Universidade Estadual Paulista
Instituto de Biociências, Letras e Ciências Exatas
Departamento de Ciência da Computação e Estatística

Luis Fernando Teixeira Silva

Um sistema para reconhecimento de comandos falados dependente do locutor

São José do Rio Preto - SP
2017

Luis Fernando Teixeira Silva

Um sistema para reconhecimento de comandos falados dependente do locutor

Monografia apresentada ao Programa de graduação em Ciência da Computação da UNESP para obtenção do título de Bacharel.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

São José do Rio Preto - SP
2017

Dedico este trabaho a todos os meus familiares, em especial aos meus pais, Nilda, Luis Carlos e a minha irmã Ana Beatriz.

Dedico também esse trabalho para a minha namorada Cristiana Luiza.

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais e à minha madrinha, pois sem o apoio deles eu nunca teria conseguido ter acesso a um ensino de qualidade como aquele que o Cursinho Alternativo me proporcionou durante todo o ano de 2012. Foi graças a essas três pessoas que pude ingressar nesta linda universidade.

Gostaria de agradecer também à minha irmã que nos momentos mais difícies da minha graduação me deu forças para continuar em frente e concluir minha formação de Bacharel em Ciência da Computação. Agradeço também a todos os meus familiares que me apoiaram ao longo desta jornada de cinco anos.

E também deixo um agradecimento especial aos meus dois grandes amigos, João Cesar Granville e Luiz Gustavo Caobianco, que tornaram estes anos na universidade mais felizes. Agradeço também minha namorada por ter me auxilidado nesses dois últimos anos de universidade e por me dar forças para concluir o curso nesta etapa final.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

WAVE *Waveform Audio File Format*

PCM *Pulse-Code Modulation*

IBM *International Business Machines*

RIFF *Rich Information File Format*

fmt *format*

IEEE *Institute of Electrical and Eletronic Engineers*

SVM *Support Vector Machine*

RNA Rede Neural Artificial

ZCR *Zero-Crossing Rate*

MFCC *Mel Frequency Cepstral Coefficients*

HMMs *Hidden Markov Models*

Sumário

Capítulo 1

Introdução

Neste Capítulo, o delineamento geral deste projeto e deste documento são esposados, provendo assim uma visão geral sobre eles.

1.1 Objetivos Específicos

Precedido por um estudo dos conceitos pertinentes, este trabalho está focado no projeto e na implementação em linguagem de programação C/C++ (1) de um algoritmo computacional para reconhecimento automático de comandos falados do Português-brasileiro em modo *off-line* e na modalidade *speaker-dependent*, isto é, com locutor pré-definido, recusando comandos proferidos por outros locutores.

1.2 Motivação e Justificativa

Automatic speech recognition (ASR) (2), ou reconhecimento automático de fala, é uma área que tem ganhado considerável destaque na atualidade, tanto no âmbito científico quanto no

industrial. Com o advento da Internet das Coisas (*Internet of Things*) (3), tem sido comum encontrar nos *veículos*, nos *smartphones* e nos *laptops*, entre outros, sistemas que fazem uso da computação para realizar o reconhecimento de comandos falados. Com os avanços da tecnologia em termos de *hardware* e de *software*, essa realidade só tende a aumentar.

Particularmente, ASR consiste na conversão de palavras faladas em texto ou, até mesmo, em comandos interpretáveis por máquinas. O potencial crescente dos computadores, principalmente de quinta geração (4), constitui um forte fator motivador para que os sistemas desenvolvidos para ASR estejam presentes em aplicações das mais variadas (6). Registre-se, ainda, que o referido tema tem merecido considerável atenção da comunidade científica, conforme é possível notar mediante uma busca realizada nas bases científicas do *Institute of the Electrical and Electronic Engineers* (IEEE) (7), do *ScienceDirect* (8) e do *Web of Science* (9), além de ter se consolidado como linha de pesquisa junto ao DCCE/IBILCE/UNESP, conforme é possível notar por meio dos trabalhos anteriores desenvolvidos na mesma área (13) (15) (16).

Em 2008, foi lançado o primeiro filme do Homem de Ferro (*Iron Man*), centrado no personagem Tony Stark. O filme além de despertar o interesse nas histórias em quadrinhos da produtora *Marvel*, também prende a atenção dos ciêntistas da computação, uma vez que o personagem principal interage com um sistema de inteligência artificial - *Jarvis* - que pode controlar a casa e a armadura do héroi. Vale ressaltar que Jarvis combina ASR com o reconhecimento do locutor Tony, aceitando apenas os seus comandos.

No ano de 2011, foi lançado a primeira versão do assistente pessoal do *iOS*, conhecido como *Siri*, que permite que o usuário execute determinadas funções do *smartphone* utilizando comandos falados. Mais recentemente, o criador do *Facebook* decidiu implementar o seu próprio assistente pessoal, que funciona também de modo dependente do locutor, para auxiliar nas tarefas domésticas.

A partir das aplicações mencionadas, surgiu a inspiração para o desenvolvimento deste tra-

balho. Adicionalmente, a Matemática, a Física e a Ciência da Computação oferecem um grande número de ferramentas e técnicas que possibilitam a implementação de sistemas de reconhecimento de padrões em sinais de voz. Técnicas para extração de características associadas aos sistemas de reconhecimento de padrões *knowledge-based* (2) têm sido alvo de pesquisas constantes. Sem dúvida, trata-se de uma área promissora de estudos, conforme recentes citações de figuras reconhecidas da computação (10). Por fim, o tema em questão, que conforme mencionado atrai fortemente a atenção do autor deste trabalho e possibilita a aplicação dos conceitos estudados durante o curso de Bacharelado em Ciência da Computação para a solução de um problema de amplo interesse.

1.3 Metodologia

Inicialmente e com base nas diretrizes traçadas pelo orientador deste trabalho, foi realizado um levantamento bibliográfico envolvendo os conceitos pertinentes para a elaboração do respectivo projeto, conforme consta no Capítulo 2. Em seguida, foram definidos, mediante livre escolha deste autor, os seguintes onze comandos específicos a serem reconhecidos pelo algoritmo a ser implementado: “Bom dia, Logan”, “Boa noite, Logan”, “Oi, Logan”, “Como está o tempo hoje ?”, “Vai chover ?”, “Abrir calculadora”, “Ver notícias”, “Pesquisar”, “Alarme”, “Calendário” e “Sair”.

Posteriormente e após a definição do locutor-padrão, pretende-se realizar a gravação digital de dez sinais de voz para cada um dos onze comandos mencionados, totalizando 110 arquivos de áudio no formato *WAVE* com taxa de amostragem $t > 8000$ amostras por segundo e quantização $q \geqslant 16$ bits. Na sequência, serão extraídos os dados brutos dos sinais, com o uso de uma biblioteca escrita em linguagem C/C++ e fornecida pelo orientador deste trabalho. O trabalho será automatizado com a implementação de *scripts* escritos na linguagem *Shell script*.

Seguindo a extração dos dados brutos, será realizada a extração das características dos sinais de voz com base no uso dos conceitos de energia, seguindo a técnica A_3 descrita em (11), e taxa de cruzamentos por zero (ZCR - *zero-crossing rates*), seguindo a técnica B_3 descrita em (12). Esta e aquela técnicas, respectivamente, permitem contabilizar a fração dos tamanhos dos sinais de voz, a partir dos seus inícios, necessárias para alcançar $K \cdot C\%$ do total de ZCRs e do montante de energia, sendo C uma porcentagem definida experimentalmente e $K = 1, 2, ...,$ $(K \cdot C < 100\%)$. Por um lado, a energia garante a análise do esforço pulmonar realizado ao longo do tempo pelo locutor que produziu o sinal de voz sob análise (2). Por outro, o ZCR captura características espectrais dos sinais, principalmente relacionadas com o período de *pitch* (F_0) (2), que constitui um dos fatores para caracterização do locutor.

Uma vez que cada sinal de voz tenha sido convertido em seu vetor de características, proceder-se-á com o estudo gráfico dessas representações visando decidir qual a melhor técnica de classificação a ser utilizada, que pode pertencer aos métodos *pattern-matching* ou *knowledge-based* (6). Naquelas técnicas estão inclusas as sistemáticas mais simples, tais como as medições de distâncias ou similaridades. Nestas, por outro lado, incluem-se metodologias mais elaboradas, a exemplo das redes neurais (5). Testes e resultados serão realizados considerando um procedimento de validação cruzada e os resultados serão contabilizados em matrizes de confusões (5).

1.4 Exequibilidade

Para a elaboração deste projeto, serão utilizadas ferramentas gratuitas como *Sublime*, *LibreOffice*, *TeXstudio*, *Audacity* e um computador pessoal. Além disso, serão também utilizados livros e artigos científicos de acesso grátis e disponíveis no *Web of Science*, IEEE, *Science Direct* e no acervo da biblioteca do Instituto de Biocências, Letras e Ciêcias Exatas da UNESP. Assim, entende-se que o projeto é exequível sem maiores dificuldades.

1.5 Organização do trabalho

Ao seu término, esta monografia estará organizada da seguinte forma:

- no Capítulo 2, apresenta-se uma série de trabalhos realizados na área de ASR tanto em nível local quanto internacional, exaltando a relevância da área no âmbito acadêmico. Além disso, é apresentado também neste capítulo, toda a fundamentação teórica do trabalho, definindo e exemplificando os principais conceitos utilizados na elaboração do mesmo;

- no Capítulo 3 apresentar-se-á todo o trabalho realizado, com detalhes. Neste momento, no qual a versão deste documento ainda está incompleta, apresenta-se uma breve descrição do estado do atual do trabalho e também um cronograma para a sua finalização;

- no Capítulo 4, serão apresentados os testes e os resultados respectivos;

- no Capítulo 5, documentar-se-ão as conclusões que serão seguidas das referências e apêndices contendo os códigos-fonte dos projeto.

Capítulo 2

Revisão Bibliográfica

2.1 A Voz Humana

A voz é uma característica única dos seres humanos, que foi desenvolvida a partir da necessidade do homem em se socializar e, assim, poder se comunicar com os seus semelhantes. Em outras palavras, a fala foi adquirida por meio de um processo evolutivo de modo que, ao longo dos séculos, a raça humana tem adaptado um conjunto de orgãos da parte superior dos aparelhos digestivo e respiratório para a produção de fonemas.

De modo geral, pode-se dividir o processo de vocalização nas seguintes partes:

1. **pulmões**: são responsáveis por gerar o fluxo de ar que é impulsionado por meio da contração do orgão pelo diafragma. Tal fluxo, funciona como um "combustível" para a voz;

2. **laringe**: local onde são encontrada as pregas vocais, que são dois músculos que têm a responsabilidade de modelar a voz com base na sua contração e/ou relaxamento. Quando o fluxo de ar que sai dos pulmões atravessa a faringe, as pregas vocais vibram com maior ou menor intensidade, logo pode-se perceber que a voz humana é formada pela interação de duas principais forças: a força com que o fluxo de ar sai dos pulmões e força muscular

da laringe;

3. **articuladores**: eles são localizados na cavidade bucal e são responsáveis por fazer a articulação do fluxo de ar vindo dos pulmões. Os principais articuladores são: lábios, língua, dentes e mandíbula, conjuntamente chamados de trato vocal.

Vale ressaltar que a partir da característica vibratória das pregas vocais, é possível determinar a frequência fundamental da voz humana (F_0), o que permite fazer a dintinção de gêneros e, com o auxílio de outras características, o reconhecimento de voz e comandos. A estrutura do sistema descrito consta na Figura 2.1.

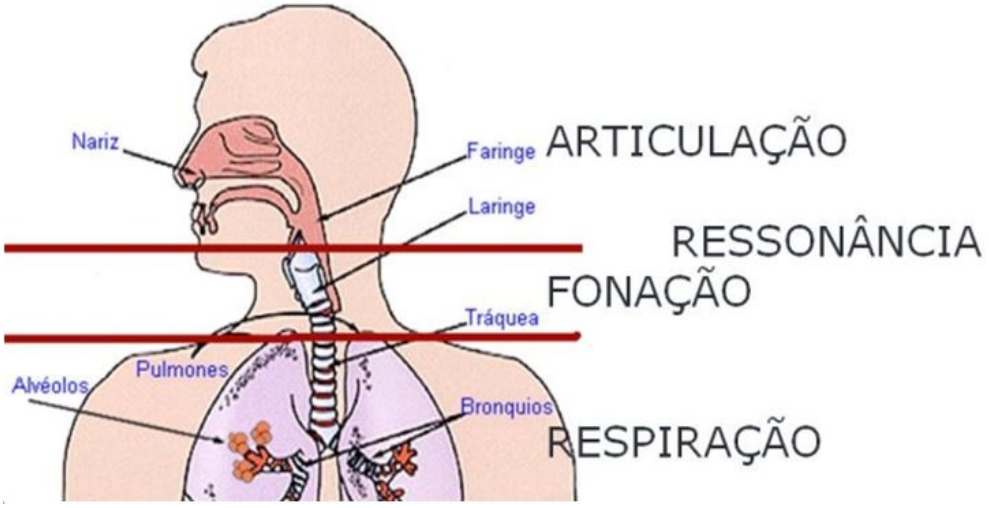

Figura 2.1 – Fisiologia da voz (extraído de (27)).

2.2 Sinais Digitais

Os dados podem ser representados de duas maneiras: analógica ou digital. Aquela, em geral, caracteriza-se por uma onda eletromagnética que pode assumir infinitos valores ao longo do tempo, sendo a voz humana como é naturalmente encontrada na natureza, um exemplo. Esta, por outro lado, corresponde a um conjunto de valores com precisão finita obtidos por um processo de amostragem temporal. Em outras palavras, o sinal digital que não seja oriundo de

um processo de síntese, corresponde ao sinal analógico discretizado no tempo e quantizado em termos de amplitudes.

Particularmente, para se realizar uma análise em um sinal analógico em um computador, é necessário realizar a sua conversão para a forma digital, uma vez que os computadores possuem uma capacidade finita de processamento, não sendo possível assim processar os infinitos pontos de um sinal analógico. Tal processo de conversão é denominado de digitalização, que é realizado em três etapas. Na figura 2.2 são apresentados os processos de amostragem e de quantização, que serão comentados a seguir.

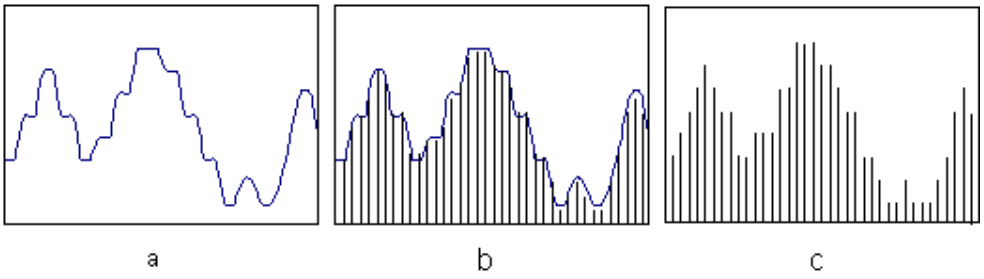

Figura 2.2 – (a) Sinal na forma analógica. (b) Amostragem. (c) Quantização (extraído de (26)).

1. **amostragem:** consiste no processo de obter amostras de um sinal analógico em instantes de tempo com espaços iguais. Vale ressaltar que para se obter sucesso nesse processo de amostragem, é necessário seguir o Teorema de Nyquist (2). Tal teorema define que, para realizar a amostragem de um sinal, é necessário no mínimo ter uma taxa de amostras igual a duas vezes a máxima frequência do sinal analógico. Caso não se obtenha tal taxa de amostragem ocorrerá um efeito chamado de *aliasing*, que consiste na medição erronea de uma frequência mais alta como sendo mais baixa. Um bom exemplo da aplicação de tal teorema é a digitalização da voz humana, que possui uma frequência máxima em torno de 4 KHz. Logo, para realizar amostragem desses sinais, é necessário no mínimo uma taxa de 8 mil amostras por segundo;

2. **quantização:** consiste na atribuição de valores discretos para as amplitudes do sinal analógico, que pertencem a um intervalo contínuo de valores. Cada amplitude é alocada ao nível de valor discreto mais próximo, o que diminui assim os erros absolutos.

Vale ressaltar que o número de níveis é definido pelo número de *bits* que serão utilizados na etapa posterior de codificação, ou seja, 2^n, sendo n o número de bits que será usado. De modo geral, nos sinais acústicos, é utilizada a quantização de 16 *bits*. Em outras palavras, as amplitudes de cada amostra podem assumir $2^{16} = 65536$ valores distintos;

3. **codificação:** etapa que consiste na modificação de um sinal para uma aplicação específica, como por exemplo o armazenamento ou transmissão de dados.

Neste trabalho, todos os sinais de voz originalmente analógicos foram digitalizados para permitir manipulação por intermédio computacional.

2.3 Arquivos Acústicos no Formato *WAVE*

Waveform audio file format, abreviado de *WAVE* ou simplesmente *WAV*, é um tipo de formato de arquivos de áudio que foi desenvolvido pela *Microsoft* em conjunto com a IBM. O formato *WAVE* é amplamente utilizado em uma variedade de trabalhos, sejam eles científicos ou profissionais, por permitir uma fiel representação dos dados digitalizados, que não sofrem compressão.

Na Tabela 2.1, a estrutura de um arquivo *WAVE* está descrita. Basicamente, o arquivo é dividido em dois grandes blocos, sendo o primeiro bloco um cabeçalho *Rich Information File Format* (RIFF) e o segundo bloco, por sua vez dividido em dois sub-blocos, um conjunto de informações referentes ao arquivo seguido dos dados brutos.

Vale ressaltar que a quantização mais comum para cada amostra de um arquivo *WAVE* é normalmente de 8 *bits* ou 16 *bits*. No primeiro caso, existem 256 opções para cada amostra do sinal, sendo 127 positivos e 128 negativos. No segundo caso, por outro lado, são 65536 possibilidades, com 32767 positivos e 32768 negativos. Particularmente para a quantização de

Tabela 2.1 – Estrutura de um arquivo *WAVE*

Classe	Posição *(bytes)*	Tamanho *(bytes)*	Descrição
Cabeçalho	0	4	identificador do cabeçalho: “RIFF”
Cabeçalho	4	4	Tamanho do arquivo
Cabeçalho	8	4	identificador “*WAVE*”
Formato	12	4	identificador do segundo bloco: “fmt”
Formato	16	4	tamanho do bloco sem o identificador
Formato	20	2	existência ou não de compressão
Formato	22	2	quantidade de canais
Formato	26	4	taxa de amostragem
Formato	30	4	taxa de *bytes*
Formato	32	2	quantidade de *bytes* para uma amostra
Formato	34	2	quantidade de *bits* para cada amostra
Dados	36	4	identificador do terceiro bloco: “*data*”
Dados	40	4	tamanho do bloco sem o identificador
Dados	44	4	sinaliza o início dos dados brutos

16 *bits*, é utilizada a codificação de complemento de 2 para representar cada valor da amplitude do sinal. Assim, o valor do *bit* mais significativo representa se o sinal é negaivo ou positivo. Neste trabalho foi utilizado o formato *WAVE* de 16 *bits Pulse-code Modulation* (PCM), ou seja, sem compressão.

2.4 Reconhecimento de Padrões e Vetores de Características

Na área de computação, reconhecer um padrão consiste, normalmente, em classificar sinais como pertencentes a determinadas classes ou categorias com base na extração de características relevantes dos mesmos. De modo geral, o processo de classificação consiste das seguintes fases:

- **extração das características:** principal etapa de todo o processo de reconhecimento de padrões, consiste na redução de dimensionalidade do sinal objetivando preservar elementos de fundamental interesse para a classificação, compondo assim o vetor de características. Caso esse processo seja mal elaborado, perder-se-ão características significativas, o que pode tornar a classificação dispendiosa. Portanto, é necessário ter um

conhecimento específico sobre o problema, para poder realizar a redução de dimensionalidade sem que ocorrá perdas de informações relevantes, presando ainda pela redução do esforço computacional;

- **classificação do sinal:** etapa que determina os procedimentos para realizar a classificação do sinal em si, agora representado pelo seu vetor de características. De modo geral, os algoritmos de classificação podem ser dos tipos *pattern-matching* e *knowledge-based*. No primeiro caso, as técnicas são mais elementares e não requerem treinamento prévio, isto é, nenhuma estatística sobre os dados a serem classificados é realizada. Por outro lado, no segundo caso, os algoritmos requerem um treinamento prévio levando em conta os padrões a serem classificados, podendo ser:

  - **treinamento supervisionado:** seu funcionamento é baseado em exemplos de entradas e saídas previamente conhecidos, o que permite assim aprender uma regra genérica para classificar as entradas posteriores;

  - **treinamento não supervisionado:** ao contrário da aprendizagem supervisionada, esta técnica não se baseia em rótulos previamente conhecidos para classificar os sinais.

Neste trabalho, o algoritmo particular a ser utilizado como classificador será determinado mediante a análise gráfica dos vetores de características. Vetores com padrões claramente distinguíveis podem fazer uso de técnicas mais modestas, isto é, *pattern-matching* tal como a Distância Euclidiana. Por outro lado, vetores com padrões similares entre as classes requerem classificadores mais elaborados, isto é, do tipo *knowledge-based*, tal como uma rede neural artificial.

2.5 O Conceito de Energia

A definição de energia está relacionada ao conceito de potencial para realizar trabalho. Neste projeto, a energia será considerada como a capacidade das estruturas vocálicas e dos pulmões de produzir um sinal acústico. Na equação 2.1, define-se a energia total $E(s[\cdot])$ de um dado sinal acústico digitalizado $s[\cdot]$ de tamanho M.

$$E(s[\cdot]) = \sum_{i=0}^{M-1} (s_i)^2 \tag{2.1}$$

Para realizar a captura das caracteríticas dos sinais, foi utilizado o método A_3 definido em (11). Tal método baseia-se em determinar comprimentos proporcionais para atingir níveis predefinidos da energia do sinal que se encontra em análise. A_3 é ideal para avaliar os níveis de energia de um sinal de voz digitalizado que foi gerado por um agente.

Vale ressaltar que A_3 requer a definição de um nível crítico de energia, C, que varia de 0 a 100%. A_3 extrai um vetor de características dividindo o sinal original em partes proporcionais ao valor de C, ou seja, $K \cdot C$, sendo $K = 1, 2, ...$ e $(K \cdot C < 100\%)$. Assim, por exemplo, se $C = 5\%$ e o vetor de características for designado por $f[\cdot]$, tem-se:

- f_0 corresponde a fração do comprimento total do sinal original, a partir do início, necessária para alcançar $1 \cdot 5 = 5\%$ da energia total do sinal;
- f_1 corresponde a fração do comprimento total do sinal original, a partir do início, necessária para alcançar $2 \cdot 5 = 10\%$ da energia total do sinal;
- f_2 corresponde a fração do comprimento total do sinal original, a partir do início, necessária para alcançar $3 \cdot 5 = 15\%$ da energia total do sinal;
- ...
- f_{18} corresponde a fração do comprimento total do sinal original, a partir do início, ne-

cessária para alcançar $19 \cdot 5 = 95\%$ da energia total do sinal.

Na figura 2.3, pode-se observar uma ilustração do funcionamento de A_3.

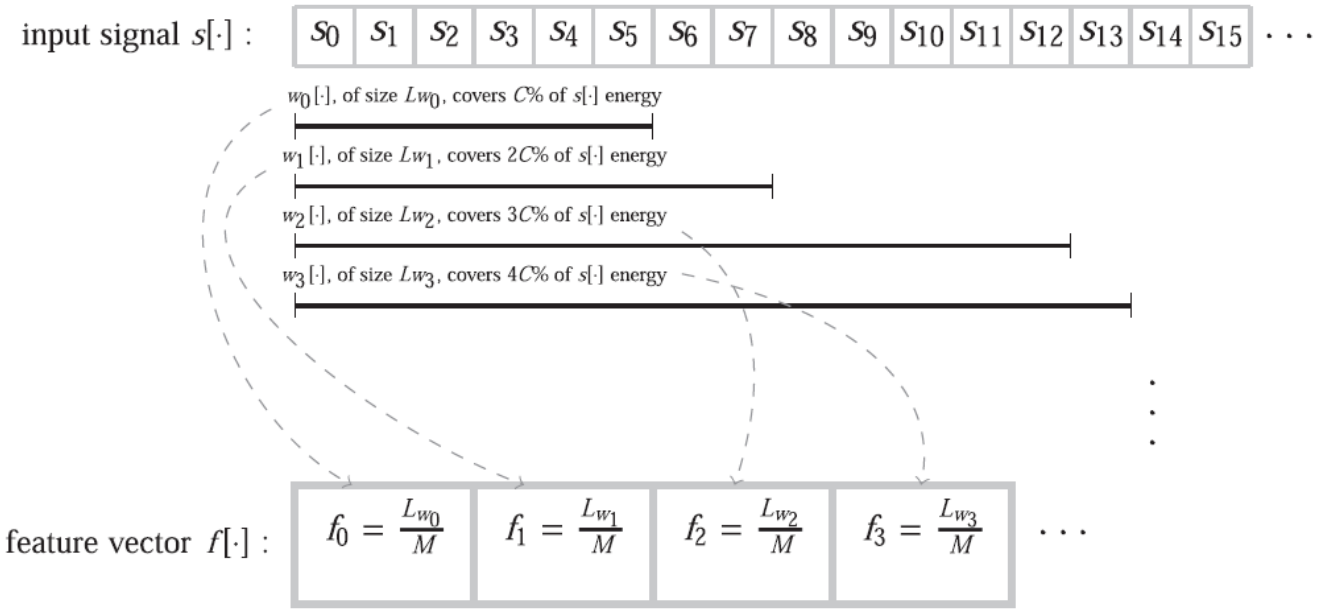

Figura 2.3 – A_3 (extraído de (11))

2.6 O Conceito de *Zero-crossing Rate* (ZCR)

A Taxa de Cruzamento por Zeros ou *Zero-Crossing Rate* - ZCR - é utilizada para definir quantas vezes a onda de um sinal cruza a amplitude zero. Pode-se também definir ZCR de acordo com a equação 2.2, onde S[.] $= \{S_0, S_1, ..., S_{M-1}\}$ é um sinal discretizado de tamanho M>1 e sign(x) $= \begin{cases} 1, & se\ x \geqslant 0; \\ -1, & caso\ contrário. \end{cases}$

Para um sinal discreto, pode-se calcular o ZCR observando a mudança do sinal de negativo para positivo, ou vice-versa, considerando as amostras vizinhas.

De modo geral, os sons vocálicos estão concentrados em baixas frequências, o que implica que se o ZCR for baixo pode-se classificar o som vocálico, caso contrário como não vocálico.

$$ZRC(S[.]) = \frac{1}{2}\sum_{j=0}^{M-2} |sign(S_j) - sign(S_{j+1})| \tag{2.2}$$

Neste trabalho o conceito de ZCR foi utilizado com base no método B_3 definido em (12). Ele consiste em determinar comprimentos ou áreas proporcionais de um sinal que são requeridos para alcançar porcentagens prédefinidas do total de ZCR, o que torna este método bem semelhante ao A_3 (11). B_3 é ideal para inspecionar a consistência na frequência de uma entidade física responsável por gerar S[.].

Assim como foi definido em A_3, B_3 requer que seja estabelecido um nível crítico de ZCR, C, que pode variar de 0 a 100% e a partir disso, pode-se definir um vetor de características $f[.]$ de tamanho T, conforme a seguir:

- f_0 equivale a fração de comprimento de S[.], a partir do seu início até alcançar C% do total de ZCR;
- f_1 equivale a fração de comprimento de S[.], a partir do seu início até alcançar $2 \cdot C\%$ do total de ZCR;
- ...
- $f_T - 1$ equivale a fração de comprimento de S[.], a partir do seu início até alcançar $T \cdot C\%$ do total de ZCR, de modo que $T \cdot C < 100\%$;

Na figura 2.4 pode-se observar uma ilustração do funcionamento do método B_3. Além disso, pode-se definir T do seguindo modo: $\mathrm{T} = \begin{cases} \frac{100}{C} - 1, & \textit{se C é múltiplo de } 100; \\ \lfloor \frac{100}{C} \rfloor, & \textit{caso contrário.} \end{cases}$

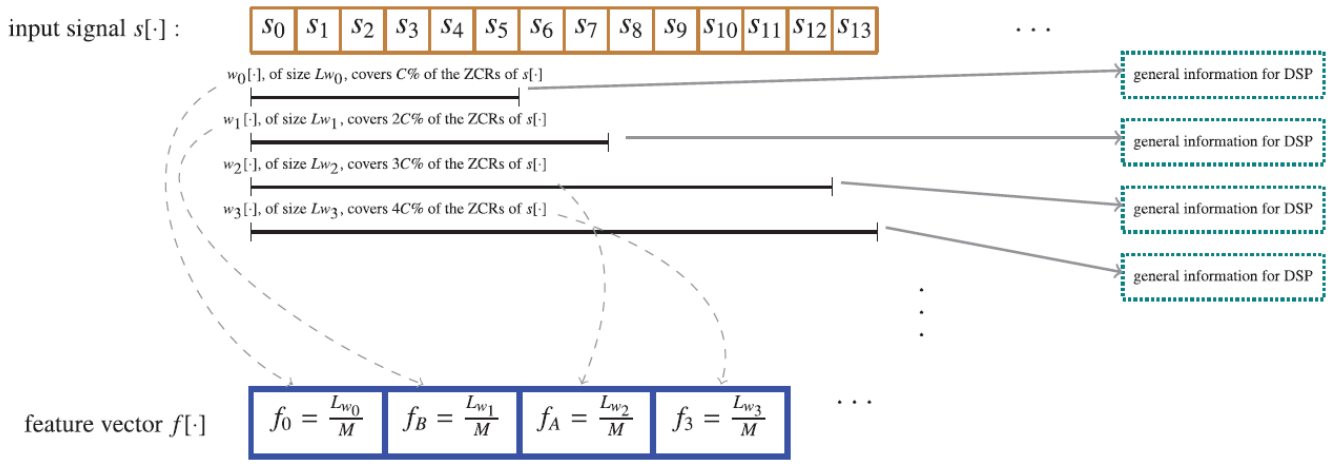

Figura 2.4 – B_3 (extraído de (12))

2.7 Classificação baseada em distâncias

Quando é necessário agrupar vetores de características, normalmente utiliza-se como quesito a similaridade ou dissimilaridade entre os mesmos. Em específico, para avaliar similaridades, uma série de métodos podem ser utilizados. No escopo das técnicas *pattern-matching*, a distância Euclidiana é uma das mais clássicas possibilidades sendo amplamente utilizada na literatura, com (13) e (15) sendo alguns dos inúmeros exemplos de trabalhos que utilizam a referida técnica. De acordo com a equação 2.3, onde P = $\{P_1, P_2, ..., P_n\}$ e Q = $\{Q_1, Q_2, ..., Q_n\}$ são dois vetores de características que distam d unidades um do outro. A similaridade entre dois objetos será máxima quando $P = Q$, implicando que $d = 0$.

$$d(P,Q) = \sqrt{(P_1 - Q_1)^2 + (P_2 - Q_2)^2 + ... + (P_n - Q_n)^2} = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2} \quad . \tag{2.3}$$

Uma variante da distância Euclidiana é a distância absoluta. Tal técnica utiliza o módulo da diferença entre dois números para realizar a medição da similaridade, conforme é mostrado na equação 2.4. Assim como no caso anterior, caso a diferença seja igual a zero, tem-se que os

dois vetores comparados são idênticos.

$$d = \sum_{i=1}^{n} |P_i - Q_i| \quad . \tag{2.4}$$

2.8 Classificação baseada em redes neurais

Uma Rede Neural Artificial (RNA) é composta por dois elementos principais que são intimamente relacionados: A arquitetura da rede e o algoritmo de aprendizado. Esse por sua vez, é tipicamente organizado em até três camadas, com unidades que podem estar ligadas às unidades da camada posterior. Na figura 2.5 é apresentada um exemplo de uma arquitetura de RNA.

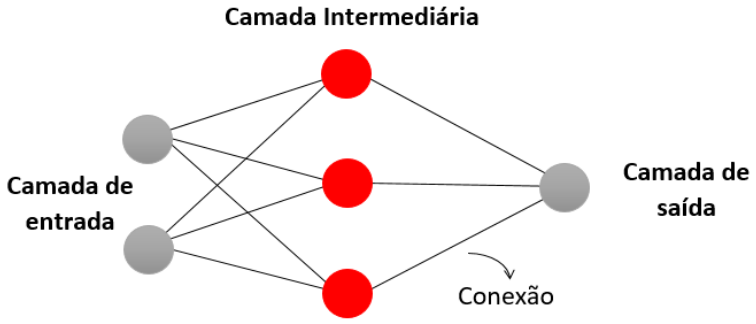

Figura 2.5 – Exemplo de uma arquitetura de uma RNA

Normalmente, as camadas são classificadas como segue:

- **Camada de entrada:** É nessa camada que os padrões serão apresentadas à rede.
- **Camada Intermediária:** É responsável por realizar a maior parte do processamento, através das conexões ponderadas.
- **Camada de saída:** Tem como funcionalidade concluir e apresentar o resultado final obtido pelo processamento.

Vale ressaltar que os nós da estrutura representado em 2.5 são considerados como neurônios, e toda informação que passa por uma conexão localizada entre aqueles leva à geração de sinap-

ses. Além disso, as RNAs funcionam de modo distribuído e paralelo, de tal forma que cada nó possui capacidade para realizar processamento.

Por fim, a outra entidade elementar de um RNA é o algoritmo de aprendizagem que, por sua vez, tem como objetivo adquirir conhecimento do ambiente. Em outras palavras, adapta os pesos das conexões aos estímulos ou entradas captadas pelo ambiente de modo iterativo. Os algoritmos de aprendizagem podem ser classificados nos seguintes paradigmas: supervionado e não supervionado, conforme mencionado na seção 2.4.

Neste trabalho, conforme já mencionado, dependendo da análise gráfica dos vetores de características, poderá a vir ser utilizada uma máquina de vetor suporte (SVM - *Support Vector Machine*) ou uma RNA para realizar as classificações e reconhecimento dos comandos.

A SVM tem como principal característica ser um classificador binário de modo que consegue realizar a distinção de duas classes por uma função induzida por meio de exemplos. Essa distinção é obtida pela procura de um hiperplano ótimo, conforme pode ser observado na figura 2.6, na qual círculos verdes e amarelos representam cada uma das classes e as linhas solidas representam as separações entre os planos.

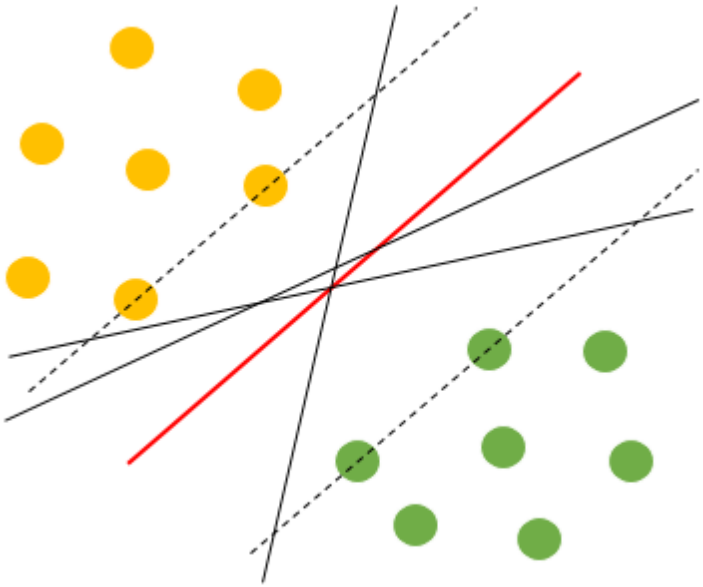

Figura 2.6 – Hiperplano ótimo

Na figura 2.6, pode-se analisar que existem vários classificadores lineares para separar as amostrar, mas há somente um (linha vermelha) que maximiza a margem - distância máxima

entre um elemento mais próximo de cada classe e o classificador. Esse por sua vez é comumente referenciado na literatura como hiperplano de separação ótimo, pois ele consegue, mesmo que intuitivamente, generalizar melhor que os demais classificadores. Já as linhas tracejadas passam por alguns pontos para ambas as classes, esses são chamados de vetores-suporte.

2.9 Matrizes de Confusão

Uma matriz de confusão é comumente utilizado para avaliar o desempenho obtido por um classificador (6), ao mostrar a quantidade de classificações corretas versus as classificações previstas para cada classe, sobre um determinado conjunto de exemplos.

Uma matriz de confusão M é uma matriz de duas dimensões, com o mesmo número de linhas e colunas (n x n), no qual cada linha i representa a classe real daquilo que está sendo classificado, e cada coluna j representa classe identificada pelo classificador (16). A quantidade de acertos, para cada classe, se encontra na diagonal principal m_{ij}, com $i = j$. Já as demais células de M representam os erros de classificação.

Tabela 2.2 – Exemplo de uma matriz de confusão

	Bom dia, Logan	Vai chover?	Pesquisar
Bom dia, Logan	8	2	0
Vai chover?	2	6	1
Pesquisar	0	0	10

A tabela 2.2 mostra um exemplo de uma possível matriz de confusão para o trabalho aqui proposto. Em 2.2, temos 3 comandos distintos a serem identificados pelo classificador. Pode-se observar que para o comando *Bom dia, Logan*, obteve-se 8 acertos e 2 erros, no qual o classificador identificou o comando em questão como *Vai chover?*. Já para o comando *Pesquisar*, obteve-se 10 acertos e nenhum erro.

Portanto, pode-se concluir que o melhor caso de uma classificação ocorre quando a diagonal

principal de M é máxima e as demais células da matriz estão zeradas, em outras palavras, o classificador obteve 100% de acurácia.

2.10 Trabalhos Correlatos

A área de ASR tem sido bastante estudada em diversas trabalhos. Nas monografias de conclusão de curso referenciadas em (13–17, 21–23) são estudados e implementadas técnicas para identificação de palavras no âmbito de vocabulário restrito. Por outro lado, nos artigos (24, 25), os esforços são voltados para o reconhecimento de um amplo vocabulário, inclusive em idiomas distintos do Português.

Particularmente em (13), o autor utiliza conceitos como energia, limiar *hard* e níveis críticos de energia, para elaborar um sistema de reconhecimento de vocabulário restrito implementado na linguagem C/C++, com um classificador baseado em distância Euclidiana.

Em (15), o autor utiliza duas abordagens diferentes para realizar o reconhecimento de voz, sendo a primeira baseada em algoritmos *pattern-matching* e a outra em *knowledge-based*. Ao final do trabalho, pode-se concluir e verificar que a segunda abordagem apresentou melhores resultados para uma base de algumas dezenas de palavras.

No trabalho referenciado em (16), o autor propõem um sistema para reconhecimento de voz independente do locutor. Para isso, o autor utiliza conceitos como ZCRs e energias para realizar o pré-processamento, extração de características e um classificador também baseado em distância Euclidiana, contando com acurácias superiores a 90% para uma base de dezenas de palavras.

Na monografia documentada em (18), o autor desenvolve um sistema de reconhecimento de

voz baseado em automação residencial via *wireless*. Para a elaboração desse projeto, o autor utiliza o dispositivo HM2007L IC, para assim realizar o reconheicmento de comandos simples, como “acender”, “apagar”, “acender a luz”, entre outros, contando com o mesmo nível de acurácia dos trabalhos anteriormente mencionados.

Em (19), o objetivo do autor é melhorar a relação homem-máquina, e com isso, desenvolve-se um método para reconhecimento de voz com o uso de características *voice-likeness*. Alcança-se uma acurácia superior a 90% considerando centenas de termos.

No trabalho (20), o autor tem como objetivo desenvolver um sistema para *smartphones*, multifuncional, que além de realizar o reconhecimento de voz, também faça a conversão do diálogo reconhecido para texto.

Em (21), o autor utiliza conceitos como *Mel frequency cepstral coefficients* (MFCC) e *Hidden Markov Models* (HMMs) para reconhecer um restrito vocabulário que contém digitos de 0 a 9, alcançando acurácia quase plena para algumas centenas de testes.

No artigo científico (24), é realizada uma comparação de desempenho de Modelos de Misturas Gaussianas com Redes Neurais Profundas para o reconhecimento de um amplo vocabulário Chinês. Os autores demonstram que as redes profundas são superiores aos modelos Gaussianos na maior parte das situações.

Em (25), com o objetivo de melhorar a eficácia do reconhecimento de um vocabulário Russo amplo, foram utilizadas técnicas *knowledge-based* para modelagem de fonemas, alcançando-se uma acurácia superior a 95% para muitas centenas de testes.

Finalmente, no trabalho a ser implementado nesta monografia, será adotada uma linha de raciocínio similar aos trabalhos (13–15), onde os autores utilizam conceitos já conhecidos e consolidados na literatura da área de reconhecimento de voz, tais como níveis críticos de ener-

gia, distância Euclidiana, entre outros. Particularmente, este trabalho diferencia-se daqueles por objetivar o reconhecimento de palavras, num vocabulário restrito, de modo **dependente**, e não independente, do locutor.

Capítulo 3

Detalhamento do Trabalho Proposto

3.1 Status do trabalho proposto e sua continuidade

No momento no qual este capítulo foi escrito, a revisão bibliográfica já havia sido realizada, assim como a coleta e elaboração do banco de sinais de voz.

Com relação aos sinais, foram definidos os 11 comandos que serão reconhecidos pelo sistema, conforme já mencionado. Os comandos foram gravados 10 vezes em diferentes dias e horários, para obter se assim, uma melhor veracidade e fidelidade à voz do locutor, pois o mesmo pode sofrer alterações significativas com base na variação do seu humor ou estado físico. A gravação dos sinais foi realizada em um ambiente fechado, para diminuir assim, a probabilidade de ruídos. Todos os arquivos foram gravados no formato *WAVE* de 16 *bits* PCM, com o auxílio do editor de áudio *Audacity*. Posteriormente, foi iniciada a extração das características de todos os sinais.

A continuidade do trabalho seguirá o seguinte cronograma proposto:

Tabela 3.1 – Cronograma para desenvolvimento do trabalho

Atividades	Julho	Agosto	Set	Out	10 Nov
Desenvolvimento					
Escrita do Capítulo 3					
Escrita do Capítulo 4					
Escrita do Capítulo 5					
Correções					
Entrega da Monografia					

Referências

1 Dr. Gary J. Bronson. A First Book of C++, 2011. 4. ed. Course Technology.

2 Dong Yu, Li Deng. Automatic Speech Recognition: A Deep Learning Approach. 1 ed. Springer, 2015.

3 Graham Meikle, Mercedes Bunz. The Internet of Things. 1 ed. Polity Press, 2017.

4 HARRIS, D.; HARRIS, S. **Digital Design and Computer Architecture**, 2.ed. Morgan Kaufmann, 2012.

5 DUDA, R.; HART, D.; STORK, D. **Pattern Classification**. 2 ed. New York: Wiley-Interscience, 2000.

6 THEODORIDIS, S.; KOUTROUMBAS, K. Pattern Recognition. 4. ed, Academic Press, 2008.

7 http://www.ieeexplore.ieee.org. Acesso em Abril de 2017.

8 http://www.sciencedirect.com. Acesso em Abril de 2017.

9 http://isiknowledge.com. Acesso em Abril de 2017.

10 https://www.cnet.com/au/news/gates-still-finding-his-voice/. Entrevista com Bill Gates. Acesso em Abril de 2017.

11 GUIDO, R. C. A tutorial on signal energy and its applications. Neurocomputting, v. 179, p.264-282, 2016.

12 GUIDO, R.C. ZCR-aided neurocomputing: a study with applications. *Knowledge-based Systems*, v. 105, pp.248-269, 2016.

13 DORDAN, M, K. Verificação de Locutores dependente do discurso baseada na Evolução do Esforço Vocálico. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2015.

14 SILVA, V. F. Identificação e Classificação de Gẽneros Musicais com Abordagens Múltiplas de Reconhecimento de Padrões. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2014.

15 CAOBIANCO, A. P. Comparação de Abordagens *PatternMatching* e *Knowledge-Based* para Reconhecimento de Locutor Dependente de Texto. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2011.

16 MASCHIO, J. V. D. Projeto e Implementação Acústico-Computacional de Palavras Isoladas. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2017.

17 LOUREIRO, W. d. F. Reconhecimento de Comandos Falados Português Brasileiro com Parâmetros de Frequência e Tempo. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2015.

18 PAUL, A.; PANJA, M.; BAGCHI, M.; DAS, N.; MAZUMDER, R. M.; GHOSH, S. Voice Recognition Based Wireless Room Automation System. In: International Conference on Intelligent Control Power and Instrumentation, 2016.

19 SAKANO, T.; KIGAWA, T.; SUGIMOTO, M.; KUSUNOKI, F.; INAGAKI, F.; MIZOGUCHI, H. Laughing Voice Recognition Using Periodic Waveforms and Voice-likeness Features. *In: Proceedings of the 2016 IEEE International Conference on Robotics and Biomimetics*. Qingdao, China, 2016.

20 CHERN, A.; TSAO, Y.; CHANG, R.; HSIU-WEN, C.; YING-HUI, L. A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Clas-

sroom. Disponível em: <http://ieeexplore.ieee.org/document/7938619/>. Acesso em: 05 jun. 2017.

21 SILVA, A. G. d. Reconhecimento de Voz para Palavras Isoladas. Monografia (Graduação em Engenharia da Computação) - UFPE, Recife, 2009.

22 LOUZADA, J. Reconhecimento Automático de Fala por Computador. Monografia(Graduação em Ciẽncia da Computação) - PUC, Goiás, 2010.

23 Piccoli, E. E. M. Reconhecimento de Padrão em Áudio no Formato Wave. IBILCE, UNESP, São José do Rio Preto (Trabalho de Conclusão de Graduação em Ciẽncia da Computação), 2014.

24 LI, X.; YANG, Y.; PANG, Z.; WU, X. A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks Based large Vocabulary Chinese Speech Recognition. Neurocomputing, v. 170, p. 251-256, 2015.

25 KARPOV, A.; MARKOV, K.; KIPYATKOVA, I.; VAZHENINA, D.; RONZHIN, A.; Large Vocabulary Russian Speech Recognition Using Syntactico-statistical language Modeling. Speech Communication, v.56, p.213-228, 2014.

26 <http://penta3.ufrgs.br/RNP/cap3/3.2%20Audio/>. Acesso em: 20 Abril 2017.

27 <http://www.koreapost.com.br/wp-content/uploads/2016/01/fisiologia-da-voz.jpg>. Acesso em: 18 Abril 2017.