Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

TESE DE DOUTORADO

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Programa de Pós-graduação em Engenharia Elétrica

Rio de Janeiro

Agosto de 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

Tese de Doutorado

Tese de Doutorado apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Abraham Alcaim

Rio de Janeiro

agosto de 2009

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Abraham Alcaim
Orientador
Centro de Estudos em Telecomunicações - PUC-Rio

Dra. Marley Maria Bernardes Rebuzzi Vellasco
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Sergio Lima Netto
COPPE/UFRJ

Prof. Fernando Gil Vianna Resende Jr.
UFRJ

Prof. Fábio Violaro
UNICAMP

Profa. Rosângela Fernandes Coelho
IME

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 24 de agosto de 2009

Vladimir Fabregas Surigué de Alencar

Graduou-se em Engenharia de Telecomunicações na UFF (Universidade Federal Fluminense) em 2003. Defendeu sua Dissertação de Mestrado em Março de 2005 pelo Departamento de Engenharia Elétrica da PUC-Rio.

Ficha Catalográfica

Alencar, Vladimir Fabregas Surigué de

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro / Vladimir Fabregas Surigué de Alencar; orientador: Abraham Alcaim. – 2009.

131 f.: il. ; 30 cm

Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Incluí referências bibliográficas.

1. Engenharia elétrica - Teses. 2. Reconhecimento de voz distribuído 3. LSF 4. LPC 5. ISF 6. HMM 7. Redes IP 8. Redes Móveis Celulares 9. ITU-T G.723.1 10. AMR-NB 11. AMR-WB. 12. Redes Neurais I. Alcaim, Abraham. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD:621.3

Para meus pais Juarez e Laura, minha irmã Tatiana
e minha namorada Daniele
pelo carinho, apoio e confiança.

Agradecimentos

Ao meu orientador, Professor Abraham Alcaim, pela oportunidade, apoio e incentivo para a realização deste trabalho.

Ao corpo docente do CETUC, pelo aprendizado proporcionado.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos.

Agradeço de forma especial aos meus pais, que mais uma vez foram fundamentais na minha vida, à minha irmã e à Daniele, que estiveram ao meu lado nesta empreitada.

Aos Professores que participaram da minha comissão examinadora.

A Audioteca Sal & Luz por todo apoio e esforço na construção e gravação da Base de vozes utilizada neste trabalho.

A todos os amigos que fiz no CETUC, que me proporcionaram não apenas momentos de aprendizagem, mas momentos de companheirismo que espero que se perpetuem.

Resumo

Alencar, Vladimir Fabregas Surigué; Alcaim, Abraham. **Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro.** Rio de Janeiro, 2009. 131p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta Tese visa explorar as oportunidades de melhoria do desempenho dos Sistemas Automáticos de Reconhecimento de voz com amplo vocabulário para o Português Brasileiro quando aplicados em um cenário distribuído (Reconhecimento de Voz Distribuído). Com esta finalidade, foi construída uma base de vozes para reconhecimento de voz contínua para o Português Brasileiro com 100 locutores, cada um falando 1000 frases foneticamente balanceadas. A gravação foi realizada em estúdio, ambiente sem ruído, com uma especificação de gravação que pudesse abranger a entrada dos diversos codificadores de voz utilizados em Telefonia Móvel Celular e IP, em particular os codecs ITU-T G.723.1, AMR-NB e AMR-WB. Para um bom funcionamento dos Sistemas Automáticos de Reconhecimento de voz é necessário que os atributos de reconhecimento sejam obtidos a uma taxa elevada, porém os codificadores de Voz para Telefonia IP e Móvel Celular normalmente geram seus parâmetros a taxas mais baixas, o que degrada o desempenho do reconhecedor. Usualmente é utilizada a interpolação linear no domínio das LSFs (*Line Spectral Frequencies*) para resolver este problema. Nesta Tese foi proposta a realização da interpolação com a utilização de um Filtro Digital Interpolador que demonstrou ter um desempenho de reconhecimento muito superior ao da interpolação linear. Foi avaliado também o uso das ISFs (*Immittance Spectral Frequencies*) interpoladas como atributo de reconhecimento, as quais se mostraram inadequadas para esta finalidade, assim como as LSFs. Outro aspecto de fundamental importância para os reconhecedores de voz distribuídos é a recuperação de perda de pacotes, que tem impacto direto no desempenho de reconhecimento. Normalmente os codificadores inserem zeros nos pacotes perdidos ou interpolam linearmente os pacotes recebidos visando restaurar estes pacotes. Foi proposta nesta tese uma nova técnica baseada em Redes Neurais que se mostrou mais eficiente na restauração destes pacotes com a finalidade da realização do reconhecimento.

Palavras-chave

Reconhecimento de Voz Distribuído; LSF; LPC; ISF; HMM; Redes IP; Redes Móveis Celulares; ITU-T G.723.1; AMR-NB; AMR-WB; Redes Neurais

Abstract

Alencar, Vladimir Fabregas Surigué; Alcaim, Abraham (Advisor). **Distributed Recognition for Continuous Speech in Large Vocabulary Brazilian Portuguese.** Rio de Janeiro, 2009. 131p. DSc. Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This Thesis aims at exploring several approaches for performance improvement of the Automatic Speech Recognition System with large vocabulary for the Brazilian Portuguese when applied in a distributed scenario (Distributed Speech Recognition). With this purpose, a speech database for continuous speech recognition for the Brazilian Portuguese with 100 speakers was constructed, each one uttering 1000 phonetic balanced sentences. The recording was carried out in a studio (environment without noise) with a specification of recording that would be able to allow the input of several speech codecs in Cellular Mobile Telephony and IP Networks, in particular the ITU-T G.723.1, AMR-NB and AMR-WB. In order to work properly, Automatic Speech Recognition Systems require that the recognition features be extracted at a high rate. However, the Speech codecs for Cellular Mobile Telephony and IP Networks normally generate its parameters at lower rates, which degrades the performance of the recognition system. Usually the linear interpolation in the LSF (*Line Spectral Frequencies*) domain is used to solve this problem. In this Thesis the accomplishment of the interpolation with the use of a Digital Filter Interpolator was proposed and demonstrated to have a higher performance than the linear interpolation in recognition systems. The use of the interpolated ISFs (*Immittance Spectral Frequencies*) was also evaluated as recognition feature, which had shown to be inadequate for this purpose, as well as the LSFs. Another very important aspect for the distributed speech recognizers is the recovery of lost packets, that has direct impact in the recognition performance. Normally the coders insert zeros in the lost packets or interpolate linearly the received packets aiming to restore them. A new technique based on Neural Networks was proposed in this thesis that showed to be more efficient in the restoration of these lost packets with the purpose of speech recognition.

Keywords

Distributed Speech Recognition; LSF; LPC; ISF; HMM; IP Networks; Cellular Mobile Networks; ITU-T G.723.1; AMR-NB; AMR-WB; Neural Networks

Sumário

1. Introdução 17

1.1. Sistemas de Reconhecimento de Voz em Ambiente Celular/Redes IP 17

1.2. Base de Voz 19

1.3. Organização da Tese 22

2. Reconhecimento de Voz Contínua em Sistemas Distribuídos 23

2.1. Extrator de Atributos em Sistemas Distribuídos 23

2.2. Reconhecimento de Voz Contínua 26

2.3. Conclusão 37

3. Codificadores de Voz em Telefonia IP e Móvel Celular 38

3.1. ITU-T G.723.1 39

3.2. Adaptive Multi-Rate Narrowband (AMR-NB) 42

3.3. Adaptive Multi-Rate Wideband (AMR-WB) 47

3.4. Conclusão 52

4. Atributos para Reconhecimento de Voz Distribuído 53

4.1. Atributos Extraídos de LPCs 54

4.1.1. Linear Predictive Coding (LPC) 54

4.1.2. LPC Cepstrum (LPCC) 58

4.1.3. Mel-Frequency LPCC (MLPCC) 59

4.2. Atributos Extraídos de LSFs 61

4.2.1. Line Spectral Frequencies (LSF) 61

4.2.2. Pseudo-Cepstral Coefficients (PCC) 63

4.2.3. Pseudo-Cepstrum (PCEP) 66

4.2.4. Mel-Frequency PCC (MPCC) 66

4.2.5. Mel-Frequency PCEP (MPCEP) 67

4.3. Atributos Extraídos de ISFs 67

4.3.1. Immittance Spectral Frequencies (ISF) 67

4.4. Atributo Extraído de Voz Reconstruída (MFCC) 69
4.5. Conclusão 75

5. Métodos de Interpolação dos Atributos 76
5.1. Interpolação Linear 76
5.2. Interpolação com Filtro Digital 77
5.3. Resultados de Simulação para o Codec ITU-T G.723.1 81
5.4. Resultados de Simulação para o Codec AMR-NB 87
5.5. Resultados de Simulação para o Codec AMR-WB 89
5.6. Conclusão 91

6. Perdas de Pacotes 93
6.1. Inserção de Zeros e Interpolação Linear 95
6.2. Redes Neurais 96
6.3. Resultados de Simulação para o Codec ITU-T G.723.1 e AMR-NB 98
6.4. Conclusão 101

7. Conclusões e Sugestões para Trabalhos Futuros 102
7.1. Conclusões 102
7.2. Sugestões para Trabalhos Futuros 106

Referências bibliográficas 107

Apêndice
A.1. Informações Técnicas da Gravação da Base 113
A.2. Publicações Relacionadas à Tese 114

Lista de figuras

Figura 1.1 – Representação gráfica da base construída 20
Figura 1.2 – Representação gráfica do cenário 1 (dependente dos 100 locutores) 21
Figura 1.3 – Representação gráfica do cenário 2 (independente do locutor com todas as frases usadas para teste e treino do sistema) 21
Figura 1.4 – Representação gráfica do cenário 3 (independente do locutor e do texto) 21
Figura 2.1 – Sistemas de Reconhecimento Distribuído – Diagrama Básico 23
Figura 2.2 – Sistema de reconhecimento de voz distribuído baseado nos parâmetros de voz do codificador 24
Figura 2.3 – Sistema de reconhecimento de voz distribuído baseado em voz decodificada 25
Figura 2.4 – Sistema de reconhecimento de voz distribuído com codificação dos atributos de reconhecimento no front-end local 26
Figura 2.5 – Diagrama em blocos de um sistema de reconhecimento automático de voz baseado em modelos estatísticos de subunidades de palavras [16] 27
Figura 2.6 – Modelo de fonema baseado em HMM 29
Figura 2.7 – Dinâmica da Busca em Feixe [18] 37
Figura 3.1 – Diagrama de blocos do codificador de voz do ITU-T G.723.1 40
Figura 3.2 – Diagrama de bloco do decodificador de voz do ITU-T G.723.1 42
Figura 3.3 – Diagrama de bloco do codificador de voz do AMR-NB 45
Figura 3.4 – Diagrama de bloco do decodificador de voz do AMR-NB 47
Figura 3.5 – Diagrama de bloco do codificador de voz do AMR-WB 49

Figura 3.6 – Diagrama de bloco do decodificador de voz do AMR-WB 51
Figura 4.1 – Percepção subjetiva da frequência fundamental de sons sonoros 70
Figura 4.2 – Magnitude do espectro dos filtros de banda crítica 71
Figura 5.1 – Representação gráfica da interpolação Linear de fator 3 78
Figura 5.2 – Representação gráfica do Sinal Original 79
Figura 5.3 – Representação gráfica do Espectro em Freqüência do Sinal Original 79
Figura 5.4 – Representação gráfica do Sinal sobre-amostrado de fator 3 79
Figura 5.5 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3 79
Figura 5.6 – Representação gráfica do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa 80
Figura 5.7 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa 80
Figura 6.1 – Modelo de Gilbert 94
Figura 6.2 – Topologia da Rede Neural 97

Lista de tabelas

Tabela 3.1 – Tabela de alocação de bits para o codificador ITU-T G.723.1 41
Tabela 3.2 – Taxa de codificação do AMR-NB e alocação de bits nas classes 43
Tabela 3.3 – Tabela de alocação de bits para o codificador AMR-NB 46
Tabela 3.4 – Taxa de codificação do AMR-WB e alocação de bits nas classes 48
Tabela 3.5 – Tabela de alocação de bits para o codificador AMR-WB 50
Tabela 4.1 – Frequências dos centros e banda crítica dos filtros utilizados para cálculo dos coeficientes mel-cepestrais 72
Tabela 5.1 – Tabela de desempenho de reconhecimento para sistema dependente de 100 locutores 84
Tabela 5.2 – Tabela de desempenho de reconhecimento para sistema independente de locutor e com as mesmas frases para teste e treino 85
Tabela 5.3 – Tabela de desempenho de reconhecimento para sistema independente de locutor e das frases 85
Tabela 5.4 – Tabela de desempenho de reconhecimento para interpolação linear e filtro digital 86
Tabela 5.5 – Tabela de desempenho de reconhecimento para o AMR-NB 89
Tabela 5.6 – Tabela de desempenho de reconhecimento para o AMR-WB 91
Tabela 6.1 – Tabela de condições de rede para o modelo de Gilbert utilizado nas simulações 94
Tabela 6.2 – Tabela de desempenho de reconhecimento para redes sem perdas de pacotes (TPP= 0% e CMR=0) 98
Tabela 6.3 – Tabela de desempenho de reconhecimento para rede com TPP= 10% e CMR=1,18 99

Tabela 6.4 – Tabela de desempenho de reconhecimento para rede com TPP= 20% e CMR=1,43 99
Tabela 6.5 – Tabela de desempenho de reconhecimento para rede com TPP= 30% e CMR=1,54 99
Tabela 6.6 – Tabela de desempenho de reconhecimento para rede com TPP= 40% e CMR=2,00 100

Lista de acrônimos

ACELP	*Algebraic Code-Excited Linear Prediction*
AMFCC	*Autocorrelation Mel-Frequency Cepstral Coefficients*
AMR-NB	*Adaptive Multi-Rate Narrowband*
AMR-WB	*Adaptive Multi-Rate Wideband*
ATK	*Application Toolkit for HTK*
CDMA	*Code Division Multiple Access*
CELP	*Code-Excited Linear Predictive*
CMR	*Comprimento Médio de Rajada*
CNPq	Conselho Nacional de Pesquisa e Desenvolvimento
CS-ACELP	*Conjugate Structure- Algebraic Code-Excited Linear Prediction*
CSR	*Continuous Speech Recognition*
DC	*Direct Current*
DP	*Dynamic Programming*
DSR	*Distributed Speech Recognition*
EVRC	*Enhanced Variable Rate Coder*
FFT	*Fast Fourier Transform*
FIR	*Finite Impulse Response*
GSM	*Global System for Mobile Communication*
GSM-EFR	*GSM-Enhanced Full Rate*
GSM-FR	*GSM-Full Rate*
GSM-HR	*GSM-Half Rate*
HMM	*Hidden Markov Model*
HTK	*HMM Toolkit*
IFFT	*Inverse Fast Fourier Transform*
IMT-2000	*International Mobile Telecommunications-2000*
IP	*Internet Protocol*
ISF	*Immittance Spectral Frequencies*
ITU-T	*International Telecommunication Union - Telecommunication Standardization Sector*
LM	*Language Mode*

LP	*Linear Prediction*
LPC	*Linear Predictive Coding*
LPCC	*LPC Cepstrum*
LSF	*Line Spectral Frequencies*
LVDCSR	*Large Vocabulary Distributed Continuous Speech Recognition*
LVR	*Large Vocabulary Recognition*
MFCC	*Mel-Frequency Cepstral Coefficients*
MLPCC	*Mel LPC Cepstral Coefficients*
MPCC	*Mel Pseudo-Cepstral Coefficients*
MPCEP	*Mel Pseudo-Cepstrum*
MP-MLQ	*Multi-pulse Maximum Likelihood Quantization*
ONG	Organização Não Governamental
PCC	*Pseudo-Cepstral Coefficients*
PCEP	*Pseudo-Cepstrum*
PCM	*Pulse-Code Modulation*
PCP	*Probabilidade Condicional de Perda*
PPI	*Probabilidade de Perda Incondicional*
PSTN	*Public switching Telecommunications Network*
PS-PA	*Pitch-Synchronous Peak Amplitude*
PSVQ	*Predictive Split Vector Quantizer*
PS-ZCPA	*Pitch-Synchronous Zero Crossings with Peak Amplitudes*
QCELP	*Qualcomm Code-Excited Linear Predictive*
RAM	*Random Access Memory*
RAV	Reconhecimento Automático de Voz
RAS	*Relative Autocorrelation Sequence*
RTT	*Round-Trip Time*
SID	*Silence Descriptor*
SMQ	*Split Matrix Quantization*
S-MSVQ	*Split-Multistage Vector Quantization*
SSCH	*Subband Spectral Centroid Histograms*
SVQ	*Split Vector Quantization*
TDMA	*Time Division Multiple Access*
TIA	*Telecommunications Industry Association*

TPP	*Taxa de Perda de Pacote*
USA	*United States of America*
VAD	*Voice Activity Detection*
VSELP	*Vector Sum Excited Linear Predictive*
WCDMA	*Wideband Code Division Multiple Access*
$\overline{WRR}$	*Average Word Recognition Rates*
ZCPA	*Zero Crossings with Peak Amplitudes*
3GPP	*3rd Generation Partnership Project*