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

Ficha catalográfica elaborada pelo Serviço de Biblioteca do IBILCE/UNESP

Luis Fernando Teixeira Silva
titulo
titulo
titulo. / fulano de tal; orientador
Rodrigo Capobianco Guido. São José do Rio Preto, 2017.
xxx p.

Monografia (TCC
TCC
TCC, 2017.

1. Processamento de sinais. 2. Reconhecimento de locutor. 3. Acústica. 4. Energia. 5. Escala *Bark*. I. Capobianco Guido, Rodrigo, orient. II. Título.

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 meus pais e minha madrinha, pois sem o apoio deles eu nunca teria conseguido ter acesso a um ensino de qualidade que o cursinho alternativo me proporcionou durante todo o ano de 2012. Foi graças a essas 3 pessoas que pude ingressar nessa linda universidade.

Gostaria de agradecer também a 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 longos dessa jornada de 5 anos.

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

"No fim tudo dá certo, e se não deu certo é porque ainda não chegou ao fim."

Fernando Sabino

Resumo

TAL, F. *titulo*. 2016. xxxp. TCC UNESP 2016.

Atualmente, ...

Palavras-chave: Processamento de sinais. Reconhecimento de locutor. Acústica. Escala *Bark*.

Abstract

TAL, F. *titulo*. 2016. xxxp. TCC UNESP 2017.

Nowadays, ...

Keywords: Signal processing. Speaker recognition. Acoustics. Bark scale.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

WAVE *Waveform Audio File Format*

PCM *Pulse-Code Modulation*

IBM *International Business Machines*

RIFF *Resource Interchange File Format*

fmt *format*

IEEE *Institute of Electrical and Eletronic Engineers*

Sumário

Capítulo 1

Introdução

1.1 Introdução

asssas

1.2 Objetivo

Este trabalho tem como objetivo implementar um algoritmo computacional desenvolvido em C/C++ para reconhecer comandos falados de modo *off-line* com locutor prédefinido, ou seja *speaker-dependent*. Esses comandos foram previamente gravados em arquivos no formato *WAVE* de 16 *bits* PCM.

1.3 justificativa

justificar e oferecer razao suficiente para a construcao desse trabalho. Responde a pergunta do que pq fazer o trabalho, procurando os antecedentes do problema e a relevancia do assunto, argumentando sobre sua importancia pratico/teorica, colocando as possiveis contribuicoes esperadas

1.4 Motivação

Em 2008, é lançado o primeiro filme do Homem de Ferro, centrado no personagem Tony Stark. O filme além de despertar o interesse nas histórias em quadrinhos da Marvel, também prende a atenção dos ciêntistas da computação, uma vez que o personagem principal interage com uma inteligência artificial - Jarvis - que podia controlar a casa e a armadura do héroi. Vale ressaltar que Jarvis além de fazer reconhecimento do locutor de modo *speaker-dependent*, também realizada síntese de voz. Já no ano de 2011, foi lançado a primeira versáo do assistente pessoal do iOS, conhecido como Siri, que permite que o usuário execute determinadas funções do *smartphone* utilizando comandos falados. E mais recente, o criador do Facebook, decidiu no ano 2016, implementar seu próprio assistente pessoal, que funciona também de modo dependente do locutor, para auxiliar nas tarefas domésticas.

FALAR SOBRE ALGUM CONTEXTO CIENTIFICO TAMBEM.

E é a partir desse contexto que surgiu a inspiração para o desenvolvimento desse trabalho. Inicialmente, o projeto tinha como proposta de conseguir auxiliar em determinadas funções de uma residência, além de executar comandos básico em um computador. Porém, o projeto necessitou de cortes em seu escopo para que se tornasse factível sua elaboração como forma de trabalho de conclusão de curso.

Assim, este projeto tem como objetivo iniciar o desenvolvimento de um futuro assitente pessoal, através da elaboração de um sistema para reconhecimento de comandos falados *speaker-dependent*.

1.5 Metodologia

Para a elaboração deste projeto foi determinado os seguintes 11 comandos:

- Bom dia, Logan;
- Bom noite, Logan;
- Oi, Logan;
- Como está o tempo hoje?;
- vai chover?;
- Abrir calculadora;
- Ver notícias;
- Pesquisar;
- Alarme;
- Calendário;
- Sair;

sendo que posteriormente foi realizada a gravação de 10 áudios para cada um dos 11 comandos referidos, totalizando 110 arquivos de áudio no formato MPEG-4.Tais arquivos foram convertidos para o formato *WAVE* de 16 *bits* PCM usando o programa *Audacity*. Vale ressaltar que

todos os áudios foram gravados em um ambiente que proporciona-se um certo grau de isolamento sonoro, para assim se obter um som com menos ruído.

A partir dessa etapa inicial foi feita a extração dos dados brutos contidos nos arquivos *WAVE*. Para isso foi utilizada uma biblioteca fornecida pelo Prof.Dr.Rodrigo Capobianco Guido do Departamento de Ciência da Computção e Estatística (DCCE), IBILCE/Unesp. Tal biblioteca, escrita em C/C++, tem a função de separar o cabeçalho dos arquivos *WAVE*. A partir desse ponto, a biblioteca foi modificada para extrair os dados brutos e guardar as amplitudes dos sinais em arquivos de texto. Foi realizada a automação de todo o processo que exigia intervenção humana para a execução do algoritmo, como a passagem de áudios como parãmetro a cada nova execução, criação de arquivos para extração dos valores das amplitudes dos sinais, entre outros. Toda essa automatização foi implementada com a utilização de *scripts* escritos na linguagem *Shell script*.

Posteriormente a etapa de extração das amplitudes dos sinais digitalizados, foi realizada a extração das características dos áudios analisados. Essa etapa consiste na utilização do método A3 - descrito com maiores detalhes no 2 - para obter assim, vetores de características. Tal processo é essencial no presente trabalho, pois os valores obtidos no processo de extração são variáveis e demasiadamente grandes, sendo que o classificador exige valores menores e com tamanho fixo.

1.6 Exequibilidade

Para a elaboração desse projeto foram utilizadas ferramentas gratuitas como *Sublime*, LibreOffice, TeXstudio, *Audacity* e um computador pessoal. Além disso, foram também utilizados artigos científicos de acesso grátis como IEEE, e os acervo disponibilizado pela biblioteca física localizada no Instituto de Biocências, Letras e Ciêcias Exatas da UNESP.

1.7 Organização do trabalho

A monografia está organizada a partir deste capítulo da seguinte forma:

- No Capítulo 2 apresenta-se uma série de trabalhos realizados na área *speaker-dependent* tanto a nível local quanto a 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 deste projeto.

- No Capítulo 3 apresenta-se uma breve descrição do estado do atual do trabalho e também um cronograma para finalização.

Capítulo 2

Revisão Bibliográfica

2.1 Fundamentação da Verificação de Locutores

A voz é uma característica única dos seres humanos, que foi desenvolvida a partir da necessidade do homem em socializar, e assim poder se comunicar entre seus semelhantes. Em outras palavras, essa nossa característica foi adquirida através de um processo evolutivo, assim sendo náo somos dotados de um aparelho específico para fala, apenas adaptamos um conjunto de orgãos da parte superior do aparelho digestivo e do respirató rio para produção de fonemas.

De modo geral, podemos dividir o processo de síntese de voz em três partes:

1. **Pulmões** - São eles os responsáveis por gerar o fluxo de ar, que é ocasionado através da contração do orgão pelo diafragma. Tal fluxo, funciona como um ”combustível para a voz.

2. **Aparelho digestivo**

    2.1. **Laringe** - Local onde são encontrada as pregas vocais, que são dois músculos que tem a responsabilidade de modelar a voz através de sua contração e/ou relaxamento. Quando o fluxo de ar que sai dos pulmões passa faringe, as cordas vocais vibram com maior ou menor intensidade, logo pode-se perceber que a voz humana é formada pela interaç de duas principais forças: A força com que o fluxo de ar sai dos

pulmóes e forç muscular da laringe.

2.2. **Articuladores** - Eles são localizados na cavidade bucal e sã o responsáveis por fazer a articulação do fluxo de ar vindo dos pulmões. Sendo que os principais articuladores são: lábios, língua, dentes e mandíbula.

Vale ressaltar que a partir da característica vibratória das cordas vocais, é possível determinar a frequência da voz humana, o que nos torna capaz de fazer a dintinção de gêneros e até mesmo reconhecimento de voz e comandos.

RECONHECIMENTO DE LOCUTORES...

2.2 Sinais Digitais

Os dados podem ser representados de duas maneiras: digital ou analógica. Sendo a primeira forma uma onda eletromagnética que pode assumir infinitos valores ao longo do tempo e um excelente exemplo desse tipo de sinal, é a voz humana. Já a segunda forma tem um número finito de valores discretos no tempo. Em outras palavras, o sinal digital assumi valores em determinados instantes de tempo e sua representação pode ser baseada na discretização de um sinal analógico.

Podemos realizar a conversão de um sinal analógico para o digital, através de 3 etapas:

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. Tal teorema define que para realizar a amostragem de um sinal, é necessário no mínimo ter uma taxa de amostra duas vezes maior que a máxima frequência do sinal analógico. Caso náo 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 aplição de tal teorema é na digitalização da voz humana, que possui uma frequência máxima de 4

mil Hertz. Logo, para realizar amostragem desse sinal, seria necessário no mínimo 8 mil amostras por segundo.

2. **Quantização:** Esse processo consiste 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 ressalvar que o número de níveis é definido pelo número de *bits* que serão utilizados na etapa posterior de codificação, que é definido como sendo 2^n, onde n é o número de bits que será usado.

   Para realizar a quantização, pode-se utilizar duas técnicas diferentes para diminuir a quantidade de erros do processo:

   2.1. **Quantização uniforme:** Essa técnica é aplicado em sinais que possuam uma pequena diferença entre sua amplitude máxima e mínima. O processo consiste em dividir de forma igualmente espaçados os intervalos de amplitude, ou seja de forma uniforme.

   2.2. **Quantização não uniforme:** É utilizada para sinais que possuam um range dinâmico alto, ou seja existe uma grande diferença entre a amplitude áxima e mínima. Esse processo consiste em encontrar valores adequados para região do sinal, o que exige diferentes quantifizadores. Um solução alternativa para esse tipo de abordagem, seria comprimir o sinal, e logo após utilizar a técnica de quantização uniforme.

   Vale ressaltar que cada técnica descrita anteriormente dependerá do sinal a ser quantificado.

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

Um sinal digitalizado possui algumas vantagens em relação ao sinal analógico. As duas principais são que os dados podem ser compressados e corrigados, caso necessário.

2.3 Arquivos Acústicos no Formato *WAVE*

Waveform audio file format é a abreviação de *WAVE* ou simplesmente *WAV*, que é um tipo de formato de arquivo 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, visto que o formato permite uma fiel representação dos dados digitalizados, uma vez que os dados digitalizados podem ser armazenados sem sofrer obrigatoriamente um processo de compressão, o que evita perdas. Porém, devido a essa característica o *WAV* ocupa muito mais espaço que os demais formatos de arquivos de áudios.

Na Tabela 2.1 pode-se ver a estrutura de um arquivo *WAVE*. Basicamente o arquivo é divido em 2 grandes blocos, sendo o primeiro bloco um cabeçalho RIFF e o segundo bloco é divido em dois sub-blocos, sendo um com informações referentes ao formato *WAVE* e o outro com os dados do áudios.

Vale ressaltar que os valores mais comuns para cada amostra de um arquivo *WAVE* pode ser 8 *bits* ou 16 *bits*. Um áudio de 8 *bits* siginifica que o valor da amplitude do sinal, de cada amostra, pode ser representado por 256 valores, sendo 127 positivos e 128 negativos. Já para um arquivo 16 *bits* a amplitude do sinal pode ser representado por 65536 valores, com 32757 positivos e 32768 negativos. Para *WAVE* de 16 *bits* é utilizada a codificação de complemento de 2 para representar o valor da amplitude do sinal. Assim, o valor do *bit* mais significativo representa se o sinal é negaivo ou positivo.

Nesse trabalho foi utilizado o formato *WAV* de 16 *bits* PCM (*Pulse-code Modulation*) que não utiliza compressão, para se obter assim uma melhor qualidade na elaboração deste projeto final. Foi fornecida uma biblioteca escrita em C/C++ pelo orientador para isolar o primeiro bloco referente ao cabeçalho RIFF e o sub-bloco de formato *WAV* dos dados brutos, que contém as amplitudes dos sinais de voz digitalizados.

.

Tabela 2.1 – Estrutura de um arquivo *WAVE*

Classe	Posição *(bytes)*	Tamanho *(bytes)*	Descrição
Cabeçalho	0	4	Apresenta o identificador do cabeçalho - ”RIFF”.
Cabeçalho	4	4	Tamanho do arquivo sem o identificado do cabeçalho.
Cabeçalho	8	4	Mostra o identificador *WAVE*.
Formato	12	4	Mostra o identificador do segundo bloco - ”fmt”.
Formato	16	4	Tamanho do bloco sem o identificador.
Formato	20	2	Mostra se o arquivo é do tipo PCM ou se tem alguma compressão.
Formato	22	2	Mostra a quantidade de canais.
Formato	26	4	Apresenta o valor da taxa de amostragem.
Formato	30	4	Apresenta a taxa de *bytes*.
Formato	32	2	Demostra a quantidade de *bytes* para uma amostra.
Formato	34	2	Demostra a quantidade de *bits* para cada amostra.
Dados	36	4	Apresenta o identificador do terceiro bloco - ”*data*”.
Dados	40	4	Mostra o tamanho do bloco sem o identificador.
Dados	44	4	Demostra os dados reais da música.

2.4 Reconhecimento de Padrões e Vetores de Características

Neste trabalho, vetores de características são vetores com tamanho fixo que armazenam valores

2.5 Energia

A definição de energia está relacionada ao conceito de conseguir realizar trabalho. Neste projeto, será considerada energia a capacidade das estruturas vocálicas e dos pulmões de produzir um sinal acústico. A equação 2.1 defini a energia total E(s[.]), de um dado sinal de áudio digitalizado s[.], de tamanho M.

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

Para realizar a captura das caracteríticas dos áudios foi utilizado o método A3, que foi desenvolvido pelo próprio orientador em (4). Para realizar a extração das características, tal método se baseia no em determinar tamanhos ou áreas proporcionais para atingir níveis predefinidos da energia do sinal que se encontra em análise. A3 é ideal para avaliar os níveis de energia de um sinal de voz digitalizado que foi gerado por um agente.

Vale ressaltar que A3 defini um nível crítico de energia que varia de 0 a 100%. Em um sinal de áudio, A3 extrai um vetor de características dividindo o áudio em partes proporcionais ao valor definido pelo nível crítico - parâmetro C - de modo que, cria uma janela de tamanho C% e extrai a característica dessa faixa de valores do sinal digitalizado, conforme é mostrado na equação 2.2, onde ϵ é uma característica extraida, E(W0[.]) a energia obtida do início do sinal até o ponto final da janela definido por C, E(Wk[.]) é a energia total do sinal. Tal processo é repetido até que se atinja o montante total de energia do áudio, de modo que, defini-se assim um vetor de características.

$$\epsilon = (E(W0[.])/(\sum_{k=0}^{T-1} E(Wk[.]))) \tag{2.2}$$

2.6 Similaridade baseada em distancias

dist absoluta dist euclidiana

Capítulo 3

Detalhamento do Trabalho Proposto

3.1 Coleta e elaboração do banco de áudios

Inicialmente, 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. Como também já mencionado, a gravação dos arquivos de áudio foi realizada em um ambiente fechado, para diminuir assim, a probabilidade de ruídos nos sinais. Todos os arquivos foram gravados no formato MPEG-4, que é o formato disponível no *software* de gravação de áudio do *Windows* 10, e posteriormente foram convertidos para o formato *WAVE* de 16 *bits* PCM, com o auxílio do editor de áudio *Audacity*. Ao total, foram gravados e convertidos 110 arquivos de áudio.

Capítulo 4

Testes e Resultados

bla bla bla...

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho, ...

Referências

1 PETRY, A. *Reconhecimento automático de locutor utilizando medidas de invariantes dinâmicas não-lineares*. 2002. 155 p. Tese (Doutorado em Ciência da Computação)-Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

2 CAMPBELL, J. P. et al. Forensic speaker recognition: a need for caution. *IEEE Signal Processing Magazine*, v. 26, n. 2, p. 95-103, 2009. doi:10.1109/msp.2008.931100.

3 ACADEFORD. Disponível em:<http://www.acadeffor.com.br/>. Acesso em: 12 ago. 2014.

numérico. São Paulo: Pearson Prentice Hall, 2007.

4 Rodrigo Capobianco Guido. A tutorial on signal energy and its applications.

Apêndice I - Gráficos das características extraídas