

Os STI's derivam dos programas de Instrução Auxiliada por Computador (*Computer Assisted Instruction* - CAI) e Instrução Auxiliada por Computador Inteligente (*Intelligent Computer Assisted Instruction* - ICAI), como ilustrado na Figura 2. Costa (2002) alega que a IA tem buscado formas plausíveis de simular, no computador, tarefas tomadas como inteligentes. Terman (1916) considera que a inteligência é a aptidão para construir conceitos e compreender o seu significado.

**Figura 2 - Domínio da Aplicação de STI's**

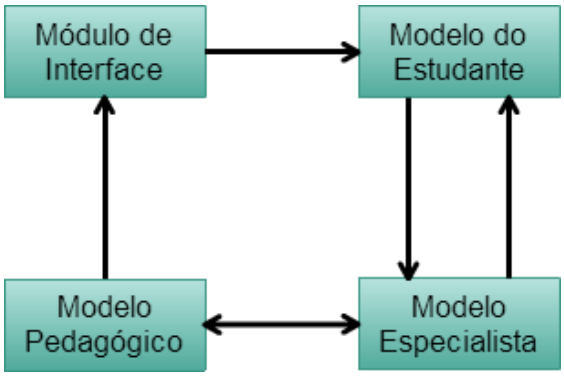



Fonte: modificada de Kearsley (1987).

Várias são as definições de STI's, dentre elas, Gamboa e Ana (2001) alegam que os STI's são softwares que dão suporte às atividades da aprendizagem. Uma outra definição, apresentada em Wenger (1987), aponta que STI's são sistemas instrucionais baseados em computador com modelos de conteúdo instrucional que especificam o 'quê' ensinar e estratégias de ensino que especificam 'como' ensinar.

Um STI é constituído de quatro módulos principais que relacionam-se entre si conforme a arquitetura clássica apresentada na Figura 3.

**Figura 3 - Arquitetura do Modelo Clássico**




Fonte: adaptado de Kaplan e Rock (1995).

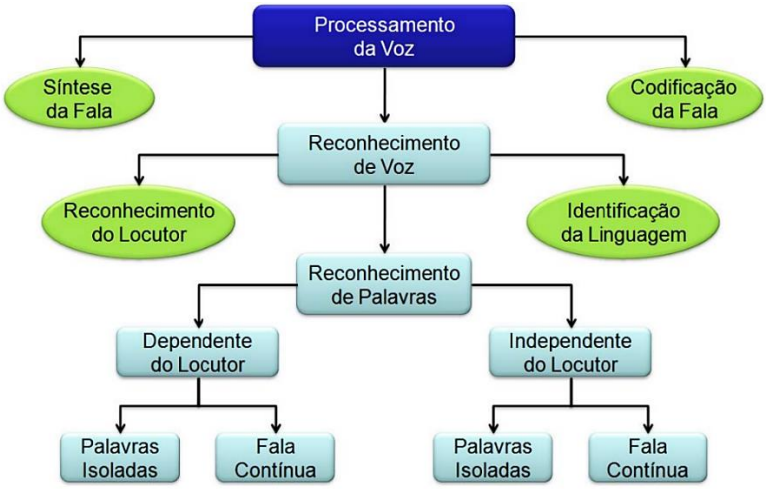
Observando a arquitetura do modelo clássico apresentada nessa figura acima, e de acordo com Costa (2002), tem-se:

- O **Modelo Pedagógico** (regras de ensino) que analisa a informação do aprendiz, decidindo quais estratégias serão empregadas e como a informação será exibida;
- Um **Modelo Especialista** (rede de conhecimento) que expõe o conhecimento de um especialista na área de domínio do sistema;
- O **Modelo do Estudante** que representa o conhecimento do aprendiz;
- Um **Módulo de Interface**, que realiza a interação de informações entre o sistema, o instrutor e o aprendiz, traduzindo toda a representação interna do sistema de modo amigável e de simples compreensão para o usuário.

## 2.2 O RECONHECIMENTO DA FALA

Os avanços tecnológicos na área de processamento da voz têm possibilitado o desenvolvimento de numerosas aplicações, compreendidas em três classificações de sistemas de comunicação humano-computador pela voz (ADAMI, 1997): Sistemas de Resposta pela Voz, Sistemas de Processamento da Fala e Sistemas de Reconhecimento. Na Figura 4 tem-se a hierarquia dos sistemas de voz.

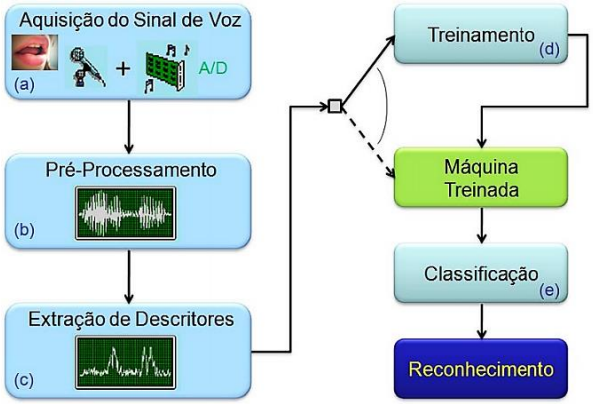
**Figura 4 - Hierarquia dos Sistemas de Processamento de Voz**




Fonte: Bresolin (2008).

O processo de reconhecimento de uma palavra, possui basicamente, duas fases: uma **Fase de Treinamento** e uma **Fase de Reconhecimento**, como ilustrado na Figura 5.

**Figura 5 - Sistema de Reconhecimento de Voz**




Fonte: Bresolin (2008).

Durante o treinamento do classificador, treina-se uma "máquina inteligente" para que aprenda a reconhecer os descritores do sinal e consequentemente o padrão de voz. Na etapa de classificação, utiliza-se a "máquina inteligente" (classificador previamente treinado) para fazer o reconhecimento através de uma decisão lógica (BRESOLIN, 2008).

Conforme Furui (1989), Nejat (1992) e Bourouba *et al*. (2006), as classificações dos sistemas de voz podem ser relacionadas da seguinte forma:

**Considerando o tipo de pronúncia**:

- **Reconhecedor de Palavras Isoladas**: cada palavra é falada de forma isolada.
- **Reconhecedor de Palavras Conectadas**: o padrão a ser reconhecido é uma sequência de palavras pertencentes a um vocabulário restrito e pronunciadas de forma contínua. Porém, essas palavras devem ser bem pronunciadas, utilizando palavras como unidades fonéticas padrões para cada palavra.
- **Reconhecedor de Fala Contínua**: capaz de reconhecer a fala na comunicação natural, podendo implicar na necessidade de segmentação do sinal de fala. Esses reconhecedores são bastante complexos, pois devem lidar com todas as características e vícios da linguagem natural, como o sotaque, a duração das palavras etc.

**De acordo com o grau de dependência do locutor**:

- **Dependente de locutor**: reconhece a fala das pessoas cujas vozes foram utilizadas para treinar o sistema.
- **Independente de locutor**: reconhece a fala de qualquer pessoa com uma taxa de acerto aceitável, sendo necessário realizar o treino do sistema com uma base que inclua diferentes pessoas com diferentes idades, sexo, sotaques, etc.

**Quanto ao tamanho do vocabulário**:

- **Vocabulário pequeno**: reconhecem até 20 palavras.
- **Vocabulário médio**: reconhecem entre 20 e 100 palavras.
- **Vocabulário grande**: reconhecem entre 100 e 1000 palavras.
- **Vocabulário muito grande**: reconhecem mais de 1000 palavras.

**Quanto ao tipo de vocabulário**:

- **Dependentes de Texto**: são usadas as mesmas locuções tanto para o treino quanto para o teste, ou ainda pode corresponder a um sistema que usa um conjunto de modelos, baseados em palavras ou nas subunidades para um vocabulário restrito.
- **Independentes de Texto:** são necessárias grandes quantidades de dados para a fase de treino, além de serem usadas locuções de tamanho maior na fase de teste.

## 2.3 A SÍNTESE DE VOZ

A síntese é um processo que produz artificialmente a fala, diminuindo a dependência do uso de arquivos de voz e consequentemente o espaço necessário para seu armazenamento, possibilitando ao computador transmitir instruções ou informações através da fala. Segundo Schroeder (1993), o objetivo principal da síntese de voz é reproduzir a fala humana a partir de uma entrada de texto em linguagem natural.

A análise do texto envolve aspectos relacionados à conversão do texto em listas manipuláveis de palavras, o processo é bastante complexo, envolvendo separação de frases e palavras, expansão de abreviaturas, conversão de símbolos, caracteres especiais, siglas e acrônimos, leitura de números, medidas e pontuação, considerando-se ainda a análise morfológica das palavras, flexões e suas derivações.

Um sistema de conversão texto-para-fala (*Text-to-Speech* - TTS) é composto por dois módulos: o primeiro corresponde ao processamento linguístico-prosódico, isto é, a análise textual e linguística; o segundo refere-se ao processamento acústico, ou seja, a geração da fala e prosódia.

Conforme Santos (2013), a última etapa para a saída da fala é a sintetização da forma de onda, podendo ser de três tipos principais:

- **Sintetizadores articulatórios**: são modelos físicos baseados em descrições detalhadas da anatomia e acústica do aparelho fonador humano, modelando mecanicamente os órgãos articuladores.

- **Sintetizadores de formantes**: *synthesis-by-rule*, isto é, sínteses por regras baseando-se no processo de reconstrução de uma onda por meio da manipulação de parâmetros. Consiste numa decomposição de filtros que modelam as ressonâncias e antirressonâncias das cavidades vocal e nasal.

- **Sintetizadores concatenativos**: segmentos fonéticos pré-gravados são concatenados e é efetuado um processamento de sinal para “suavizar” a transição entre as unidades fonéticas.

## 2.4 PROCESSAMENTO DO SINAL DE VOZ

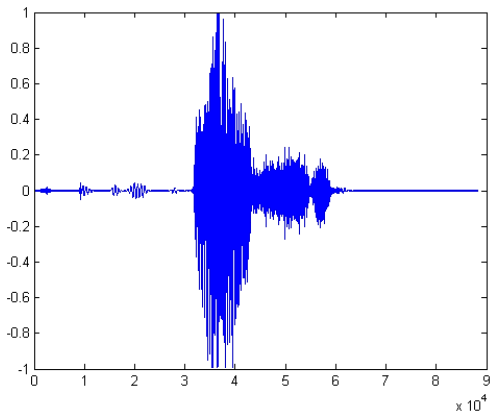
As etapas do processamento do sinal de voz são apresentadas a seguir, nas quais são mostrados os passos a serem adotados durante todo o processo.

### 2.4.1 Aquisição da fala

A primeira etapa consiste em realizar a aquisição do sinal de voz através da conversão das ondas sonoras em sinais elétricos a partir de um transdutor[6], filtragem do sinal e conversão analógico-digital (SILVA, 2009). Uma vez obtido, realiza-se a Detecção de Atividade de Voz (*Voice Activity Detection* - VAD), no qual o sinal de voz é identificado baseado em um *threshold* (limiar) previamente estabelecido, sendo de suma importância ter ciência do momento em que uma pessoa está falando ou não, seja para economizar processamento, para um emprego eficiente da memória disponível, desabilitando um processo enquanto não houver discurso, bem como, abstrair regiões sonoras e mudas da locução. A seguir, são apresentadas duas características importantes de um sistema de aquisição de dados (BRESOLIN, 2003):

- **O Sinal**: adquirido pelo sistema, condicionado e convertido em dados que um computador pode ler (Figura 6) para se extrair informações significantes.

**Figura 6 - Sinal de Voz do Acrônimo "UECE"**




Fonte: Elaborada pelo autor.

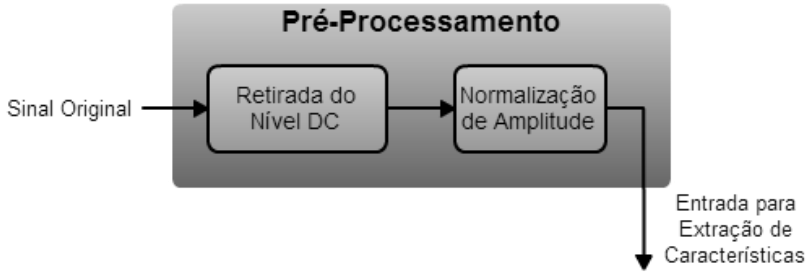
- **Os Dados**: uma vez processado, o sinal transforma-se em dados no computador que são convertidos em sinal analógico e enviados para um atuador.

[6] Dispositivo que converte a energia de entrada em outra forma de energia de saída. O microfone é um sensor que converte energia do som (na forma de pressão) em energia elétrica.

### 2.4.2 Pré-processamento do Sinal

O reconhecimento normalmente é dificultado por características que refletem o ambiente de gravação e o canal de comunicação, como ruídos de alta frequência e distância do microfone. Assim, o sinal deve passar por um pré-processamento a fim de deixá-lo mais próximo da fala pura. A metodologia utilizada no pré-processamento pode ser visualizada no diagrama de blocos da Figura 7.

**Figura 7 - Diagrama de Blocos da Fase de Pré-processamento**




Fonte: Elaborada pelo autor.

O tratamento de ruídos pode ser decisivo no resultado final, pois um áudio sem ruídos e sem perdas das características originais aumenta a robustez do sistema. Os filtros antirruídos visam a minimizar os ruídos aditivos, que são aqueles sons que surgem no ambiente e se somam ao sinal de voz do locutor (televisão, rádio, automóveis, vento, pessoas conversando) e ruídos convolucionais, que caracterizam-se como distorções do canal. Como o ruído varia de um aparelho ou de um ambiente para outro, se ele não for removido, a fase de extração não só extrairá dados da palavra pronunciada, como também dados do ruído, confundindo a fase de classificação e interferindo no resultado final (SILVA, 2010).

Os sinais de voz apresentam, muitas vezes, uma componente contínua que atrapalha a comparação em valores absolutos. É necessária, então, a retirada desse desvio de compensação, a fim de deixar todas as amostras oscilando em torno do valor zero. Portanto, calcula-se a média aritmética das amplitudes[7] do sinal, subtraindo-a de cada amplitude. Após a aplicação dos filtros supressores de ruídos ao sinal de voz, realiza-se a normalização da amplitude, fazendo com que todos os valores de amplitudes de todos os sinais estejam na mesma faixa (por exemplo, entre $-1$ e 1), como ilustrado na Figura 6. Para isto, dividiu-se o valor de cada amostra do sinal pelo maior valor de amplitude, garantindo assim que todos os sinais sejam processados igualmente com relação ao volume da voz, ou seja, sons mais baixos e mais altos serão processados igualmente (SILVA, 2009).

[7] Uma quantidade sem unidades, que pode ser positiva ou negativa (WEEKS, 2012).

### 2.4.3 Extração de Características do Sinal

Durante a fase de extração de características são gerados os dados que serão utilizados para modelagem da palavra ou para efetuar a comparação com algum modelo armazenado. Quando os dados gerados são utilizados para modelar uma palavra, então o sistema está em fase de treinamento, na qual os dados capturados são enviados pelo extrator de características que gera ou treina um modelo matemático para representar aquele conjunto (SILVA, 2010).

O processo de extração consiste em obter parâmetros distintivos que possam ser utilizados na classificação (ADAMI, 1997), que modelem o formato do trato vocal humano. Deve-se, portanto, gerar um modelo com base nas características extraídas, tipicamente obtidas a partir de técnicas de análise espectral, como a transformada rápida de Fourier (*Fast Fourier Transform* - FFT), descrita a seguir.

#### 2.4.3.1 A Transformada de Fourier

O pesquisador Jean-Baptiste Joseph Fourier, foi responsável pela investigação sobre a decomposição de funções periódicas em séries trigonométricas convergentes, chamadas de Séries de Fourier. A análise de Fourier constitui a base do processamento de sinais (BRESOLIN, 2008).

A Transformada de Fourier para funções contínuas representa qualquer função integrável $x(t)$ como a soma de exponenciais complexas conforme a Equação 1:

$$F(\omega) = \int_{-\infty}^{+\infty} x(t).e^{-i\omega t}dt, \tag{1}$$

onde $x(t)$ é um sinal periódico, $\omega$ é a frequência angular, $i = \sqrt{-1}$ e $e^{-i\omega t} = \cos(\omega) - isen(\omega)$.

A Transformada permite analisar características não percebidas diretamente no domínio original do sinal. Por meio da decomposição espectral do sinal é possível obter as frequências presentes no mesmo.

### 2.4.3.2 A Transformada Discreta de Fourier

Quando deseja-se utilizar a Transformada de Fourier em computadores é preciso discretizar o sinal $x(t)$ em um sinal $x_k$ (SMITH, 2007). A Transformada Discreta de Fourier (*Discret Fourier Transform* (DFT)), é muito usada no estudo do espectro[8] de sinais. Um problema da transformada é a sua complexidade computacional, sendo necessário $O(n^2)$ operações. Para resolver este problema, foi desenvolvido o algoritmo FFT que reduz a complexidade para $O(n\ log\ n)$. Uma definição da transformada é apresentada na Equação 2 (COOLEY e TUKEY, 1965):

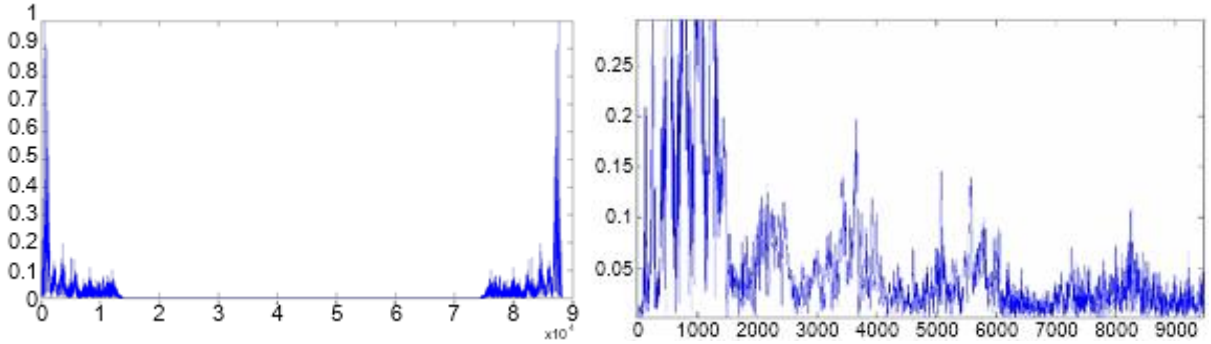
$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{\frac{i2\pi kn}{N}}, \qquad n = 0, 1, \ldots, N-1 \tag{2}$$

onde $N$ corresponde ao número de amostras, $n$ é o tamanho da amostra considerada, $x_n$ representa o valor do sinal, $k$ é a frequência atual que está sendo analisada (0 Hertz até $N-1$ Hertz) e $X_k$ é a quantidade de frequência $k$ no sinal. A razão $n/N$ indica o percentual do tempo que já passou. O produto $2\pi k$ corresponde à velocidade em radianos/segundo e o movimento de conversão no caminho circular é indicado por $e^{-i}$.

### 2.4.3.3 A Transformada Rápida de Fourier

A FFT foi utilizada como mecanismo de extração das características por tratar-se de um método eficiente de reagrupar os cálculos dos coeficientes de uma DFT com menor esforço computacional. A Figura 8 mostra a faixa normalizada de frequência completa na parte esquerda e uma visão ampliada na parte direita.

**Figura 8 - Resposta de Magnitude em Frequência do Acrônimo "UECE"**




Fonte: Elaborada pelo autor.

[8] Nome atribuído ao traçado de frequência de um sinal (WEEKS, 2012).

A FFT permite, a partir de um sinal no domínio do tempo, obter o sinal correspondente no domínio da frequência, usando as funções de seno e cosseno. A fala é um sinal real, mas sua FFT tem componentes reais e imaginários, porém apenas os valores absolutos são utilizados para este fim (SMITH, 1997).

2.4.3.4 O Algoritmo da FFT

Escrevendo $F(u)$ na forma:

$$\mathrm{F}(u) = \frac{1}{\mathrm{N}} \sum_{x=0}^{N-1} f(x) W_N^{ux} \ , \tag{3}$$

onde $W_N = \exp\left[\frac{-i2\pi}{N}\right]$ e $N$ é uma potência de 2, ou seja, $N = 2^n$.

Assim, $N$ pode ser escrito também como $N = 2M$. Substituindo na expressão original teremos:

$$\mathrm{F}(u) = \frac{1}{2\mathrm{M}} \sum_{x=0}^{2M-1} f(x) W_{2M}^{ux} = \frac{1}{2}\left[\frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x) W_{2M}^{u(2x)} + \frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x+1) W_{2M}^{u(2x+1)}\right]. \tag{4}$$

Da definição de $W$ é fácil ver que $W_{2M}^{2ux} = W_M^{ux}$. Assim a expressão acima pode ser reescrita como:

$$\mathrm{F}(u) = \frac{1}{2}\left[\frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x) W_M^{ux} \ + \ \frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x+1) W_M^{ux} W_{2M}^{u}\right] , \tag{5}$$

para

$$\mathrm{F}_{par}(u) = \frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x) W_M^{ux} \qquad \text{e} \qquad \mathrm{F}_{\text{ímpar}}(u) = \frac{1}{\mathrm{M}} \sum_{x=0}^{M-1} f(2x+1) W_{2M}^{ux} \, . \tag{6,7}$$

Assim,

$$\mathrm{F}(u) = \frac{1}{2}\left[\mathrm{F}_{par}(u) + \mathrm{F}_{\text{ímpar}}(u) W_{2M}^{u}\right] , \tag{8}$$

mas, como $W_M^{u+M} = W_M^u$ e $W_{2M}^{u+M} = -W_{2M}^u$, podemos escrever:

$$\mathrm{F}(u + \mathrm{M}) = \frac{1}{2}\left[\mathrm{F}_{par}(u) + \mathrm{F}_{\text{ímpar}}(u) W_{2M}^{u}\right] \tag{9}$$

Para computar uma transformada de $N$ pontos, pode-se decompor o cálculo em duas metades, como visto em (8) e (9). A primeira metade (8) demanda duas transformadas com $N/2$ pontos cada (6 e 7). Os valores são substituídos em (8) para obter a primeira metade de $\mathrm{F}(u)$ para $u = 0, 1, \ldots, (N/2 - 1)$. A segunda metade é obtida de (9) sem nova transformada.

2.4.3.5 A Transformada de Fourier de Tempo Curto

As funções senos e cossenos têm um suporte infinito e são bem adaptadas para analisar sinais estacionários[9]. Porém, não são apropriados para descrever sinais não-estacionários (transientes). Nenhuma informação de frequência está disponível no domínio do tempo do sinal e nenhuma informação de tempo está disponível no sinal transformado (domínio da frequência) (SANCHES, 2001).

Dennis Gabor adaptou a Transformada de Fourier com a utilização de janelamento do sinal. Conhecida como *Short-Time Fourier Transform* (STFT), a adaptação de Gabor põe o sinal em função de duas dimensões, tempo e frequência (RIOUL e VETTERLI, 1991). A STFT foi utilizada na extração de características do classificador DTW e pode ser obtida pela Equação 10, em que tem-se a Transformada de Fourier de um sinal $x(t)$, previamente limitada por uma função janela $g(t - \tau)$ centrado em $\tau$, definida por:

$$STFT(\tau, f) = \int_{-\infty}^{\infty} [x(t) \,.\, g(t - \tau)] \,.\, e^{-i2\pi f t} dt, \qquad (10)$$

A STFT é uma solução para obter uma melhor localização no tempo e frequência na decomposição de um sinal (COHEN *et al.*, 1995). A STFT é uma versão da transformada de Fourier que utiliza janelas no tempo e seus respectivos deslocamentos, como bases para a transformada. Em análise de sinais, existem várias escolhas possíveis para a função janela $g(t)$, destacando-se aquelas que possuem suporte compacto e regularidade razoável (SANCHES, 2001).

[9] Sinais cuja a resposta em frequência não varia no tempo.

2.4.3.6 O Algoritmo da STFT

No processo de obtenção do espectrograma[10] do sinal, inicialmente, extraem-se as características do sinal usando a STFT, que fornece um espectrograma do sinal especificado em um vetor para uma matriz. Por padrão, o vetor é dividido em 8 segmentos com 50% de *overlap* (sobreposição), em que cada segmento está com uma janela de *Hamming* (RABINER e ALLEN, 1977) apresentada na Equação 11.

$$w_i = \left(0.54 - 0.46 * \cos\left(\frac{2\pi i}{N}\right)\right), \tag{11}$$

onde $N$ indica o número total de pontos no sinal e o janelamento é utilizado para suavizar as extremidades das ondas, fazendo com que tornem-se mais próximas de zero. O número de pontos de frequência utilizadas para calcular a DFT é igual ao valor máximo de 256 ou a próxima potência de 2 maior que o tamanho de cada segmento do vetor. Caso o vetor não possa ser dividido exatamente em 8 segmentos, então ele será truncado conforme seu comprimento.

### 2.4.4 A Classificação dos Padrões

Na classificação de padrões existe um módulo de decisão que elege uma palavra a partir do reconhecimento de padrões, medindo a similaridade entre as características do modelo gerado a partir da entrada e algum modelo armazenado. Campbell (1997) proporciona um estudo sobre métodos de extração de características em um sinal.

Para os classificadores que utilizaram o valor médio, desvio padrão e covariância dos sinais, a extração das características foi realizada através da aplicação da FFT ao sinal de voz normalizado, e a classificação foi obtida pela diferença entre os sinais, cada um segundo o tipo de classificador utilizado, no qual aquele que obtiver o menor erro dentre todos, equivale à palavra pronunciada. Na classificação por DTW foi utilizado a STFT, que, por meio da geração do espectrograma, torna possível a obtenção do caminho de menor custo na matriz de distâncias locais.

[10] Espectrogramas são representações bidimensionais (tempo e frequência) de um sinal unidimensional.

2.4.4.1 O Erro Médio

Como primeiro método de classificação, foi utilizado o erro médio ( $\mu$ ) do sinal. Uma vez que o sinal tenha sido processado, obtém-se as diferenças entre o sinal de entrada e cada um dos sinais armazenados na base de dados. Em seguida, procede-se o cálculo do valor médio do sinal, no qual foram somados todos os valores dos sinais ($X_i$) e o resultado foi dividido pelo tamanho total da população ($N$), conforme a Equação 12 a seguir:

$$\mu = \sum_{i=1}^{N} X_i \Big/ N \, . \tag{12}$$

2.4.4.2 O Desvio Padrão

Como segundo método de classificação, foi utilizado o desvio padrão do sinal, representando a dispersão da população. Uma vez que o sinal tenha sido processado, obtém-se as diferenças entre o sinal de entrada e cada um dos sinais armazenados na base de dados. Em seguida, procede-se o cálculo do desvio padrão do sinal, podendo ser considerado como uma medida de variabilidade dos dados de uma distribuição de frequências, isto é, a dispersão dos valores individuais em torno da média, conforme a seguinte equação:

$$\sigma = \sqrt{\left[\frac{(\sum_{i=1}^{N}(X_i - \mu)^2)}{N-1}\right]} . \tag{13}$$

2.4.4.3 A Covariância

Como terceiro método de classificação, foi utilizada a covariância do sinal. Uma vez que o sinal tenha sido processado, obtém-se as diferenças entre o sinal de entrada e cada um dos sinais armazenados na base de dados. Em seguida, procede-se o cálculo da covariância do sinal, no qual, para duas séries de dados, $X(X_1, X_2, \ldots)$ e $Y(Y_1, Y_2, \ldots)$, a covariância fornece uma medida não padronizada do grau no qual as séries movem-se juntas. A relação entre as séries é dada pelo sinal da covariância, caso positivo, as séries movem-se juntas, seguindo uma mesma direção e caso seja negativo, movem-se em direções contrárias. Uma covariância grande indica uma forte relação, caso seja pequena, então a relação é fraca. É estimada pelo produto da dispersão em relação à média ($\mu$) para cada variável em cada período, conforme a Equação 14:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(X_i - \mu_x).(Y_i - \mu_y)}{N-1} \quad . \tag{14}$$

2.4.4.4 A Similaridade Cosseno

A similaridade cosseno dada pela Equação 15 é uma medida de similaridade entre dois vetores em que mede-se o cosseno do ângulo entre eles. Para isto, calcula-se uma matriz de similaridade entre os espectrogramas como matrizes de características A e B, visando à aquisição da distância cosseno entre as magnitudes da STFT dos sinais. Um dos motivos para a popularidade da similaridade cosseno se dá pela sua eficiência de avaliação, especialmente vetores esparsos, uma vez que consideram-se apenas as dimensões diferentes de zero.

$$Similaridade = \cos(\theta) = \frac{A \,.\, B}{\|A\|\|B\|} = \left(\frac{\sum_{i=1}^{n} A_i \,.\, B_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \;.\; \sqrt{\sum_{i=1}^{n}(y_i)^2}}\right). \qquad (15)$$

Para ver se duas palavras são as mesmas, calcula-se a distância das características das locuções de referência. As palavras serão consideradas iguais, caso a distância entre as características destas sejam as menores dentre todas as demais amostras analisadas. No entanto, se o comprimento de $A$ é diferente de $B$, então não podemos utilizar medidas de distâncias como a similaridade cosseno. Em vez disso, precisamos de um método mais flexível, capaz de encontrar o melhor mapeamento de elementos em $A$ para aqueles em $B$. Trata-se de um método de programação dinâmica que será abordado a seguir.

2.4.4.5 Dynamic Time Warping (DTW)

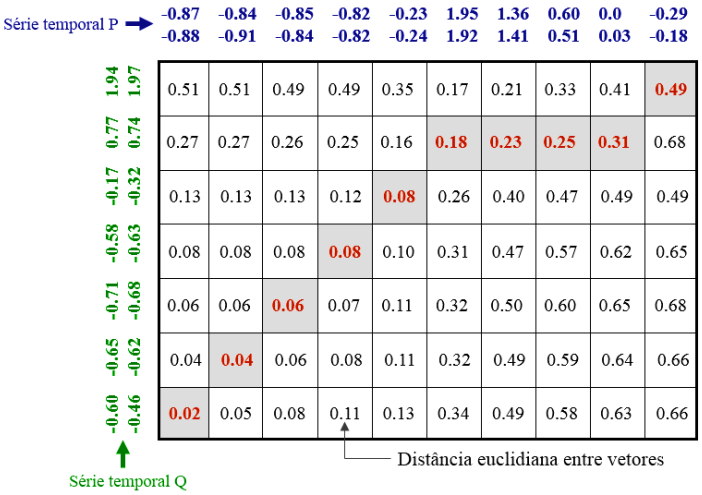
A fala é um processo dependente do tempo e pronúncias de uma mesma palavra terão diferentes durações e, ainda que possuam a mesma duração, serão diferentes, devido a várias partes das palavras que estão sendo ditas em frequências diferentes. Para obter a distância global entre dois padrões de fala, torna-se necessário um alinhamento temporal. O primeiro algoritmo para o reconhecimento de palavras conectadas, proposto por Vintsyuk (1968), mostrou como as técnicas de programação dinâmica poderiam ser utilizadas para descobrir a sequência de palavras ótima que combina com uma dada locução (YNOGUTI, 1999). O método DTW busca por um alinhamento que minimize a distorção causada pelos efeitos da fala, podendo ser obtido por meio do cálculo da distância entre os sinais (ADAMI, 1997). Na prática, as características são vetores e a distância entre eles é usualmente tomada como alguma medida de distância, como a Similaridade Cosseno apresentada na Equação 15.

Existem diversas variações do algoritmo DTW, nas quais utilizam-se diferentes métricas de distorção, caminhos permitidos e procedimentos de buscas, com complexidade de $O(N^2)$ (ADAMI, 1997). O DTW trata-se de uma técnica de minimização de custos, no qual o sinal de entrada é estirado ou comprimido de acordo com um modelo de referência.

### 2.4.4.6 O Algoritmo DTW

Em um sistema de reconhecimento de palavras isoladas, o DTW começa e termina nos pontos limites de cada amostra (ADAMI, 1997), ou próximo deles. Para entender o DTW, dois conceitos precisam ser tratados: as **características**, representado as informações que modelam cada sinal, e as **distâncias**, alguma métrica deve ser usada para obter um caminho de correspondência, podendo ser de dois tipos: **distância local**, que corresponde à diferença entre uma característica de um sinal e uma característica de outro; e a **distância global**, que diz respeito à diferença global entre um sinal inteiro e um outro sinal de comprimento, possivelmente diferente ou não. O mapeamento do caminho entre dois sinais utiliza-se de uma matriz de distorção (Figura 9).

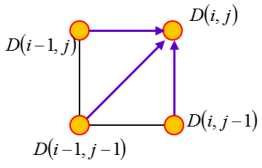
**Figura 9 - Matriz de Distorção de Duas Séries Temporais**




Fonte: modificada de Tsiporkova (2012).

O número de caminhos a serem considerados durante a busca pelo melhor caminho de correspondência entre uma entrada e um modelo é exponencial em relação ao comprimento da entrada. A Figura 10 mostra três possíveis direções a partir do ponto $(i, j)$.

**Figura 10 - Os Três Sentidos Adotados**




Fonte: Jang (2005).

Tendo os vetores de referência $A$ e de teste $B$ com comprimentos $M$ e $N$, respectivamente, para encontrar o melhor alinhamento entre eles, é necessário localizar o caminho através da matriz, adotando as restrições a seguir:

- **Condição de Monotonicidade**: Para qualquer ponto $(i, j)$ no caminho, as possíveis direções se restringem a $(i-1, j), (i, j-1), (i-1, j-1)$. Esta restrição local garante que o caminho de mapeamento é monotonicamente não-decrescente em seus primeiro e segundo argumentos, ou seja, o caminho não voltará sobre si mesmo, tanto os índices $i$ e $j$ ou permanecem o mesmo, ou aumentam, mas nunca diminuem. Além disso, para qualquer elemento em $A$, devemos ser capazes de encontrar pelo menos um elemento correspondente em $B$, e vice-versa.

- **Condição de Continuidade**: O caminho avança um passo de cada vez, isto indica que não se pode ignorar qualquer elemento de $A$ e $B$, evitando o salto para um próximo índice.

- **Condição Limite (fronteira)**: o caminho começa no canto inferior esquerdo (âncora inicial) e termina no canto superior direito (âncora final), isto é, $(a_1, b_1) = (1, 1)$ e $(a_k, b_k) = (M, N)$, garantindo que o alinhamento não considere de forma parcialmente uma das sequências.

- **Janela de Distorção (*warping*):** Em um bom alinhamento, é pouco provável que percorra muito distante da diagonal. Garante que o alinhamento não tentará pular características diferentes e ficar preso em características semelhantes.

- **Restrição de Inclinação (*slope*)**: O caminho não deve ser muito vertical ou muito horizontal. A condição é expressa como uma relação de $n/m$ em que $m$ é o número de passos na direção $a$ e $n$ é o número de passos na direção $b$. Após $m$ etapas em $a$, deve-se avançar em $b$ e vice-versa.

Ao aplicar essas restrições, pode-se restringir os movimentos que podem ser feitos a partir de qualquer ponto do caminho, e assim limita-se o número de percursos que devem ser considerados. Na Equação 16, tem-se $D(i, j)$ representando a distância global até $(i, j)$, e a distância local em $(i, j)$ é dada por $d(i, j)$.

$$D(i, j) = min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \ . \qquad (16)$$

2.4.4.7 Modelos Ocultos de Markov

Os processos markovianos são aqueles para os quais a propriedade de Markov é satisfeita, ou seja, a probabilidade de um passo depende unicamente do estado atual do sistema (RABINER, 1989). Esses processos, são um tipo especial de processos estocásticos com aplicabilidade quase universal. Uma Cadeia de Markov é um processo que consiste num número de estados de probabilidades associadas às transições entre os estados. O Modelo Oculto de Markov ou *Hidden Markov Model* (HMM) é utilizado na modelagem de sistemas com comportamentos discretos e dependentes do tempo (ADAMI, 1997).

Em uma HMM, as observações ($v_i$) são símbolos emitidos por estados não observáveis ($S_i$) de acordo com determinadas funções probabilísticas, em que cada sequência de estados é uma cadeia de Markov de primeira ordem.

Tendo $N$ como o número de estados do modelo, o conjunto de estados individuais é representado por $S = \{S_1, \dots, S_N\}$ e o estado no tempo $t$ pelo símbolo $q_t$. O número de símbolos distintos observáveis é representado por $M$, e o conjunto de símbolos individuais é representado por $V = \{v_1, \dots, v_M\}$ (RABINER, 1989). Ainda segundo o autor, uma HMM básica é constituída de 3 conjuntos de parâmetros:

1. As probabilidades de emissão de símbolo $B = \{b_{jk}\}$: a probabilidade do símbolo $v_k$ ser emitido pelo estado $S_j$:

$$b_{jk} = P\left(v_k \, em \, t \middle| q_t = S_j\right), 1 \leq j \leq N \, e \, 1 \leq k \leq M. \quad (17)$$

2. As probabilidades de transição de estado $A = \{a_{ij}\}$: a probabilidade de estar no estado $S_j$ no instante de tempo subsequente, dado que o estado atual é $S_i$:

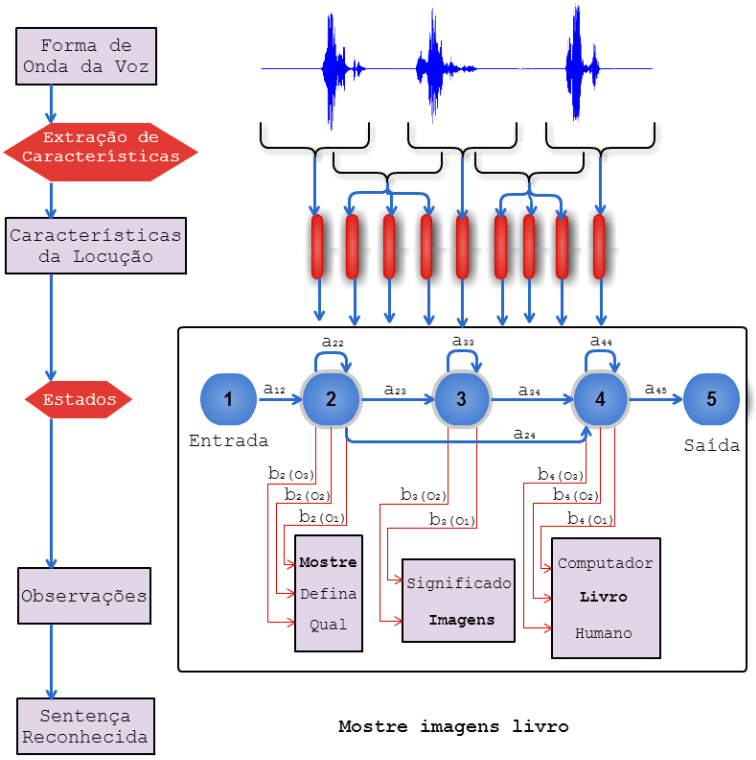
$$a_{ij} = P\left(q_{t+1} = S_j \middle| q_t = S_i\right), 1 \leq i, j \leq N. \quad (18)$$

3. A distribuição de probabilidade a priori $\pi = \{\pi_i\}$ do sistema estar em um dado estado $S_i$ no instante inicial de tempo:

$$\pi_i = P(q_1 = S_i), 1 \leq i \leq N. \quad (19)$$

Uma HMM pode ser definida como uma máquina de estados finita onde as transições entre os estados são dependentes da ocorrência de algum símbolo. Associado com cada transição de estado, há uma distribuição de probabilidade de saída, a qual descreve a probabilidade com o que um símbolo ocorrerá durante a transição e uma probabilidade de transição indicando a probabilidade dessa transição (ADAMI, 1997). Uma ilustração desse processo de classificação no STI é mostrada na Figura 11.

**Figura 11 - Representação da Classificação por HMM**

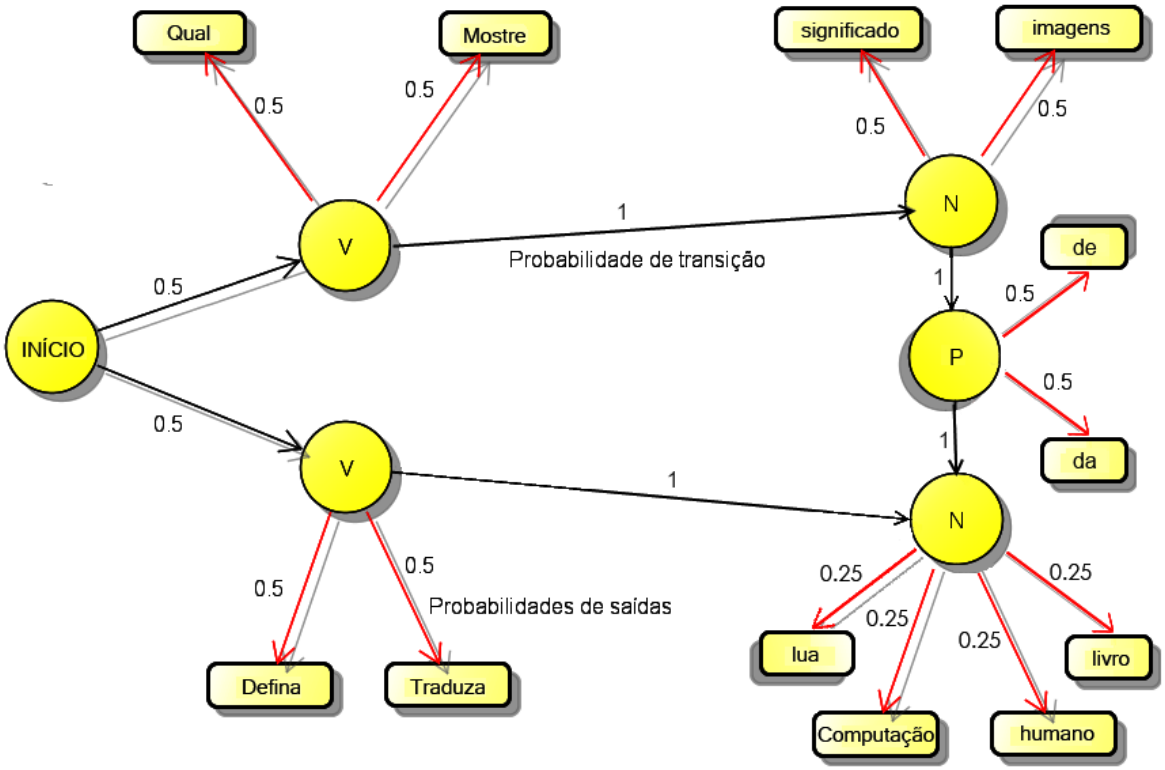



Fonte: Elaborada pelo autor.

Na Figura 11, têm-se cinco estados com suas probabilidades de transição de estado $(a_{ij})$ e um conjunto de símbolos observáveis $(O_k)$ com suas respectivas probabilidades de emissão de símbolo $(b_{ij})$. Se o sistema pode passar de um estado $i$ para outro $j$, então $a_{ij} > 0$. Caso esses dois estados não estejam conectados, então $a_{ij} = 0$.

A adoção da HMM visa a uma formalização das sequências de sentenças lógicas a serem consideradas verdadeiras, estabelecendo uma restrição sintática/semântica durante a concatenação das palavras. Considerando um conjunto de palavras, uma cadeia de Markov pode ser criada para identificar a probabilidade de determinada palavra ser pronunciada após outra, seguindo a estrutura na Figura 12, em que $V$ indica uma ação, $P$ uma preposição e $N$ um nome.

**Figura 12 - Máquina de Estados do Reconhecimento por HMM**




Fonte: Elaborada pelo autor.

Cada estado envolve a produção de uma palavra (observação). A Cadeia de Markov é usada para identificar probabilisticamente o próximo estado, dado o estado atual e um estímulo externo (detecção de atividade de voz). A partir desse estímulo, o processo inicia e as transições entre os estados da Cadeia de Markov são baseadas unicamente nas probabilidades, já a transição para uma observação baseia-se também na medida de verossimilhança entre a locução e as observações.

## 2.5 CONSIDERAÇÕES

Com base nos dados de treinamento, são gerados os modelos de referência, para os quais são atribuídos rótulos que identificam cada padrão. Na fase de reconhecimento, através dos dados de testes, são obtidos padrões que serão comparados com os modelos gerados durante o treino e, utilizando-se uma regra de decisão lógica, identifica-se o modelo que mais se assemelha ao padrão de entrada desconhecido (RABINER e SCHAFER, 1978; SHAUGHNESSY, 2000; DELLER , PROAKIS e HANSEN, 1993).

# 3 TRABALHOS RELACIONADOS

Este capítulo apresenta alguns trabalhos relacionados, sob alguns aspectos, a esta pesquisa. O mesmo foi dividido em duas partes: a primeira aponta pesquisas relativas a sistemas de reconhecimento da fala ou síntese de voz; a segunda assinala trabalhos relativos aos STI's, descrevendo suas principais características. Em ambas as partes, são mostrados trabalhos frutos da pesquisa, cada um segundo sua natureza de aplicação.

## 3.1 RECONHECIMENTO AUTOMÁTICO DA FALA E SÍNTESE DE VOZ

Várias são as áreas e aplicações que envolvem técnicas de processamento digital de sinais. A seguir, são feitas breves explanações sobre trabalhos que se correlacionam com a presente dissertação nesse domínio de conhecimento.

Em Bresolin (2003), foi realizado um estudo do reconhecimento de fala para um grupo de 10 palavras (números de zero a nove) e acionamento de equipamentos elétricos. Nele, o autor buscou desenvolver um sistema de reconhecimento de voz, intitulado "*Parlato*", que fosse capaz de acionar um equipamento elétrico via microcomputador, utilizando a correlação das magnitudes dos sinais por meio da FFT, densidade espectral e wavelets. O sistema elétrico controlado trata-se de um robô educativo. O autor afirma que, se o sistema desenvolvido é capaz de acionar um robô do tipo didático, então seria capaz de acionar, por comando de voz, qualquer outro sistema elétrico, desde que observadas suas características.

Santos (2013) apresenta uma interface multimodal de Interação Humano-Computador (IHC) para um sistema de Recuperação de Informação (RI) em português baseada em voz e texto, focando na tarefa a ser executada. O autor propõe a verificação dos possíveis benefícios provenientes e a viabilidade do uso da Interação Humano-Computador Multimodal (IHCM) em uma interface computacional baseada em voz artificial, vinculada a um mecanismo de RI, visando a melhoria do diálogo homem-máquina nas operações de troca de informações.

O trabalho de Michael e Lawrence (1982) propõe um procedimento alternativo para a implementação do algoritmo DTW em comparação com o método padrão. O procedimento proposto apresenta um menor tempo de computação. Porém, acarreta numa maior sobrecarga de processamento.

Lee, Chen e Jang (2005) apresentam uma proposta para incremento da velocidade na utilização do DTW para o reconhecimento de melodias de forma a manter taxas de acurácias razoáveis e reduzir eficazmente o cálculo do caminho de menor custo na matriz de distâncias locais.

Em Rabiner (1989), é realizado um estudo do reconhecimento de palavras conectadas utilizando as HMM's. Partindo da premissa de que o reconhecimento de palavras conectadas é baseado nos modelos de palavras individuais, então, o problema de reconhecimento trata-se de encontrar a sequência ótima (concatenação) de modelos de palavras que melhor combinem com uma sentença desconhecida de palavras conectadas, utilizando *level building* (determinando a posição de uma palavra em uma *string*) e o algoritmo de busca Viterbi.

Já em Ynoguti (1999) é abordado o problema de reconhecimento da fala contínua por HMM. O autor investiga a influência de alguns conjuntos de subunidades fonéticas, e dos modelos de duração e de linguagem no desempenho do sistema. Também são propostos alguns métodos de redução do tempo de processamento para os algoritmos de busca utilizados no reconhecimento das sentenças.

O artigo de Juang (1984) fornece uma visão teórica unificada das técnicas DTW e HMM para problemas de reconhecimento de fala. Discute-se no trabalho a aplicação dos modelos ocultos de Markov no reconhecimento de fala. Mostra-se ainda que o método DTW com medições obtidas por predição linear é implicitamente associado a uma classe específica de modelos de Markov e é equivalente aos procedimentos de maximização de probabilidade para funções autoregressivas de probabilidade Gaussiana multivariada. Essa visão unificada oferece *insights* sobre a eficácia dos modelos probabilísticos em aplicações de reconhecimento de fala.

Ravinder (2010) apresenta o desenvolvimento de um reconhecedor de palavras isoladas da língua regional indiana Punjabi, no modo dependente de locutor, em tempo real. Utilizando as técnicas de HMM's e DTW, o trabalho enfatiza a abordagem baseada em codificação preditiva linear com cálculo de programação dinâmica e quantização vetorial com reconhecedores baseados em HMM nas tarefas de reconhecimento de palavras isoladas.

## 3.2 SISTEMAS TUTORES INTELIGENTES

A seguir são apresentados trabalhos que de alguma forma abordaram a criação de interfaces inteligentes voltadas para área educacional de interação com seres humanos.

Em Clancey (1986), é apresentado o GUIDON, um sistema tutorial para o ensino de diagnóstico de doenças infecciosas do sangue. O GUIDON foi desenvolvido para uso em faculdades de medicina no treinamento de estudantes e médicos. Possui representação do conhecimento por meio de regras em um ambiente reativo com interações estruturais. A principal deficiência do sistema é o fato de pressupor que o aluno tenha o entendimento dos termos técnicos utilizados.

Um STI para ensino da linguagem LOGO é apresentada em Miller (1982). O SPADE objetiva desenhar uma figura utilizando os desenhos primitivos do LOGO, tal como a decomposição do problema. Baseia-se na teoria de planejamento, que contém ações executáveis pelo aluno em um ambiente reativo com treinamento. Espera-se que o aluno possa interagir com o sistema através de uma sequência de comandos que compõem a solução do problema.

A Tabela 1 apresenta um sumário dos trabalhos relacionados neste capítulo, incluindo o presente trabalho. Pela observação de alguns aspectos listados na tabela, nota-se que o STI desenvolvido para ser utilizado como subsídio da pesquisa, situa-se na intersecção dos demais trabalhos.

**Tabela 1 – Sumário de Trabalhos Relacionados**

| Trabalhos | Realiza Síntese | Reconhece a Fala | Caso de Uso | Disponibiliza Base de Dados | Disponibiliza Código Fonte | Tipo de Ambiente |
|---|---|---|---|---|---|---|
| *Michael e Lawrence, 1982* | - | ✓ | - | - | - | Não se aplica |
| *Juang, 1984* | - | ✓ | - | - | - | Não se aplica |
| *Rabiner, 1989* | - | ✓ | - | Não se aplica | Não se aplica | Não se aplica |
| *Ynoguti, 1999* | - | ✓ | ✓ | ✓ | - | Não se aplica |
| *Bresolin, 2003 (Parlato)* | - | ✓ | ✓ | - | - | Reativo com interações |
| *Lee, Chen e Jang, 2005* | - | ✓ | - | - | - | Não se aplica |
| *Ravinder, 2010* | - | ✓ | - | - | - | Não se aplica |
| *Santos,2013* | ✓ | - | ✓ | - | - | Reativo com interações |
| *STI DeVoice 2015* | ✓ | ✓ | ✓ | ✓ | ✓ | Reativo com interações |

Fonte: Elaborada pelo autor.

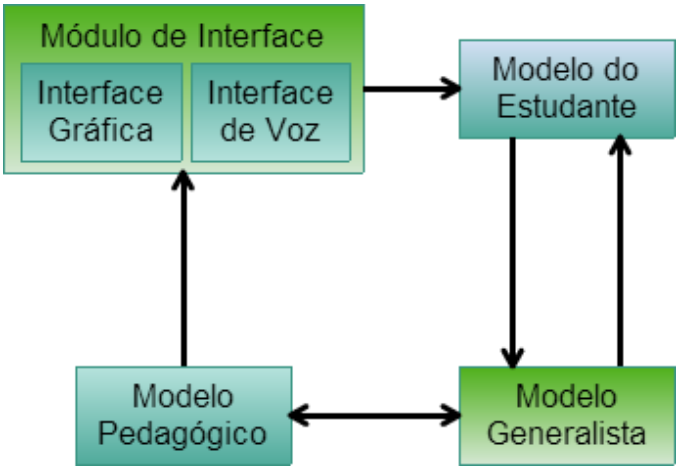
# 4 METODOLOGIA E CENÁRIO DO STI

Este capítulo apresenta a modelagem do STI, no qual é apresentada a arquitetura da interface desenvolvida, além do processo de confecção da base de dados e dos métodos usados para a construção dos mecanismos responsáveis pelo reconhecimento da fala e síntese da voz.

## 4.1 A ARQUITETURA UTILIZADA

A arquitetura implementada (Figura 13) propôs a substituição do modelo especialista em função da adoção de um modelo generalista com domínio sobre "qualquer área"[11]. Como consequência, perde-se em relação ao conhecimento profundo de um determinado contexto, mas, ganha-se uma expansão das áreas de domínios abrangidas.

**Figura 13 - Arquitetura do Modelo de STI Implementado**



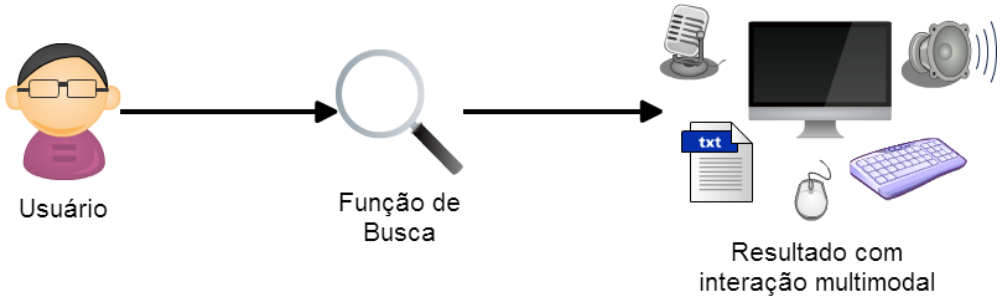

Fonte: Elaborada pelo autor.

A arquitetura sugeriu ainda, uma modificação no módulo de interface presente na arquitetura clássica (Figura 3), por meio da divisão do mesmo, possibilitando a adoção e implementação de um módulo de Reconhecimento Automático de Voz (**Interface de Voz**) e um Ator Sintético (AS) (**Interface Gráfica**) em substituição do instrutor humano, como exposto na Figura 13, associada a um mecanismo de RI, aperfeiçoando e automatizando o diálogo entre o STI e o aprendiz.

[11] A expressão "qualquer área" corresponde a qualquer assunto (nomes de cidades, objetos, animais, profissões, etc.) de qualquer domínio do conhecimento humano, desde que tal assunto componha a base de palavras reconhecidas pelo STI e a mesma também seja constituída de significado prosódico na base de conhecimento.

Para o STI, foi desenvolvida uma IHCM, adicionando além do teclado, mouse e monitor, considerados componentes tradicionais, elementos da voz artificial e do texto, como ilustrado na Figura 14.

**Figura 14 - Representação do Fluxo da IHCM**




Fonte: baseado em Santos (2013).

O motor de pesquisa foi projetado para procurar palavras-chave na base de dados prosódica (conhecimento cognitivo) localizada na Web e retornar uma lista de referências que combinem com o termo informado. No sistema desenvolvido, ele é utilizado como uma "*interface*" (*front-end*) para o motor de busca da base de dados *online*. O sistema torna-se mais poderoso alcançando uma certa robustez, devido à habilidade de adaptar-se ao vocabulário, permitindo o "aprendizado" (com alguns inconvenientes[12]) do significado de novas palavras em tempo de execução.

A concepção de um projeto piloto foi um dos primeiros passos no desenvolvimento da pesquisa. As integrações de soluções tecnológicas, tal como a utilização de uma interface de voz para a realização da síntese de fala, uma solução factível encontrada e adotada, foi a utilização da SAPi5 Raquel 4.0[13], corroborando na produção de um sistema que subsidiasse a pesquisa, atendendo as expectativas de forma satisfatória, bem como alguns objetivos específicos. A interface projetada, batizada de "DeVoice", foi desenvolvida pensando nos indivíduos com problemas motores ou limitações de visão que necessitem de uma solução tecnológica que os assista. Para isso, foi necessário a implementação de um sistema de RI no ambiente MatLab®[14], que não apenas fosse capaz de responder às solicitações submetidas ao motor de busca desenvolvido, mas também fosse apto a emitir respostas audíveis referentes ao conteúdo retornado pelo mecanismo de busca, bem como notificar o usuário quando necessário durante o diálogo humano-computador.

[12] Cabe ao usuário informar a transcrição fônica (grava a pronúncia) e gráfica (digita a palavra) do novo termo.
[13] Uma *Speech Application Programming Interface* (SAPI) da *Microsoft* em português do Brasil, disponibilizada gratuitamente pela *Next-Up* e *ScanSoft*.
[14] Criado pela *MathWorks*® Icn, o *MATrix LABoratory* (MATLAB) trata-se de um software que permite a manipulação de matrizes, criação de gráficos de funções e de dados, criação e execução de algoritmos, etc.

## 4.2 BASE DE DADOS

O vocabulário de um sistema de reconhecimento de fala é a unidade que define o universo de palavras que podem ser reconhecidas. Em termos gerais, quanto maior e mais abrangente o vocabulário, mais flexível é o sistema, embora o reconhecimento torne-se cada vez mais difícil à medida em que o vocabulário cresce. Em sistemas de vocabulário pequeno, é comum utilizar-se as palavras como unidades fundamentais. Para que um sistema seja útil, não é necessário um vocabulário muito grande. Existem sistemas que possuem um vocabulário de apenas duas palavras: "sim" e "não" (YNOGUTI, 1999). Para o STI DeVoice, foi realizada a aquisição de 50 locuções.

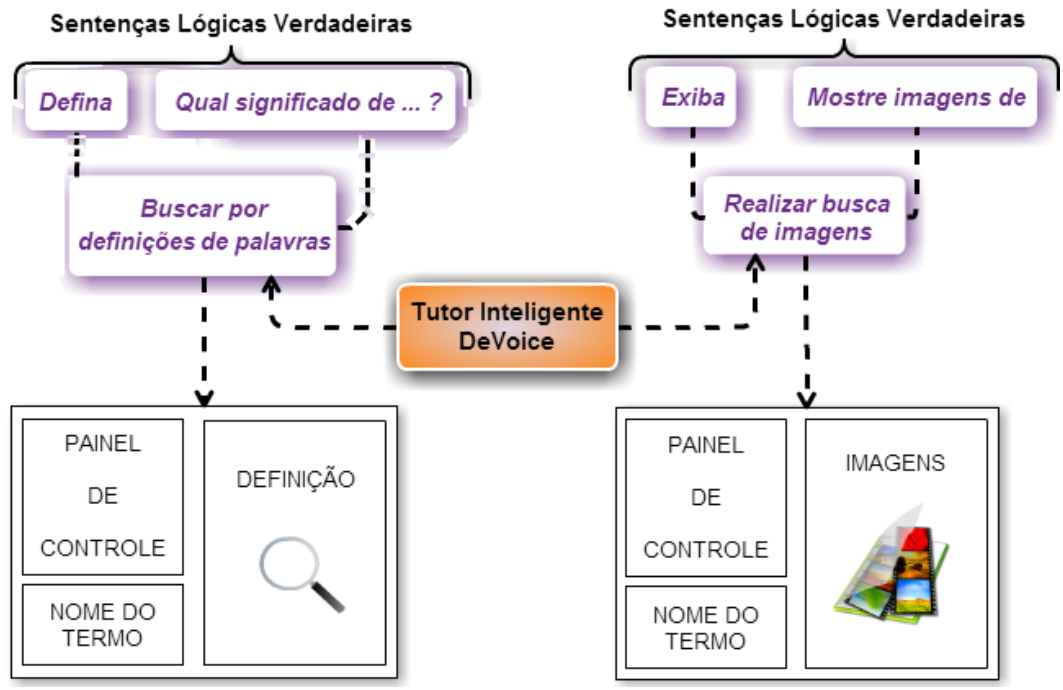
A constituição da base de locuções se deu pelo uso de dados coletados localmente por um locutor, uma vez que não existe um acervo público na língua portuguesa que servisse para os propósitos desta pesquisa, tornando-se necessário confeccionar uma base de dados própria, constituída de duas etapas: a **escolha das palavras**, selecionadas e categorizadas segundo uma análise que buscou atender os requisitos funcionais do STI conforme a necessidade de formação das sentenças para as tarefas que desejava-se que o STI fosse capaz de executar; e a **gravação das locuções**, realizadas em ambiente silencioso, com microfone omnidirecional no modo mono a 75% do volume máximo. A aquisição dos dados se deu através de uma placa de som Realtek ALC662, em um microcomputador. A taxa de amostragem[15] utilizada foi de 11,025 kHz e resolução de 16 bits por amostra (65.536 níveis). O áudio foi armazenado em formato Windows com modulação de pulsos (*Pulse-Code Modulation* - PCM) WAV[16], que utiliza um método de armazenamento de áudio não comprimido (sem perda), visando a qualidade máxima do áudio. O tamanho da amostra foi mantido constante com duração de dois segundos. A base foi dividida em duas partes: uma **base de conhecimento cognitivo**, correspondendo à ciência que o STI possui sobre o significado denotativo das palavras, isto é, a correspondência semântica das locuções reconhecidas e uma **base de *corpus***, constituída pelas locuções a serem utilizadas como padrões durante a classificação. Segundo Silva (2009), a criação da base de dados para a realização do treinamento representa uma etapa decisiva para uma boa performance do sistema de reconhecimento de voz.

[15] Amostrar um sinal significa pegar um "instantâneo" do sinal de um sensor em um tempo discreto, convertendo o sinal analógico em uma representação digital (WEEKS, 2012).
[16] WAV (ou WAVE), é a forma curta de *WAVEform audio format*, é um formato padrão de arquivo de áudio da Microsoft e IBM, para armazenamento de áudio em computadores.

## 4.3 FUNCIONAMENTO DO STI DEVOICE

O DeVoice é um sistema baseado no tipo de entrada, atuando como um sistema de palavras isoladas e conectadas, que são convertidas em comandos e posteriormente traduzidos em ações, como observado na Figura 15.

**Figura 15 - Fluxograma do Tutor Inteligente**




Fonte: Elaborada pelo autor.

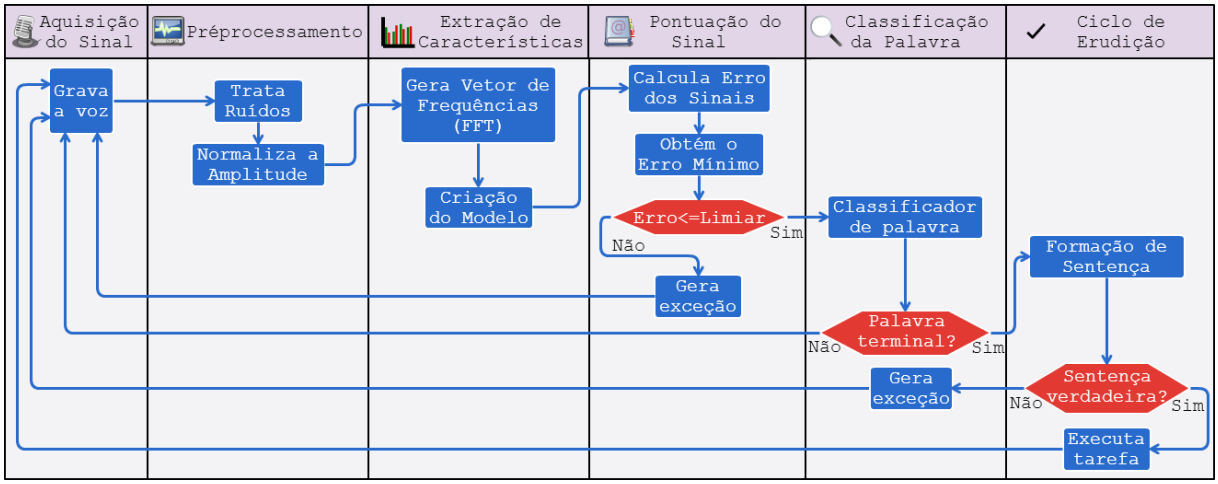
A separação das palavras é realizada por meio de pausas entre as locuções proferidas e o processamento ocorre a cada palavra pronunciada, atuando assim, também, como um sistema de palavras conectadas por meio da combinação de palavras (concatenação), simulando uma conversa natural. Uma visão geral do processo de treinamento e reconhecimento das sentenças é apresentada nas figuras 16, 17, 18 e 19.

Durante o reconhecimento, caso as características extraídas sejam consideradas insuficientes/excessivas para a geração de um modelo, o sinal é classificado como silêncio/barulho, respectivamente, e será descartado. O processo de aquisição é então reiniciado.

Uma vez gerado, o modelo é classificado segundo uma lista em que cada posição contém o erro (pontuação) de cada sinal e quanto maior a similaridade, maior a pontuação. A palavra será aceita caso essa pontuação seja maior que um limiar estabelecido. Caso contrário, será recusada.

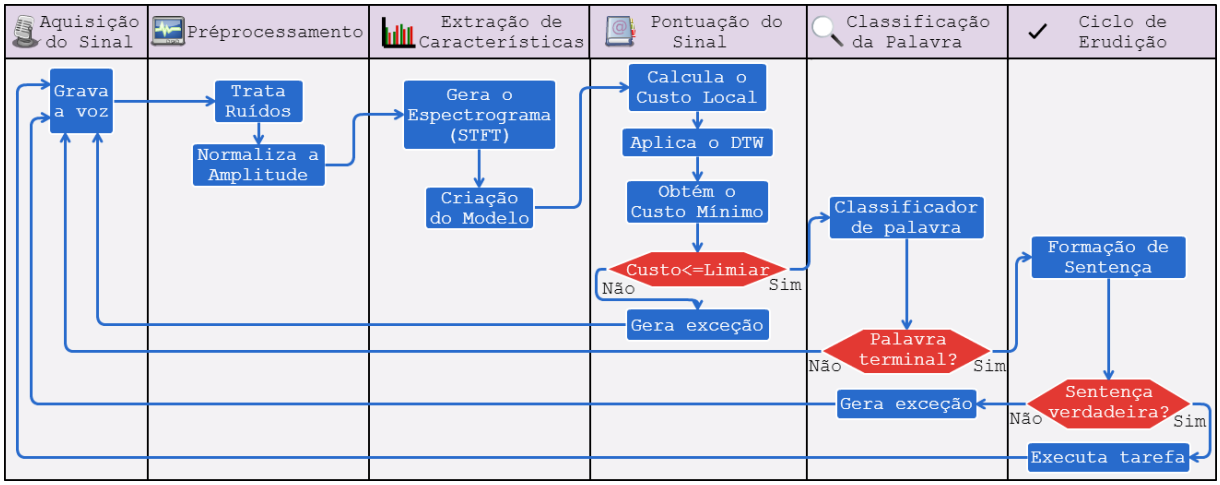
Após essa etapa, a palavra é analisada pelo avaliador de sentenças lógicas, que fará a classificação da palavra. Como resultado, a palavra é processada no ciclo de erudição, visando a aprovação da sentença de acordo com sua categoria.

**Figura 16 - Fluxo Geral do Processo de Reconhecimento por Menor Erro**




Fonte: Elaborada pelo autor.

**Figura 17 - Fluxo Geral do Processo de Reconhecimento por DTW**



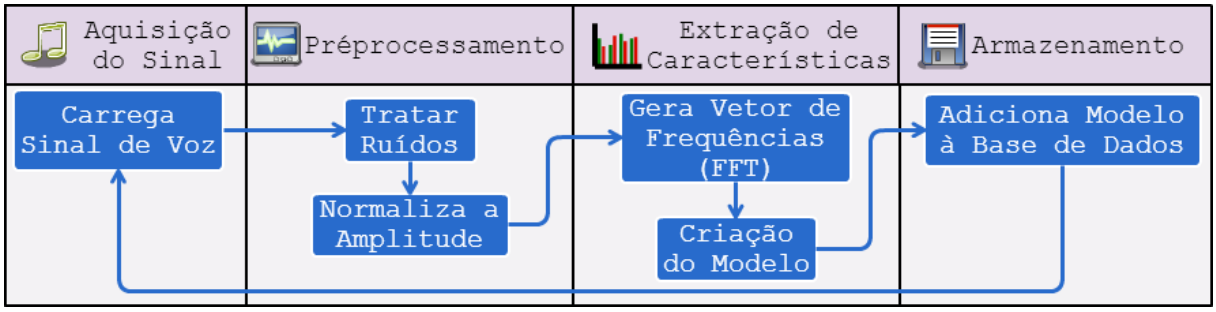

Fonte: Elaborada pelo autor.

Todo o processo é repetido até que uma palavra terminal (palavra que deseja-se obter o *feedback*) seja pronunciada. O resultado é retornado de forma audível e/ou visual (dependendo da sentença formada), facilitando o acesso à informação por qualquer pessoa[17].

[17] Para o caso em que o usuário seja portador de deficiência da voz e o mesmo não conseguir proferir palavra alguma ou com bastante dificuldade, a pesquisa também pode ser realizada via teclado.

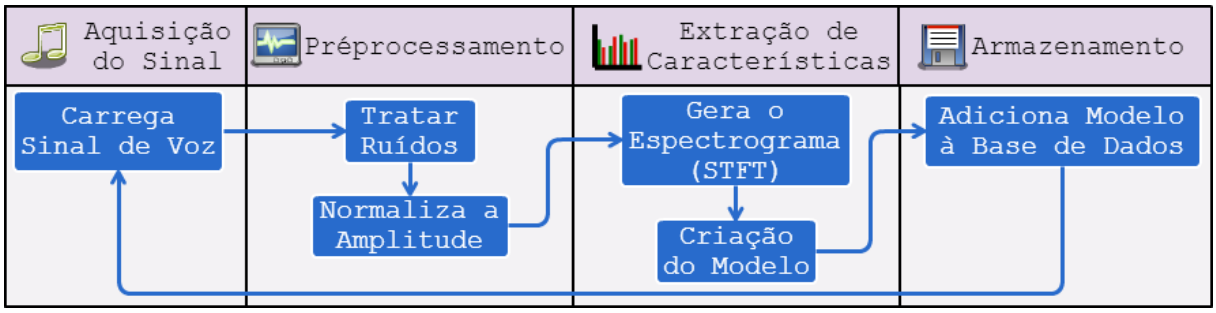
No treinamento, por meio da extração das características do sinal, gera-se um modelo da palavra a ser armazenado na base, conforme fluxograma das figuras 18 e 19.

**Figura 18 - Fluxo Geral do Processo de Treinamento por FFT**




Fonte: Elaborada pelo autor.

**Figura 19 - Fluxo Geral do Processo de Treinamento por STFT**




Fonte: Elaborada pelo autor.

## 4.4 DESENVOLVIMENTO DO ATOR SINTÉTICO: DANDO VOZ AO STI

Para gerar ilusão de vida, é preciso expressar e controlar a personalidade, emoções e atitudes. Através do mecanismo de percepção do mundo, o AS capta e reage às alterações no ambiente, fornecendo respostas mais próximas do real e condizentes com as locuções interpretadas no mecanismo de reconhecimento de voz.

Há diferentes arquiteturas para sistemas de diálogo cujos conjuntos de componentes que são incluídos no sistema, bem como a forma como esses componentes dividem as responsabilidades, diferem de sistema para sistema. O sistema de diálogo é a parte do sistema destinada a conversar com um humano de forma coerente, empregando texto, discurso, gráficos, gestos e outros meios de comunicação sobre a entrada e canal de saída (LESTER, BRATING e MOTT, 2004).

A iniciativa do diálogo pode ser tomada tanto pelo humano como pelo AS, possibilitando um ganho afetivo computacional que pode ser percebido e sentido durante o diálogo, o qual transmite toda sua emoção pela síntese da fala. O ciclo de atividades no sistema de diálogo do DeVoice contém as seguintes fases:

1. O usuário fala e a entrada é convertida para texto;
2. Um *parser* realiza a análise sintático/semântico do texto reconhecido;
3. Finalmente, a saída é processada usando uma *engine* TTS como gerador de linguagem natural e um mecanismo de *layout*.

Para efetivação da IHCM no STI DeVoice, foi necessário inserir uma função pedagógica no AS, passando então a ser considerado um ator pedagógico[18] com função de tutor. Sendo responsável, portanto, de monitorar as ações do aluno, expor conteúdo segundo modelo, fornecer *feedback* às ações do aluno através de mensagens escritas e/ou narradas, conduzi-lo durante a interação, auxiliar em circunstâncias críticas e motivá-lo a aprender, provendo interatividade ao STI.

## 4.5 CONSIDERAÇÕES

Este capítulo apresentou a modelagem da arquitetura de um sistema de reconhecimento de voz e sua integração ao STI, visando uma evolução e automação de um STI clássico. Foi exposto o processo de confecção da base de palavras, além de uma explanação sobre o funcionamento da interação no STI entre o AS e o usuário.

Assim, a partir do modelo de STI apresentado e implementado neste capítulo, foi possível atingir alguns dos objetivos específicos relatados.

[18] Atores que estão inseridos em um ambiente de ensino e aprendizagem que visam a comunicação com o aluno.

# 5 ANÁLISE DOS RESULTADOS

Esta etapa buscou analisar a interface desenvolvida, visando a sua validação e avaliação do mecanismo de reconhecimento de voz. O conjunto de palavras utilizadas nos experimentos encontram-se nos testes a seguir e nos apêndices A (palavras consideradas terminais), B (palavras indicando ações executadas pelo STI) e C (números). O STI em questão e os métodos envolvidos, foram todos implementados e testados no software Matlab® R2014a de 32 bits.

Utilizou-se um limiar para a determinação de sensibilidade do sistema, a fim de desconsiderar palavras não existentes no vocabulário, isto é, as elocuções que não possuírem semelhança maior ou igual a este limiar serão consideradas palavras desconhecidas, o que gera uma maior robustez ao sistema. Foram então escolhidos valores razoáveis para que as palavras existentes pudessem ainda assim ser reconhecidas pelos classificadores.

Durante os testes, foi utilizado como transdutor para a captação do sinal, um microfone externo com as seguintes especificações:

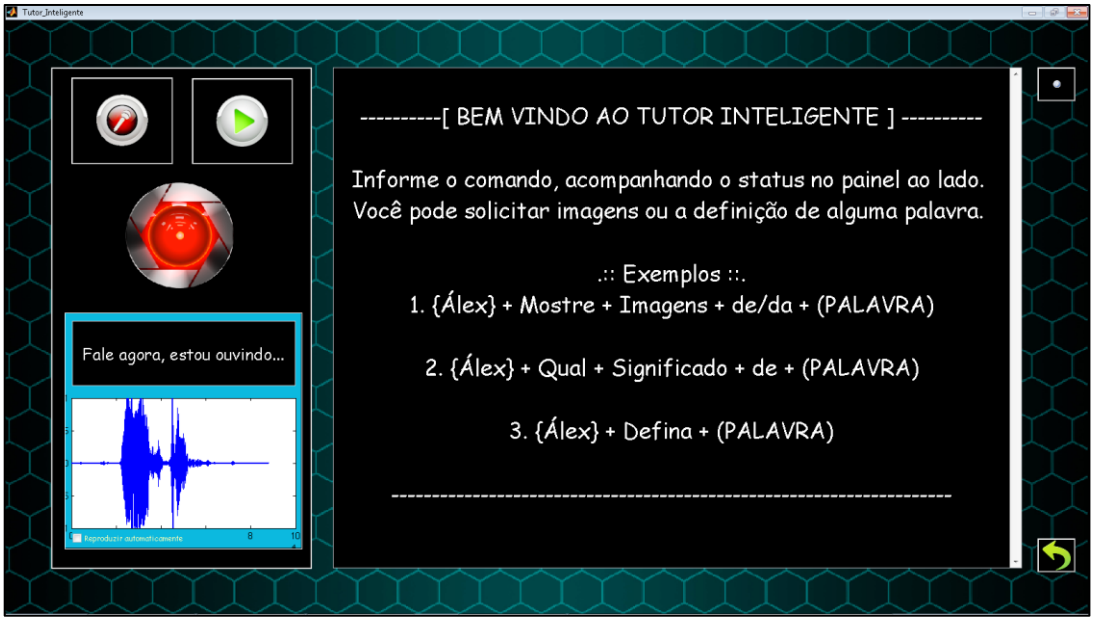
- Direcionamento: Omnidirecional;
- Frequência de Resposta: 20Hz ~ 20KHz;
- Sensibilidade: -58 decibéis +/- 3 decibéis em 1KHz (acrescido de +30 decibéis);
- Relação Sinal/Ruído: < 40 decibéis;
- Voltagem: 3 Volts.

## 5.1 DEMONSTRAÇÃO DE USABILIDADE DO STI

A interface do STI é autoexplicativa, possuindo ícones com *feedbacks* sonoros que descrevem sua respectiva função no sistema por meio da síntese de voz. Tudo o que aparece na tela do STI e cada etapa do processo é narrada pelo AS, exibido na área destinada à transcrição gráfica e no painel de *status*, permitindo uma usabilidade mais acessível pela utilização de mais de um sentido de percepção humana. O STI inicia então a execução[19] da tarefa correspondente às locuções. O Usuário deverá informar o que deseja que o STI realize. São pronunciadas palavras, que isoladamente serão processadas, e cada palavra reconhecida deve pertencer à sua respectiva categoria de palavra, na sequência correta de formação da sentença, como nos exemplos descritos na Figura 20.

[19] Para ter êxito no acesso à base de dados cognitiva, o computador deve estar conectado à internet.

**Figura 20 - *Screenshot* da Tela Inicial do STI DeVoice**



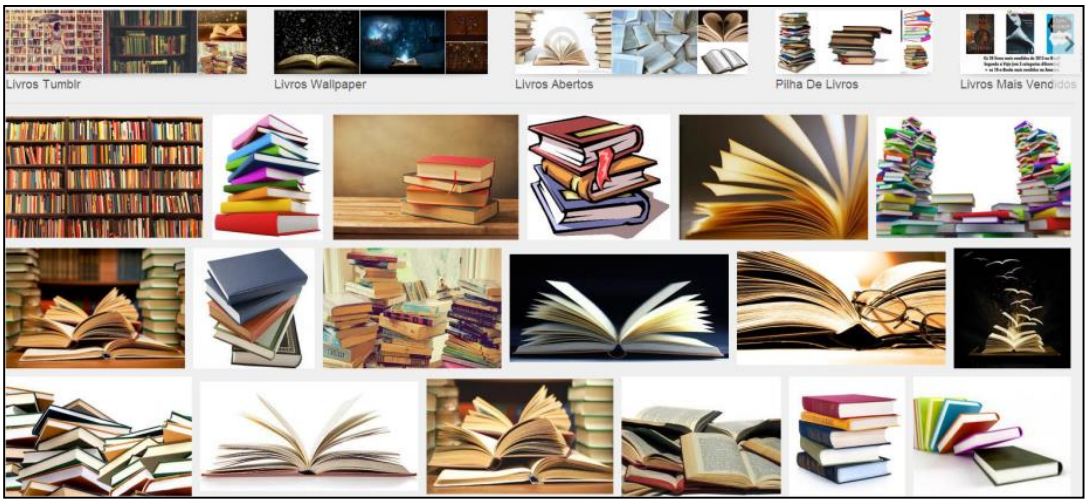

Fonte: Elaborada pelo autor.

A seguir são apresentados casos de uso do STI DeVoice visando a demonstração da eficácia do reconhecimento da fala e da síntese de voz empregada no sistema, na qual a *priori*, é demonstrado um cenário em que é executada uma tarefa correspondente ao reconhecimento da sentença "*Álex, Mostre imagens de livros*" (Figura 21) e no segundo caso (Figura 22), foi solicitado ao STI que retorna-se o significado de computação.

### 5.1.1 Caso de Uso 1: Pesquisando por Imagens

O usuário pronuncia a sentença "*Álex, mostre imagens de livro*" e o STI providencia imagens relacionas à sentença, no caso, livros, conforme a Figura 21.

**Figura 21 - Resultado da Pesquisa por Imagens de Livro**

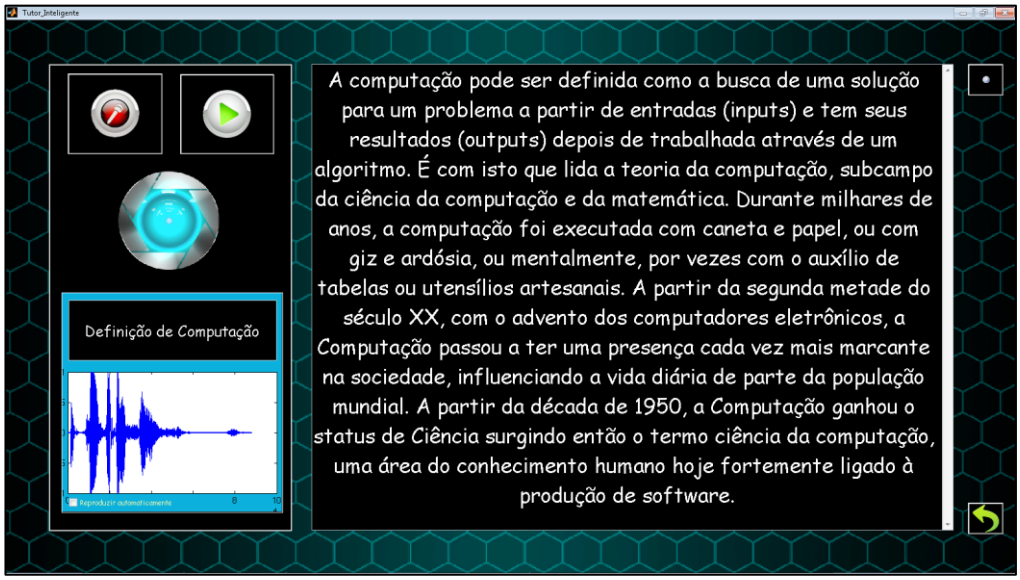



Fonte: Elaborada pelo autor.

### 5.1.2 Caso de Uso 2: Informando Definições

O usuário pronuncia a sentença "*Álex, qual significado de computação?*", como *feedback* à esta requisição, a definição de computação é exibida na tela, paralelamente à transcrição fônica do mesmo por síntese de voz, o qual pronuncia perfeitamente cada uma das palavras transcritas na tela (Figura 22), bem como qualquer dado presente na página web.

**Figura 22 - Resultado da Pesquisa pela Definição de Computação**




Fonte: Elaborada pelo autor.

Avaliando os casos de usos ilustrados anteriormente e disponibilizados[20] percebe-se a presença de alterações de pronúncia apenas em algumas palavras estrangeiras. As demais palavras do idioma Português foram perfeitamente sintetizadas pelo STI.

## 5.2 AVALIAÇÃO DO MECANISMO DE RECONHECIMENTO DE VOZ

Tendo em vista a validação deste trabalho, limitou-se o foco ao reconhecimento de palavras isoladas e conectadas no modo dependente de locutor, com vocabulário médio dependente do texto. As taxas de acerto no reconhecimento de palavras são utilizadas como medida de desempenho. Sendo assim, os testes a seguir visam a avaliação de desempenho dos classificadores, demonstrando sua eficácia no reconhecimento de palavras. Nos testes, também são mensurados o tempo necessário para o processamento do sinal, visando a obtenção de uma estimativa de custo computacional.

[20] Acesse vídeos demonstrativos da interface do STI e dos casos de uso em https://sites.google.com/site/leinylson/home/projetos/devoice, pelos quais tornam-se possíveis perceber um alto grau de entendimento das pronúncias obtidas pela síntese de voz.

A avaliação da acurácia pode ser adquirida através de coeficientes de concordância, podendo serem expressos como concordância total ou para classes individuais. Diversas técnicas de avaliação de acurácia têm sido discutidas na literatura. A maioria dos métodos quantitativos utiliza a matriz de confusão ou de contingência proveniente dos conjuntos de dados de classificação e referência (Story e Congalton, 1986; Foody, 2002).

Spiegel (1993) alega que cada frequência observada na matriz corresponde também a uma frequência esperada, e esta é medida sob uma determinada hipótese segundo as regras da probabilidade. Ainda segundo o autor, a diagonal da matriz ($X_{ii}$) exibe a frequência observada e representa a concordância entre o esperado e observado para cada classe.

Para avaliação de desempenho do processo de classificação utilizou-se os índices de **Exatidão Global**, proposta por Hellden *et al*. (1980) e o **Coeficiente Kappa**, proposto por Cohen (1960). Obtidos a partir das matrizes de contingência, possibilitando identificar não somente o erro global da classificação para cada classe, mas também, como se deram as confusões entre estas.

De acordo com Congalton e Green (1999), a estimativa de acurácia adquirida pela Exatidão Global ($E_{Global}$), é dada pela razão entre o número total de palavras corretamente classificadas ($X_{ii}$) e o número total de amostras ($N$), indicando a proporção de predições corretas, desconsiderando o que é positivo e o que é negativo, conforme Equação 20:

$$E_{Global} = \sum_{i=1}^{N} X_{ii} \Big/ N \, . \tag{20}$$

Devido à característica da Exatidão Global ser altamente suscetível a desbalanceamentos do conjunto de dados e podendo facilmente induzir a uma conclusão errada sobre o desempenho do sistema, optou-se por utilizar também o Coeficiente Kappa ($k$), que leva em conta os erros de omissão e comissão.

O Coeficiente Kappa é uma medida amplamente empregada na acurácia de classificação. Ele considera todos os elementos da matriz de contingência, diferentemente dos que consideram somente aqueles que se situam na diagonal principal da matriz, estimando assim a soma da coluna e linha marginais (Cohen, 1960).

Rosenfield e Fitzpatrick-Lins (1986) apontam o Coeficiente Kappa como uma proporção de acerto após a supressão do acerto casual, isto é, depois que a concordância imposta à casualidade é retirada de consideração, indicando a concordância esperada a posteriori, expresso na Equação 21:

$$k_i = \frac{N.X_{ii} - X_{i+}.X_{+i}}{N.X_{i+} - X_{i+}.X_{+i}}, \tag{21}$$

onde, $N$ é o total da amostragem, $X_{ii}$ são os sinais corretamente classificados e $X_{+i}$ e $X_{i+}$ indicam os erros de omissão e comissão, respectivamente.

Landis e Koch (1977) associam estimações de Coeficiente Kappa à qualidade da classificação, satisfazendo à uma avaliação de integridade que corrige por concordância o acaso, em que, Coeficiente Kappa > 80% é considerado excelente; Coeficiente Kappa entre 60% e 80% é considerado bom; Coeficiente Kappa entre 40% e 60% é considerado regular e Coeficiente Kappa < 40% é considerado ruim. O valor positivo de Coeficiente Kappa indica que o valor observado de concordância é maior que a concordância ao acaso esperada. O valor $k = 1$ incide quando houver total concordância entre os pontos de referência e as classes classificadas (COHEN 1960).

A avaliação da acurácia individual pode ser realizada através da análise dos **erros de comissão**, quando ocorre a inclusão de um objeto na classe à qual ele não pertence e dos **erros de omissão**, quando um objeto é excluído da classe a que pertence (Congalton e Green, 1999). Na matriz de contingência essas estimações são adquiridas através da **exatidão do usuário** ($e_u$), expressa pela razão do número de elementos distribuídos corretamente em uma classe ($X_{ii}$), pelo número total de elementos classificados na mesma ($X_{i+}$). Esta medida reflete os erros de comissão na classificação, indicando a probabilidade de um elemento amostral agrupado em uma determinada classe realmente pertencer à mesma. E a **exatidão do produtor** ($e_p$), expressa pela razão entre o número de elementos classificados corretamente em uma determinada classe ($X_{ii}$), pelo número de elementos de referência amostrados para a mesma classe ($X_{+i}$), refletindo, assim, os erros de omissão (Lillesand e Kiefer, 1994). Estas medidas são expressas como:

$$e_u = \frac{X_{ii}}{X_{i+}} \qquad e \qquad e_p = \frac{X_{ii}}{X_{+i}}, \tag{22, 23}$$

Segundo Brites *et al.* (1996), a Exatidão Global apresenta os maiores valores por considerar somente a diagonal principal da matriz de contingência. Por outro lado o Coeficiente Kappa, ao calcular a concordância casual, abrange os elementos da diagonal principal fazendo com que seja superestimada, reduzindo o valor do coeficiente. Baseado nessa afirmação (constatada durantes os testes), a acurácia final adotada na análise foi obtida pela média aritmética de ambos os índices.

Logo após a matriz de contingência de cada teste, é apresentada uma tabela com formatação condicional de gradiente, numa escala de cores variando do valor mais baixo (amarelo) ao valor mais alto (vermelho), indicando o tempo de processamento, isto é, o custo computacional de reconhecimento para cada uma das dez amostras de cada palavra do grupo[21] analisado durante os testes.

Para uma melhor compreensão da acuracidade dos métodos utilizados, os resultados alcançados podem ser observados nos gráficos cujos valores de omissão, comissão, Exatidão Global e Coeficiente Kappa, são utilizados para comparar os classificadores.

### 5.2.1 Teste com Palavras Isoladas

O processo do funcionamento do teste de reconhecimento de palavras isoladas, está resumidamente descrito na Tabela 2, onde cada palavra analisada foi testada 10 vezes.

**Tabela 2 - Funcionamento do Teste de Reconhecimento de Palavras**

| **Entradas** | | | **Resultados esperados** | |
|---|---|---|---|---|
| ***Nº*** | ***Pré-condições*** | ***Descrição da entrada*** | ***Pós-condições*** | ***Saídas*** |
| 1 | Uma palavra deve ser pronunciada. | É pronunciada a palavra que será processada no STI | Após a análise e classificação da palavra, segundo o método utilizado, é registrado seu tempo de reconhecimento. | O sistema mostra a palavra reconhecida e o tempo necessário para reconhecê-la. |

[21] Por uma questão de praticidade, no corpo deste trabalho, encontram-se apenas as análises referentes à um dos grupos de palavras analisados: o grupo A. As demais análises dos grupos restantes, podem ser verificadas nos relatórios de análise dos apêndices A, B e C.

5.2.1.1 Análise da Classificação de Palavra por Erro Médio

A matriz de contingência na Tabela 3 equivale ao experimento de análise de reconhecimento por erro médio do Grupo A de palavras, pertencentes ao *corpus* do DeVoice. O valor do limiar utilizado para determinar a sensibilidade do classificador quanto ao aceite ou recusa da palavra foi de 25.

**Tabela 3 - Matriz de Contingência da Classificação por Erro Médio do Grupo A**

| Erro Médio (FFT) Grupo A | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Alfabetização | Amor | Brasil | Cachorro | Casa | Computação | Computador | Fome | Gato | Humano | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Alfabetização | 8 | 2 | | | | | | | | | 10 | 80% | 20% | 78% |
| | Amor | | 10 | | | | | | | | | 10 | 100% | 0% | 100% |
| | Brasil | | | 7 | 1 | | | | 2 | | | 10 | 70% | 30% | 68% |
| | Cachorro | | | | 10 | | | | | | | 10 | 100% | 0% | 100% |
| | Casa | | | | | 10 | | | | | | 10 | 100% | 0% | 100% |
| | Computação | | | | | | 7 | 3 | | | | 10 | 70% | 30% | 68% |
| | Computador | | | | | | | 6 | 4 | | | 10 | 60% | 40% | 56% |
| | Fome | | | | | | | | 10 | | | 10 | 100% | 0% | 100% |
| | Gato | | 1 | | | 1 | | | | 8 | | 10 | 80% | 20% | 78% |
| | Humano | | | | | | | | | | 10 | 10 | 100% | 0% | 100% |
| Omissão | $\sum x_{+i}$ | 8 | 13 | 7 | 11 | 11 | 7 | 9 | 16 | 8 | 10 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 77% | 100% | 91% | 91% | 100% | 67% | 63% | 100% | 100% | 89% | 86% | 14% | 85% |
| | Erro | 0% | 23% | 0% | 9% | 9% | 0% | 33% | 38% | 0% | 0% | 11% | Acurácia Final: | | 85% |

Fonte: Elaborada pelo autor.

Como observado na matriz de contigência acima, as palavras Amor, Cachorro, Casa, Fome e Humano, obtiveram 100% de acerto durante os experimentos. A menor taxa de acerto registrada foi a da palavra Computador, atingindo 60%. A acurácia final foi de 85%.

Na Tabela 4, tem-se os tempos de processamento, dado em segundos, das palavras do Grupo A. As células vazias indicam que a amostra não foi reconhecida. No experimento, verificou-se que o maior tempo de processamento (destacado de vermelho) foi o da amostra de número 1 da palavra Computador e os menores tempos (destacados de amarelo) couberam às amostras de números 5 e 9 da palavra Humano.

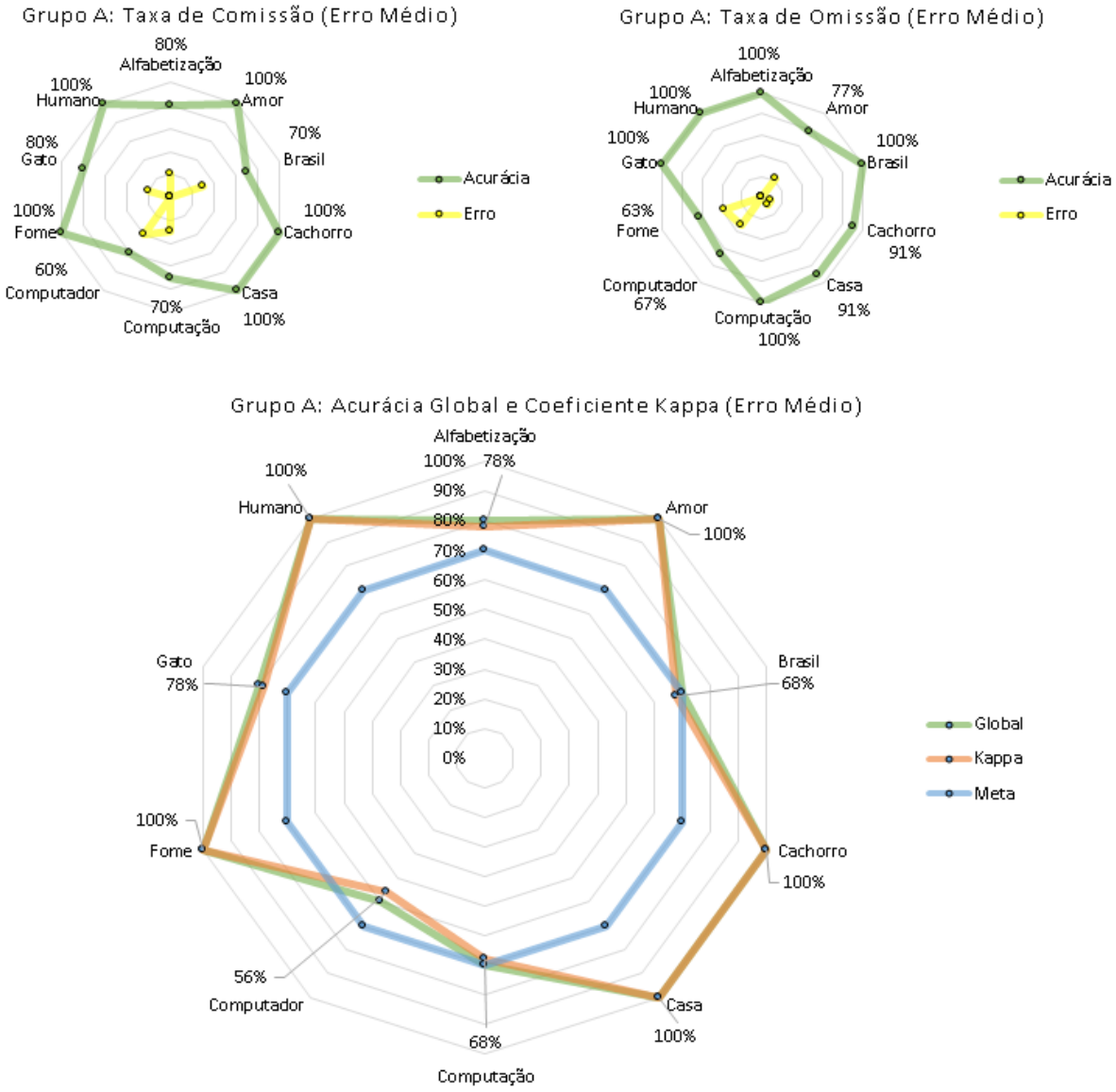
**Tabela 4 - Tempo de Processamento por Erro Médio do Grupo A**

| Erro Médio Grupo A | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Alfabetização | 0,19 | 0,18 | 0,18 | 0,19 | | 0,18 | 0,19 | 0,18 | | 0,18 | 0,18 | 1,84 |
| Amor | 0,18 | 0,18 | 0,18 | 0,17 | 0,16 | 0,18 | 0,18 | 0,18 | 0,17 | 0,18 | 0,17 | |
| Brasil | 0,18 | 0,18 | | 0,18 | 0,17 | 0,18 | 0,18 | | | 0,18 | 0,18 | |
| Cachorro | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,17 | 0,18 | 0,18 | 0,18 | 0,18 | |
| Casa | 0,20 | 0,17 | 0,18 | 0,18 | 0,18 | 0,19 | 0,18 | 0,19 | 0,18 | 0,18 | 0,18 | Tempo Total do Treinamento |
| Computação | 0,21 | 0,21 | 0,21 | 0,21 | 0,21 | | 0,21 | | 0,21 | | 0,21 | |
| Computador | 0,34 | 0,18 | | 0,18 | | 0,18 | | 0,18 | | 0,18 | 0,21 | 0,23 |
| Fome | 0,19 | 0,18 | 0,19 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | |
| Gato | 0,19 | 0,17 | | | 0,20 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | |
| Humano | 0,17 | 0,17 | 0,17 | 0,17 | 0,15 | 0,16 | 0,17 | 0,17 | 0,15 | 0,17 | 0,16 | |

Fonte: Elaborada pelo autor.

O gráfico na Figura 23 foi gerado a partir dos dados contidos na matriz de contingência anteriormente apresentada na Tabela 3. Na parte superior, podem ser observados os níveis das taxas de comissão e omissão e na parte inferior uma sobreposição dos índices utilizados para medir a acurácia da classificação por erro médio, segundo a meta estipulada.

**Figura 23 - Gráfico da Acurácia da Classificação por Erro Médio do Grupo A**




Fonte: Elaborada pelo autor.

Como observado no gráfico da Figura 23, apenas o reconhecimento da palavra Computador ficou abaixo da meta de 70% de acurácia final e os erros de omissão e comissão foram de apenas 11% e 14%, respectivamente, não apresentando índices elevados.

5.2.1.2 Análise da Classificação de Palavra por Desvio Padrão

A matriz de contingência na Tabela 5 equivale ao experimento de análise de reconhecimento por desvio padrão do Grupo A de palavras, pertencentes ao *corpus* do DeVoice. O valor do limiar utilizado para determinar a sensibilidade do classificador quanto ao aceite ou recusa da palavra foi de 75.

**Tabela 5 - Matriz de Contingência da Classificação por Desvio Padrão do Grupo A**

| Desvio Padrão (FFT) Grupo A | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Alfabetização | Amor | Brasil | Cachorro | Casa | Computação | Computador | Fome | Gato | Humano | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Alfabetização | 9 | 1 | | | | | | | | | 10 | 90% | 10% | 89% |
| | Amor | | 8 | | 2 | | | | | | | 10 | 80% | 20% | 77% |
| | Brasil | | | 2 | 5 | 1 | | | 2 | | | 10 | 20% | 80% | 16% |
| | Cachorro | | | | 8 | | | | | 2 | | 10 | 80% | 20% | 75% |
| | Casa | | | | | 9 | | | | 1 | | 10 | 90% | 10% | 89% |
| | Computação | | | 1 | | | 4 | 5 | | | | 10 | 40% | 60% | 34% |
| | Computador | | | 2 | 1 | | 3 | 4 | | | | 10 | 40% | 60% | 34% |
| | Fome | | 1 | | 2 | | | | 7 | | | 10 | 70% | 30% | 67% |
| | Gato | | | | 2 | 1 | | | | 7 | | 10 | 70% | 30% | 67% |
| | Humano | | 4 | | | | 2 | | | | 4 | 10 | 40% | 60% | 38% |
| Omissão | $\sum x_{+i}$ | 9 | 14 | 5 | 20 | 11 | 9 | 9 | 9 | 10 | 4 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 57% | 40% | 40% | 82% | 44% | 44% | 78% | 70% | 100% | 66% | 62% | 38% | 58% |
| | Erro | 0% | 43% | 60% | 60% | 18% | 56% | 56% | 22% | 30% | 0% | 34% | Acurácia Final: | | 60% |

Fonte: Elaborada pelo autor.

Como observado na matriz de contigência acima, as palavras Alfabetização e Casa obtiveram 90% de acerto durante os experimentos. A menor taxa de acerto registrada foi a da palavra Brasil, atingindo somente 20%. A acurácia final foi de 60%.

Na Tabela 6, tem-se os tempos de processamento dado em segundos das palavras do Grupo A. As células vazias indicam que a amostra não foi reconhecida. No experimento, verificou-se que o maior tempo de processamento (destacado de vermelho) foi o da amostra de número 7 da palavra Computação, e o menor tempo (destacado de amarelo) coube ao da amostra de número 7 da palavra Cachorro.

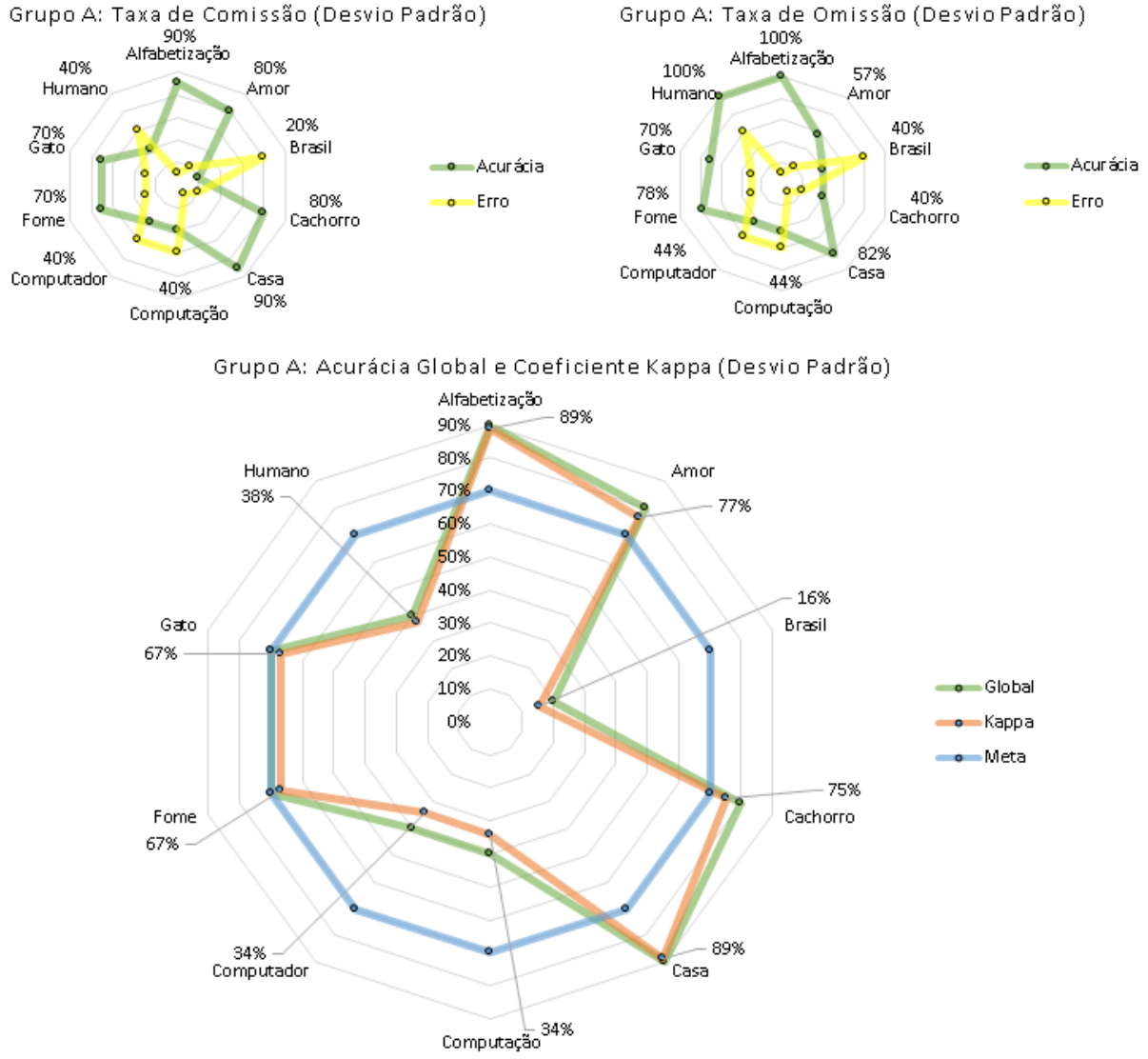
**Tabela 6 - Tempo de Processamento por Desvio Padrão do Grupo A**

| Desvio Padrão Grupo A | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Alfabetização | 0,20 | 0,21 | 0,21 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | | 0,20 | 0,20 | 2,06 |
| Amor | | 0,21 | 0,19 | 0,19 | 0,20 | 0,19 | 0,20 | 0,20 | | 0,20 | 0,20 | |
| Brasil | 0,23 | | | | | | | 0,22 | | | 0,22 | |
| Cachorro | 0,22 | | 0,16 | 0,19 | 0,21 | 0,20 | 0,15 | | 0,20 | 0,20 | 0,19 | |
| Casa | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | | 0,20 | Tempo Total do Treinamento |
| Computação | | 0,23 | 0,22 | | | | 0,24 | | 0,22 | | 0,23 | |
| Computador | | | 0,22 | | | 0,22 | 0,22 | | | 0,22 | 0,22 | 0,23 |
| Fome | | 0,20 | 0,20 | 0,20 | | | 0,20 | 0,20 | 0,21 | 0,20 | 0,20 | |
| Gato | 0,18 | 0,19 | | | 0,20 | 0,20 | | 0,20 | 0,20 | 0,21 | 0,20 | |
| Humano | | | | | 0,19 | 0,19 | 0,19 | | 0,19 | | 0,19 | |

Fonte: Elaborada pelo autor.

O gráfico na Figura 24 foi gerado a partir dos dados contidos na matriz de contingência anteriormente apresentada na Tabela 5. Na parte superior, podem ser observados os níveis das taxas de comissão e omissão e na parte inferior uma sobreposição dos índices utilizados para medir a acurácia da classificação por desvio padrão, segundo a meta estipulada.

**Figura 24 - Gráfico da Acurácia da Classificação por Desvio Padrão do Grupo A**




Fonte: Elaborada pelo autor.

Como observado no gráfico da Figura 24, apenas o reconhecimento das palavras Alfabetização, Amor, Cachorro, Casa, Fome e Gato ficaram acima da meta de 70% de acurácia final e os erros de omissão e comissão foram de 34% e 38%, respectivamente.

5.2.1.3 Análise da Classificação de Palavra por Covariância

A matriz de contingência na Tabela 7 equivale ao experimento de análise de reconhecimento por covariância do Grupo A de palavras, pertencentes ao *corpus* do DeVoice. O valor do limiar utilizado para determinar a sensibilidade do classificador quanto ao aceite ou recusa da palavra foi de 5000.

**Tabela 7 - Matriz de Contingência da Classificação por Covariância do Grupo A**

| Covariância (FFT) Grupo A | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Alfabetização | Amor | Brasil | Cachorro | Casa | Computação | Computador | Fome | Gato | Humano | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Alfabetização | 9 | 1 | | | | | | | | | 10 | 90% | 10% | 89% |
| | Amor | | 8 | | 2 | | | | | | | 10 | 80% | 20% | 77% |
| | Brasil | | | 2 | 5 | 1 | | | 2 | | | 10 | 20% | 80% | 16% |
| | Cachorro | | | | 8 | | | | | 2 | | 10 | 80% | 20% | 75% |
| | Casa | | | | | 9 | | | | 1 | | 10 | 90% | 10% | 89% |
| | Computação | | | 1 | | | 4 | 5 | | | | 10 | 40% | 60% | 34% |
| | Computador | | | 2 | 1 | | 3 | 4 | | | | 10 | 40% | 60% | 34% |
| | Fome | | 1 | | 2 | | | | 7 | | | 10 | 70% | 30% | 67% |
| | Gato | | | | 2 | 1 | | | | 7 | | 10 | 70% | 30% | 67% |
| | Humano | | 4 | | | | 2 | | | | 4 | 10 | 40% | 60% | 38% |
| Omissão | $\sum x_{+i}$ | 9 | 14 | 5 | 20 | 11 | 9 | 9 | 9 | 10 | 4 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 57% | 40% | 40% | 82% | 44% | 44% | 78% | 70% | 100% | 66% | 62% | 38% | 58% |
| | Erro | 0% | 43% | 60% | 60% | 18% | 56% | 56% | 22% | 30% | 0% | 34% | Acurácia Final: | | 60% |

Fonte: Elaborada pelo autor.

Como observado na matriz de contigência acima, as palavras Alfabetização e Casa obtiveram 90% de acerto durante os experimentos. A menor taxa de certo registrada foi a da palavra Brasil, atingindo somente 20%. A acurácia final foi de 60%.

Na Tabela 8, tem-se os tempos de processamento, dado em segundos, das palavras do Grupo A. As células vazias indicam que a amostra não foi reconhecida. No experimento, verificou-se que os maiores tempos de processamento (destacado de vermelho) foram das amostras de número 1 da palavra Brasil; 3, 7 e 9 da Palavra Computação e 3 da palavra Computador. Já o menor tempo (destacado de amarelo), coube ao da amostra de número 6 da palavra Humano.

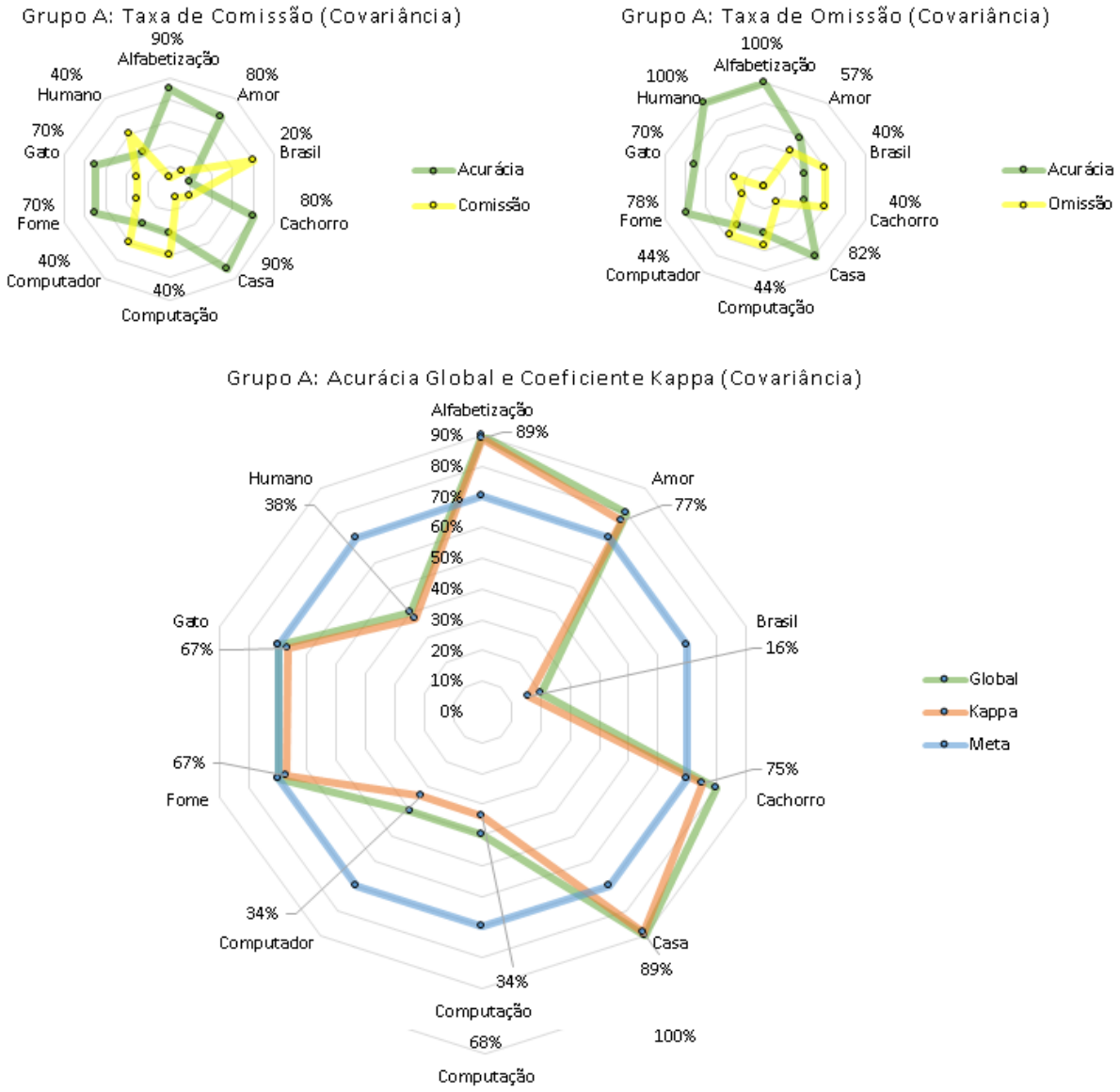
**Tabela 8 - Tempo de Processamento por Covariância do Grupo A**

| Covariância Grupo A | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Alfabetização | 0,21 | 0,21 | 0,22 | 0,21 | 0,20 | 0,20 | 0,22 | 0,21 | | 0,21 | 0,21 | 2,10 |
| Amor | | 0,21 | 0,19 | 0,22 | 0,21 | 0,19 | 0,20 | 0,21 | | 0,22 | 0,21 | |
| Brasil | 0,23 | | | | | | | 0,22 | | | 0,23 | |
| Cachorro | 0,21 | | 0,21 | 0,21 | 0,21 | 0,21 | 0,20 | | 0,20 | 0,22 | 0,21 | |
| Casa | 0,21 | 0,21 | 0,21 | 0,20 | 0,21 | 0,20 | 0,21 | 0,21 | 0,20 | | 0,21 | Tempo Total do Treinamento |
| Computação | | 0,22 | 0,23 | | | | 0,23 | | 0,23 | | 0,23 | |
| Computador | | | 0,23 | | | 0,22 | 0,22 | | | 0,22 | 0,22 | 0,23 |
| Fome | | 0,20 | 0,20 | 0,21 | | | 0,21 | 0,19 | 0,21 | 0,20 | 0,20 | |
| Gato | 0,19 | 0,19 | | | 0,20 | 0,20 | | 0,20 | 0,20 | 0,22 | 0,20 | |
| Humano | | | | | 0,19 | 0,18 | 0,19 | 0,19 | | | 0,19 | |

Fonte: Elaborada pelo autor.

O gráfico na Figura 25 foi gerado a partir dos dados contidos na matriz de contingência anteriormente apresentada na Tabela 7. Na parte superior, podem ser observados os níveis das taxas de comissão e omissão e na parte inferior uma sobreposição dos índices utilizados para medir a acurácia da classificação por covariância, segundo a meta estipulada.

**Figura 25 - Gráfico da Acurácia da Classificação por Covariância do Grupo A**




Fonte: Elaborada pelo autor.

Como observado no gráfico da Figura 25, apenas o reconhecimento das palavras Alfabetização, Amor, Cachorro, Casa, Fome e Gato ficaram acima da meta de 70% de acurácia final e os erros de omissão e comissão foram de 34% e 38%, respectivamente.

5.2.1.4 Análise da Classificação de Palavra por DTW

Como último método de classificação, foi aplicado um alinhamento dinâmico no sinal por meio do DTW do tipo 2 que leva em consideração as direções 0º, 45º e 90º. A matriz de contingência na Tabela 9, equivale ao experimento de análise da classificação por DTW do Grupo A de palavras, pertencentes ao *corpus*. O valor do limiar de sensibilidade foi de 10.

**Tabela 9 - Matriz de Contingência da Classificação DTW do Grupo A**

| DTW (STFT) Grupo A | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Alfabetização | Amor | Brasil | Cachorro | Casa | Computação | Computador | Fome | Gato | Humano | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Alfabetização | 8 | 1 | | | | 1 | | | | | 10 | 80% | 20% | 77% |
| | Amor | | 10 | | | | | | | | | 10 | 100% | 0% | 100% |
| | Brasil | | | 6 | | | | | 2 | | 2 | 10 | 60% | 40% | 57% |
| | Cachorro | | | | 5 | | | | 5 | | | 10 | 50% | 50% | 44% |
| | Casa | 2 | 3 | | 1 | 2 | 2 | | | | | 10 | 20% | 80% | 18% |
| | Computação | | | | | | 10 | | | | | 10 | 100% | 0% | 100% |
| | Computador | 1 | | | 1 | | | 6 | 2 | | | 10 | 60% | 40% | 57% |
| | Fome | | | | | | | | 6 | | 4 | 10 | 60% | 40% | 52% |
| | Gato | 1 | | 1 | 2 | | | | 2 | 4 | | 10 | 40% | 60% | 38% |
| | Humano | | | | 1 | | 1 | | | | 8 | 10 | 80% | 20% | 77% |
| Omissão | $\sum x_{+i}$ | 12 | 14 | 7 | 10 | 2 | 14 | 6 | 17 | 4 | 14 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 67% | 71% | 86% | 50% | #### | 71% | 100% | 35% | 100% | 57% | 74% | 65% | 35% | 62% |
| | Erro | 33% | 29% | 14% | 50% | 0% | 29% | 0% | 65% | 0% | 43% | 26% | Acurácia Final: | | 64% |

Fonte: Elaborada pelo autor.

Como observado na matriz de contigência acima, as palavras Amor e Computação obtiveram 100% de acerto durante os experimentos. A menor taxa de acerto registrada foi a da palavra Casa, alcançando apenas 20%. A acurácia final foi de 64%.

Na Tabela 10, tem-se os tempos de processamento dado em segundos das palavras do Grupo A. As células vazias indicam que a amostra não foi reconhecida. No experimento, verificou-se que o maior tempo de processamento (destacado de vermelho) foi o da amostra de número 1 da palavra Computador, e o menor tempo (destacado de amarelo) coube ao da amostra de número 7 da palavra Computação.

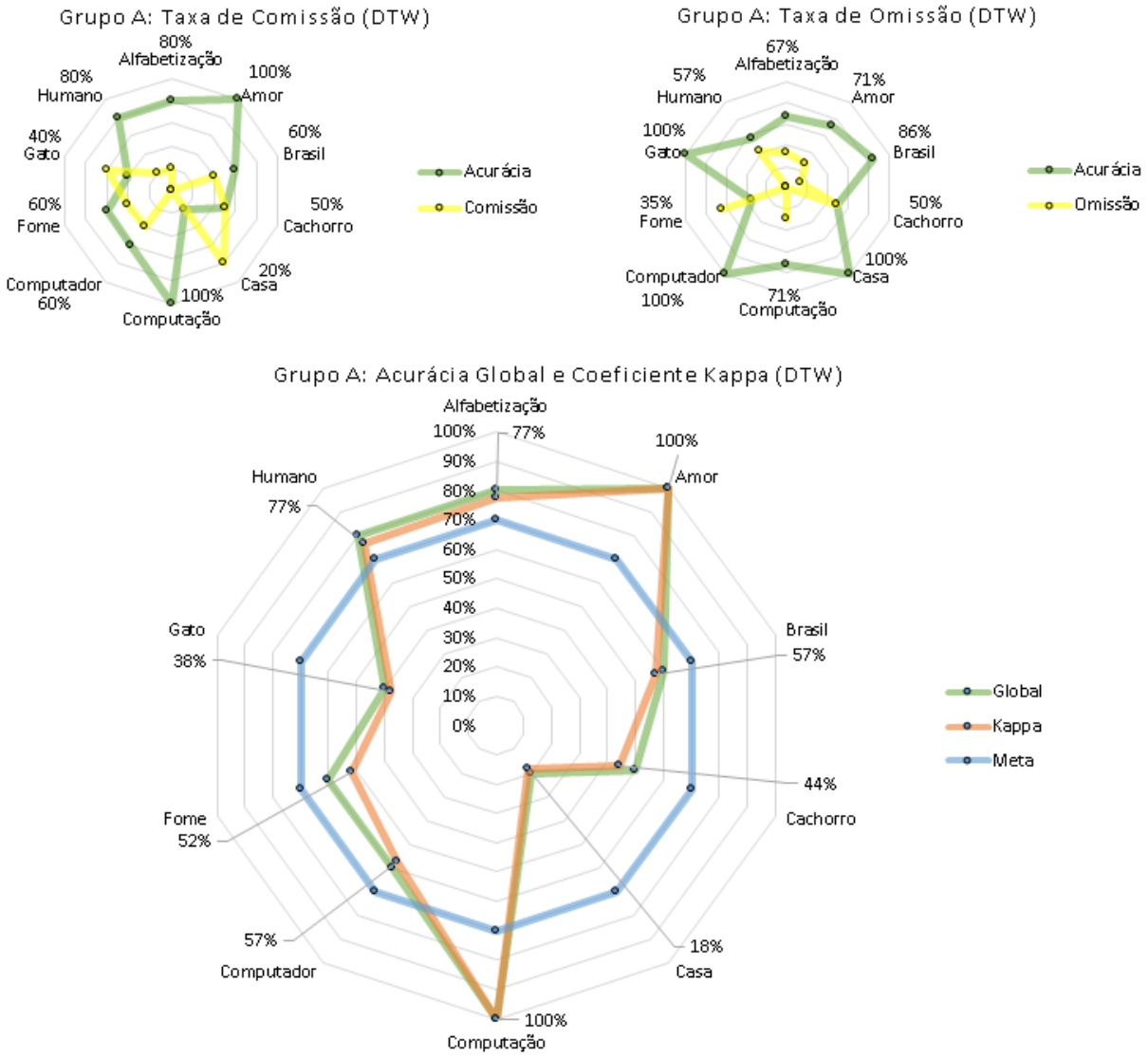
**Tabela 10 - Tempo de Processamento por DTW do Grupo A**

| DTW Grupo A | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Alfabetização | 0,32 | 0,39 | 0,30 | 0,31 | | 0,30 | 0,31 | 0,32 | | 0,37 | 0,33 | 3,55 |
| Amor | 0,30 | 0,36 | 0,33 | 0,37 | 0,34 | 0,39 | 0,34 | 0,34 | 0,35 | 0,40 | 0,35 | |
| Brasil | 0,35 | 0,30 | 0,31 | | | 0,29 | | | 0,30 | 0,35 | 0,32 | |
| Cachorro | 0,38 | | 0,40 | | | | | 0,35 | 0,40 | 0,39 | 0,38 | |
| Casa | 0,35 | | | | | | | | | 0,34 | 0,34 | Tempo Total do Treinamento |
| Computação | 0,37 | 0,32 | 0,32 | 0,30 | 0,30 | 0,32 | 0,28 | 0,34 | 0,29 | 0,29 | 0,31 | |
| Computador | 0,71 | 0,35 | | 0,33 | | 0,38 | | 0,32 | | 0,35 | 0,41 | 0,54 |
| Fome | | | 0,38 | | 0,39 | 0,35 | 0,37 | 0,38 | 0,37 | | 0,37 | |
| Gato | | 0,37 | | | | 0,37 | | 0,37 | 0,38 | | 0,37 | |
| Humano | 0,36 | 0,34 | 0,36 | | 0,37 | 0,36 | 0,37 | 0,38 | 0,36 | | 0,36 | |

Fonte: Elaborada pelo autor.

O gráfico na Figura 26 foi gerado a partir dos dados contidos na matriz de contingência anteriormente apresentada na Tabela 9. Na parte superior, podem ser observados os níveis das taxas de comissão e omissão, e na parte inferior uma sobreposição dos índices utilizados para medir a acurácia da classificação segundo a meta estipulada.

**Figura 26 - Gráfico da Acurácia da Classificação DTW do Grupo A**



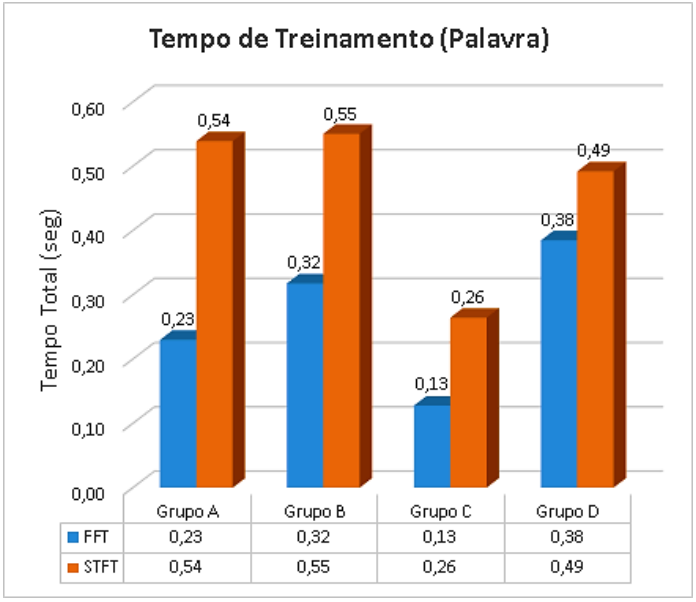

Fonte: Elaborada pelo autor.

Como observado no gráfico da Figura 26, apenas o reconhecimento das palavras Alfabetização, Amor, Computação e Humano ficaram acima da meta de 70% de acurácia final e os erros de omissão e comissão foram de 26% e 35%, respectivamente.

### 5.2.1.5 Comparativo do Custo Computacional

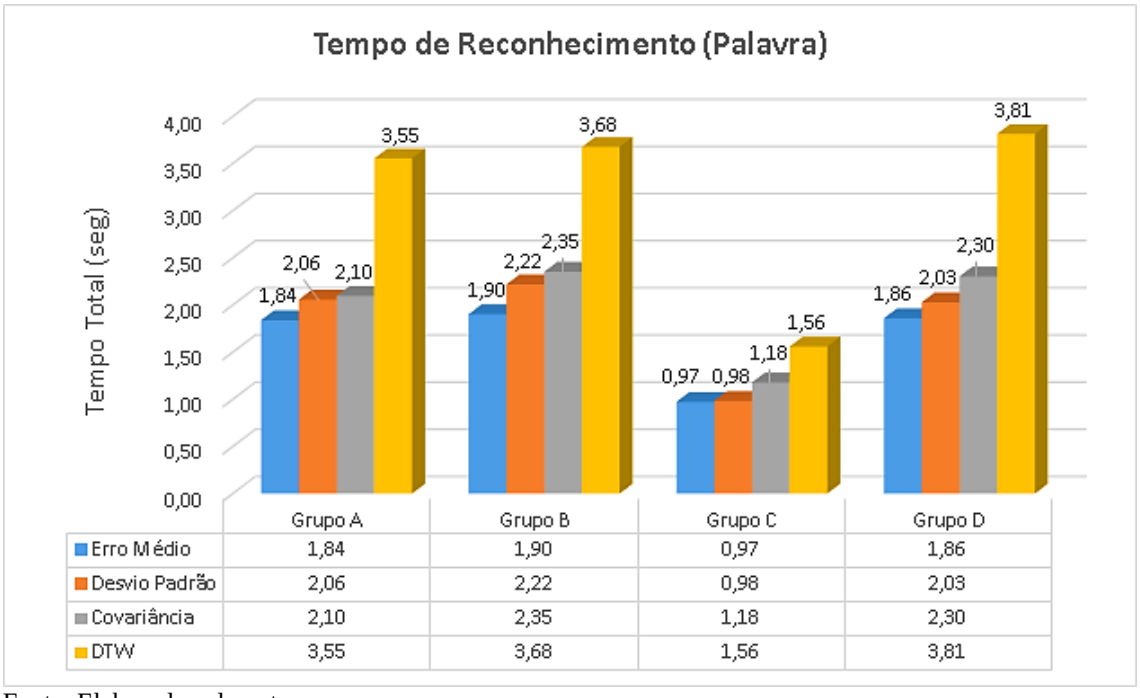
Objetivando comparar o custo computacional necessário para o treinamento e reconhecimento de cada grupo de palavra, segundo o método utilizado, foram gerados os gráficos das figuras 27 e 28. Analisando o gráfico, percebe-se que em termos de desempenho computacional, o STFT apresenta-se menos eficiente, visto que exigiu quase que o dobro do tempo para completar seu processamento, quando comparado à FFT.

**Figura 27 - Custo Computacional de Treinamento das Palavras**




Fonte: Elaborada pelo autor.

**Figura 28 - Custo Computacional de Reconhecimento de Palavras**



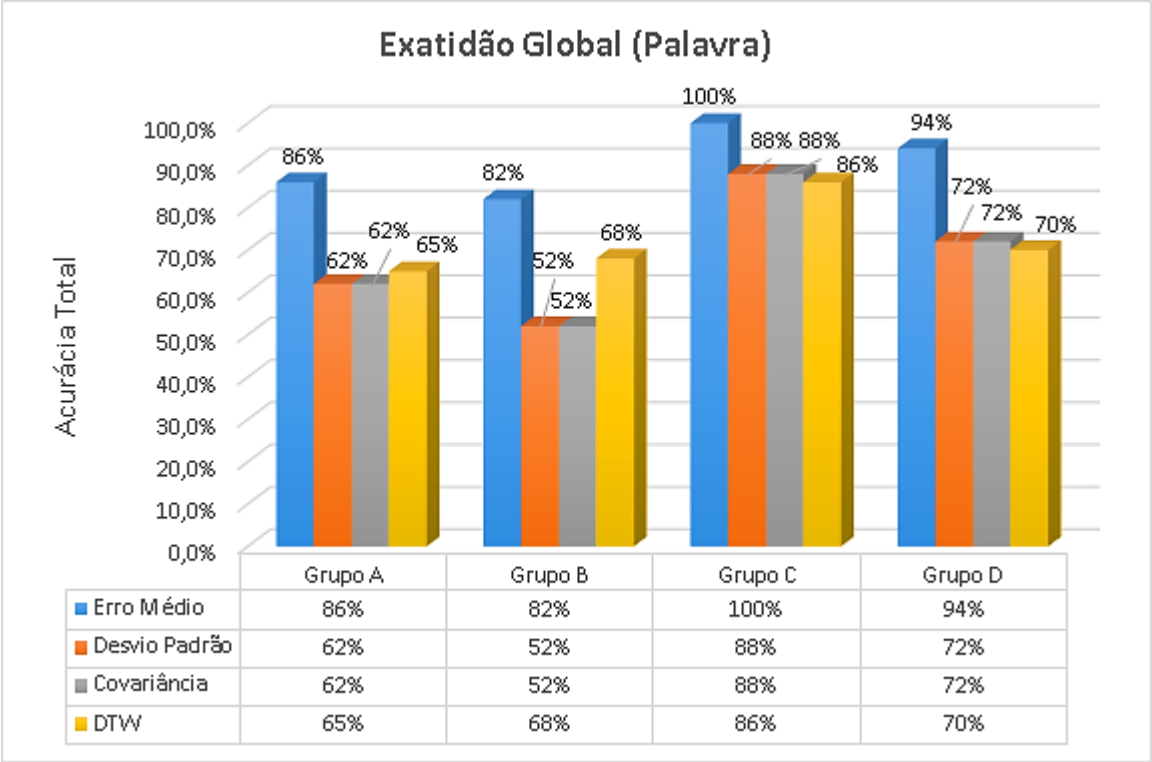

Fonte: Elaborada pelo autor.

Como observado na figuras 27 e 28, para este teste específico, o classificador por erro médio mostrou-se superior aos demais classificadores implementados, consumindo apenas 0,19 segundos em média para processar as locuções. Por outro lado, assim como aconteceu com o custo computacional para treinamento do banco, o classificador por DTW apresentou desempenho inferior, consumindo 0,35 segundos em média.

### 5.2.1.6 Comparativo da Acurácia de Reconhecimento de Palavras

O nível de Exatidão Global de reconhecimento e o Coeficiente Kappa de cada grupo de palavras segundo o método utilizado na classificação podem ser claramente visualizados nos gráficos das figuras 29 e 30 a seguir.

**Figura 29 - Comparação de Exatidão Global dos Classificadores**




| | Grupo A | Grupo B | Grupo C | Grupo D |
|---|---|---|---|---|
| Erro Médio | 86% | 82% | 100% | 94% |
| Desvio Padrão | 62% | 52% | 88% | 72% |
| Covariância | 62% | 52% | 88% | 72% |
| DTW | 65% | 68% | 86% | 70% |

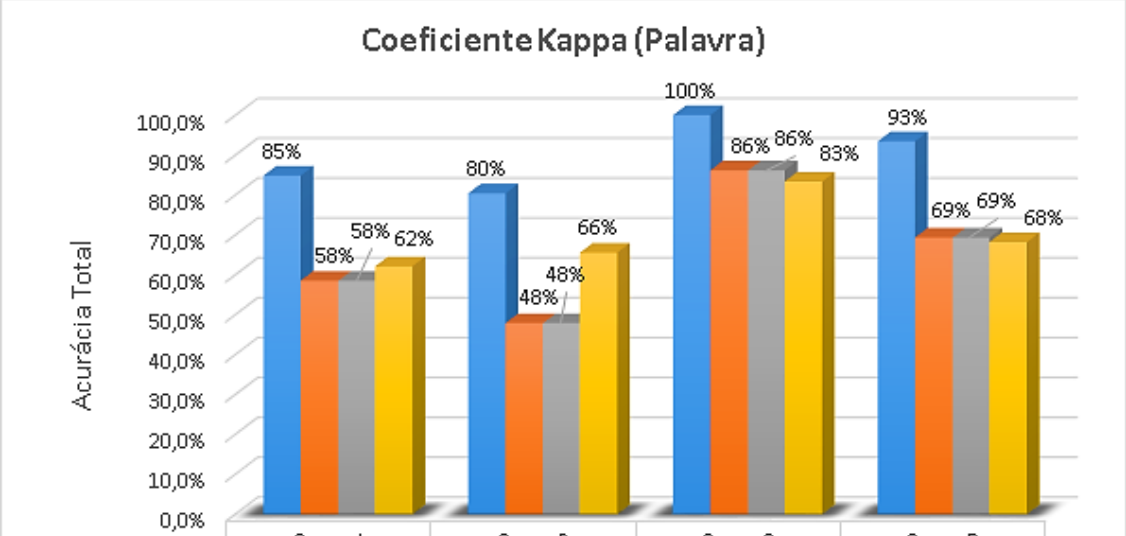
Fonte: Elaborada pelo autor.

Como observado na Figura 29, o classificador por erro médio alcançou níveis de Exatidão Global superiores aos demais classificadores analisados, ultrapassando a meta estipulada nos quatro grupos de palavras e atingindo em média 90,5% de Exatidão Global. Os classificadores por desvio padrão e covariância alcançaram níveis de acurácia muito próximos da meta, ultrapassando-a em dois dos quatro grupos de palavras analisados e atingindo em média 68,5% de Exatidão Global.

Os maiores níveis de acurácia ocorreram na análise do Grupo C, no qual a classificação por erro médio alcançou 100% de acerto e a classificação por DTW apresentou acurácia inferior aos demais classificadores, atingindo 86% de Exatidão Global. Os menores níveis de acurácia ocorreram na análise do Grupo B, no qual a classificação por erro médio alcançou 82% de acerto, a classificação por DTW apresentou 68% e os classificadores por desvio padrão e covariância apenas 52% cada.

Observando a Figura 30, nota-se que o classificador por erro médio alcançou valores de Coeficiente Kappa superiores aos demais classificadores, ultrapassando a meta estipulada nos quatro grupos de palavras e atingindo em média 89,5%. Os classificadores por desvio padrão e covariância alcançaram baixos níveis, ultrapassado a meta em apenas um dos quatro grupos de palavras analisados e atingindo em média 65,25%.

**Figura 30 - Comparação de Integridade dos Classificadores**




| | Grupo A | Grupo B | Grupo C | Grupo D |
|---|---|---|---|---|
| Erro Médio | 85% | 80% | 100% | 93% |
| Desvio Padrão | 58% | 48% | 86% | 69% |
| Covariância | 58% | 48% | 86% | 69% |
| DTW | 62% | 66% | 83% | 68% |

Fonte: Elaborada pelo autor.

Os maiores valores de Coeficiente Kappa ocorreram no Grupo C, cuja classificação por erro médio alcançou 100% e a classificação por DTW apresentou índice inferior aos demais classificadores, atingindo 83%. Diferentemente da análise de Exatidão Global, quando analisado os Coeficientes Kappa dos classificadores, a classificação por DTW apresentou índices inferiores em dois dos quatro grupos de palavras analisadas. Os menores valores de Coeficiente Kappa ocorreram na análise do Grupo B, no qual a classificação por erro médio alcançou 80%, a classificação por DTW apresentou 66% e os classificadores por desvio padrão e covariância somente 48% cada.

### 5.2.2 Teste com Palavras Concatenadas (frases)

No teste de reconhecimento de frases o processamento ocorreu a cada palavra pronunciada, tal processo está resumidamente descrito na Tabela 11, onde cada frase analisada foi testada 10 vezes.

**Tabela 11 - Funcionamento do Teste de Reconhecimento de Frases**

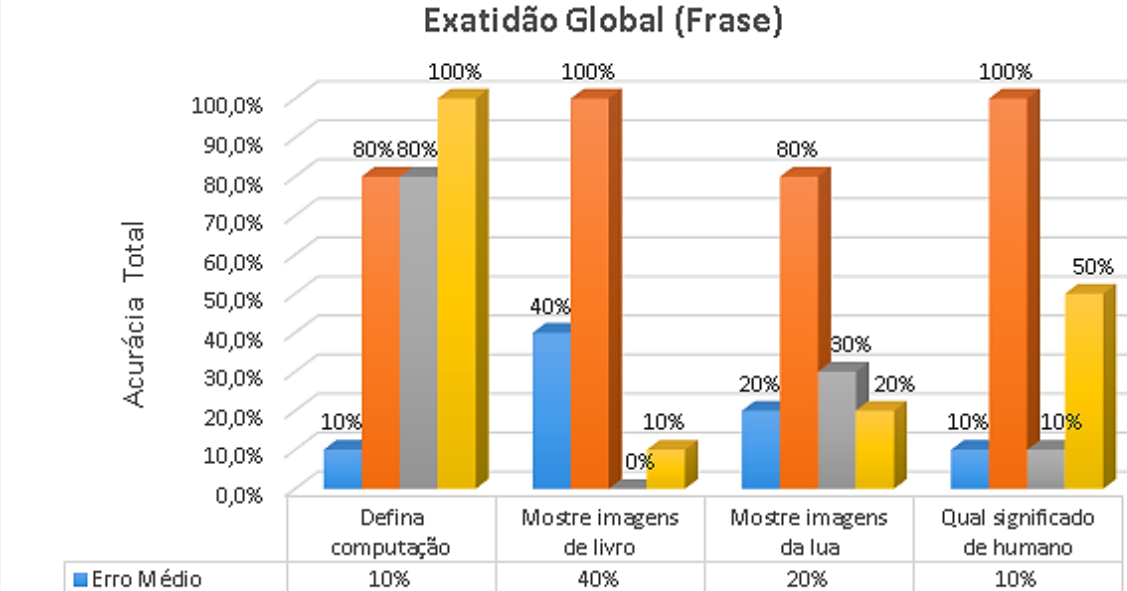
| **Entradas** | | | **Resultados esperados** | |
|---|---|---|---|---|
| ***Nº*** | ***Pré-condições*** | ***Descrição da entrada*** | ***Pós-condições*** | ***Saídas*** |
| 1 | Uma palavra chave deve ser pronunciada. | É pronunciada a palavra-chave '*Álex*', que isoladamente, será processada pelo STI, dando início ao processo de formação da sentença. | Após a análise e classificação da palavra chave, é registrado seu tempo de reconhecimento. | O STI exibe o tempo necessário para o processamento da palavra-chave. |
| 2 | Uma 1ª palavra que irá compor a sentença deve ser pronunciada. | É pronunciada uma palavra, que isoladamente, será processada pelo STI. | Após a análise e classificação da palavra, é registrado seu tempo de reconhecimento. | O STI exibe o tempo necessário para o processamento da palavra. |
| 3 | Uma 2ª palavra que irá compor a sentença deve ser pronunciada. | É pronunciada uma palavra, que isoladamente, será processada pelo STI. | Após a análise e classificação da palavra, é registrado seu tempo de reconhecimento. | O STI exibe o tempo necessário para o processamento da palavra. |
| 4 | Dependendo da 2ª palavra, uma 3ª palavra que irá compor a sentença deverá ser pronunciada. | Caso a 2ª palavra não seja uma palavra terminal, então é pronunciada uma 3ª palavra, que isoladamente, será processada pelo STI. | Após a análise e classificação da palavra, é registrado seu tempo de reconhecimento. | O STI exibe o tempo necessário para o processamento da palavra. |
| 5 | Dependendo da 2ª palavra, uma 4ª palavra que irá compor a sentença deverá ser pronunciada. | Caso a 2ª palavra não seja uma palavra terminal e a 3ª palavra tenha sido pronunciada, então é pronunciada uma 4ª palavra, que isoladamente, será processada pelo STI. | Após a análise e classificação da palavra, é registrado seu tempo de reconhecimento. | O STI exibe o tempo necessário para o processamento da palavra. |

Fonte: Elaborada pelo autor.

### 5.2.2.1 Comparativo da Acurácia de Reconhecimento de Frases

Objetivando comparar visualmente os níveis de acurácia alcançados no reconhecimento de cada frase[22], segundo o método de classificação utilizado, foi gerado o gráfico da Figura 31.

**Figura 31 - Acurácia de Reconhecimento de Frases**




| | Defina computação | Mostre imagens de livro | Mostre imagens da lua | Qual significado de humano |
|---|---|---|---|---|
| Erro Médio | 10% | 40% | 20% | 10% |
| Erro Médio + HMM | 80% | 100% | 80% | 100% |
| DTW | 80% | 0% | 30% | 10% |
| DTW + HMM | 100% | 10% | 20% | 50% |

Fonte: Elaborada pelo autor.

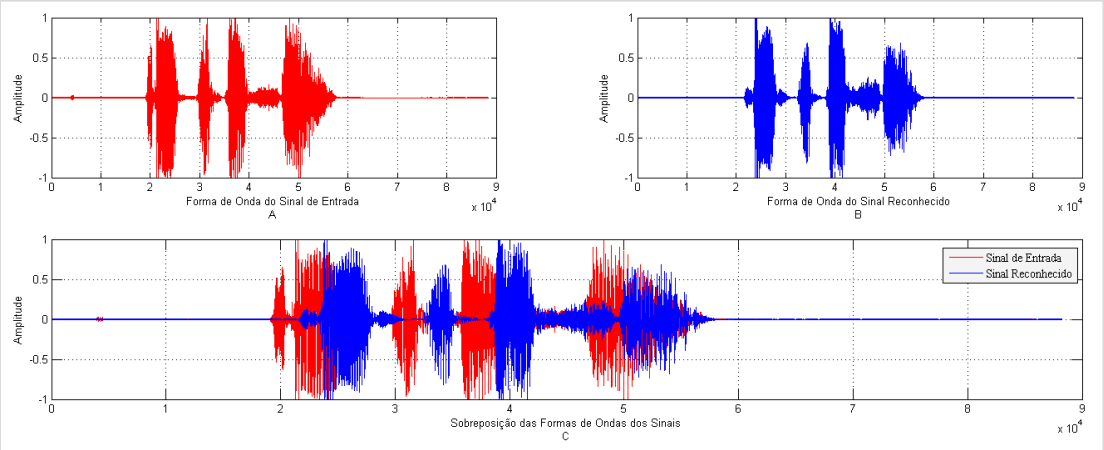
Como observado no gráfico acima, a configuração do classificador por erro médio + HMM alcançou níveis de Exatidão Global superiores às demais, atingindo níveis mínimos de 80% e alcançando 100% de acurácia em dois dos quatro testes realizados, atingindo em média 90% de Exatidão Global no reconhecimento das frases. A classificação por DTW apresentou acurácia inferior às demais configurações, atingindo taxas baixíssimas, na qual em apenas um dos casos, apresentou resultado satisfatório com 80% de Exatidão Global, não chegando à 50% nos três casos de testes restantes. Percebe-se, então, que fazendo uso das HMM's na classificação, o desempenho destes classificadores alcançaram uma acentuada melhora em comparação com as classificações obtidas sem a utilização das Cadeias de Markov, isto foi possível graças à restrição sintática estabelecida pelas HMM's durante a formação das sentenças a serem analisadas.

[22] O teste de reconhecimento de frases utilizando HMM, baseou-se na máquina de estados de reconhecimento de sentenças apresentada anteriormente na Figura 12 da página 32.

### 5.2.3 Análise de Sinais no DeVoice

Nas figuras que se seguem, têm-se as plotagens referentes à locução da palavra **COMPUTAÇÃO** de um sinal de referência e seu respectivo padrão reconhecido, ambos plotados não somente em planos distintos, mas também em um mesmo plano unidimensional, permitindo assim, uma melhor interpretação e visualização sobre as mais variadas perspectivas utilizadas na análise e comparação de sinais no DeVoice. Na Figura 32, tem-se as amplitudes do sinal de entrada (32.A) e do sinal reconhecido (32.B), respectivamente, além da sobreposição (32.C) dos sinais analisados no domínio do tempo.

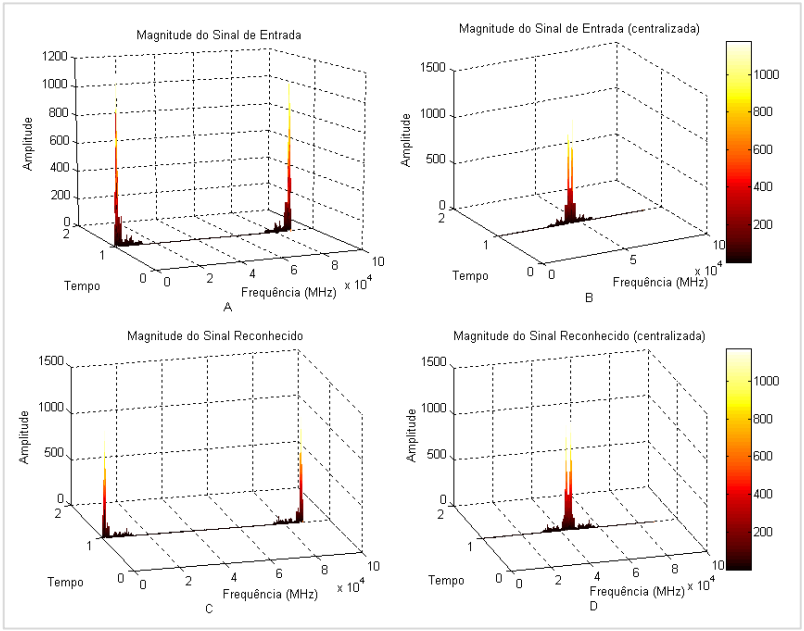
**Figura 32 - Sobreposição das Formas de Onda Geradas no DeVoice**




Fonte: Elaborada pelo autor.

Na Figura 33, estão representados as amplitudes dos coeficientes em função das frequências. Cada elemento do vetor de frequências é um componente da série de Fourier.

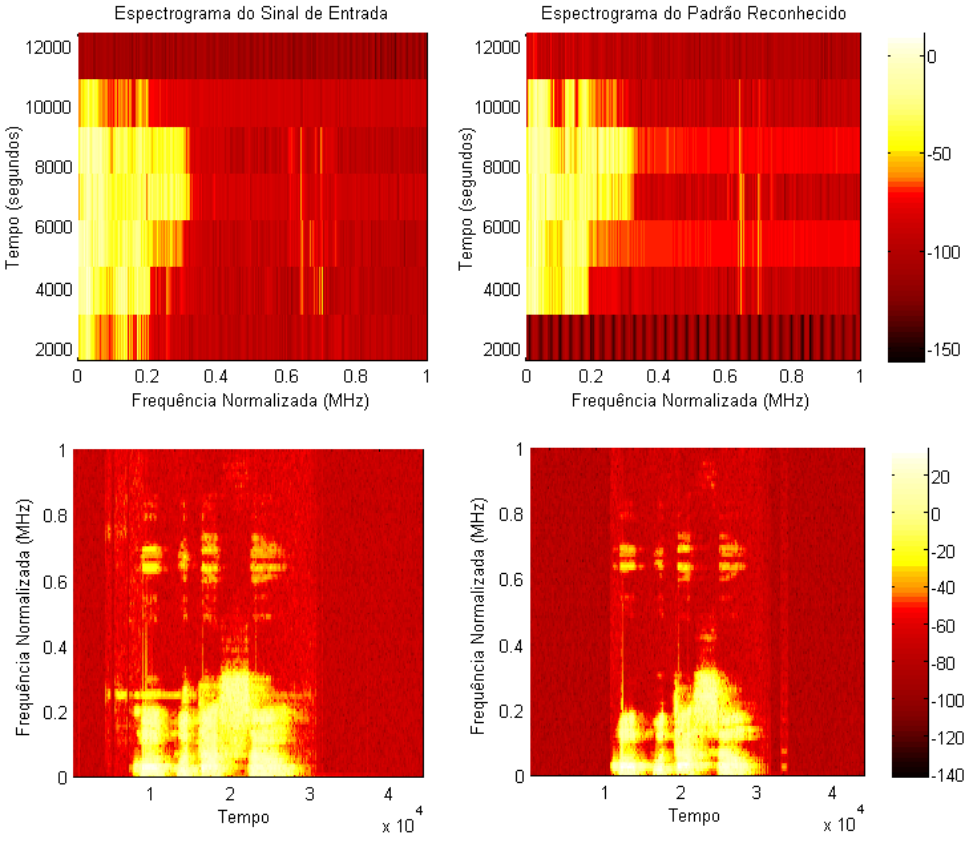
**Figura 33 - Amplitude de Dois Sinais Obtida pela FFT no DeVoice**




Fonte: Elaborada pelo autor.

A Figura 34 exibe dois espectrogramas obtidos a partir da STFT de dois sinais dados como iguais durante uma análise no DeVoice. Um espectrograma é um gráfico em que a altura de cada ponto é representada por uma cor ou tonalidade diferente. Na figura, as marcações escuras (vermelho) representam partes do sinal de voz em que a fala não é produzida, e as partes claras (amarelo) representam a intensidade do sinal de voz produzido.

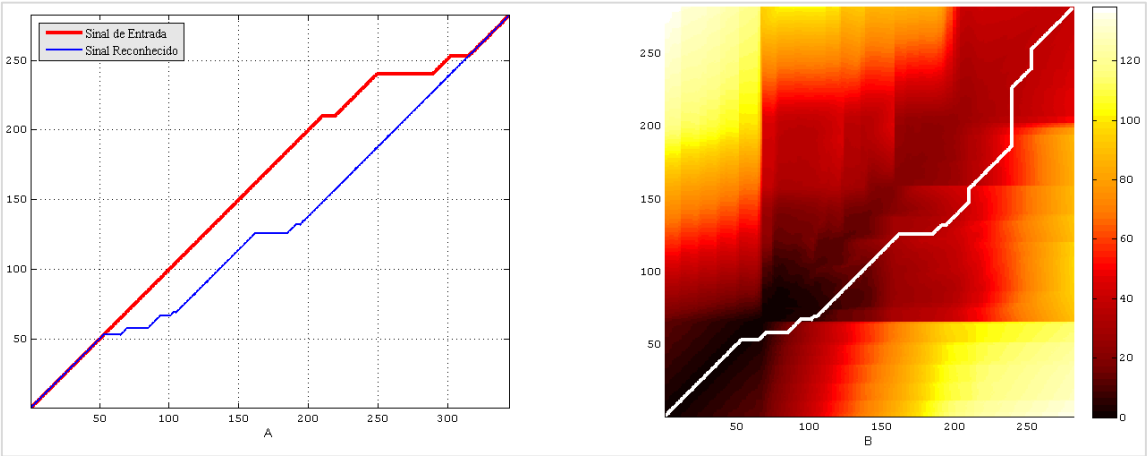
**Figura 34 - Representação Espectral Gerada pela STFT no DeVoice**




Fonte: Elaborada pelo autor.

Na Figura 35, é possível observar os sinais originais desalinhados (35.A) e uma sobreposição do melhor caminho (35.B) resultante do alinhamento temporal das séries.

**Figura 35 - *Bestpath* de dois Sinais no Classificador DTW do DeVoice**




Fonte: Elaborada pelo autor.

# 6 CONSIDERAÇÕES FINAIS

Neste capítulo, serão feitas as considerações finais desta dissertação, onde são apresentadas as principais contribuições da pesquisa, bem como algumas limitações. Em seguida, são apontadas algumas direções futuras, visando trabalhos posteriores. Finalmente, é apresentada a conclusão do trabalho.

## 6.1 CONTRIBUIÇÕES E LIMITAÇÕES DA PESQUISA

Como contribuições principais deste trabalho, podem-se citar:

- ✓ Nesta dissertação foi avaliado o desempenho do mecanismo utilizado no reconhecimento dos sinais da fala em português por meio de uma análise de custo computacional e estimativa do nível de acurácia para cada grupo de palavras que constituem a base.
- ✓ Este trabalho contribuiu com a disponibilização[23] de uma ferramenta capaz de "dialogar" com o usuário devido à sua característica de tornar a interação mais natural através do reconhecimento automático da fala em português e da geração de voz por meio de síntese;
- ✓ Fornecimento de uma pesquisa descritiva e aplicação dos conceitos e teorias envolvidas no processo de reconhecimento da voz, observando todas as etapas, desde a gravação e transcrição da base de dados, aquisição, treinamento e reconhecimento do sinal;
- ✓ Confecção e disponibilização da base de locuções de palavras utilizadas nos experimentos.

Como limitações pode-se citar:

- A habilidade de aprendizagem de novas palavras por parte do STI fica sujeita à informação desta pelo usuário via teclado, ou seja, existe uma vulnerabilidade a erros sintáticos e há possibilidade de que mesmo reconhecendo a nova palavra, o STI não "saiba" o seu significado, em virtude da disponibilização ou não desta na sua base de conhecimento;

[23] A interface projetada, a base de locuções, o código fonte, além de vídeos demonstrativos da utilização do STI DeVoice, encontram-se disponíveis em https://sites.google.com/site/leinylson/home/projetos/devoice.

- Faz-se necessário uma base de locuções padronizada em Português do Brasil, que propicie confrontar metodologias e resultados. Um trabalho pioneiro foi feito por Ynoguti (1999), porém, torna-se sem utilidade para os fins desta pesquisa, em virtude da forma como foram tratadas as palavras, uma vez que, nessa dissertação, optou-se por ter a palavra como unidade fundamental e não fonemas.

- A característica interdisciplinar de um sistema de reconhecimento de voz, a larga variedade de conhecimentos necessários e sua complexidade de implementação, implicam no fato de que, para o desenvolvimento de um STI apto ao reconhecimento e interpretação perfeita do diálogo humano, torna-se indispensável o empenho conjunto de uma equipe de pesquisadores com ciência nas mais distintas áreas abrangidas, além de investimentos na aquisição de equipamentos de qualidade.

## 6.2 TRABALHOS FUTUROS

O aperfeiçoamento e a continuidade, além das próprias limitações apresentadas nesta dissertação, constituem oportunidades para trabalhos futuros, que incluem:

- Utilização de outros métodos para extração, como o *Mel-Frequency Cepstral Coefficient* (MFCC) e para a mensuração do vetor de características, variações do DTW e a wavelet Daubechies de 10 níveis, buscando sempre a manutenção de níveis computacionais satisfatórios;

- Adaptação e ampliação da base de dados para diferentes ambientes e locutores;

- Utilizar diferentes configurações nos reconhecedores, como por exemplo, mudando o número de estados nas HMM’s;

- Aumentar o número de amostras de treinamento, visando a uma obtenção de amostras mais homogêneas e representativas;

- Realização de uma análise mais minuciosa dos valores escolhidos como limiares de aceite ou recusa da locução;

- Anseia-se aprimorar a imersão do usuário no ambiente do STI por meio da modelagem de um avatar em duas dimensões com aspecto tridimensional, humanizando a interação e aumentando as percepções e ações do AS através da renderização de imagens.

- Pretende-se utilizar os mecanismos de reconhecimento de voz e de síntese, para outros propósitos[24], tais como: simular um ambiente inteligente (Domótica) para acionamento de dispositivos eletrônicos; cadastro e identificação de usuário por meio de pronúncia de senha; controle de um veículo robô construído sobre a plataforma Arduíno e uma assistente virtual capaz de executar músicas e vídeos online, informar a previsão do tempo, notícias e horário local, imprimir documentos, acionar webcam, teclado virtual, enviar e-mail, dentre outras tarefas.

## 6.3 CONCLUSÃO

Com relação ao desempenho dos classificadores, verificou-se que o classificador por erro médio apresentou melhor acurácia de uma forma geral, mostrando-se uma abordagem simples, rápida e eficiente, tanto para o reconhecimento de palavras isoladas, bem como de palavras conectadas, quando combinado com as HMM's, obtendo taxa de acerto bem superior aos demais classificadores analisados. Em todos os testes, a classificação por desvio padrão e covariância apresentaram resultados idênticos, diferenciando-se apenas quanto ao custo computacional.

A classificação com menor tempo de processamento foi realizada pelo classificador por erro médio, consumindo apenas 0,19 segundos em média. Por outro lado, a classificação por DTW apresentou desempenho inferior aos demais classificadores, tanto para o treinamento como para o reconhecimento das locuções, chegando a 0,35 segundos em média. Pela observação dos tempos individuais de processamento de cada palavra, nota-se que os tempos médios são inferiores a 0,5 segundos, logo, mesmo a transformada de Fourier sendo de alta complexidade computacional, isso não afetou de forma considerável o desempenho do DeVoice quanto à agilidade do processo de reconhecimento de palavras.

O STI apresentou resultados satisfatórios quanto ao seu mecanismo de reconhecimento automático de voz, como exposto nos casos de uso e nas análises. O enfoque adotado na dissertação demonstrou sua eficácia e resultou não só em um bom embasamento teórico, mas prático, possibilitando o desenvolvimento de uma plataforma inicial sobre a qual pesquisas posteriores possam ser mais facilmente realizadas, visto a possibilidade da geração

[24] *Screenshots* das demais interfaces projetadas e imagens do carro utilizado nos experimentos até a presente data, podem ser visualizadas em https://sites.google.com/site/leinylson/home/projetos/devoice. Todas as funcionalidades foram testadas, mas ainda não passaram por validação.

de gráficos no STI para uma análise comparativa mais profunda dos espectros, magnitudes e similaridades entre os sinais, como ilustrado nas figuras 32, 33, 34 e 35.

Analisando a base de locuções criada e disponibilizada, foi possível concluir que as palavras de menor porte foram melhor classificadas quando realizou-se a classificação por erro médio, desvio padrão e covariância. Por outro lado, os sinais correspondentes às locuções de maior tamanho foram melhor reconhecidas pelo classificador DTW, que por sinal, obteve o maior custo computacional tanto para o treinamento como para o reconhecimento das locuções, isto se deve em grande parte ao tempo e memória necessários para a gerar e processar os espectrogramas, bem como computar o caminho de menor custo na matriz.

Verificou-se também que, a forma como foi realizada a alocação das palavras em seus respectivos grupos, produziu diferentes taxas de acurácia por grupo de palavras, ainda que analisadas por um mesmo classificador. Isso pode ter ocorrido devido à um desbalanceamento de palavras foneticamente similares presentes mais em alguns grupos do que em outros. Com relação às diferenças no reconhecimento das sentenças, constatou-se que, sentenças com menos palavras terão uma maior taxa de acerto, no caso, a sentença: “Defina computação”. Somado a isso, tem-se as condições do ambiente de aquisição das locuções utilizadas no treinamento, assim como, no momento em que foram realizados os testes de reconhecimento, pois devido as diferentes influências percebidas nos mais variados tipos de ambientes de gravação, esses estão susceptíveis aos mais variados tipos de ruídos. Conclui-se então que a exatidão de um reconhecedor está diretamente ligada à qualidade da base de dados a ser utilizada no reconhecimento e, para evitar uma degradação do desempenho, deve-se não somente aplicar filtros supressores de ruídos, visando sua redução ou eliminação, mas também gravar a base de dados no mesmo local onde vai ser utilizado o sistema, ou nas condições mais próximas possíveis.

Através das demonstrações de usabilidade do STI e levando em conta as dificuldades encontradas, foi possível verificar que um STI com interface baseada em voz para sistemas de RI segundo o modelo exposto e implementado, corrobora ser viável em termos de uma aplicação real e as observações e resultados encontrados nos testes, mostram que, de um modo geral, o STI encontra-se apto ao reconhecimento de palavras, retornando o significado, sinônimos, antônimos, definições gramaticais, contextualização da palavra em uma frase e ilustrações de cada termo reconhecido.

# REFERÊNCIAS BIBLIOGRÁFICAS


ADAMI, A. G. **Sistema de Reconhecimento de Locutor Utilizando Redes Neurais Artificiais**. (Dissertação de Mestrado). Curso de Pós-Graduação em Ciência da Computação (CPGCC) - Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 1997.

BAKER, J. K. **The Dragon System - an Overview**. IEEE Transactions on Acoustics, Speech and Signal Processing - ASSP, Fev., 1975.

BOUROUBA, E-H., BEDDA, M., DJEMILI, R. **Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM**. Informatica, An International Journal of Computing and Informatics, Vol. 30, n. 3, pp. 373-384, 2006.

BRESOLIN, A. D. A. **Estudo do Reconhecimento de Voz para o Acionamento de Equipamentos Elétricos via Comandos em Português**. Programa de Pós-Graduação em Automação Industrial (PGAI) Universidade do Estado de Santa Catarina (UDESC) - Departamento de Engenharia Elétrica (DEE), Joinville, 2003.

BRESOLIN, A. D. A. **Reconhecimento de voz Através de Unidades Menores do que a Palavra, Utilizando Wavelet Packet e SVM, em uma Nova Estrutura Hierárquica de Decisão**. (Tese de Doutorado). Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal do Rio Grande do Norte - Centro de Tecnologia, Natal, 2008.

BRITES, R. S. *et al*. **Comparação de Desempenho entre Três Índices de Exatidão Aplicados a Classificações de Imagens Orbitais**. VIII Simpósio Brasileiro de Sensoriamento Remoto, INPE, p. 813-821, Salvador, 1996.

CAMPBELL, J. P. **Speaker recognition: a tutorial**. Proceedings of the IEEE, v. 85, n. 9, p. 1437-1462, ISSN 0018-9219, set. 1997.

CLANCEY, W. J. **From GUIDON to NEOMYCIN and HERACLES in Twenty Short Lessons: ORN Final Report 1979-1985**. AI Magazine, v. 7, 1986.

COHEN, F. M. et al. **Wavelets and their Applications in Computer Graphics**. ACM: SIGGRAPH 95 Conference, 1995.

COHEN, J. A. **Coefficient of Agreement for Nominal Scales**. Educational and Psychological Measurement, v. 20, n.1, p. 37-46, 1960.

CONGALTON, R.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. CRC Press, Danvers, EUA, 1999.

COOLEY, J. W.; TUKEY, J. W. **An Algorithm for the Machine Calculation of Complex Fourier Series**. In: Mathematics of Computation. v. 19, Cap. 90, p. 297-301, 1965. Disponível em: <http://links.jstor.org/sici?sici=0025-5718%28196504%2919%3A90%3C297%3AAAFTMC%3E2.0.CO%3B2-7>. Acesso em: 03 ago. 2014.

COSTA, R. J. M. **Sistemas Tutores Inteligentes**. Mestrado de Informática Aplicada à Educação - Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, jul., 2002. Disponível em: <http://www.nce.ufrj.br/ginape/publicacoes/trabalhos/t_2002_raimundo_ose_macario_costa/index.htm>. Acesso em: 02 mar. 2014.

DELLER , J. R.; PROAKIS, J. G.; HANSEN, J. H. L. **Discrete-time Processing of Speech Signals**. New York: Macmillan, 1993.

FISCHETTI, E.; GISOLFI, A. **From Computer-Aided Instruction to Intelligent**, Lawrence Lipsitz (Ed.). Publisher Educational Technology Publications Englewood Cliffs, NJ, USA. ISSN: 0013-1962, v. 30, Issue 8, p. 7-17, Ago, 1990.

FOODY, G. M. **Status of land cover classification accuracy assessment**. Remote Sensing of Environment, v. 80, p. 185-201, 2002.

FURUI, S. **Digital Speech Processing, Synthesis and Recognition**. Marcel Dekker, Inc., 1989.

GAMBOA, H.; ANA, F. **Designing Inteliggent Tutoring System: a Bayesian Approach**. 3rd International Conference on Enterprise Information Systems (ICEIS), 2001.

HELLDEN, U. **A test of Landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya**. Sweden, Lund University Natural Geography Institute Report, v. 47. 1980.

JANG, J. S. R. **DTW for Speech Recognition**. MIR Lab. National Taiwan University, Taiwan, 2005. Disponível em: <http://www.cs.nthu.edu.tw/~jang>. Acesso em: 23 abr. 2014.

JUANG, B. H. **On the hidden Markov model and dynamic time warping for speech recognition - A unified view**. AT&T Bell Laboratories Technical Journal. [S.l.]: Alcatel-Lucent. p. 1213-1243, 1984.

KAPLAN, R.; ROCK, D. **New Directions for Intelligent Tutoring Systems**. AI Expert, 1995.

KEARSLEY, G. **Artificial Intelligence and Instruction - Applications and Methods**. [S.l.]: Addison Wesley, 1987.

LANDIS, J.R.; KOCH, G.G. **The measurement of observer agreement for categorical data**. Biometrics, v. 33, n.1, p. 159-174, 1977.

LEE, H.-R.; CHEN, C.; JANG, R. J. S. **Approximate Lower-Bounding Functions for the Speedup of DTW for Melody Recognition**. Computer Science Department, National Tsing Hua University. Taiwan: IEEE. p. 178-181, 2005.

LESTER, J.; BRATING, K.; MOTT, B. "**Conversational Agents**". The Practical Handbook of Internet Computing, 2004.

LILLESAND, T.M.; KIEFER, R.W. **Remote Sensing and Image Interpretation**. 3rd. ed. John Wiley & Sons: New York. ISBN: 0471 305 758, 750 p, 1994.

LINO, N. D. L.; TEDESCO, P.; ROUSY, D. **Modelo de Percepção de Agentes Baseados em Emoções**. Universidade Federal de Pernambuco, Recife, 2006.

MICHAEL, K. B.; LAWRENCE, R. R. **An Adaptive, Ordered, Graph Search Technique for Dynamic Time Warping for Isolated Word Recognition**. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1982.

MILLER, J. R. **Foundations of Intelligent Tutoring Systems Interacting with Computers Series**. [S.l.]: Psychology Press, 1982.

NEJAT, A. **Digital Speech Processing, Speech Coding, Synthesis and Recognition**, Publisher Springer US, v. 155. SSN: 0893-3405. ISBN: 978-1-4757-2148-5. DOI: 10.1007/978-1-4757-2148-5, 1992.

RABINER, L. R. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**. Proceedings. IEEE, v. 77, n. 2, 1989.

RABINER, L. R.; ALLEN, J. B. **A unified approach to short-time Fourier analysis and synthesis**. Proceedings IEEE, v. 65. p. 1558-1564, 1977.

RABINER, L. R.; JUANG, H. B. **Fundamentals of Speech Recognition**. New Jersey: Prentice Hall, 1993.

RABINER, R. L.; SCHAFER, W. R. **Digital Processing of Speech Signals**. New Jersey: Prentice Hall, 1978.

RAVINDER, K. **Comparison of HMM and DTW for Isolated Word Recognition System of Punjabi Language**. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. [S.l.]: Springer Berlin Heidelberg. p. 244-252, 2010.

RIOUL, O.; VETTERLI, M. **Wavelet and signal processing**. IEEE Signal Processing Magazine, v. 8, n. 4, Out, 1991.

ROSENFIELD, G.H., FITZPATRICK-LINS, K.A. **A coefficient of agreement as a measure of thematic classification accuracy**. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 52 (2), p. 223-227, 1986.

SANCHES, I. J. **Compressão Sem Perdas de Projeções de Tomografia Computadorizada Usando a Transformada Wavelet**. (Dissertação de Mestrado). Curso de Pós-Graduação em Informática - Universidade Federal do Paraná, Curitiba, fev. 2001. Disponível em: <http://www.dainf.ct.utfpr.edu.br/~ionildo/wavelet/cap3.htm>. Acesso em: 05 ago. 2014.

SANTOS, M. A. D. **Interface Multimodal de Interação Humano-Computador em Sistema de Recuperação de Informação Baseado em Voz e Texto em Português**. (Dissertação de Mestrado). Pós-Graduação em Ciência da Informação, 2013.

SCHROEDER, M. **A brief history of synthetic speech**. Speech Communication. p. 231-237, 1993.

SHAUGHNESSY, D. O. **Speech Communications, Human and machine**. New York: IEEE Press, 2000.

SILVA, A. G. D. **Reconhecimento de Voz para Palavras Isoladas**. Graduação em Engenharia da Computação - Universidade Federal de Pernambuco (UFPE) - Centro de Informática, Recife, 2009.

SILVA, S. M. **Biometria de Voz: Aspectos Teóricos e Práticos**. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual de Londrina Londrina, Paraná, 2010.

SMITH, J. O. I. **Mathematics of the Discrete Fourier Transform (DFT): with Audio Applications**. W3K Publishing, California, 2007. ISBN-10: 097456074x, ISBN-13: 9780974560748. Disponível em: <http://www.dsprelated.com/freebooks/mdft/Discrete_Fourier_Transform_DFT.html>. Acesso em: 11 abr. 2013.

SMITH, S. W. **The Scientist and Engineer's Guide to Digital Signal Processing**. 1. ed. California Technical, 1997. ISBN-13: 978-0966017632. Disponível em: <http://www.dspguide.com/ch12.htm>. Acesso em: 12 jul. 2014.

SPIEGEL, M. R. **Estatística**. McGrawhill, 3ºedição, 1993.

STORY, M.; CONGALTON, R. G. **Accuracy assessment: A user's perspective**. Photogrammetric Engineering and Remote Sensing, v. 52, p. 397−399, 1986.

TERMAN, L. M. **The Measurement of Intelligence An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale**, 1916. Disponível em: <http://www.gutenberg.org/files/20662/20662-h/20662-h.htm>. Acesso em: 22 mar. 2015.

TSIPORKOVA, E. **An Integrative DTW-Based Imputation Method for Gene Expression Time Series Data**. Intelligent Systems (IS) - IEEE International Conference, Sofia, n. 6, p. 258-263, 6-8, ISSN: 978-1-4673-2276-8, 2012.

VINTSYUK, T. K. **Speech Discrimination by Dynamic Programming**, n.1, p. 81-88, 1968.

WEEKS, M. **Processamento Digital de Sinais: Utilizando Matlab e Wavelets**. Tradução de Edson Tanaka. 2ª. ed. Rio de Janeiro: LTC, 2012.

WENGER, E. **Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communications of Knowledge**. Los Altos, CA: Morgam Kaufmann Publishers, 1987.

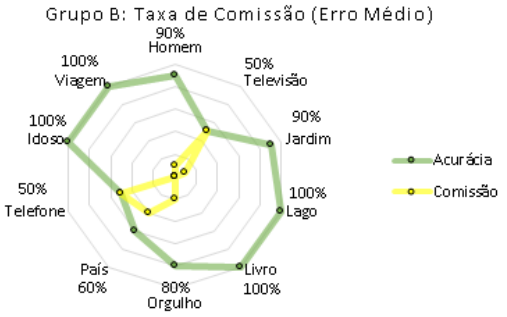
YNOGUTI, C. A. **Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov**. (Tese de Doutorado). Universidade Estadual de Campinas - Faculdade de Engenharia Elétrica e de Computação, Campinas, 1999.

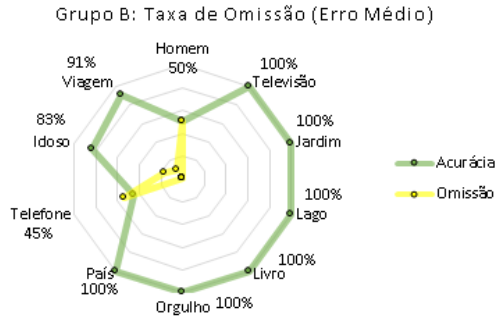
# APÊNDICES

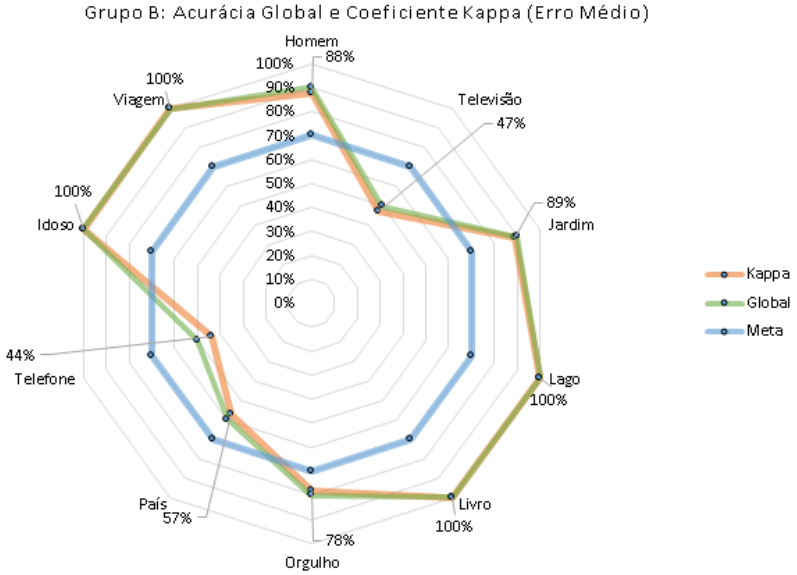
# APÊNDICE A - Relatório de Análise do Grupo B

| Erro Médio (FFT) Grupo B | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Homem | Televisão | Jardim | Lago | Livro | Orgulho | País | Telefone | Idoso | Viagem | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Homem | 9 | | | | | | | 1 | | | 10 | 90% | 10% | 88% |
| | Televisão | | 5 | | | | | | 5 | | | 10 | 50% | 50% | 47% |
| | Jardim | | | 9 | | | | | | | 1 | 10 | 90% | 10% | 89% |
| | Lago | | | | 10 | | | | | | | 10 | 100% | 0% | 100% |
| | Livro | | | | | 10 | | | | | | 10 | 100% | 0% | 100% |
| | Orgulho | | | | | | 8 | | | 2 | | 10 | 80% | 20% | 78% |
| | País | 4 | | | | | | 6 | | | | 10 | 60% | 40% | 57% |
| | Telefone | 5 | | | | | | | 5 | | | 10 | 50% | 50% | 44% |
| | Idoso | | | | | | | | | 10 | | 10 | 100% | 0% | 100% |
| | Viagem | | | | | | | | | | 10 | 10 | 100% | 0% | 100% |
| Omissão | $\sum x_{+i}$ | 18 | 5 | 9 | 10 | 10 | 8 | 6 | 11 | 12 | 11 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 50% | 100% | 100% | 100% | 100% | 100% | 100% | 45% | 83% | 91% | 87% | 82% | 18% | 80% |
| | Erro | 50% | 0% | 0% | 0% | 0% | 0% | 0% | 55% | 17% | 9% | 13% | Acurácia Final: | | 81% |

| Erro Médio Grupo B | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Homem | 0,18 | 0,18 | 0,18 | 0,55 | 0,18 | 0,19 | 0,19 | 0,18 | | 0,18 | 0,22 | 1,90 |
| Televisão | 0,18 | 0,18 | 0,18 | | | | 0,18 | | 0,18 | | 0,18 | |
| Jardim | 0,19 | 0,18 | 0,18 | 0,18 | | 0,17 | 0,18 | 0,15 | 0,18 | 0,16 | 0,17 | |
| Lago | 0,18 | 0,18 | 0,18 | 0,18 | 0,16 | 0,14 | 0,18 | 0,15 | 0,18 | 0,19 | 0,17 | |
| Livro | 0,18 | 0,19 | 0,17 | 0,18 | 0,18 | 0,16 | 0,18 | 0,18 | 0,18 | 0,17 | 0,18 | Tempo Total do Treinamento |
| Orgulho | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,16 | | 0,21 | 0,18 | | 0,18 | |
| País | 0,18 | 0,19 | 0,20 | 0,15 | | | 0,18 | | 0,18 | | 0,18 | 0,32 |
| Telefone | 0,18 | | 0,18 | | | | 0,18 | 0,18 | 0,18 | | 0,18 | |
| Idoso | 0,18 | 0,17 | 0,15 | 0,20 | 0,18 | 0,21 | 0,21 | 0,20 | 0,29 | 0,19 | 0,20 | |
| Viagem | 0,18 | 0,56 | 0,19 | 0,20 | 0,23 | 0,20 | 0,20 | 0,21 | 0,21 | 0,20 | 0,24 | |



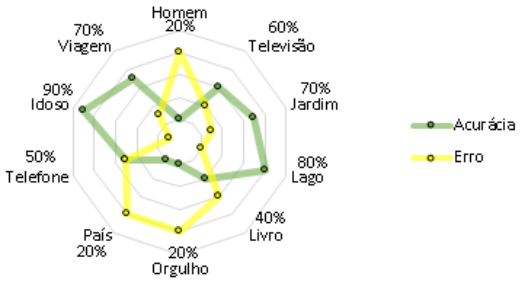



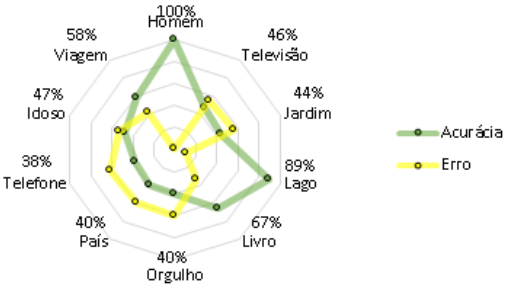
| Desvio Padrão (FFT) Grupo B | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Homem | Televisão | Jardim | Lago | Livro | Orgulho | País | Telefone | Idoso | Viagem | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Homem | 2 | 2 | 1 | | | | | 4 | | 1 | 10 | 20% | 80% | 18% |
| | Televisão | | 6 | | | | | | 3 | | 1 | 10 | 60% | 40% | 54% |
| | Jardim | | | 7 | | | | 3 | | | | 10 | 70% | 30% | 64% |
| | Lago | | | 1 | 8 | 1 | | | | | | 10 | 80% | 20% | 78% |
| | Livro | | | | | 4 | 2 | | | 3 | 1 | 10 | 40% | 60% | 36% |
| | Orgulho | | | | | | 2 | | | 6 | 2 | 10 | 20% | 80% | 16% |
| | País | | | 7 | 1 | | | 2 | | | | 10 | 20% | 80% | 16% |
| | Telefone | | 5 | | | | | | 5 | | | 10 | 50% | 50% | 43% |
| | Idoso | | | | | | 1 | | | 9 | | 10 | 90% | 10% | 88% |
| | Viagem | | | | | 1 | | | 1 | 1 | 7 | 10 | 70% | 30% | 66% |
| Omissão | $\sum x_{+i}$ | 2 | 13 | 16 | 9 | 6 | 5 | 5 | 13 | 19 | 12 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 46% | 44% | 89% | 67% | 40% | 40% | 38% | 47% | 58% | 57% | 52% | 48% | 48% |
| | Erro | 0% | 54% | 56% | 11% | 33% | 60% | 60% | 62% | 53% | 42% | 43% | Acurácia Final: | | 50% |

| Desvio Padrão Grupo B | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Homem | | 0,24 | | | 0,22 | | | | | | 0,23 | 2,22 |
| Televisão | 0,23 | 0,22 | 0,24 | | 0,24 | 0,22 | | | | 0,22 | 0,23 | |
| Jardim | 0,22 | 0,23 | 0,23 | 0,23 | 0,23 | 0,22 | | 0,22 | | | 0,23 | |
| Lago | 0,20 | 0,20 | | | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | |
| Livro | | 0,22 | | | | | 0,22 | | 0,22 | 0,23 | 0,22 | Tempo Total do Treinamento |
| Orgulho | | | | 0,23 | | | 0,24 | | | | 0,24 | |
| País | | | 0,23 | | | | 0,22 | | | | 0,23 | 0,32 |
| Telefone | 0,20 | | | | | | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | |
| Idoso | 0,20 | 0,20 | | 0,27 | 0,23 | 0,23 | 0,22 | 0,22 | 0,22 | 0,24 | 0,22 | |
| Viagem | | 0,26 | | 0,23 | | 0,23 | 0,22 | 0,23 | 0,22 | 0,22 | 0,23 | |

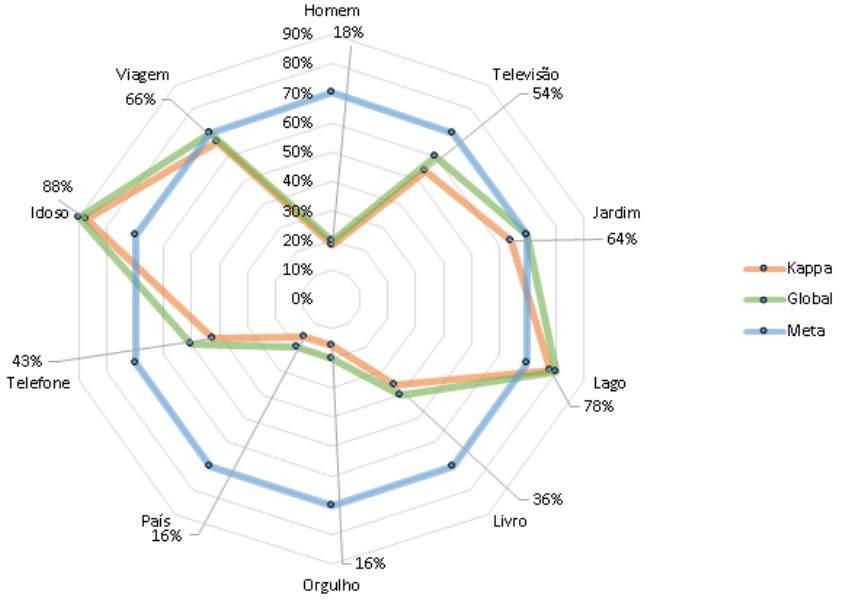
Grupo B: Taxa de Comissão (Desvio Padrão)




Grupo B: Taxa de Omissão (Desvio Padrão)

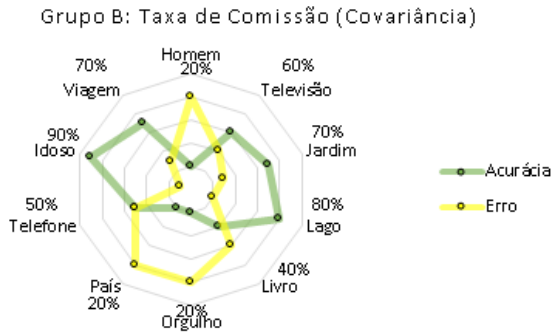


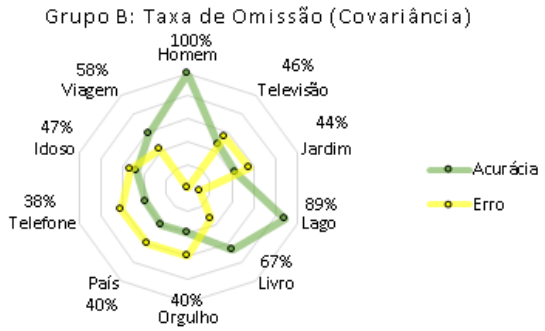

Grupo B: Acurácia Global e Coeficiente Kappa (Desvio Padrão)

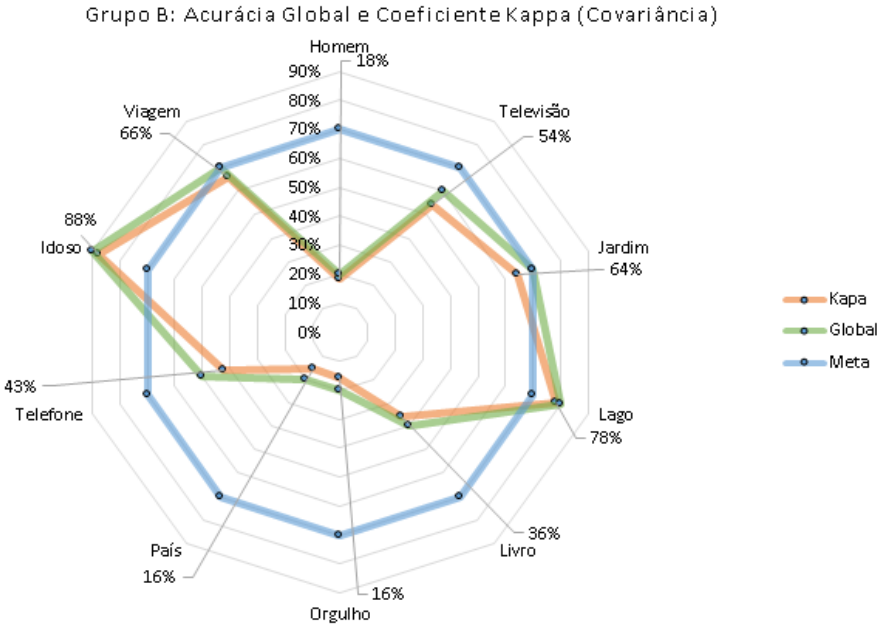


| Covariância (FFT) Grupo B | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Homem | Televisão | Jardim | Lago | Livro | Orgulho | País | Telefone | Idoso | Viagem | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Homem | **2** | 2 | 1 | | | | | 4 | | 1 | **10** | 20% | 80% | 18% |
| | Televisão | | **6** | | | | | | 3 | | 1 | **10** | 60% | 40% | 54% |
| | Jardim | | | **7** | | | | 3 | | | | **10** | 70% | 30% | 64% |
| | Lago | | | 1 | **8** | 1 | | | | | | **10** | 80% | 20% | 78% |
| | Livro | | | | | **4** | 2 | | | 3 | 1 | **10** | 40% | 60% | 36% |
| | Orgulho | | | | | | **2** | | | 6 | 2 | **10** | 20% | 80% | 16% |
| | País | | | 7 | 1 | | | **2** | | | | **10** | 20% | 80% | 16% |
| | Telefone | | 5 | | | | | | **5** | | | **10** | 50% | 50% | 43% |
| | Idoso | | | | | | 1 | | | **9** | | **10** | 90% | 10% | 88% |
| | Viagem | | | | | 1 | | | 1 | 1 | **7** | **10** | 70% | 30% | 66% |
| Omissão | $\sum x_{+i}$ | **2** | **13** | **16** | **9** | **6** | **5** | **5** | **13** | **19** | **12** | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 46% | 44% | 89% | 67% | 40% | 40% | 38% | 47% | 58% | **57%** | **52%** | **48%** | **48%** |
| | Erro | 0% | 54% | 56% | 11% | 33% | 60% | 60% | 62% | 53% | 42% | **43%** | Acurácia Final: | | **50%** |

| Covariância Grupo B | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Homem | | 0,26 | | | 0,24 | | | | | | **0,25** | **2,35** |
| Televisão | 0,22 | 0,22 | 0,23 | | 0,23 | 0,24 | | | | 0,24 | **0,23** | |
| Jardim | 0,22 | 0,23 | 0,23 | 0,24 | 0,23 | 0,20 | | 0,26 | | | **0,23** | |
| Lago | 0,20 | **0,19** | | | 0,20 | 0,20 | 0,20 | 0,20 | **0,19** | 0,20 | **0,20** | |
| Livro | | **0,19** | | | | | 0,23 | | 0,23 | 0,24 | **0,23** | **Tempo Total do Treinamento** |
| Orgulho | | | | 0,24 | | | 0,22 | | | | **0,23** | |
| País | | | 0,23 | | | | 0,23 | | | | **0,23** | **0,32** |
| Telefone | 0,21 | | | | | | 0,22 | 0,21 | 0,22 | 0,21 | **0,21** | |
| Idoso | 0,20 | 0,25 | | 0,24 | 0,23 | 0,24 | 0,26 | 0,23 | 0,24 | 0,24 | **0,24** | |
| Viagem | | **0,68** | | 0,26 | | 0,24 | 0,24 | 0,23 | 0,23 | 0,24 | **0,30** | |



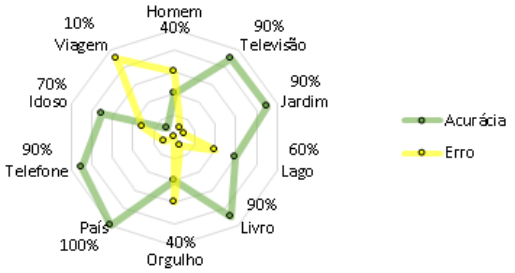



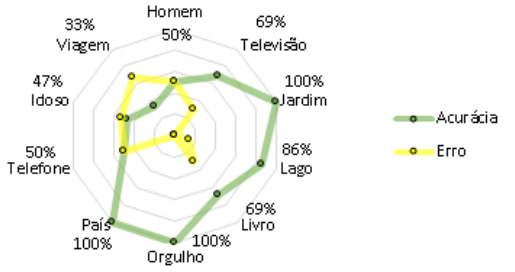
| DTW (STFT) Grupo B | | Classes Preditas (Referência $+i$ ) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Homem | Televisão | Jardim | Lago | Livro | Orgulho | País | Telefone | Idoso | Viagem | $\sum x_{+i}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Homem | **4** | 2 | | | | | | 4 | | | **10** | 40% | 60% | 35% |
| | Televisão | | **9** | | | | | | 1 | | | **10** | 90% | 10% | 89% |
| | Jardim | | | **9** | | | | | 1 | | | **10** | 90% | 10% | 89% |
| | Lago | | | | **6** | 2 | | | 1 | | 1 | **10** | 60% | 40% | 57% |
| | Livro | | | | | **9** | | | | 1 | | **10** | 90% | 10% | 89% |
| | Orgulho | | | | | | **4** | | | 6 | | **10** | 40% | 60% | 38% |
| | País | | | | | | | **10** | | | | **10** | 100% | 0% | 100% |
| | Telefone | | | | 1 | | | | **9** | | | **10** | 90% | 10% | 88% |
| | Idoso | | | | | 2 | | | | **7** | 1 | **10** | 70% | 30% | 65% |
| | Viagem | 4 | 2 | | | | | | 2 | 1 | **1** | **10** | 10% | 90% | 7% |
| Omissão | $\sum x_{+i}$ | **8** | **13** | **9** | **7** | **13** | **4** | **10** | **18** | **15** | **3** | **100** | **Exatidão Global** | **Erro Global** | **Kappa Global** |
| | **Acurácia** | 50% | 69% | 100% | 86% | 69% | 100% | 100% | 50% | 47% | 33% | **70%** | **68%** | **32%** | **66%** |
| | **Erro** | 50% | 31% | 0% | 14% | 31% | 0% | 0% | 50% | 53% | 67% | **30%** | **Acurácia Final:** | | **67%** |

| DTW Grupo B | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Homem | | 0,35 | 0,36 | **0,30** | 0,37 | | | | | | **0,34** | **3,68** |
| Televisão | 0,35 | 0,33 | 0,38 | 0,35 | 0,34 | 0,34 | 0,36 | 0,34 | | 0,33 | **0,35** | |
| Jardim | 0,37 | 0,38 | 0,36 | 0,37 | 0,37 | 0,43 | 0,38 | 0,36 | | 0,32 | **0,37** | |
| Lago | 0,36 | 0,37 | 0,34 | | | | 0,37 | 0,38 | 0,38 | | **0,37** | |
| Livro | 0,32 | 0,36 | 0,32 | 0,32 | 0,33 | | 0,35 | 0,43 | 0,37 | 0,36 | **0,35** | **Tempo Total do Treinamento** |
| Orgulho | 0,38 | 0,34 | | 0,36 | | | | | 0,32 | | **0,35** | |
| País | 0,38 | 0,37 | 0,37 | 0,40 | 0,37 | 0,37 | 0,42 | 0,37 | 0,41 | 0,36 | **0,38** | **0,55** |
| Telefone | 0,43 | 0,39 | 0,36 | 0,36 | | 0,36 | 0,41 | 0,41 | 0,42 | 0,42 | **0,39** | |
| Idoso | | | | 0,43 | 0,40 | 0,41 | **0,47** | 0,41 | 0,42 | 0,45 | **0,43** | |
| Viagem | | 0,34 | | | | | | | | | **0,34** | |

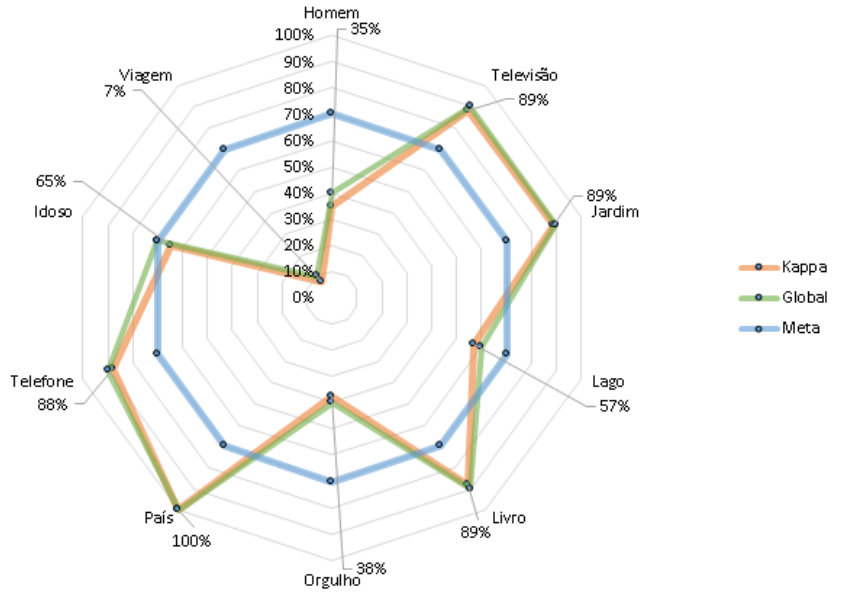
Grupo B: Taxa de Comissão (DTW)




Grupo B: Taxa de Omissão (DTW)



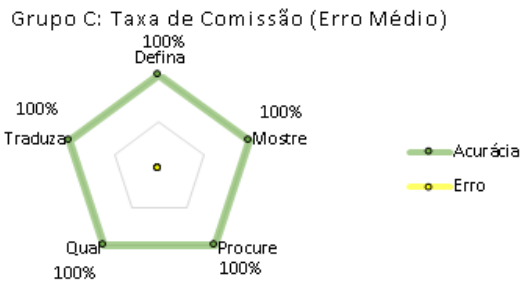

Grupo B: Acurácia Global e Coeficiente Kappa (DTW)

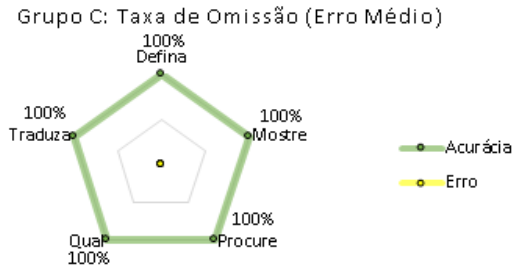


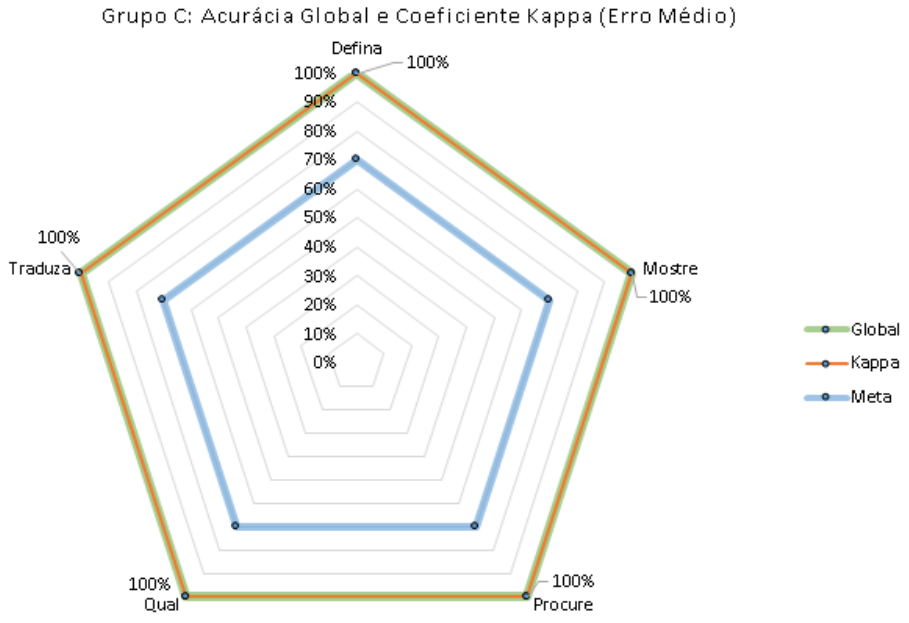
## APÊNDICE B - Relatório de Análise do Grupo C

| Erro Médio (FFT) Grupo C | | Classes Preditas (Referência $+i$) | | | | | Comissão | | Kappa |
|---|---|---|---|---|---|---|---|---|---|
| | | Defina | Mostre | Procure | Qual | Traduza | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Defina | 10 | | | | | 10 | 100% | 0% | 100% |
| | Mostre | | 10 | | | | 10 | 100% | 0% | 100% |
| | Procure | | | 10 | | | 10 | 100% | 0% | 100% |
| | Qual | | | | 10 | | 10 | 100% | 0% | 100% |
| | Traduza | | | | | 10 | 10 | 100% | 0% | 100% |
| Omissão | $\sum x_{+i}$ | 10 | 10 | 10 | 10 | 10 | 50 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 0% | 100% |
| | Erro | 0% | 0% | 0% | 0% | 0% | 0% | Acurácia Final: | | 100% |

| Erro Médio Grupo C | Tempo de Processamento de cada Amostra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Defina | 0,55 | 0,17 | 0,17 | 0,17 | 0,17 | 0,17 | 0,17 | 0,17 | 0,26 | 0,17 | 0,22 | 0,97 |
| Mostre | 0,17 | 0,17 | 0,17 | 0,14 | 0,18 | 0,15 | 0,17 | 0,17 | 0,17 | 0,20 | 0,17 | |
| Procure | 0,17 | 0,17 | 0,17 | 0,17 | 0,46 | 0,19 | 0,21 | 0,19 | 0,20 | 0,17 | 0,21 | Tempo Total do Treinamento |
| Qual | 0,18 | 0,18 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,13 |
| Traduza | 0,19 | 0,18 | 0,17 | 0,19 | 0,18 | 0,19 | 0,19 | 0,21 | 0,19 | 0,19 | 0,19 | |

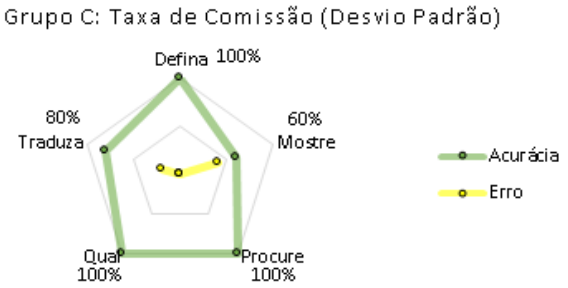


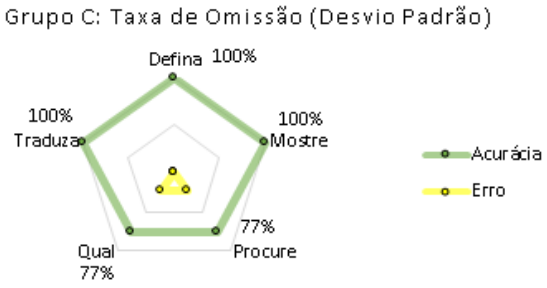



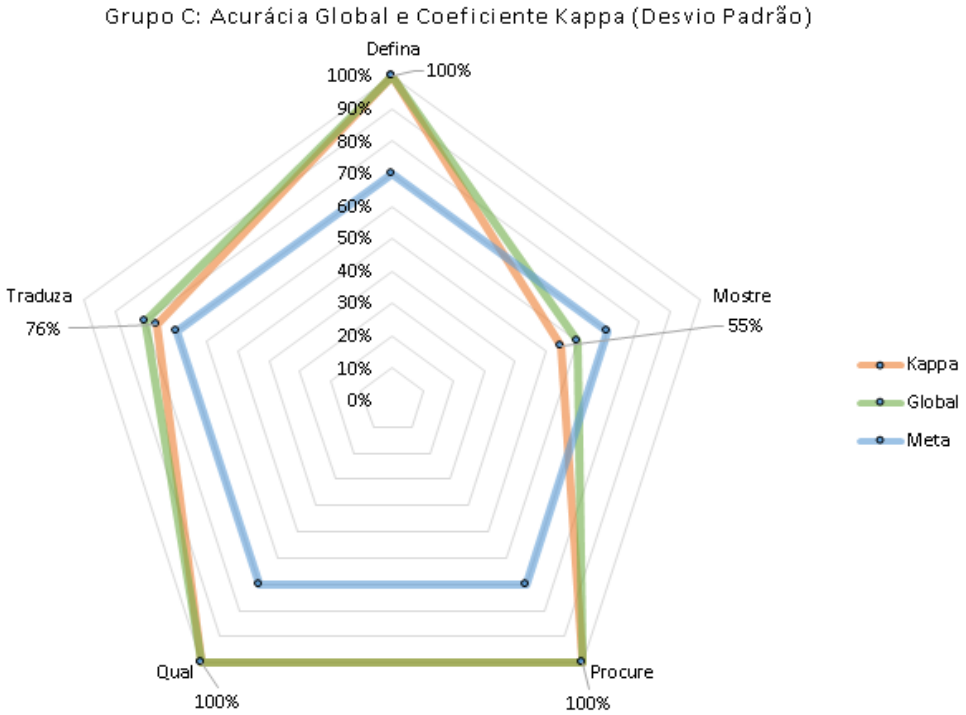



| Desvio Padrão (FFT) Grupo C | | Classes Preditas (Referência $+i$) | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Defina | Mostre | Procure | Qual | Traduza | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Defina | 10 | | | | | 10 | 100% | 0% | 100% |
| | Mostre | | 6 | 1 | 3 | | 10 | 60% | 40% | 55% |
| | Procure | | | 10 | | | 10 | 100% | 0% | 100% |
| | Qual | | | | 10 | | 10 | 100% | 0% | 100% |
| | Traduza | | | 2 | | 8 | 10 | 80% | 20% | 76% |
| Omissão | $\sum x_{+i}$ | 10 | 6 | 13 | 13 | 8 | 50 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 100% | 77% | 77% | 100% | 91% | 88% | 12% | 86% |
| | Erro | 0% | 0% | 23% | 23% | 0% | 9% | Acurácia Final: | | 87% |

| Desvio Padrão Grupo C | Tempo de Processamento de cada Amostra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Defina | 0,24 | 0,19 | 0,18 | 0,17 | 0,18 | 0,25 | 0,18 | 0,19 | 0,18 | 0,18 | 0,19 | 0,98 |
| Mostre | | | 0,18 | 0,18 | 0,18 | 0,19 | | 0,18 | 0,18 | | 0,18 | |
| Procure | 0,18 | 0,18 | 0,19 | 0,18 | 0,24 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | Tempo Total do Treinamento |
| Qual | 0,21 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,21 | 0,20 | 0,20 | 0,20 | 0,13 |
| Traduza | 0,20 | 0,21 | 0,21 | 0,20 | | 0,20 | 0,20 | 0,20 | 0,20 | | 0,20 | |

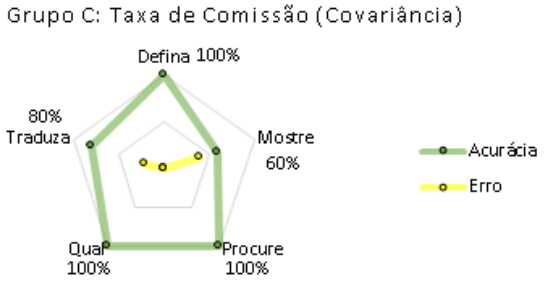


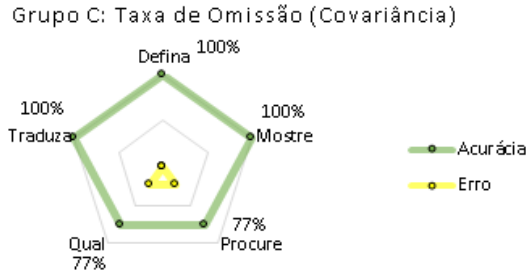



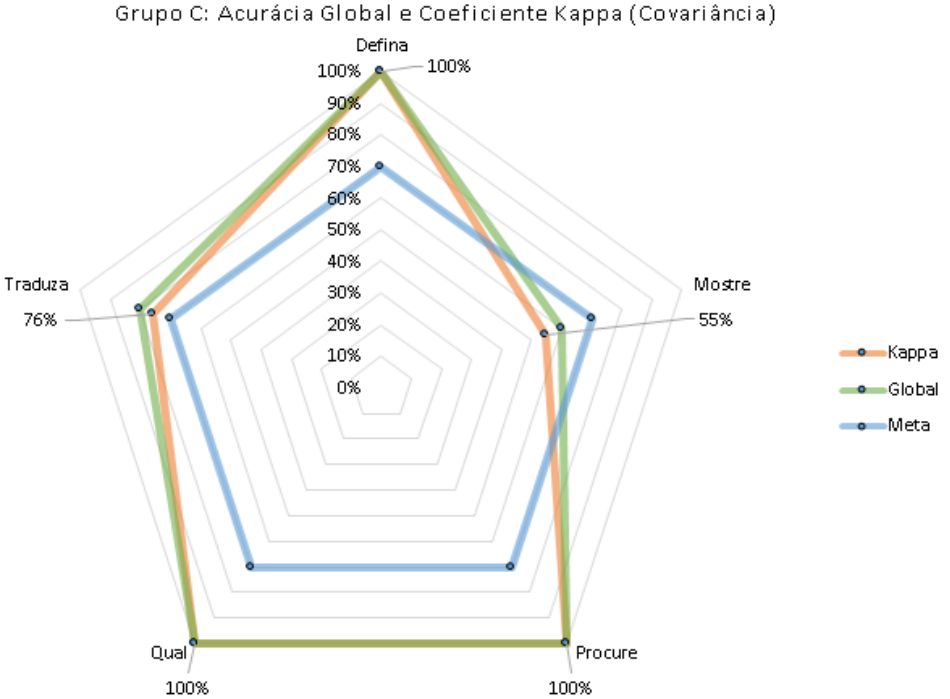



| Covariância (FFT) Grupo C | | Classes Preditas (Referência $+i$) | | | | | Comissão | | Kappa |
|---|---|---|---|---|---|---|---|---|---|
| | | Defina | Mostre | Procure | Qual | Traduza | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Defina | 10 | | | | | 10 | 100% | 0% | 100% |
| | Mostre | | 6 | 1 | 3 | | 10 | 60% | 40% | 55% |
| | Procure | | | 10 | | | 10 | 100% | 0% | 100% |
| | Qual | | | | 10 | | 10 | 100% | 0% | 100% |
| | Traduza | | | 2 | | 8 | 10 | 80% | 20% | 76% |
| Omissão | $\sum x_{+i}$ | 10 | 6 | 13 | 13 | 8 | 50 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 100% | 100% | 77% | 77% | 100% | 91% | 88% | 6% | 86% |
| | Erro | 0% | 0% | 23% | 23% | 0% | 9% | Acurácia Final: | | 87% |

| Covariância Grupo C | Tempo de Processamento de cada Amostra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Defina | 0,63 | 0,19 | 0,19 | 0,19 | **0,18** | 0,75 | **0,18** | 0,19 | 0,19 | **0,18** | 0,29 | 1,18 |
| Mostre | | | **0,18** | 0,22 | **0,18** | 0,19 | | 0,19 | 0,20 | | 0,19 | |
| Procure | 0,19 | **0,18** | 0,20 | 0,19 | **0,78** | 0,22 | 0,22 | 0,22 | 0,23 | 0,23 | 0,26 | Tempo Total do Treinamento |
| Qual | 0,21 | 0,20 | 0,21 | 0,21 | 0,21 | 0,21 | 0,22 | 0,21 | 0,20 | 0,21 | 0,21 | 0,13 |
| Traduza | 0,21 | 0,22 | 0,22 | 0,21 | | 0,21 | 0,28 | 0,22 | 0,22 | | 0,22 | |



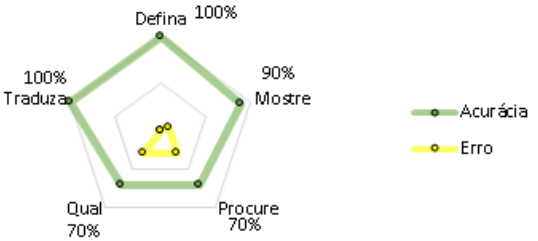



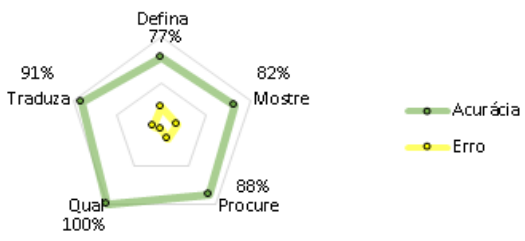
| DTW (STFT) Grupo C | | Classes Preditas (Referência $+i$) | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Defina | Mostre | Procure | Qual | Traduza | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Defina | 10 | | | | | 10 | 100% | 0% | 100% |
| | Mostre | | 9 | | | 1 | 10 | 90% | 10% | 87% |
| | Procure | 3 | | 7 | | | 10 | 70% | 30% | 64% |
| | Qual | | 2 | 1 | 7 | | 10 | 70% | 30% | 65% |
| | Traduza | | | | | 10 | 10 | 100% | 0% | 100% |
| Omissão | $\sum x_{+i}$ | 13 | 11 | 8 | 7 | 11 | 50 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 77% | 82% | 88% | 100% | 91% | 87% | 86% | 14% | 83% |
| | Erro | 23% | 18% | 13% | 0% | 9% | 13% | Acurácia Final: | | 85% |

| DTW Grupo C | Tempo de Processamento de cada Amostra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Defina | 0,75 | 0,31 | 0,26 | 0,28 | 0,29 | 0,81 | 0,27 | 0,26 | 0,26 | 0,26 | 0,38 | 1,56 |
| Mostre | 0,28 | 0,27 | 0,26 | 0,27 | 0,26 | 0,26 | | 0,25 | 0,26 | 0,23 | 0,26 | |
| Procure | 0,29 | 0,25 | | | 0,28 | 0,27 | 0,28 | 0,28 | 0,28 | | 0,28 | Tempo Total do Treinamento |
| Qual | | 0,32 | 0,32 | 0,31 | 0,33 | | 0,32 | | 0,30 | 0,33 | 0,32 | 0,26 |
| Traduza | 0,33 | 0,36 | 0,32 | 0,33 | 0,30 | 0,32 | 0,36 | 0,32 | 0,33 | 0,31 | 0,33 | |

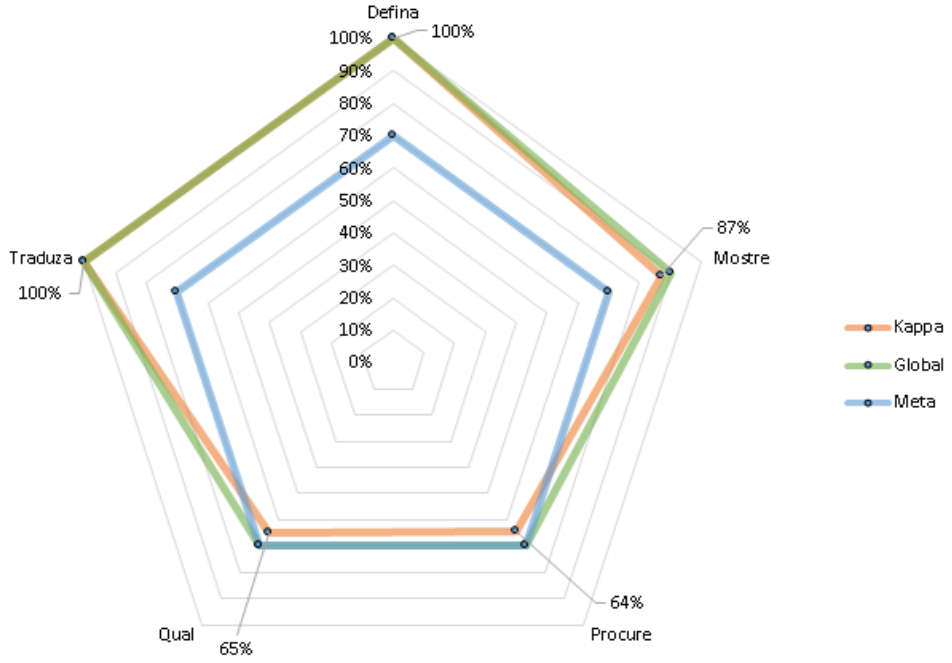
Grupo C: Taxa de Comissão (DTW)




Grupo C: Taxa de Omissão (DTW)



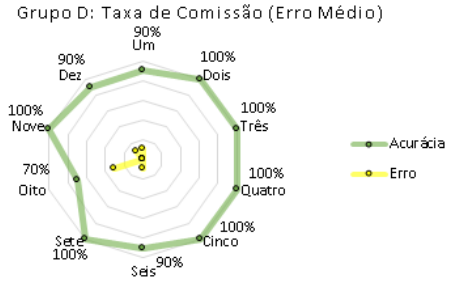

Grupo C: Acurácia Global e Coeficiente Kappa (DTW)

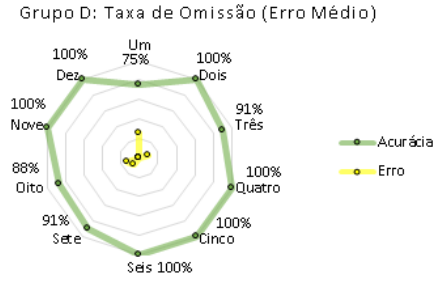


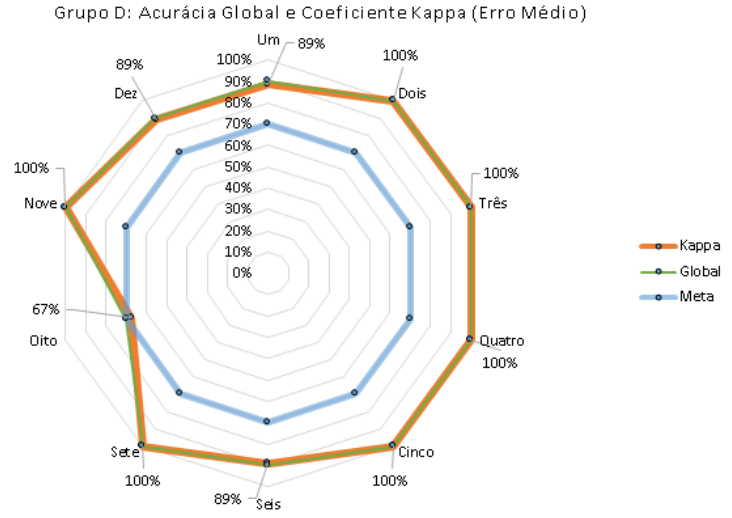
## APÊNDICE C - Relatório de Análise do Grupo D

| Erro Médio (FFT) Grupo D | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Um | Dois | Três | Quatro | Cinco | Seis | Sete | Oito | Nove | Dez | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Um | 9 | | | | | | | 1 | | | 10 | 90% | 10% | 89% |
| | Dois | | 10 | | | | | | | | | 10 | 100% | 0% | 100% |
| | Três | | | 10 | | | | | | | | 10 | 100% | 0% | 100% |
| | Quatro | | | | 10 | | | | | | | 10 | 100% | 0% | 100% |
| | Cinco | | | | | 10 | | | | | | 10 | 100% | 0% | 100% |
| | Seis | | | 1 | | | 9 | | | | | 10 | 90% | 10% | 89% |
| | Sete | | | | | | | 10 | | | | 10 | 100% | 0% | 100% |
| | Oito | 3 | | | | | | | 7 | | | 10 | 70% | 30% | 67% |
| | Nove | | | | | | | | | 10 | | 10 | 100% | 0% | 100% |
| | Dez | | | | | | | 1 | | | 9 | 10 | 90% | 10% | 89% |
| Omissão | $\sum x_{+i}$ | 12 | 10 | 11 | 10 | 10 | 9 | 11 | 8 | 10 | 9 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 75% | 100% | 91% | 100% | 100% | 100% | 91% | 88% | 100% | 100% | 94% | 94% | 6% | 93% |
| | Erro | 25% | 0% | 9% | 0% | 0% | 0% | 9% | 13% | 0% | 0% | 6% | Acurácia Final: | | 94% |

| Erro Médio Grupo D | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Um | 0,19 | 0,19 | 0,19 | | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 0,19 | 1,86 |
| Dois | 0,20 | 0,19 | 0,19 | 0,21 | 0,19 | 0,19 | 0,20 | 0,19 | 0,20 | 0,19 | 0,19 | |
| Três | 0,15 | 0,58 | 0,18 | 0,18 | 0,18 | 0,17 | 0,18 | 0,18 | 0,18 | 0,18 | 0,22 | |
| Quatro | 0,14 | 0,17 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,19 | 0,18 | |
| Cinco | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | 0,19 | 0,18 | 0,18 | 0,18 | 0,18 | 0,18 | Tempo Total do Treinamento |
| Seis | 0,18 | 0,18 | 0,18 | 0,18 | | 0,18 | 0,19 | 0,19 | 0,20 | 0,14 | 0,18 | |
| Sete | 0,18 | 0,18 | 0,18 | 0,18 | 0,15 | 0,18 | 0,19 | 0,20 | 0,18 | 0,18 | 0,18 | 0,38 |
| Oito | 0,18 | 0,18 | 0,18 | 0,18 | | | 0,18 | 0,18 | | 0,18 | 0,18 | |
| Nove | 0,18 | 0,18 | 0,18 | 0,19 | 0,18 | 0,19 | 0,18 | 0,19 | 0,18 | 0,18 | 0,18 | |
| Dez | 0,18 | 0,18 | 0,18 | 0,19 | 0,18 | 0,18 | | 0,18 | 0,19 | 0,18 | 0,18 | |



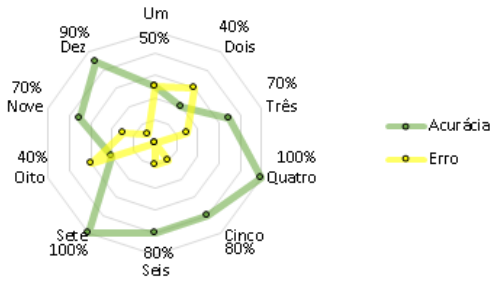



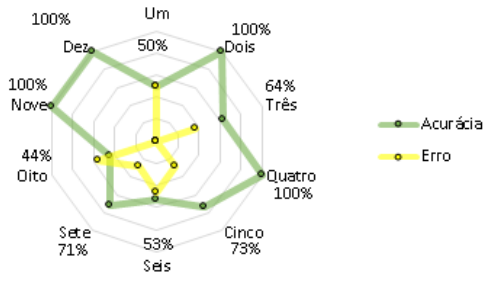
| Desvio Padrão (FFT) Grupo D | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Um | Dois | Três | Quatro | Cinco | Seis | Sete | Oito | Nove | Dez | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Um | 5 | | | | 1 | 1 | | 3 | | | 10 | 50% | 50% | 44% |
| | Dois | | 4 | | | 2 | 2 | | 2 | | | 10 | 40% | 60% | 38% |
| | Três | | | 7 | | | 3 | | | | | 10 | 70% | 30% | 66% |
| | Quatro | | | | 10 | | | | | | | 10 | 100% | 0% | 100% |
| | Cinco | | | 2 | | 8 | | | | | | 10 | 80% | 20% | 78% |
| | Seis | | | 2 | | | 8 | | | | | 10 | 80% | 20% | 76% |
| | Sete | | | | | | | 10 | | | | 10 | 100% | 0% | 100% |
| | Oito | 5 | | | | | 1 | | 4 | | | 10 | 40% | 60% | 34% |
| | Nove | | | | | | | 3 | | 7 | | 10 | 70% | 30% | 68% |
| | Dez | | | | | | | 1 | | | 9 | 10 | 90% | 10% | 89% |
| Omissão | $\sum x_{+i}$ | 10 | 4 | 11 | 10 | 11 | 15 | 14 | 9 | 7 | 9 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 50% | 100% | 64% | 100% | 73% | 53% | 71% | 44% | 100% | 100% | 76% | 72% | 28% | 69% |
| | Erro | 50% | 0% | 36% | 0% | 27% | 47% | 29% | 56% | 0% | 0% | 24% | Acurácia Final: | | 71% |

| Desvio Padrão Grupo D | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Um | 0,21 | 0,21 | | | | | | 0,21 | 0,21 | | 0,17 | 2,03 |
| Dois | | | 0,21 | 0,23 | | | | 0,22 | 0,22 | | 0,22 | |
| Três | 0,21 | 0,27 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | | | | 0,21 | |
| Quatro | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | 0,19 | 0,20 | 0,20 | 0,21 | 0,20 | |
| Cinco | 0,21 | 0,20 | 0,20 | 0,20 | 0,20 | 0,20 | | | 0,20 | 0,20 | 0,20 | Tempo Total do Treinamento |
| Seis | 0,20 | 0,20 | 0,20 | 0,20 | | 0,20 | | 0,20 | 0,20 | 0,20 | 0,20 | |
| Sete | 0,20 | 0,21 | 0,21 | 0,20 | 0,20 | 0,20 | 0,21 | 0,21 | 0,20 | 0,21 | 0,20 | 0,38 |
| Oito | 0,22 | 0,20 | | | | 0,22 | | | 0,22 | | 0,22 | |
| Nove | 0,20 | 0,20 | 0,20 | | 0,20 | | 0,20 | 0,20 | | 0,20 | 0,20 | |
| Dez | 0,20 | 0,21 | 0,21 | 0,21 | 0,20 | 0,20 | | 0,20 | 0,20 | 0,20 | 0,20 | |

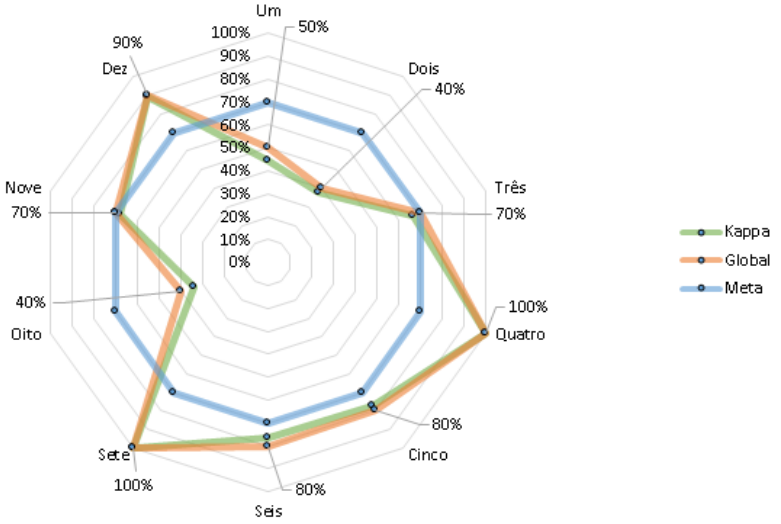
Grupo D: Taxa de Comissão (Desvio Padrão)




Grupo D: Taxa de Omissão (Desvio Padrão)



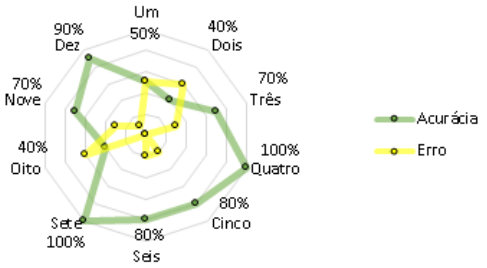

Grupo D: Acurácia Global e Coeficiente Kappa (Desvio Padrão)



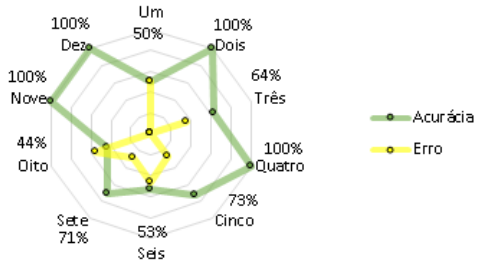
| Covariância (FFT) Grupo D | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Um | Dois | Três | Quatro | Cinco | Seis | Sete | Oito | Nove | Dez | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Um | **5** | | | | 1 | 1 | | 3 | | | 10 | 50% | 50% | 44% |
| | Dois | | **4** | | | 2 | 2 | | 2 | | | 10 | 40% | 60% | 38% |
| | Três | | | **7** | | | 3 | | | | | 10 | 70% | 30% | 66% |
| | Quatro | | | | **10** | | | | | | | 10 | 100% | 0% | 100% |
| | Cinco | | | 2 | | **8** | | | | | | 10 | 80% | 20% | 78% |
| | Seis | | | 2 | | | **8** | | | | | 10 | 80% | 20% | 76% |
| | Sete | | | | | | | **10** | | | | 10 | 100% | 0% | 100% |
| | Oito | 5 | | | | | 1 | | **4** | | | 10 | 40% | 60% | 34% |
| | Nove | | | | | | | 3 | | **7** | | 10 | 70% | 30% | 68% |
| | Dez | | | | | | | 1 | | | **9** | 10 | 90% | 10% | 89% |
| Omissão | $\sum x_{+i}$ | 10 | 4 | 11 | 10 | 11 | 15 | 14 | 9 | 7 | 9 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 50% | 100% | 64% | 100% | 73% | 53% | 71% | 44% | 100% | 100% | **76%** | **72%** | **28%** | **69%** |
| | Erro | 50% | 0% | 36% | 0% | 27% | 47% | 29% | 56% | 0% | 0% | **24%** | Acurácia Final: | | **71%** |

| Covariância Grupo D | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Um | 0,23 | 0,21 | 0,22 | | | | | 0,21 | 0,21 | | **0,22** | **2,30** |
| Dois | | | 0,23 | 0,22 | | | | 0,22 | 0,23 | | **0,23** | |
| Três | 0,23 | **0,81** | **0,20** | **0,20** | 0,21 | 0,21 | 0,21 | | | | **0,30** | |
| Quatro | **0,20** | 0,21 | **0,20** | 0,21 | 0,21 | 0,21 | **0,20** | 0,21 | 0,21 | 0,22 | **0,21** | |
| Cinco | **0,20** | 0,21 | 0,21 | 0,21 | **0,20** | **0,20** | | | 0,21 | 0,22 | **0,21** | **Tempo Total do Treinamento** |
| Seis | 0,21 | **0,20** | 0,21 | 0,21 | | | 0,31 | 0,21 | 0,22 | 0,22 | **0,22** | |
| Sete | **0,20** | 0,21 | 0,22 | 0,65 | 0,21 | 0,21 | 0,22 | 0,22 | **0,20** | 0,21 | **0,26** | **0,38** |
| Oito | 0,22 | 0,23 | | | | 0,24 | | | 0,22 | | **0,23** | |
| Nove | 0,21 | 0,21 | **0,20** | | 0,22 | | 0,21 | 0,21 | | 0,21 | **0,21** | |
| Dez | 0,21 | 0,22 | 0,21 | 0,22 | 0,37 | 0,21 | | 0,21 | **0,20** | 0,23 | **0,23** | |

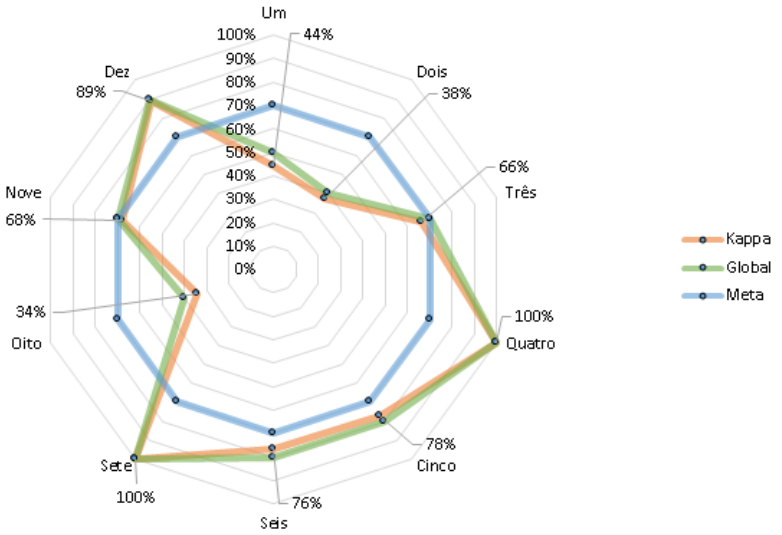
Grupo D: Taxa de Comissão (Covariância)




Grupo D: Taxa de Omissão (Covariância)




Grupo D: Acurácia Global e Coeficiente Kappa (Covariância)



| DTW (STFT) Grupo D | | Classes Preditas (Referência $+i$) | | | | | | | | | | Comissão | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Um | Dois | Três | Quatro | Cinco | Seis | Sete | Oito | Nove | Dez | $\sum x_{i+}$ | Acurácia | Erro | Individual |
| Classes Reais (Classificação $i+$) | Um | 10 | | | | | | | | | | 10 | 100% | 0% | 100% |
| | Dois | 1 | 5 | | | | | | | | 4 | 10 | 50% | 50% | 46% |
| | Três | 4 | 2 | 3 | | | | | | | 1 | 10 | 30% | 70% | 28% |
| | Quatro | 4 | | | 5 | | | | 1 | | | 10 | 50% | 50% | 47% |
| | Cinco | 2 | | | | 8 | | | | | | 10 | 80% | 20% | 78% |
| | Seis | 2 | | | | | 8 | | | | | 10 | 80% | 20% | 78% |
| | Sete | | | | | | | 9 | | | 1 | 10 | 90% | 10% | 89% |
| | Oito | 4 | | | | | | | 6 | | | 10 | 60% | 40% | 57% |
| | Nove | 3 | | | | | | | | 6 | 1 | 10 | 60% | 40% | 57% |
| | Dez | | | | | | | | | | 10 | 10 | 100% | 0% | 100% |
| Omissão | $\sum x_{+i}$ | 30 | 7 | 3 | 5 | 8 | 8 | 9 | 7 | 6 | 17 | 100 | Exatidão Global | Erro Global | Kappa Global |
| | Acurácia | 33% | 71% | 100% | 100% | 100% | 100% | 100% | 86% | 100% | 59% | 85% | 70% | 30% | 68% |
| | Erro | 67% | 29% | 0% | 0% | 0% | 0% | 0% | 14% | 0% | 41% | 15% | Acurácia Final: | | 69% |

| DTW Grupo D | Tempo de Processamento de cada Amostra por Palavra | | | | | | | | | | Tempo Médio de Reconhecimento | Tempo Total do Reconhecimento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| Um | 0,42 | 0,39 | 0,40 | 0,35 | 0,41 | 0,35 | 0,37 | 0,39 | 0,39 | 0,36 | 0,38 | 3,81 |
| Dois | | | 0,34 | | | 0,38 | 0,40 | 0,40 | | 0,39 | 0,38 | |
| Três | | 0,34 | | 0,34 | | | | | | 0,36 | 0,35 | |
| Quatro | | | | 0,41 | 0,40 | 0,39 | | | 0,38 | 0,39 | 0,39 | |
| Cinco | 0,44 | 0,40 | 0,41 | 0,38 | | 0,34 | | 0,36 | 0,41 | 0,40 | 0,39 | Tempo Total do Treinamento |
| Seis | 0,43 | 0,39 | 0,38 | 0,41 | | 0,33 | | 0,38 | 0,41 | 0,39 | 0,39 | |
| Sete | 0,41 | 0,39 | 0,41 | 0,37 | | 0,40 | 0,43 | 0,41 | 0,41 | 0,41 | 0,40 | 0,49 |
| Oito | 0,39 | | | | 0,36 | | 0,40 | 0,35 | 0,35 | 0,38 | 0,37 | |
| Nove | | 0,39 | 0,37 | | 0,38 | 0,35 | | 0,38 | | 0,39 | 0,38 | |
| Dez | 0,38 | 0,40 | 0,39 | 0,38 | 0,40 | 0,39 | 0,35 | 0,39 | 0,37 | 0,35 | 0,38 | |