



**Jan Krueger Siqueira**

**Reconhecimento de Voz Contínua com Atributos MFCC,  
SSCH e PNCC, Wavelet Denoising e Redes Neurais**

**DISSERTAÇÃO DE MESTRADO**

Dissertação apresentada como requisito parcial para  
obtenção do título de Mestre pelo Programa de Pós-  
Graduação em Engenharia Elétrica do Departamento de  
Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Abraham Alcaim

Rio de Janeiro  
Setembro de 2011



**Jan Krueger Siqueira**

**Reconhecimento de Voz Contínua com Atributos MFCC,  
SSCH e PNCC, Wavelet Denoising e Redes Neurais**

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

**Prof. Abraham Alcaim**

Orientador

Centro de Estudos em Telecomunicações - PUC-Rio

**Fernando Gil Vianna Resende Junior**

UFRJ

**Prof. Marco Antonio Grivet Mattoso Maia**

Centro de Estudos em Telecomunicações - PUC-Rio

**Prof. José Eugênio Leal**

Coordenador Setorial do

Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 2 de setembro de 2011

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

### **Jan Krueger Siqueira**

Graduou-se em Engenharia Elétrica nas áreas de Eletrônica e Telecomunicações pela PUC-Rio em 2008. Atua na empresa SmartSolutions Ltda, onde desenvolve sistemas para dispositivos móveis e para a web. Também ministra aulas de Matlab na PUC-Rio.

#### Ficha Catalográfica

Siqueira, Jan Krueger

Reconhecimento de voz contínua com atributos MFCC, SSCH e PNCC, wavelet denoising e redes neurais / Jan Krueger Siqueira ; orientador: Abraham Alcaim. – 2011.

85 f. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2011.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Reconhecimento de voz contínua. 3. Ruído aditivo. 4. MFCC. 5. SSCH. 6. PNCC. 7. Wavelet denoising. 8. Rede neural. I. Alcaim, Abraham. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

A todos os professores que se dedicam de coração aos seus alunos.

## Agradecimentos

Ao CNPq, por financiar a pesquisa.

Ao professores Abraham Alcaim e Marley Maria Bernardes Rebuzzi Vellasco, pelo ensino, inspiração e orientação neste projeto.

A Apple, Google, Mathworks e University of Cambridge, por fornecerem a tecnologia necessária.

À minha mãe, pelo amor e pelos ensinamentos desde que nasci.

A Paula Maurício Nunes, por ter sido uma grande companheira durante todo o mestrado. Mesmo não estando mais juntos, eu serei eternamente grato pelos dois anos de namoro e pela amizade.

Aos meus grandes amigos da PUC-Rio: Bruno Baère Pederassi Lomba de Araújo, Clarissa Costalonga e Gandour, Larissa Figueiredo Terra de Faria Reis, Vinícius Costa Villas Bôas Segura. Espero que a gente mantenha essa amizade especial que atravessou inúmeras aulas, almoços e cinemas.

Aos alunos, bolsistas e professores da minha turma de salsa do Jaime Aroxa. Os sábados continuam sendo especiais por causa de vocês.

A todos os familiares, pelo carinho de sempre.

A quem mais tenha facilitado, ainda que muito indiretamente, a conclusão deste trabalho.

## Resumo

Siqueira, Jan Krueger; Alcaim, Abraham. **Reconhecimento de Voz Contínua com Atributos MFCC, SSCH e PNCC, Wavelet Denoising e Redes Neurais**. Rio de Janeiro, 2011. 85 p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Um dos maiores desafios na área de reconhecimento de voz contínua é desenvolver sistemas robustos ao ruído aditivo. Para isso, este trabalho analisa e testa três técnicas. A primeira delas é a extração de atributos do sinal de voz usando os métodos MFCC, SSCH e PNCC. A segunda é a remoção de ruído do sinal de voz via wavelet denoising. A terceira e última é uma proposta original batizada de feature denoising, que busca melhorar os atributos extraídos usando um conjunto de redes neurais. Embora algumas dessas técnicas já sejam conhecidas na literatura, a combinação entre elas trouxe vários resultados interessantes e inéditos. Inclusive, nota-se que o melhor desempenho vem da união de PNCC com feature denoising.

## Palavras-chave

Reconhecimento de voz contínua; ruído aditivo; MFCC; SSCH; PNCC; wavelet denoising; rede neural.

## Abstract

Siqueira, Jan Krueger; Alcaim, Abraham. **Continuous Speech Recognition with MFCC, SSCH and PNCC Features, Wavelet Denoising and Neural Networks**. Rio de Janeiro, 2011. 85 p. M. Sc. Dissertation – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

One of the biggest challenges on the continuous speech recognition field is to develop systems that are robust to additive noise. To do so, this work analyses and tests three techniques. The first one extracts features from the voice signal using the MFCC, SSCH and PNCC methods. The second one removes noise from the voice signal through wavelet denoising. The third one is an original one, called feature denoising, that seeks to improve the extracted features using a set of neural networks. Although some of these techniques are already known in the literature, the combination of them brings many interesting and new results. In fact, it is noticed that the best performance comes from the union of PNCC and feature denoising.

## Keywords

Continuous speech recognition; additive noise; MFCC; SSCH; PNCC; wavelet denoising; neural network.

# Sumário

1	Introdução	16
2	Reconhecimento de Voz	18
2.1	Reconhecimento de Palavras Isoladas	18
2.2	Reconhecimento com Vocabulário Amplo	23
2.3	Reconhecimento de Voz Contínua	25
2.4	Parâmetros do HMM	30
2.4.1.	Escolha da Função para $b_i(O_t)$	30
2.4.2.	Otimização dos Parâmetros	32
2.4.3.	Descasamento entre Treino e Teste	33
3	Extração de Atributos	34
3.1	Pré-Ênfase	34
3.2	Divisão em Quadros	35
3.3	Transformada de Fourier	37
3.4	Informações do Espectro	38
3.4.1.	MFCC	39
3.4.2.	SSCH	41
3.4.3.	PNCC	43
3.5	Transformada Discreta do Cosseno (DCT)	44
3.6	Coeficientes Delta e de Aceleração	46
4	Wavelet Denoising	48
4.1	Limpeza do Sinal	48
4.2	Transformada Wavelet Discreta	49
4.3	Limiar (Thresholding)	52
4.4	Estimação do Limiar	55
5	Feature Denoising	56
5.1	Rede Neural Artificial	56
5.2	Configuração da Rede Neural	60



6	Condições Experimentais e Resultados	62
6.1	Banco de Dados para Treino e Teste	62
6.1.1.	Banco de Dados de Voz	62
6.1.2.	Banco de Dados de Ruído	63
6.2	Ferramentas Computacionais	64
6.3	Parâmetros Escolhidos	66
6.3.1.	Parâmetros do Reconhecimento de Voz	66
6.3.2.	Parâmetros da Extração de Atributos	66
6.3.3.	Parâmetros do Wavelet Denoising	67
6.3.4.	Parâmetros do Feature Denoising	67
6.4	Testes e Resultados	68
6.4.1.	Teste dos Métodos de Extração de Atributos	68
6.4.2.	Teste do Wavelet Denoising	69
6.4.3.	Teste do Feature Denoising	70
6.4.4.	Teste do Sistema Completo	71
7	Conclusões e Sugestões para Trabalhos Futuros	72
8	Bibliografia	74
9	Apêndice: Demonstrações	77
9.1	Probabilidade de Observação num HMM	77
9.2	Algoritmo Forward-Backward	78
9.3	Probabilidade de Ocorrência de Palavras	79
9.4	Algoritmo de Viterbi	80
9.5	Algoritmo de Baum-Welch	81
9.6	Criação dos Filtros do Método MFCC	84

## Lista de Figuras

Figura 1: Etapas do sistema proposto para reconhecimento de voz.	17
Figura 2: Forma de onda das palavras “mar” (voz masculina), “mel” (voz masculina) e “mar” (voz feminina).	18
Figura 3: Esquema geral de um reconhecedor de voz para comandos do tipo SIM/NÃO.	19
Figura 4: Possível representação da palavra “mar” no Modelo de Markov Escondido e suas transições de estado.	19
Figura 5: Possível representação da palavra “mar” no Modelo de Markov Escondido e algumas características importantes para cada estado.	20
Figura 6: Representação matemática do Modelo de Markov Escondido.	20
Figura 7: Esquema mais detalhado de um reconhecedor de voz de palavras isoladas, para M possíveis palavras.	22
Figura 8: Conexão dos HMMs de fones para a formação de HMMs de palavras.	24
Figura 9: Conexão dos HMMs de bifones e trifones, para a formação de palavras.	25
Figura 10: Conexão cíclica das palavras para a formação de frases. O final de cada palavra é ligado ao começo de todas.	26
Figura 11: Conexão dos modelos de palavras para a formação da frase $w = \text{“Hoje está muito quente”}$ .	27
Figura 12: Rede de trigramas, para o caso de apenas duas palavras.	28
Figura 13: Transições de estado representados numa treliça.	29
Figura 14: Esboço das densidades de probabilidade gaussianas da frequência fundamental para homens e para mulheres.	31
Figura 15: Esboço da densidade de probabilidade unificada da frequência fundamental para homens e mulheres.	31

Figura 16: Extração dos vetores $O_t$ para o treinamento de cada fonema.	32
Figura 17: Formação de um sinal corrompido através da adição de ruído ao sinal de voz.	33
Figura 18: Esquema geral da extração de atributos.	34
Figura 19: Entrada e saída do bloco de Pré-Ênfase.	35
Figura 20: Entrada e saída do bloco da Divisão em Quadros.	35
Figura 21: Divisão em quadros consecutivos, causando perda de informação.	36
Figura 22: Divisão em quadros com superposição.	36
Figura 23: Suavização de um quadro através da multiplicação por uma função janela.	36
Figura 24: Sinal sonoro no domínio do tempo das vogais “a” e “e”, respectivamente.	37
Figura 25: Espectro das vogais “a” e “e”, respectivamente.	38
Figura 26: Entrada e saída do bloco correspondente à transformada de Fourier.	38
Figura 27: Entrada e saída do bloco correspondente às técnicas específicas.	39
Figura 28: Detalhamento do bloco “técnicas gerais” para o espectro de um quadro.	39
Figura 29: Detalhamento da Figura 28 para o método MFCC.	40
Figura 30: Filtros triangulares do método MFCC.	40
Figura 31: Comparação entre os espectros de um sinal limpo e do mesmo sinal com ruído branco de razão sinal-ruído de 10 dB.	41
Figura 32: Centróide de um trecho limpo do espectro comparado com o centróide desse mesmo trecho com ruído.	41
Figura 33: Comparação geral entre os métodos MFCC e SSCH.	42
Figura 34: Criação do histograma.	43
Figura 35: Comparação geral entre os métodos MFCC e PNCC.	43
Figura 36: Filtros gammatone igualmente espaçados na escala ERB.	44
Figura 37: Entrada e saída do bloco correspondente à transformada discreta do cosseno.	45

Figura 38: Exemplo da compressão de informação através da DCT.	46
Figura 39: Entrada e saída do bloco correspondente aos coeficientes delta.	46
Figura 40: Restauração de um sinal corrompido, antes da extração de atributos.	48
Figura 41: À esquerda, espectros do sinal e do ruído ocupando regiões diferentes. À direita, o resultado de sua adição.	49
Figura 42: À esquerda, superposição dos espectros do sinal e do ruído. À direita, o resultado de sua adição.	49
Figura 43: Exemplos de duas funções mãe $\Psi(n)$ : Daubechie-10 e Symlets 2.	51
Figura 44: Efeito dos parâmetros a e b na função mãe $\Psi(n)$ .	51
Figura 45: Representação dos valores da transformada de Fourier e da transformada Wavelet.	52
Figura 46: Valores wavelet do ruído e do sinal de voz.	53
Figura 47: Comparação entre o gráfico da função hard thresholding com a da função soft thresholding.	54
Figura 48: Exemplo da aplicação do wavelet denoising.	54
Figura 49: Restauração dos atributos extraídos, antes do reconhecimento de voz.	56
Figura 50: Exemplo da criação de um modelo matemático para classificar animais.	57
Figura 51: Esquema de um neurônio biológico.	58
Figura 52: Esquema de um neurônio artificial.	58
Figura 53: Rede neural <i>feedforward</i> , formada por conexões entre neurônios artificiais.	59
Figura 54: Configuração da rede neural com todos os atributos restaurados do quadro na saída.	60
Figura 55: Configuração de várias redes neurais, cada uma com um único atributo na saída.	61
Figura 56: Exemplos de frases da base TIMIT.	63
Figura 57: Amostra do dicionário de palavras da base TIMIT.	63
Figura 58: Exemplos de comandos da ferramenta HTK.	65
Figura 59: Exemplos de janelas do programa Matlab.	65

Figura 60: Configuração dos HMMs utilizados.	66
Figura 61: Exemplo do cálculo da taxa de acerto para uma frase.	68
Figura 62: Divisão da escala mel em intervalos de mesmo comprimento e a conversão dessas fronteiras para a escala em Hz.	85

## Lista de Tabelas

Tabela 1: Taxas de acerto utilizando apenas extração de atributos	69
Tabela 2: Taxas de acerto utilizando wavelet denoising e extração de atributos	70
Tabela 3: Taxas de acerto utilizando extração de atributos e feature denoising	70
Tabela 4: Taxas de acerto utilizando wavelet denoising, extração de atributos e feature denoising	71

*Your time is limited, so don't waste it living someone else's life. Don't be trapped by dogma — which is living with the results of other people's thinking. Don't let the noise of others' opinions drown out your own inner voice. And most important, have the courage to follow your heart and intuition. They somehow already know what you truly want to become. Everything else is secondary.*

**Steve Jobs**

*It's in Apple's DNA that technology alone is not enough. That it's technology married with liberal arts, married with the humanities, that yields us the result that makes our hearts sing.*

**Steve Jobs**

*Dance bem, dance mal, dance sem parar!*

**Nelson Motta / Rubens Queiroz**