Aprimoramento De Ferramentas De Reconhecimento E Comandos Por Voz Para Português Do Brasil

Ildeberto G. Bugatti

FATEC Garça - bugatti.fatec@gmail.com

Gustavo Luiz da C. Rosa

UNIVEM - gustavoluizcr@gmail.com

Pedro Henrique Bugatti

UTFPR-Cornélio-Procópio - phbugatti@gmail.com

Edio R. Manfio

FATEC Garça - prof.ediorobertomanfio@gmail.com

Resumo

A necessidade de organizar as atividades diárias através da otimização do tempo disponível para efetuá-las, gera a necessidade de utilizar ferramentas automatizadas para auxiliar a execução de tarefas repetitivas. Esse fato motiva o desenvolvimento de equipamentos automatizados que auxiliem a execução de atividades de forma mais eficiente e organizada. Para esse fim existe uma grande gama de processos automatizados aplicados às mais diversas áreas do conhecimento. No entanto, a utilização efetiva dessa automação está diretamente relacionada à qualidade da interface disponibilizada para o usuário. Para otimizar e gerar interfaces mais amigáveis, o trabalho em tela, propõe a introdução comandos de voz nos processos de controle de equipamentos automatizados. Neste contexto, o projeto propõe o estudo de viabilidade e a implementação de ferramentas de reconhecimento de voz que possibilitará gerar um banco de dados contendo conjuntos de comandos de voz, contemplando as características fonéticas regionais da língua portuguesa falada no Brasil e, aplicar esses comandos em sistemas de controle de processos.

Palavras-chave: Reconhecimento de voz. Comandos de voz. Controle de processos. Microcontroladores.

Voice Command Recognition and Enhancement Tools for Brazilian Portuguese

Abstract

The need to organize daily activities by optimizing the time available to modify them, makes it necessary to use automated tools to assist in performing daily tasks. This fact motivates the development of automated equipment that help performing daily tasks more efficient and organized

manner. To this end there is a wide range of automated processes applied to various fields of knowledge. However, the effective use of this automation is directly related to the quality of available user interface. To optimize and generate more user-friendly interfaces, work on screen, it proposes the introduction of voice commands in automated equipment control processes. In this context, the project proposes the feasibility study and the implementation of speech recognition tools, which will enable to generate a database containing sets of voice commands, contemplating the features regional phonetic of the Portuguese language spoken in Brazil and apply these commands in process control systems.

***Keywords*:** Voice recognition. Voice commands. Process control. Microcontroller.

---

1. INTRODUÇÃO

A tecnologia de semicondutores aplicada a sistemas computacionais evolui de forma muito rápida. Atualmente é possível utilizar essa tecnologia em uma diversidade de aplicações. Dentre as áreas de pesquisa computacional a interface Homem-Máquina está revolucionando o acesso a sistemas computacionais tornando cada vez mais cômoda e acessível essa comunicação. O recurso de comando de voz possibilita a construção das técnicas de interface que possuem alto grau de naturalidade e proximidade do homem.

Atualmente há várias tecnologias disponíveis no mercado para reconhecimento de voz aplicada a sistemas de controle de processos e automação. Entre estas aplicações, as mais comuns estão sendo utilizadas na indústria automobilística; possibilitando a execução de várias funções substituindo os sistemas de controles tradicionais por comandos de voz. Há também comandos de voz para controlar funções em aparelhos portáteis como computadores, smartphones, celulares dentre muitos outros.

No entanto, de forma geral, poucas dessas aplicações utilizam técnicas de reconhecimento de voz aplicada para Língua Portuguesa falada no Brasil, gerando problemas quando da utilização desses comandos de voz: veículos comercializados na região sul do país não devem levar em consideração apenas a variante linguística local, mas avaliar ocorrências e variantes linguísticas de todas regiões do país. Um sistema de controle que utiliza comandos de voz deve reconhecer os comandos inserido no sistema em todo seu território. Os sistemas de reconhecimento de voz utilizados pelas indústrias estão geralmente associados a sistemas fonéticos da língua inglesa gerando falhas e problemas quando aplicadas a outro idioma, entre eles o português do Brasil. Isto se deve ao fato de que no Brasil existem variantes típicas de cada região.

O projeto visa estudar algoritmos de reconhecimento de voz existentes, verificar a eficiência dos mesmos e implementar programas de reconhecimento de voz para um conjunto de comandos a ser definido utilizando técnicas de detecção de fonemas específicos da língua

portuguesa falada no Brasil, incluindo as variantes regionais. Para tanto contará com efetiva participação de profissionais das áreas de Linguística, Engenharia Elétrica e Informática.

O objetivo geral do projeto é estudar e aprimorar ferramentas de reconhecimento e comandos de voz para o português falado no Brasil. Para alcançar o objetivo foi proposto um estudo de caso que abrange um sistema de controle automatizado de um veículo contendo cinco comandos de voz.

2. METODOLOGIA

Durante o desenvolvimento do projeto foram estudados e implementados algoritmos de reconhecimento de voz existentes, para verificar a eficiência dos mesmos e gerar resultados para subsidiar o desenvolvimento de ferramentas e técnicas de reconhecimento de voz aplicadas ao controle de processos e interfaces naturais. Os estudos iniciais abrangeram um conjunto mínimo de fonemas, definidos como necessários, para gerar o controle de uma diversidade de equipamentos. Foram estudadas e utilizadas um conjunto de técnicas de reconhecimento de voz aplicadas ao conjunto mínimo de fonemas escolhidos. Os resultados obtidos com essas técnicas foram comparados e contribuíram para determinar as técnicas mais adequadas para realizar o reconhecimento de voz aplicada ao controle de processos.

O reconhecimento dos fonemas estudados além de contemplar variantes regionais da língua portuguesa falada no Brasil deve também considerar outros aspectos relevantes para o reconhecimento de voz, tais como: sexo, voz masculina, voz feminina, faixa etária (crianças, adolescentes, adultos e, idosos), entre outros.

A análise eficiente de todos esses fatores exige conhecimentos multidisciplinares envolvendo diversas áreas do conhecimento, tais como: Processamento de Sinais, Ciência da Computação, Reconhecimento de Padrões, Inteligência Artificial, Neurofisiologia, Teoria das Comunicações e Linguística. Como consequência demandam profissionais de várias áreas, tais como: Linguística, Mecatrônica e Computação.

Além dos parâmetros já listados, os sistemas de reconhecimento da voz, em geral, devem ser aptos a funcionar em condições desfavoráveis, que envolvem a existência de ruídos nos ambientes onde os sistemas de reconhecimento de voz podem ser utilizados. Esse fato, exige o estudo de técnicas extras para conseguir eficiência e robustez do sistema. Essa multidisciplinaridade motivou o desenvolvimento e a implementação do projeto que caracteriza um grupo de pesquisas apto a ser formalizado junto à instituição ou órgãos de fomento e pesquisa.

A fase de estudos e definição das ferramentas mais adequadas para o desenvolvimento efetivo do projeto foi realizada a contento. As ferramentas utilizadas estão descritas em conjunto com suas funcionalidades. Foi gerado um conjunto de fonemas que foram capturados em laboratório gerando um conjunto de amostras satisfatório para aplicar as ferramentas estudadas gerando resultados de eficiência e robustez que foram compilados e comparados. Gerando subsídios para trabalhos futuros.

Nas próximas etapas as atividades necessárias para a implementação de algoritmos em ambiente de hardware, para obter maior eficiência e reconhecimento de voz em tempo real.

3. DESENVOLVIMENTO

Para obter os objetivos estipulados, foi definido um estudo de caso que demandou o estudo de ferramentas adequadas para o desenvolvimento de todas as etapas do projeto. Foi então definido um sistema de controle automatizado utilizando comandos de voz para controlar um veículo envolvendo cinco comandos de voz que utilizaram os seguintes fonemas: “direita”; “esquerda”; “frente”; “trás” e “para”. Esses cinco fonemas foram considerados como o conjunto mínimo de comandos para controlar um veículo de forma natural e eficiente.

Definido o conjunto de fonemas foram necessárias desenvolver uma sequência de atividades para a execução do projeto:

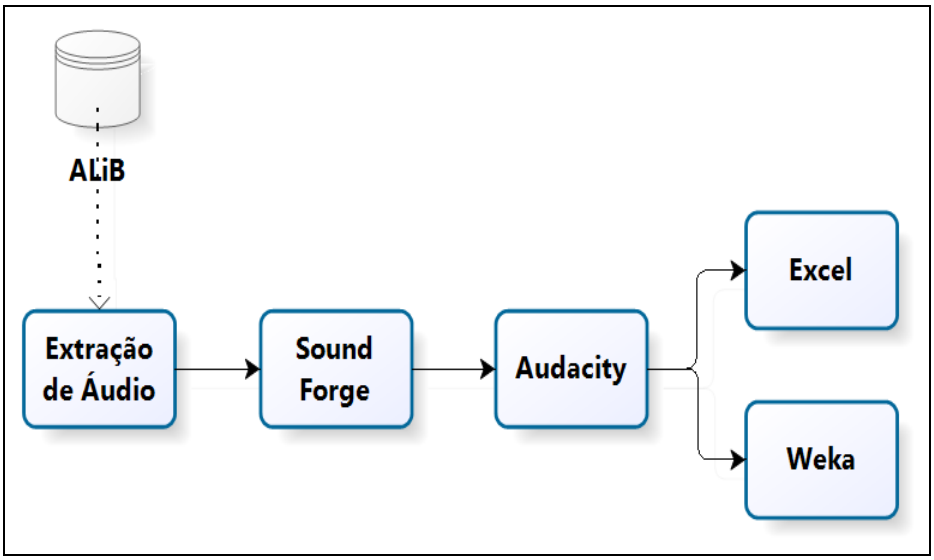
- Primeiramente foi necessário definir ferramentas e ambientes adequados e suficientes para realizar a aquisição de amostras reais dos fonemas de controle;
- Definição das ferramentas para extração de áudio;
- Estudo e utilização de ferramentas e algoritmos eficientes para reconhecimento de voz;
- Compilação dos resultados e definição dos algoritmos mais eficientes entre a gama de algoritmos estudados.

Para executar essas atividades foram estudados o escolhido o seguinte conjunto de ferramentas: conceitos do Atlas Linguístico Brasileiro (ALiB); tecnologias de extração e edição de áudio junto de ferramentas para o reconhecimento de imagens.

Nessa fase além do ALiB, foram utilizadas as seguintes ferramentas: Sound Forge, Audacity, e Weka.

A figura 1 mostra o fluxograma do desenvolvimento do projeto e cada tópico será discutido nos demais capítulos.

Figura 1- Ferramentas utilizadas nas Etapas do projeto



3.1- Utilização do Atlas Linguístico do Brasil (ALiB)

Para avaliar ferramentas existentes e iniciar os estudos e a implementação deste projeto, foi necessário contar com alguns dados do Atlas Linguístico do Brasil (ALiB), que oferece um estudo bastante completo de possíveis variantes da Língua Portuguesa do Brasil. Paralelamente a isso, foram selecionados cinco comandos que representam os vetores de direção/posicionamento ‘direita’, ‘esquerda’, ‘frente’, ‘trás’ e ‘para’. A razão da escolha por esses comandos/vetores – que chamaremos apenas de vetores a partir de agora – é simples: com eles é possível movimentar qualquer veículo sobre rodas, para ficar apenas em um exemplo.

Os cinco vetores foram capturados em peças de áudio individuais e considerando as principais variantes do Português do Brasil relativas à região Centro-Oeste Paulista.

Foi realizada a aquisição dos cinco comandos de voz a partir de 66 informantes no envolvendo alunos da FATEC Garça e do UNIVEM, além funcionários das instituições. Os métodos para aquisição dessas peças de áudio contaram com critérios bastante cuidadosos: estúdio de gravação livre de ruídos, microfone profissional de boa qualidade, sistema ‘anti-puf’, software de gravação específico entre outros.

Ao final, para o processamento, estiveram disponíveis 345 peças de áudio representando os cinco vetores definidos e denominados: DIREITA, ESQUERDA, FRENTE, TRÁS e PARA.

A Tabela 1, definida por Cardoso *2014[4],* ilustra bem a distribuição das diferentes variantes da língua Portuguesa falada no Brasil.

Tabela 1: Distribuição de variâncias - Fonte: Cardoso 2014[5]

Região	Direita	Esquerda	Trás	Frente	Para
Centro Oeste de São Paulo e Norte do Paraná	direita 'a'	esquerda 'a'	trás 'a'	frente 'a'	para 'a'
	direita 'c'				
Cidade de São Paulo		esquerda 'c'			
Santa Catarina e Rio Grande do Sul	direita 'b'	esquerda 'b'	trás 'b'	frente 'b'	para 'b'
Cidade do Rio de Janeiro e parte do Estado do Rio de Janeiro		esquerda 'd'	trás 'c'		
		esquerda 'e'			

Analisando as informações contidas na tabela 1 verifica-se que na região Centro-Oeste do estado de São Paulo, existem duas variações para a palavra *Direita*, e uma única variação para os demais comandos (*Esquerda*, *Trás*, *Frente* e *Para*). Já para a cidade de São Paulo, possui uma variação apenas na palavra *Esquerda*. Com os estados de Santa Catarina e Rio Grande do Sul, possui uma variação para os comandos de *Direita*, *Esquerda*, *Trás*, *Frente* e *Para*. No Rio de Janeiro existem duas variações para a palavra *esquerda*, e uma para a palavra *trás*.

Com os vetores disponíveis, o próximo passo foi encontrar metodologias e formas de comparar cada vetor e descobrir as similaridades entre as diferentes variantes em relação ao mesmo comando e entre sexos diferentes.

3.2. Extração das informações dos vetores através do Audacity

Para extrair as informações dos vetores, foi utilizado o programa AUDACITY 2.0.6, um software *open source* de edição de áudio digital. Neste, existe a possibilidade de extrair as informações dos vetores de maneiras diferentes. Nas que utilizamos neste projeto foram quatro extrações diferentes: “Sample Data Export”, “Análise da Frequência utilizando algoritmo Espectro com tamanho de 512”, “Análise da Frequência utilizando algoritmo Espectro com tamanho de 1024” e “Análise da Frequência utilizando algoritmo Cepstro com tamanho de 512”.

Com todas as informações necessárias já extraídas, o próximo passo foi fazer definitivamente a comparação das informações. Tentamos efetuar a comparação e busca de similaridade das informações através de planilhas eletrônicas (MICROSOFT OFFICE EXCEL) e nos deparamos com uma dificuldade, pois utilizando funções do próprio software não foi possível encontrar similaridades entre as informações disponíveis.

3.3. Software Weka, resultados obtidos na avaliação de desempenho

Após tentativas frustrantes da busca pela similaridade software EXCEL, foi estudado o software WEKA 3.6, um software *open source* que tem como objetivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas

O Weka possibilita o reconhecimento de voz através da análise do espectro de frequência através da utilização de 18 diferentes algoritmos. Os critérios para avaliar o desempenho dos 18 algoritmos foram: eficácia (avaliação de assertividade) e eficiência (tempo de resposta).

A utilização do Weka possibilitou testar diversos algoritmos que poderiam auxiliar na busca por similaridade. Definimos três objetivos diferentes para identificar quais os melhores algoritmos para cada situação, entre elas: *verificar qual algoritmo é o mais indicado para distinguir qual a direção falada pelo locutor, independente do sexo e da variância da voz do locuto*r; *verificar qual algoritmo é o mais indicado para distinguir o sexo do locutor* e por fim, *verificar qual algoritmo é o mais indicado para distinguir a direção informada e qual o sexo do locutor*. Vale destacar que a avaliação foi feita com dois tipos de extrações, um com o algoritmo de Espectro com o tamanho de 512 e outro com o algoritmo de Cepstro com o tamanho de 512.

A tabela 2 mostra os 18 algoritmos estudados, eles estão subdivididos em três classes: Trees, Bayes e Lazy.

Tabela 2 – relação de algoritmos disponibilizados no software Weka

Classe	Algoritmo
Trees	BFTree
	DecisionStump
	FT
	J48
	J48graft
	LADTree
	LMT
	NBTree
	RandomForest
	RandomTree
	REPTree
	SimpleCart

Classe	Algoritmo
Bayes	NaiveBayes

Classe	Algoritmo
Lazy	IBk -1
	IBk -2
	IBk -3
	IBk -4
	IBk -5

Porém, antes de obter estes resultados, foi necessário entregar ao WEKA 3.6, um conjunto de amostras reais para que os diversos algoritmos fossem aplicados de forma objetiva. Para isso, foi montado um conjunto com todas os 30 vetores. Para preparar o conjunto, foi necessário aprender a criar o mesmo conforme o **modelo arff** do software pede:

- Declarar o nome do projeto;
- Declarar quais atributos serão considerados;
- Declarar o conjunto de amostras;

3.3.1. Análise e comparação dos resultados apresentados pelos algoritmos do Weka

Foram realizados várias simulações e tratamentos de dados através da utilização dos algoritmos Espectro e Cepstro. Com a análise dos dados obtidos pelo processamento dos arquivos do Weka, foi possível a análise dos dados de acordo com dois critérios de avaliação:

- Eficácia: porcentagem de acertos do algoritmo;
- Eficiência: Que será o tempo de resposta do algoritmo.

Para a execução dos algoritmos, a arquitetura utilizada para a realização dos testes foi:

- Sistema Operacional: Windows 7 Professional
- Memória RAM: 8GB
- Processador: Core i7 2670QM 2.2 GHz (até 3.1 GHz)

Com esses passos definidos, o projeto foi separado em três tipos de reconhecimentos: Reconhecimento do sexo do falante; Reconhecimento da Direção falada; e por fim o reconhecimento do sexo do falante e da direção falada.

Para facilitá-la a separação, as análises serão divididas nos tópicos de algoritmos utilizados no Audacity (Cepstro e o Espectro). Após testes exaustivos de todas as amostras de vetores separadas por sexo e diferentes comandos foram obtidos resultados que contribuíram para escolher os melhores algoritmos para reconhecimento dos cinco comandos de voz escolhidos.

Os itens que seguem mostram, de forma resumida, os melhores algoritmos utilizando análise Cepstro:

- **Identificação do Sexo,** algoritmo **IBk (3 e 5)** – 96,23 % de eficácia e 0,00 segundos para execução;
- **Identificação do Comando,** algoritmo **Randon Forest** – 68,7 % de eficácia e 0,57 segundos para execução;
- **Identificação do Comando e do Sexo,** algoritmo **NaiveBayes** – 64,06 % de eficácia e 0,08 segundos para execução.

Os itens que seguem mostram, de forma resumida, os melhores algoritmos utilizando análise Espectro:

- **Identificação do Sexo,** algoritmo **FT** – 97,39 % de eficácia e 0,5 segundos para execução;
- **Identificação do Comando,** algoritmo **FT**– 85,80 % de eficácia e 1,31 segundos para execução.
- **Identificação do Comando e do Sexo,** algoritmo **FT** – 89,99 % de eficácia e 3,39 segundos para execução.

Os resultados obtidos foram compilados e os melhores algoritmos avaliados segundo os critérios de eficácia e eficiência foram:

- O algoritmo **FT** é o mais indicado quando o *objetivo for identificar o sexo do locutor*, por ser um dos algoritmos que teve uma alta porcentagem de assertividade (97,39%) e por ser executado em apenas 0,50 segundos;
- O algoritmo **RandomForest** é o mais indicado quando o *objetivo for identificar a direção informada pelo locutor*, por ser um dos algoritmos que teve uma alta porcentagem de assertividade (79,71%) e por ser executado em apenas 0,47 segundos.

- O algoritmo **RandomForest** também é o mais indicado quando o *objetivo for identificar a direção informada e o sexo do locutor*, por ser o algoritmo que teve uma alta porcentagem de assertividade (65,51%) e por ser executado em apenas 0,69 segundos.

4. RESULTADOS E ANÁLISE DE ABRANGENCIA

Esse projeto teve como metas, estudar e definir a eficiência de algoritmos de reconhecimento de voz utilizando técnicas de similaridade e, criar um protótipo físico, com a funcionalidade de reconhecer, a princípio, cinco (5) comandos: “esquerda”, “direita”, “cima”, “trás” e “para”. Esse conjunto de comandos foi escolhido porque possibilita a sua utilização em uma grande quantidade de aplicações tais como controle de veículos, deslocamento de robôs e/ou controle de movimentos de qualquer dispositivo sobre rodas utilizando apenas uma interface de voz.

A aplicabilidade de interfaces de comando de voz é ampla e pode ser utilizada de forma generalizada. Para sua utilização de forma ampliada, o banco de dados de comandos deve ser também ampliado. As características das aplicações também interferem na forma e nos algoritmos que podem ser utilizados. Assim devem ser também estudados e definidos a relação entre as características das aplicações e dos algoritmos de forma integrada.

A utilização do controle de comando de voz pode também se estender à navegação em bancos de dados. Essa aplicação, dentre outras, está diretamente relacionada aos objetivos e capacitação do egresso da área de Sistemas de Informação, Ciência da Computação e Mecatrônica, caracterizando uma área com aplicação nas áreas de abrangência desses cursos.

5. CONCLUSÓES

A tecnologia de reconhecimento de voz possibilita a construção de interfaces naturais entre o homem e a máquina. Além disso, possibilita a utilização de equipamentos e sistemas de controle e automação de processos liberando as mãos para outras tarefas - *hands free*. A definição e construção de comandos de voz aplicada a língua portuguesa falada no Brasil, através da utilização do ALiB, possibilita o desenvolvimento de aplicações eficazes, gerando como resultado o desenvolvimento e comercialização de equipamentos voltados para o mercado brasileiro e, dessa forma, contribuir para o desenvolvimento e aplicação dessa tecnologia pelas indústrias regionais e nacionais.

6. REFERENCIAS BIBLIOGRÁFICAS

BEDO, Marcos Vinícius Naves **Incluindo funções de distância e extratores de características para suporte a consultas por similaridade**. 2013. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08112013-160506/>. Acesso em: 14 mar. 2015.

BUGATTI, Pedro H.; TRAINA, Agma J.M; TRAINA JR, Caetano. Assessing the best integration between distance-function and image-feature to answer similarity queries. In: Proceedings of the 2008 ACM symposium on Applied computing. ACM, 2008. p. 1225-1230.

MANFIO, Edio Roberto. Processamento de Linguagem Natural, Processamento de Sinais da Fala, Geolinguística e um Naco de Humor. In: *Anais do X Seminário de Iniciação Científica Estudos Linguísticos e Literários - Sóletras*. UENP – Jacarezinho, 2013. Disponível em: <http://www.cj.uenp.edu.br/index.php/institucional/eventos/1-soletras/event_details>. Acesso em: 21 jun. 2014.

MELO, Bruno Mariani de. **Áudio sobre IP**. 2011. Monografia (Especialização - Pós-Graduação Lato Sensu em Sistemas de Telecomunicações) Escola Superior Aberta do Brasil –Esab, Vila Velha, Espírito Santo, 2011. Disponível em: <http://www.esab.edu.br/arquivos/monografias/monografia_brunomariani.pdf>. Acesso em: 20 nov. 2014.

CARDOSO, Suzana Alice Marcelino da Silva et al. *Atlas Linguístico do Brasil*: Cartas Linguísticas I. Vol. 2. Londrina: Eduel, 2014b.