UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DISSERTAÇÃO DE MESTRADO

ESTIMATIVAS DE COMPORTAMENTO VOCÁLICO DE LOCUTORES E UM NOVO SISTEMA DE SEPARAÇÃO SILÁBICA

ELDA LIZANDRA FERNANDES DA SILVA

RECIFE, 28 DE MAIO DE 2012

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ESTIMATIVAS DE COMPORTAMENTO VOCÁLICO DE LOCUTORES E UM NOVO SISTEMA DE SEPARAÇÃO SILÁBICA

POR

ELDA LIZANDRA FERNANDES DA SILVA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para obtenção do grau de Mestre em Engenharia Elétrica.

ORIENTADOR: PROF. HÉLIO MAGALHÃES DE OLIVEIRA

Recife, 28 de Maio de 2012.

Catalogação na fonte
Bibliotecário Marcos Aurélio Soares da Silva, CRB-4 / 1175

S586e Silva, Elda Lizandra Fernandes da.
Estimativas de comportamento vocálico de locutores e um novo sistema de separação silábica / Elda Lizandra Fernandes da Silva. - Recife: O Autor, 2012.
xii, 150 folhas, il., gráfs., tabs.

Orientador: Profº Drº. Hélio Magalhães de Oliveira.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Elétrica, 2012.
Inclui Referências e Anexos.

1. Engenharia Elétrica. 2. Caracterização de Voz. 3. Processamento da Fala. 4.Sons Vocálicos. I. Oliveira, Hélio Magalhães de (Orientador). II. Título.

621.3 CDD (22. ed.) UFPE
BCTG/2013-003

Universidade Federal de Pernambuco
Pós-Graduação em Engenharia Elétrica

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE DISSERTAÇÃO DO MESTRADO ACADÊMICO DE

ELDA LIZANDRA FERNANDES DA SILVA

TÍTULO

"ESTIMATIVAS DE COMPORTAMENTO VOCÁLICO DE LOCUTORES E UM NOVO SISTEMA DE SEPARAÇÃO SILÁBICA"

A comissão examinadora composta pelos professores: RICARDO MENEZES CAMPELLO DE SUZA, DES/UFPE, JULIANO BANDEIRA LIMA, DM/ UFPE e ADRIÃO DUARTE DÓRIA NETO, EC/UFRN sob a presidência do primeiro, consideram a candidata **ELDA LIZANDRA FERNANDES DA SILVA** **APROVADA**.

Recife, 28 de maio de 2012.

CECÍLIO JOSÉ LINS PIMENTEL
Coordenador do PPGEE

RICARDO MENEZES CAMPELLO DE SOUZA
Membro Titular Interno

ADRIÃO DUARTE DÓRIA NETO
Membro Titular Externo

JULIANO BANDEIRA LIMA
Membro Titular Externo

AGRADECIMENTOS

Gostaria de agradecer primeiramente a meus pais, Maria José da Silva e Edmilson Fernandes da Silva, pelo amor, dedicação, incentivo e paciência durante todo esse longo percurso. A minha irmã Melba Lizania Fernandes Borba pela compreensão e companheirismo, em todos os momentos.

Aos colegas do LACRI e DES por suas colaborações neste trabalho, em especial a: Gilson Jerônimo Júnior, Paulo Roberto Lima Martins, Paulo Hugo Espírito Santo e a Raimundo C. de Oliveira por valiosas sugestões teóricas e pelo auxílio durante o processo de implementação do código.

A todos os funcionários do DES, em especial a secretária do programa de pós-graduação de engenharia elétrica, Andrea Tenório Pinto.

Gostaria, principalmente, de agradecer ao meu orientador, professor Hélio Magalhães de Oliveira, pela oportunidade e pela dedicação.

Finalmente, gostaria de agradecer ao CNPq pelo apoio financeiro e ao PPGEE-UFPE.

RESUMO DA DISSERTAÇÃO APRESENTADA À UFPE COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

ESTIMATIVAS DE COMPORTAMENTO VOCÁLICO DE LOCUTORES E UM NOVO SISTEMA DE SEPARAÇÃO SILÁBICA

Elda Lizandra Fernandes da Silva

MAIO DE 2012

ORIENTADOR: PROF. DR. HÉLIO MAGALHÃES DE OLIVEIRA, *DOCTEUR*

ÁREA DE CONCENTRAÇÃO: TELECOMUNICAÇÕES/PROCESSAMENTO DE VOZ.

PALAVRAS-CHAVE: CARACTERIZAÇÃO DE VOZ, *PITCH*, SONS VOCÁLICOS. DIVISÃO SILÁBICA, LÍNGUA PORTUGUESA, PROCESSAMENTO DE FALA, CONVERSÃO FALA-PARA-TEXTO.

NÚMERO DE PÁGINAS: 150

Nesta dissertação um método simples para a estimação automática do comportamento espectral de trechos vocálicos de locutores é proposto. Uma implementação computacional em Matlab® é apresentada e sua validação é conduzida comparando os resultados com uma identificação realizada manualmente, empregando o Audacity 1.3®. Locutores (masculinos e femininos) foram considerados e os testes foram conduzidos para sete diferentes sons vocálicos da língua portuguesa (a, é, ê, i, ó, ô, u). A abordagem é potencialmente útil em modelos de trato vocal, na melhoria da qualidade de sintetizadores de voz ou em algoritmos de reconhecimento automático de locutor. Em uma segunda parte, um novo algoritmo para divisão silábica automática de arquivos de voz na língua portuguesa é proposto, com base na envoltória do sinal de voz. Uma implementação computacional em Matlab® é apresentada, a qual encontra-se disponibilizada na URL http://www2.ee.ufpe.br/codec/divisao_silabica.html. Trechos longos contendo mais de uma sílaba e identificados com uma mesma envoltória são chamados de supersilabas e são separados posteriormente. Os resultados identificam as amostras correspondentes ao início e o fim de cada sílaba detectada. Foram realizados testes preliminares com meia centena de palavras, com uma taxa de identificação de cerca de 70%, porém melhorias podem ser incorporadas para tratar fonemas nos quais o envelope não é o principal parâmetro na identificação. Este algoritmo também pode ser particularmente útil em sistemas com comandos de voz ou como ferramenta de apoio no ensino da língua portuguesa ou para pacientes em tratamento fonoaudiológico.

ABSTRACT OF DISSERTATION PRESENTED TO UFPE AS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER IN ELECTRICAL ENGINEERING

ESTIMATES OF VOWEL SPEAKER BEHAVIOR AND A NEW SYLLABIC SEPARATION SYSTEM

Elda Lizandra Fernandes da Silva

MAY 2012

ADVISOR: PROF. HÉLIO MAGALHÃES DE OLIVEIRA, DOCTEUR.

AREA OF CONCENTRATION: TELECOMMUNICATIONS/VOICE PROCESSING

KEYWORDS: SPEECH CHARACTERIZATION, PITCH, VOWEL SOUNDS. SYLLABIC DIVISION, PORTUGUESE LANGUAGE, SPEECH PROCESSING, VOICE-TO-TEXT.

NUMBER OF PAGES: 150

In this thesis a new automatic method for estimating the spectral behavior of vowel excerpts of speakers is proposed. A computational implementation in Matlab™ is presented and the validation of the approach is conducted by comparing the results with an identification manually performed using Audacity 1.3™. Speakers (both male and female) were considered and the tests were conducted for seven different vowel sounds of the Portuguese language (namely, a, é, ê, i, o, ô, u). The approach may be useful in models of the vocal tract, for improving the quality of speech synthesizers or to be used in algorithms for automatic speaker recognition. Also, a new algorithm for automatic syllabic splitting in the Portuguese language is proposed, which is based on the envelope of the speech signal of an audio file. A computational implementation in Matlab™ is presented and made available at the URL http://www2.ee.ufpe.br/codec/divisao_silabica.html. Voice excerpts containing more than one syllable and identified by the same envelope are named as super-syllables and they are separated subsequently. The results indicate which samples correspond to the beginning and end of each detected syllable. Preliminary tests were performed to fifty words at an identification rate circa 70% (further improvements may be incorporated to treat particular phonemes). This algorithm can also be particularly useful in voice command systems, as a tool in the teaching of Portuguese language or even for patients with speech pathology.

Sumário

LISTA DE FIGURAS ..... **vii**
LISTA DE TABELAS ..... **xii**
1. INTRODUÇÃO **14**
1.1. OBJETIVOS **15**
1.2. ORGANIZAÇÃO DO TRABALHO **15**
2. SOM, VOZ E SUAS RESPECTIVAS CARACTERÍSTICAS **18**
2.1. O SOM **18**
2.1.1. CARACTERÍSTICAS **18**
2.1.2. O OUVIDO HUMANO **19**
2.1.3. O OUVIDO EXTERNO **19**
2.1.4. O OUVIDO MÉDIO **20**
2.1.5. O OUVIDO INTERNO **21**
2.2. VOZ **22**
2.2.1. PRODUÇÃO DA VOZ **23**
2.2.2. SONS VOCÁLICOS E NÃO VOCÁLICOS **24**
3. TÉCNICAS DE PROCESSAMENTO DO SOM **26**
3.1. DIGITALIZAÇÃO DO SINAL DE VOZ **26**
3.1.1. AMOSTRAGEM **26**
3.1.2. QUANTIZAÇÃO **27**
3.1.2.1. ERROS DE QUANTIZAÇÃO **28**
3.2. MODELO DO SISTEMA DE PRODUÇÃO VOCAL **30**
3.3. ALGORITMOS DE DETECÇÃO DE PITCH **32**
4. ESTIMATIVA DO COMPORTAMENTO VOCÁLICO DE LOCUTORES **34**
4.1. AQUISIÇÃO DE VOZ E JANELAMENTO **34**
4.2. PREENCHIMENTO COM ZEROS E NORMALIZAÇÃO ESPECTRAL **41**
4.3. A DERIVADA BINÁRIA E IDENTIFICAÇÃO DE PICOS **41**
4.4. RESULTADOS EXPERIMENTAIS **43**
5. IMPLEMENTAÇÃO DE UM ALGORITMO DE DIVISÃO SILÁBICA AUTOMÁTICA EM ARQUIVOS DE VOZ NA LÍNGUA PORTUGUESA **48**
5.1. AQUISIÇÃO E PRÉ–PROCESSAMENTO DE VOZ **48**
5.2. RETIFICAÇÃO DA ONDA, VALOR RMS E DESCARGA LINEAR **49**

5.3. LOCALIZADOR SILÁBICO 53
5.4. IDENTIFICAÇÃO DE SUPERSÍLABAS E QUEBRA 55
5.5. AGLUTINAÇÃO DE SÍLABAS SEPARADAS INDEVIDAMENTE 58
5.6. DESEMPENHO DO DIVISOR SILÁBICO 61
6. CONCLUSÕES 67
6.1. SOBRE O ESTIMADOR VOCÁLICO 67
6.2. SOBRE O DIVISOR SILÁBICO 67
7. TRABALHOS FUTUROS 69
7.1. POTENCIAIS MELHORIAS NO ESTIMADOR VOCÁLICO 69
7.2. POTENCIAIS MELHORIAS NO DIVISOR SILÁBICO 69
ANEXO A – CÓDIGO FONTE DOS ALGORITMOS 71
ANEXO B – TABELAS E GRÁFICOS 80
ANEXO C – CURVAS DE CORRELAÇÃO DE CADA VOGAL PARA CADA LOCUTOR OBTIDOS PELO AUDACITY 1.3® E PELO ESTIMADOR. 93
ANEXO D – SEPARAÇÃO SILÁBICA DAS PALAVRAS TESTADAS 118
ANEXO E – ANÁLISE DA DIVISÃO SILÁBICA EFETUADA PELO ALGORITMO DE SEPARAÇÃO SÍLABICA PARA POESIA DE MANUEL BANDEIRA 141
REFERÊNCIAS BIBLIOGRÁFICAS 146

LISTA DE FIGURAS

Figura 2.1: Ilustração do esquema do sistema auditivo (VERLAG, 2012). 20
Figura 2.2: Ilustração esquema do ouvido externo (COLEMAN, 2012). 20
Figura2.3: Ilustração do esquema do ouvido médio (COLEMAN, 2012). 22
Figura 2.4: Esquema do ouvido interno (BISTAFA, 2006). 23
Figura 2.5: Ilustração do esquema do sistema produtor de voz (BOUMAN, 2012). 24
Figura 3.1: Efeitos da amostragem: (a) espectro de Fourier do sinal original; (b) espectro de Fourier do sinal amostrador; (c) amostragem a uma taxa superior à taxa de Nyquist; (d) amostragem a uma taxa inferior à taxa de Nyquist. 28
Figura 3.2: Características de entrada-saída de um quantizador uniforme. (SOTERO, 2009). 29
Figura 3.3: Função de compressão de um quantizador não uniforme típico (SOTERO, 2009). 30
Figura 3.4: Modelo de um gerador de voz (fonte-filtro) (VASEGHI, 2007). 31
Figura 4.1: Ilustração do sinal pré-processado correspondente a um sinal de voz da vogal "a" sendo repetida por 7 segundos pelo locutor Alessandra. 35
Figura 4.2: Interface do Audacity 1.3® para análise de frequência com 128 amostras para a vogal a, sendo repetida por 7 segundos pelo locutor Alessandra.(a) Espectro em escala frequencial linear. (b) Espectro em escala frequencial logarítmica. 37
Figura 4.3: Ilustração para um sinal amostrado com N amostras. a) Amostras no domínio do tempo. b) Amostras no domínio da frequência. 38
Figura 4.4: Interface do Audacity 1.3® para análise de frequência com 512 amostras para a vogal "a", sendo repetida por 7 segundos pelo locutor Alessandra. O aumento do comprimento da DFT conduz a explicitar os picos relacionados com os sons harmônicos. 39
Figura 4.5: Interface do Audacity 1.3 ® para análise de frequência com 1024 amostras para a vogal "a' sendo repetida por 7 segundos pelo locutor Alessandra". 39
Figura 4.6: Interface do Audacity 1.3® para análise de frequência com 2048 amostras para a vogal "a" sendo repetida por 7 segundos pelo locutor Alessandra.Verifica-se uma estabilização no formato espectral. 40

Figura 4.7: Ilustração da Interface gráfica em Matlab®, do estimador de Pitch para a vogal "a" (Alessandra). Seleciona-se o arquivo extensão .wav (canto direito da tela) 19 picos são mostrados para esta vogal. 44

Figura 4.8: Correlação entre picos identificados pelo aplicativo (ajuste linear) para: a) locutor Alessandra, pronunciando longamente o som vocálico "a". b) locutor Ricardo, pronunciando longamente o som vocálico "a". Equação de regressão e coeficiente de determinação indicados. Ajustes de regressão linear com n pontos, 9 <n <37, dependendo do som. 47

Figura 5.1: Ilustração da interface do programa Matlab® para a palavra "departamento" após o pré-processamento. 49

Figura 5.2: Ilustração da Interface gráfica do Matlab® para o sinal de áudio da palavra "batata" após retificação de meia-onda. 50

Figura 5.3: Envoltória mal ajustada, por uso de taxa de descarga muito elevada (9702amostras, o que excede até a janela padrão de 2048 amostras). O sinal de voz corresponde a um trecho do sinal "batata". 51

Figura 5.4: Forma de onda da envoltória demodulada corretamente para a sílaba "BA" da palavra "batata". A descarga completa da envoltória pode ocorrer em até 22 ms. 51

Figura 5.5: Forma de onda da envoltória recuperada usando uma constante de tempo muito pequena, 2,2 ms, para a silaba "BA" da palavra "batata" ilustrada na Figura 5.2. 52

Figura 5.6: Ilustração da Interface gráfica do Matlab® para a envoltória do sinal de áudio da palavra "batata" referente a Figura 5.2. 53

Figura 5.7: Forma de onda correspondente à identificação de sílabas para a palavra "departamento". São identificadas 4 sílabas sendo a terceira uma supersílaba. 57

Figura 5.8: Forma de onda correspondente à identificação de sílabas para a palavra "departamento". Após a "quebra" de uma supersílaba são identificadas corretamente as 5 sílabas. 58

Figura 5.9: Forma de onda da palavra vale com a separação inicial composta por três "sílabas": va-l-le. 60

Figura 5.10: Divisor silábico para a palavra vale após a aplicação do procedimento de aglutinação: as sílabas identificadas correspondem a "va-le". 60

Figura 5.11: Forma de onda para a palavra hoje com a separação inicial composta por três “sílabas”: “ho-j-je”. 61
Figura 5.12: Divisor silábico para a palavra hoje após a aplicação do procedimento de aglutinação: as sílabas identificadas correspondem a “ho-je”. 61
Figura 5.13: Formas de onda envolvida (as cores são disponibilizadas na versão eletrônica)s: a) em cor verde, sinal de áudio referente à palavra “departamento”; b) em cor vermelha, a onda retificada em meia-onda; c) em cor azul, o envelope do sinal obtido com envoltória de descarga linear. 64
Figura 5.14: Forma de onda com a respectiva separação silábica obtida por limiar aplicado ao envelope do sinal de voz para a palavra “departamento”. 65
Figura 5.15: Separação silábica no sinal original: note a separação clara dos fonemas “DE-PAR-TAMEN-TO”. A supersílaba “TAMEN” foi corretamente subdividida.65
Figura 5.16: Divisão silábica completada. Trechos são indicados; há acesso ao comprimento em amostras e duração das sílabas e ao trecho de áudio isolado de cada silaba para a palavra “departamento”. 66
Figura A.1: Diagrama em blocos do código fonte do estimador vocálico. 72
Figura A.2: Diagrama em blocos do código fonte do separador silábico. 75
Figura C.1: Curvas de correlação das vogais “A” e “Ê” para a locutora Alessandra. 94
Figura C.2: Curvas de correlação das vogais “É” e “I” para a locutora Alessandra. 95
Figura C.3: Curvas de correlação das vogais “Ô” e “Ó” para a locutora Alessandra. ... 96
Figura C.4: Curvas de correlação da vogal “U” para a locutora Alessandra. 97
Figura C.5: Curvas de correlação das vogais “A” e “Ê” para a locutora Lidiane. 98
Figura C.6: Curvas de correlação das vogais “É” e “I” para a locutora Lidiane. 99
Figura C.7: Curvas de correlação das vogais “Ô” e “Ó” para a locutora Lidiane. 100
Figura C.8: Curvas de correlação da vogal “U” para a locutora Lidiane. 101
Figura C.9: Curvas de correlação das vogais “A” e “Ê” para a locutora Lizandra. 102
Figura C.10: Curvas de correlação das vogais “É” e “I” para a locutora Lizandra. 103
Figura C.11: Curvas de correlação das vogais “Ô” e “Ó” para a locutora Lizandra.... 104
Figura C.12: Curvas de correlação da vogal “U” para a locutora Lizandra. 105
Figura C.13: Curvas de correlação das vogais “A” e “Ê” para o locutor Paulo Freitas. 106
Figura C.14: Curvas de correlação das vogais “É” e “I” para o locutor Paulo Freitas. 107

Figura C.15: Curvas de correlação das vogais "Ô" e "Ó" para o locutor Paulo Freitas. 108
Figura C.16: Curvas de correlação da vogal "U" para o locutor Paulo Freitas. 109
Figura C.17: Curvas de correlação das vogais "A" e "Ê" para o locutor Paulo Martins. 110
Figura C.18: Curvas de correlação das vogais "É" e "I" para o locutor Paulo Martins. 111
Figura C.19: Curvas de correlação das vogais "Ô" e "Ó" para o locutor Paulo Martins. 112
Figura C.20: Curvas de correlação da vogal "U" para o locutor Paulo Martins. 113
Figura C.21: Curvas de correlação das vogais "A" e "Ê" para o locutor Ricardo. 114
Figura C.22: Curvas de correlação das vogais "É" e "I" para o locutor Ricardo. 115
Figura C.23: Curvas de correlação das vogais "Ô" e "Ó" para o locutor Ricardo. 116
Figura C.24: Curvas de correlação da vogal "U" para o locutor Ricardo. 117
Figura D.1: Separação silábica da palavra "ABACATE". 119
Figura D.2: Separação silábica da palavra "ABACAXI". 119
Figura D.3: Separação silábica da palavra "ASSADO". 120
Figura D.4: Separação silábica da palavra "AVIÃO". 120
Figura D.5: Separação silábica da palavra "BATATA". 121
Figura D.6: Separação silábica da palavra "BERIMBOCA". 121
Figura D.7: Separação silábica da palavra "BONITA". 122
Figura D.8: Separação silábica da palavra "BRASILEIRO". 122
Figura D.9: Separação silábica da palavra "BUTANTÃ". 123
Figura D.10: Separação silábica da palavra "CABEÇA". 123
Figura D.11: Separação silábica da palavra "CABELO". 124
Figura D.12: Separação silábica da palavra "CAFÉ". 124
Figura D.13: Separação silábica da palavra "CAMPUS". 125
Figura D.14: Separação silábica da palavra "CIRCUNFERÊNCIA". 125
Figura D.15: Separação silábica da palavra "COMPLEXO". 126
Figura D.16: Separação silábica da palavra "COMPUTADOR". 126
Figura D.17: Separação silábica da palavra "CORPO". 127
Figura D.18: Separação silábica da palavra "DEPARTAMENTO". 127
Figura D.19: Separação silábica da palavra "DUZENTOS". 128

Figura D.20: Separação silábica da palavra “ECONOMIA”. 128
Figura D.21: Separação silábica da palavra “ELETRÔNICA”. 129
Figura D.22: Separação silábica da palavra “ENGENHARIA”. 129
Figura D.23: Separação silábica da palavra “FARMÁCIA”. 130
Figura D.24: Separação silábica da palavra “HOJE”. 130
Figura D.25: Separação silábica da palavra “HISTÓRIA”. 131
Figura D.26: Separação silábica da palavra “MATEMÁTICA”. 131
Figura D.27: Separação silábica da palavra “MINUTO”. 132
Figura D.28: Separação silábica da palavra “MISTÉRIO 132
Figura D.29: Separação silábica da palavra “MÚSICA”. 133
Figura D.30: Separação silábica da palavra “OFICINA”. 133
Figura D.31: Separação silábica da palavra “PERNAMBUCO”. 134
Figura D.32: Separação silábica da palavra “PITOCO”. 134
Figura D.33: Separação silábica da palavra “RECIFE”. 135
Figura D.34: Separação silábica da palavra “ROUPA”. 135
Figura D.35: Separação silábica da palavra “SEMICONDUTOR”. 136
Figura D.36: Separação silábica da palavra “SIRI”. 136
Figura D.37: Separação silábica da palavra “SOLTEIRO”. 137
Figura D.38: Separação silábica da palavra “TELEVISÃO”. 137
Figura D.39: Separação silábica da palavra “UNIVERSIDADE”. 138
Figura D.40: Separação silábica da palavra “UVA”. 138
Figura D.41: Separação silábica da palavra “VALE”. 139
Figura D.42: Separação silábica da palavra “VESTIBULAR”. 139
Figura D.43: Separação silábica da palavra “ZEBRA”. 140

LISTA DE TABELAS

Tabela 4.1:: DFT com relação à respectiva resolução. 40
Tabela 4.2: Exemplo para o vetor derivada binária 42
Tabela 4.3: Picos de frequências, em Hz, para os 19 picos significativos, obtidos pelo Audacity 1.3® e pelo Identificador para a vogal "a" pronunciada por um locutor do sexo feminino. 46
Tabela 5.1: Matriz de Contagem Obtida Via Matlab® para o Arquivo de Trecho de Áudio Pré-processado Referente à Palavra "batata". 55
Tabela 5.2: Dados do separador silábico para a palavra "departamento". Notar a identificação de uma supersílaba (terceira sílaba). 56
Tabela 5.3: Dados do separador silábico para a palavra "departamento", após a correta divisão da supersílaba identificada. 57
Tabela 5.4: Sílabas, amostra inicial e amostra final para a palavra "vale" antes da aglutinação 59
Tabela 5.5: Sílabas, amostra inicial, amostra final e durações da palavra "vale" após aglutinação 60
Tabela 5.6: Sílabas Separadas no Arquivo de Voz Contendo a Palavra "departamento". Índice da Amostra Inicial e da Amostra Final e Duração Estimada da Sílaba (exemplo de saída de dados). 62
Tabela 5.7: Lista de palavras com as respectivas sílabas separadas corretamente pelo algoritmo de divisão silábica (23 Palavras) 63
Tabela 5.8: Lista de palavras com as respectivas sílabas separadas de forma parcial pelo algoritmo de divisão silábica (20 palavras) 64
Tabela B.1: Raias espectrais iniciais e respectivos passos para os seis locutores, para as vogais "a","e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz). 81
Tabela B.2: Comparação da estimativa de pitch para os seis locutores, para as vogais a, e , incluindo acentuação (grave/agudo), de acordo com o programa de estimativa espectral vocálica (Nesta Dissertação) e algoritmo de identificação de pitch (XUEJING, 2002) (valores em Hz). 81
Tabela B.3: Raias espectrais de Alessandra, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz). 82

Tabela B.4: Raias espectrais de Lidiane, para as vogais “a”, “e” ,incluindo acentuação (grave/agudo), “i”, “o”, “u” (valores em Hz). **83**
Tabela B.5: Raias espectrais de Lizandra, para as vogais “a”, “e” ,incluindo acentuação (grave/agudo), “i”, “o”, “u” (valores em Hz). **85**
Tabela B.6: Raias espectrais de Paulo Freitas, para as vogais “a”, “e” ,incluindo acentuação (grave/agudo), “i”, “o”, “u” (valores em Hz). **87**
Tabela B.7: Raias espectrais de Paulo Martins, para as vogais “a”, “e”, incluindo acentuação (grave/agudo), “i”, “o”, “u” (valores em Hz) **89**
Tabela B.8: Raias espectrais de Ricardo, para as vogais “a”, “e”, incluindo acentuação (grave/agudo), “i”, “o”, “u” (valores em Hz). **91**

1. INTRODUÇÃO

Neste capítulo é realizado um resumo sobre alguns elementos abordados nesta dissertação. Inicia-se essa abordagem tratando da classificação dos sons dentro da língua portuguesa, diferenciando os vocálicos dos não vocálicos. O estudo do comportamento dos sons vocálicos deu origem ao primeiro estudo deste trabalho (Uma estimativa do comportamento vocálico de locutores). Em seguida é feita uma breve apresentação sobre as recentes técnicas empregadas em sistemas fala-texto e é introduzido o segundo estudo da dissertação (Implementação de um algoritmo de divisão silábica automática em arquivos de voz na língua portuguesa).

Os sons relacionados à fala podem ser classificados como vocálicos e não vocálicos (HOLMES&HOLMES, 2001, RABINER&SCHAFER, 2007). Os vocálicos são obtidos quando o ar que vem dos pulmões passa pela glote e não sofre interrupção parcial ou total por: língua, lábios, dentes, etc., provocando vibrações quase periódicas. Já os sons não-vocálicos são obtidos pela interrupção parcial ou total do ar, durante o percurso dos pulmões até sua saída pelas cavidades oral e nasal (HOLMES&HOLMES, 2001).

A maior parte dos sons emitidos na língua portuguesa é vocálica e, portanto, de comportamento espectral caracterizado por frequências mais bem definidas (picos/raias espectrais). Este trabalho propõe investigar o comportamento espectral de sons vocálicos na língua Portuguesa.

Com o advento de um processamento mais eficiente, e com a disponibilização crescente de novas técnicas (VASEGNHI, 2007, HOLMES&HOLMES, 2001, OPPENHEIM&SCHAFER, 2010), o desenvolvimento de novos aplicativos envolvendo síntese de voz, reconhecimento de locutor, tradutores, conversão (texto-fala e fala-texto) vem se proliferando (TAYLOR, 2003, dos SANTOS, 1997, SOTERO&de OLIVEIRA, 2009).

As aplicações de sistemas envolvendo conversão de voz acústica em texto de língua portuguesa já são promissoras (FRAGA, 2001, SILVA *et* al., 2008, SILVA&KLAUTAU, 2009), mas os sistemas atuais ainda carecem de melhorias (NETO, 2005). Alguns sistemas de conversão fala-texto para o Português já foram propostos (e.g. FRAGA, 2001) e eles tradicionalmente envolvem três etapas: uma segmentação (sub)silábica, a posterior conversão de fonemas segmentados em texto, e verificação ortográfica e gramatical das palavras e sentenças identificadas (HUANG&ACERO, 2001). A ênfase e foco deste trabalho estão na

etapa de segmentação dos sinais de voz, com vocabulário ilimitado, para uma divisão silábica automática (GOUVEIA *et* al., 2000). ROSENBERG *et* al.(1983) propuseram um sistema de reconhecimento de palavras através da concatenação de "meia-sílabas", protótipos usando modelos de referência (*templates*), com taxa de erro no reconhecimento de subsílabas na faixa 18 – 33%. Outros sistemas para reconhecimento de palavras isoladas foram propostos com base em empenamento temporal dinâmico (LIPEIKA *et* al., 2002). Frequentemente, a segmentação automática de fala emprega modelos nos quais as subunidades fonéticas (dependentes de contexto) são representadas com Modelos Ocultos de Markov-HMM (SELMINI, 2008, dos SANTOS&ALCAIM, 2001). Embora as taxas de acerto sejam aceitáveis, os algoritmos empregados são computacionalmente intensivos, especialmente para aplicações em tempo real, ou desenvolvimento em sistemas embarcados (iPods, celulares etc.). A proposta aqui é introduzir uma técnica alternativa bastante simples (quando comparada à HMM ou técnicas envolvendo análise Cepstral, redes neuronais etc.) que pode funcionar como um passo preliminar a ser incorporado nos algoritmos existentes. Não se trata de uma concorrência ou alternativa direta para os sistemas (nem as taxas de acerto, nem as complexidades estão no mesmo patamar), mas uma maneira de concretizar um pré-processamento, para só então, após a aplicação desta técnica inicial, incorporar outras estratégias. Apesar do objetivo mais modesto, isto potencialmente pode aumentar a velocidade dos sistemas fala-texto (para a língua portuguesa) já propostos.

1.1. OBJETIVOS

Este trabalho possui como objetivos:

- Um método simples para a estimação automática do comportamento vocálico de locutores.
- E um novo algoritmo automático para divisão silábica de arquivos de voz para língua portuguesa, com base na envoltória deste sinal de voz.

1.2. ORGANIZAÇÃO DO TRABALHO

A seguir é apresentada a estruturação e uma breve descrição da dissertação, sua divisão em capítulos e anexos.

Capítulo 1: o primeiro capítulo traz uma breve introdução sobre os sistemas propostos, além dos objetivos desta dissertação e a organização do trabalho.

Capítulo 2: este capítulo apresenta as características do som e da voz. É realizada, neste capítulo, uma breve descrição sobre as características que definem o som e como este se propaga pelo sistema auditivo humano. Além disso, neste capítulo, é descrito como as ondas sonoras são produzidas no corpo humano e se transformam em voz e uma breve explanação do que são sons vocálicos e não vocálicos.

Capítulo 3: o terceiro capítulo aborda as técnicas de processamento do som, amostragem e quantização, além disto, o capítulo apresenta o modelo do sistema de produção vocal mais usado e faz uma breve descrição de alguns algoritmos de detecção de *pitch*.

Capítulo 4: este capítulo descreve o algoritmo desenvolvido, neste trabalho, para estimar o comportamento vocálico de locutores. Neste capítulo está descrito todas as etapas do processo em subseções como se segue: aquisição de voz e janelamento, preenchimento com zeros e normalização, derivada binária e identificação de picos e os resultados experimentais.

Capítulo 5: este capítulo descreve o método proposto para a separação silábica de arquivos de áudio de palavras. Neste capítulo está descrito todas as etapas do processo em subseções como se segue: aquisição de voz e pré-processamento; retificação da onda, valor *RMS* e descarga linear; localizador silábico; identificação de supersílabas e quebra; aglutinação de sílabas separadas indevidamente e o desempenho do divisor silábico.

Capítulo 6: o sexto capítulo apresenta as conclusões desta dissertação para ambos os métodos.

Capítulo 7: este capítulo apresenta os trabalhos futuros desta dissertação para ambos os métodos.

ANEXO A: Código fonte dos algoritmos do estimador vocálico e do separador silábico.

ANEXO B: Tabelas e gráficos

ANEXO C: Curvas de correlação de cada vogal para cada locutor obtidas pelo Audacity 1.3® e pelo estimador.

ANEXO D: Separação silábica das palavras testadas.

ANEXO E: Análise da divisão silábica efetuada pelo algoritmo de separação silábica para a poesia de Manuel Bandeira “Vou-me embora para Pasárgada”.

2. SOM, VOZ E SUAS RESPECTIVAS CARACTERÍSTICAS

Este capítulo apresenta uma introdução sobre aspectos referentes à produção da fala humana. Inicialmente é realizado um sumário sobre acústica e características dos sons de fala. Além disso, tem-se uma breve descrição do ouvido humano, que se inicia com o ouvido externo, passa pelo ouvido médio e vai até o ouvido interno. Em seguida, há uma explanação sobre a fala e sua respectiva produção e finalmente, tem-se, ainda, entre as características da fala, a distinção entre sons os vocálicos e os não-vocálicos.

2.1. O SOM

O som é uma onda mecânica e, para se propagar, necessita de um meio. O meio mais comum de propagação do som é o ar, no entanto, em outros meios tais como sólidos, líquidos e outros gases ela também ocorre. Para dar origem a uma onda sonora é necessário que haja uma fonte. Ao vibrar, essa fonte ocasiona ao seu redor, variações de densidade e pressão ao longo da direção de propagação (RUMSEY&MCCORMICK, 2006). As rarefações e compressões geradas por uma fonte acústica dão origem a ondas sonoras, que são captadas pelo ouvido e posteriormente transformadas em impulsos elétricos e transmitidas ao cérebro. As ondas sonoras apresentam três características principais que são examinadas a seguir.

2.1.1. CARACTERÍSTICAS

Com base em (SMITH, 2003), o som tem suas características definidas por:

- **Altura** (*pitch*) – É a propriedade que permite classificar um som como grave ou agudo. Essa propriedade está relacionada diretamente com a frequência do sinal, isto é, quanto mais grave um sinal sonoro, menor sua respectiva frequência; em contrapartida, quanto mais agudo, maior a frequência. *Pitch* é a frequência fundamental de uma onda sonora, ou seja, é a frequência mais importante da onda, a partir da qual a forma de onda se repete.
- **Volume** – É a medida da intensidade de uma onda sonora.
- **Timbre** – É uma propriedade que permite diferenciar os indivíduos e instrumentos musicais através da onda sonora que estes produzem. O timbre está relacionado

com as amplitudes das componentes harmônicas de uma onda sonora e suas respectivas distribuições ao longo dessa onda, ou seja, o timbre está relacionado com a forma do espectro de frequência dessa onda.

Como o órgão responsável por receber as ondas sonoras e convertê-las em impulsos elétricos é o ouvido, torna-se necessária a compreensão do mecanismo de funcionamento desse sistema. Para isso, uma breve explanação sobre a anatomia do sistema auditivo humano é descrita nas seções subsequentes.

2.1.2. O OUVIDO HUMANO

O ouvido é um órgão complexo que possui a função de captar os estímulos sonoros do ambiente. Como visto, esses estímulos são inicialmente variações de pressão que ao passarem pelo ouvido são convertidos em impulsos elétricos e em seguida enviados ao cérebro, responsável por realizar o processo de decodificação desses impulsos (CHU, 2003).

Nem sempre o mesmo estímulo sonoro é percebido da mesma maneira por diferentes pessoas. Isto se deve ao fato de que fatores ambientais e emocionais podem influenciar a percepção de cada indivíduo de maneira única (CASIERRA, 2009).

De maneira a tornar mais simples o estudo do aparelho auditivo, dividiu-se o ouvido em três partes:

- Ouvido externo;
- Ouvido médio;
- Ouvido interno.

A ilustração das três partes do ouvido pode ser vista na Figura 2.1.

2.1.3. O OUVIDO EXTERNO

O ouvido externo, como mostra a ilustração da Figura 2.2, é constituído por: pavilhão auricular ou orelha, conduto auditivo externo e membrana timpânica. Devido as suas saliências e curvaturas, a orelha é capaz de não só distinguir a origem da fonte sonora como também amplificar ou amortecer esses sons, direcionando-os para o canal auditivo externo. Quando a onda sonora chega ao canal auditivo, ela sofre ressonâncias em algumas frequências. Essas ressonâncias acontecem devido ao fato do conduto auditivo ser um órgão de comprimento considerável para suas dimensões, cerca de 30 mm, e ser fechado pela membrana timpânica na outra extremidade (BISTAFA, 2006). Por este motivo as

ondas sonoras apresentam uma amplificação de até 20 dB (CASIERRA, 2009), em que dB é a abreviatura em decibel. É uma unidade logarítmica utilizada para medir a intensidade do som.

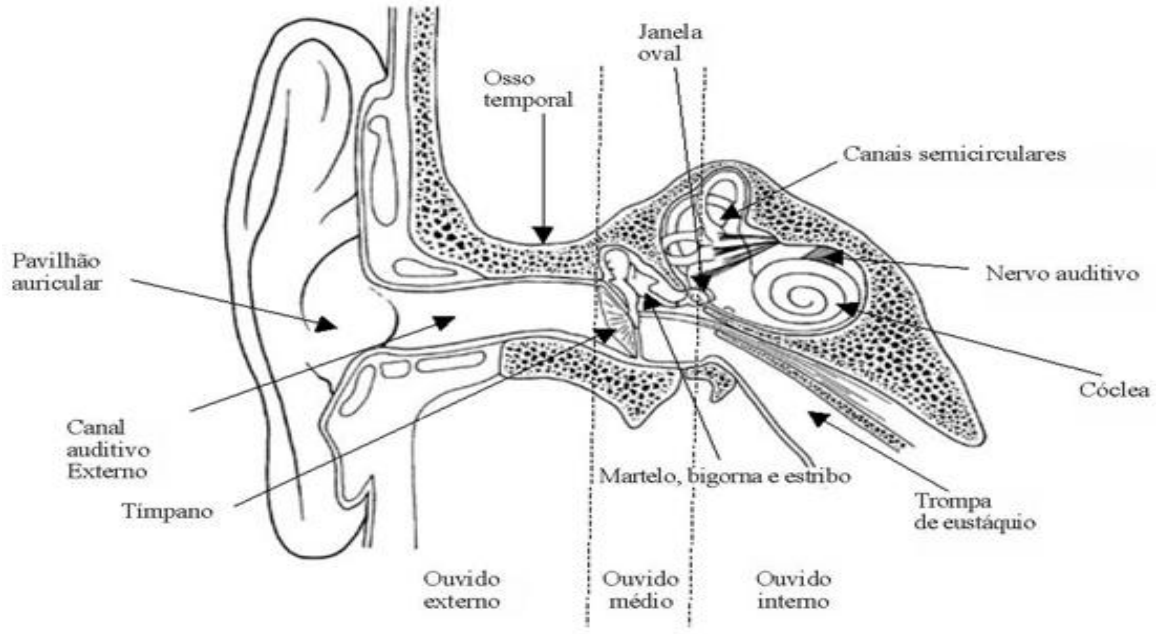


Figura 2.1: *Ilustração do esquema do sistema auditivo (VERLAG, 2012).*

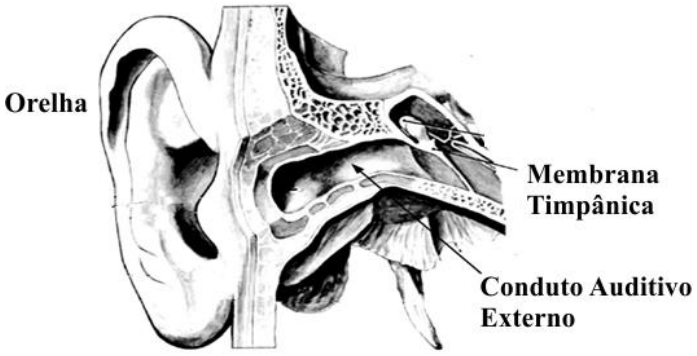


Figura 2.2: *Ilustração esquema do ouvido externo (COLEMAN, 2012).*

2.1.4. O OUVIDO MÉDIO

É constituído por três ossículos, que podem ser vistos na Figura 2.3: martelo, bigorna e estribo, além do tímpano, pois, essa membrana separa o ouvido externo do ouvido médio. Aqueles três pequenos ossos estão contidos dentro da cavidade timpânica e

esta, por sua vez, está conectada as fossas nasais pelo tubo de Eustáquio. Essa conexão permite que a pressão ambiente seja transmitida para o tímpano de maneira a não ocasionar o seu rompimento, em momentos em que não ocorram variações de pressão, sobre a membrana.

A função do ouvido médio é a de amplificar a onda sonora vinda do ouvido externo. Para isso, as vibrações resultantes das ondas sonoras vindas do ouvido externo, sofridas pela membrana timpânica são transmitidas para o martelo. Em seguida, essas vibrações são entregues à bigorna, que repassa para o estribo e que, por fim, as transfere à janela oval que está conectada ao ouvido interno através da cóclea. As amplificações ocorridas dentro do ouvido médio são necessárias, pois a janela oval consistente de um pequeno orifício dividindo o ouvido médio do ouvido interno, elas refletem parte da intensidade destas ondas sonoras, antes que estas cheguem ao ouvido interno.

O ganho do nível de pressão entre o ouvido externo e o médio é da ordem de 27 dB para frequências em torno de 1 kHz, porém esta intensidade diminui para frequências menores (BISTAFA, 2006). A unidade de medida Hz corresponde ao número de ciclos por segundo em um período da onda. A unidade Hz é uma abreviação para Hertz, que é uma homenagem ao físico alemão Heinrich Rudolf Hertz que efetuou significativas descobertas na área do eletromagnetismo (BARBOSA, 2009).

2.1.5.O OUVIDO INTERNO

O ouvido interno é composto por: labirinto ósseo (ilustrado em tonalidade cinza-escuro na Figura 2.4), que são cavidades e canais dentro do osso temporal e pelo labirinto membranáceo (ilustrado em tonalidade preta na Figura 2.4), constituído por vesículas comunicantes e dutos alojados no labirinto ósseo. [A ilustração do ouvido interno pode ser vista na Figura 2.4]. No labirinto membranáceo encontra-se a cóclea, órgão responsável por detectar e codificar as ondas sonoras que são enviadas ao cérebro. Os nervos vestibulares cocleares são nervos que conectam a cóclea ao cérebro, esses nervos são responsáveis por transmitirem os estímulos de equilíbrio do corpo ao cérebro.

É na cóclea que acontece o processo de transformação das ondas sonoras em estímulos elétricos. O processo consiste basicamente no envio das ondas sonoras vindas do ouvido médio através da perilinfa (líquido contido entre o labirinto ósseo e o labirinto membranáceo), como se fossem pressões hidráulicas. Essas pressões atravessam as escalas vestibulares e timpânicas fazendo vibrar a membrana basilar e por fim excita o Corti

(órgão responsável por transformar as vibrações mecânicas em impulsos elétricos) (BISTAFA, 2006).

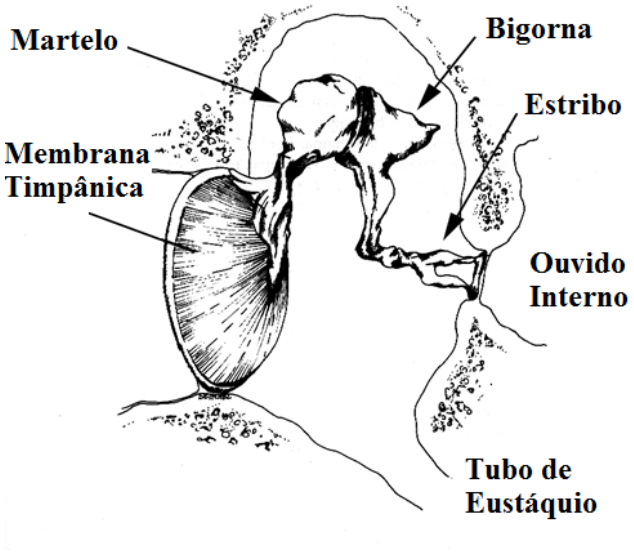


Figura2.3*: Ilustração do esquema do ouvido médio (COLEMAN, 2012).*

A membrana basilar possui uma faixa de excitação de 20 Hz a 20 kHz, que corresponde à faixa do audível do ouvido humano dentro do espectro sonoro. O limitante inferior causa excitações nos pontos localizados mais próximos ao ápice da cóclea, enquanto o limitante superior causa excitações nos pontos mais próximos da base da cóclea, para tons com apenas uma frequência. Para sons multifrequenciais, a resposta à excitação da cóclea é dada para cada frequência individualmente.

2.2. VOZ

A voz é um dos principais meios pelos quais os seres humanos realizam sua comunicação. A voz é formada pelas variações de pressão do ar que saem dos pulmões e chegam ao trato vocal dando origem a ondas sonoras. Isoladamente, sinais sonoros não transmitem qualquer informação, porém, quando produzidos em uma determinada sequência por pessoas de um mesmo grupo, produzem um conjunto de mensagens e consequentemente uma comunicação.

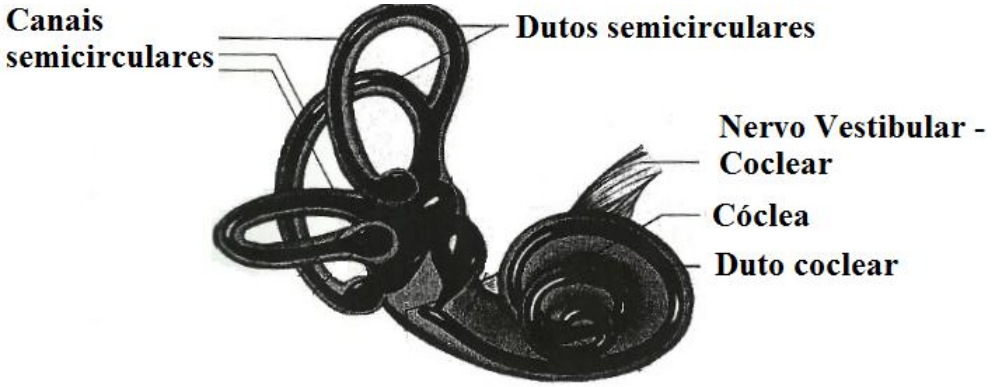


Figura 2.4: *Esquema do ouvido interno (BISTAFA, 2006).*

Para uma melhor compreensão do tema abordado nessa dissertação, faz-se necessária uma abordagem mais detalhada da voz, bem como de seu mecanismo de produção, suas classificações e seus respectivos modelos de produção.

2.2.1. PRODUÇÃO DA VOZ

O fonema é a menor unidade sonora que relaciona a grafia das letras de um alfabeto com o respectivo som produzido por esses elementos dentro de uma língua. Um fonema, quando pronunciado sozinho, geralmente não traz qualquer informação, porém, fonemas pronunciados em sequência dão origem a elementos de mensagens que servem de elo de comunicação entre o locutor e o ouvinte. Essas unidades sonoras são produzidas pelo ar expelido pelos pulmões, por meio do diafragma, que segue pelas cordas vocais, onde são produzidas as vibrações sonoras e finalmente saem pela boca (CHU, 2003).

O sistema vocal é composto por: diafragma, pulmões, traqueia, laringe, faringe, cordas vocais, fossas nasais e cavidade oral. A Figura 2.5 ilustra os órgãos que compõem o sistema vocal.

O ar, antes de produzir sinais sonoros, percorre um longo caminho. Esse percurso começa com a respiração, absorção do ar, que entra pelas fossais nasais, percorre a traqueia e chega aos pulmões. Em seguida, o ar retorna novamente pela traqueia, atravessando dessa vez, as cordas vocais, que estão localizadas na laringe. Dentro das cordas vocais, localiza-se a glote, órgão responsável pela produção do som. Na verdade, a glote é uma pequena estrutura muscular que abre e fecha rapidamente durante a saída de ar. Os sons produzidos durante esse processo de passagem pela glote soam de forma quase constante e possuem uma frequência única para cada indivíduo. Por fim, o ar passa pela laringe e é

emitido através pela boca e cavidade nasal (CHU, 2003). Os espaços das cavidades ósseas, fossas nasais, boca, traqueia, garganta, e laringe geram as ressonâncias, ou seja, o timbre de cada pessoa. A definição de timbre foi apresentada na Seção 2.1.1.

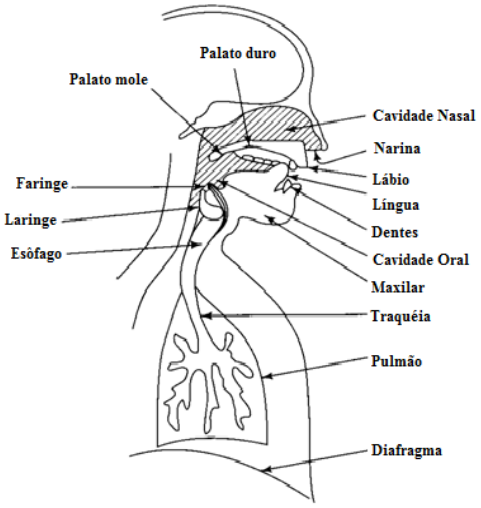


Figura 2.5: *Ilustração do esquema do sistema produtor de voz (BOUMAN, 2012).*

Como visto, as variações que ocorrem na glote modificam as vibrações do ar que a atravessam, porém, essa movimentação da glote só ocorre durante a saída de ar, enquanto, na entrada o ar percorre essa musculatura sem encontrar variações ou interrupções (HOLMES&HOLMES, 2001). Além disso, as formas como estão interligados e distribuídos os órgãos articuladores e as cavidades do trato vocal de cada indivíduo fazem com que as ondas ressonantes produzidas sejam únicas. Na próxima seção é realizada a classificação da voz.

2.2.2. SONS VOCÁLICOS E NÃO VOCÁLICOS

Vogais e consoantes são as letras que compõem um alfabeto de uma determinada língua (HOLMES&HOLMES, 2001). Os sons relacionados à fala podem ser classificados como: vocálicos e não vocálicos.

Os sons vocálicos são obtidos quando ar que vem dos pulmões passa pela glote e não sofre interrupção (parcial ou total) por parte da língua, lábios, dentes, etc., provocando vibrações quase periódicas. Os sons vocálicos são os sons relacionados às vogais de um alfabeto. Já os sons não vocálicos são obtidos pela interrupção parcial ou total do ar, durante o percurso dos pulmões até sua saída pelas cavidades orais e nasais. Esse tipo de

som é produzido quando se pronunciam as consoantes (HOLMES&HOLMES, 2001). As consoantes, dependendo de suas características sofrem bloqueios, os quais são responsáveis pela "letra" a ser pronunciada. Dentre os vários tipos de consoantes existentes pode-se citar (CIPRO&INFANTE, 2009):

I. As Fricativas – são obtidas pelo bloqueio parcial do ar que sofre algum tipo de fricção durante sua saída. Entre elas: f, v, s, z, e j.
II. As Nasais – são sons obtidos pela ressonância das ondas sonoras nas cavidades nasais e orais. São elas: "m" e "n".
III. As oclusivas – são aquelas em que a passagem do ar é bloqueada momentaneamente pela boca e, no momento de ser pronunciado, o som sofre uma explosão, devido a corrente de ar que ficou aprisionada. Entre elas: p, b, t e d.

3.TÉCNICAS DE PROCESSAMENTO DO SOM

Este capítulo descreve técnicas de processamento do som, especialmente, a amostragem e a quantização. Em seguida é descrito modelo de produção de fala mais utilizado e por fim apresenta algumas das técnicas frequentemente utilizadas na detecção de *pitch*.

3.1. DIGITALIZAÇÃO DO SINAL DE VOZ

Atualmente, grande parte dos sistemas de comunicação é digital. Por isso, antes de transmitir, armazenar ou efetuar qualquer modificação em um sinal contínuo, torna-se necessário digitalizá-lo. Digitalizar um sinal consiste primeiramente em amostrar o sinal e em seguida quantizá-lo. A amostragem uniforme é o processo em que amostras são coletadas, de maneira uniformemente espaçada, ao longo de um sinal no domínio do tempo (OPPENHEIM&SCHAFER, 2010). Já a quantização é a conversão dos valores da amplitude do sinal analógico em valores discretos por um número finito de elementos (OPPENHEIM&SCHAFER, 2010). Esses elementos fazem parte das palavras pertencentes ao dicionário de um código, cuja palavra-código possui comprimento de n *bits* (abreviatura para "*binary digit* – digito binário", correspondente à menor unidade de armazenamento e transmissão de uma informação). Os processos descritos anteriormente são detalhados a seguir.

3.1.1. AMOSTRAGEM

O processo de amostragem consiste na tomada de amostras de um sinal contínuo, por exemplo, no tempo, formando um conjunto finito de amostras discretas.

Em diversas aplicações, os sinais analógicos são convertidos em sinais digitais para em seguida, serem processados e só então convertidos em sinais analógicos novamente, se assim houver necessidade (OPPENHEIM&SCHAFER, 2010). Todo esse processo torna-se possível, desde que as amostras coletadas satisfaçam o Teorema da amostragem (OPPENHEIM&WILLSKY, 2010), apresentado a seguir.

Teorema 3.1 [Capítulo 7 -(OPPENHEIM&WILLSKY, 2010)]- Considere $x(t)$ um sinal contínuo, banda limitada, ou seja, para $x(t) \leftrightarrow X(j\omega)$ tem-se que $|X(j\omega)| = 0 \; para \; \omega \geq W$, em que, W é a máxima frequência presente no sinal. O Teorema de Shannon-Nyquist

(Teorema da Amostragem) afirma que para um sinal ser recuperado sem perdas, a partir de suas amostras, é necessário que este seja amostrado a uma taxa constante duas vezes maior ou igual à máxima frequência W do sinal (OPPENHEIM&WILLSKY, 2010). Demonstração (KONDOZ, 2004):

Um sinal analógico $s_c(t)$ amostrado pode ser representado por:

$$s[n] := s_c(nT), \quad \text{com } -\infty < n < \infty, \tag{3.1}$$

em que $s[n]$ representa o sinal amostrado, n é o número da amostra e T, com unidades em segundos, é o período de amostragem (que corresponde ao intervalo de tempo entre duas amostras consecutivas). Por meio do período de amostragem é encontrada a frequência de amostragem (Hz), aqui definida como w_s. Com base no Teorema 3.1, tem-se que, se um sinal $s_c(t)$, banda limitada, possui Transformada de Fourier, banda limitada, dada por:

$$S_c(j\omega) := \int_{-\infty}^{+\infty} s_c(t) e^{-j\omega t}\, dt, \tag{3.2}$$

de modo que $S_c(j\omega) = 0$ para $|\omega| \geq 2\pi W$, então, o sinal original pode ser reconstruído sem perdas, provido que $T \leq \frac{1}{2W}$, em que $2W$ representa a taxa de *Nyquist*, logo $w_s \geq 2W$.

A Figura 3.1a ilustra a Transformada de Fourier do sinal $x(t)$. A Figura 3.1b mostra o espectro de Fourier do sinal amostrador, enquanto que a Figura 3.1c ilustra o sinal $X(j\omega)$ amostrado com frequência maior do que a taxa de *Nyquist* ($w_s - W > W$). Quando a taxa de amostragem for inferior à taxa de *Nyquist*, ou seja, quando o critério de *Nyquist* não é atendido ($w_s - W < W$), ocorre uma sobreposição do espectro do sinal (*aliasing*). Nesse cenário, o sinal pode ser recuperado, porém com perdas (OPPENHEIM&WILLSKY, 2010) é o que ilustra a Figura 3.1 d.

3.1.2. QUANTIZAÇÃO

A quantização é o processo que aproxima os valores amostrados de um sinal contínuo, que podem assumir quaisquer valores, para um conjunto finito de valores pré-determinados. No processo de quantização, as amostras coletadas são aproximadas para o valor mais próximo da respectiva amplitude. A quantização pode ser representada por:

$$\hat{x}[n] = Q(x[n]), \tag{3.3}$$

em que $x[n]$ é a representação da amostra do sinal de entrada, $Q(.)$ o quantizador e $\hat{x}[n]$ a amostra quantizada na Equação 3.3. Essas aproximações geram desvios em relação ao valor original dando origem aos chamados erros de quantização (KONDOZ, 2004).

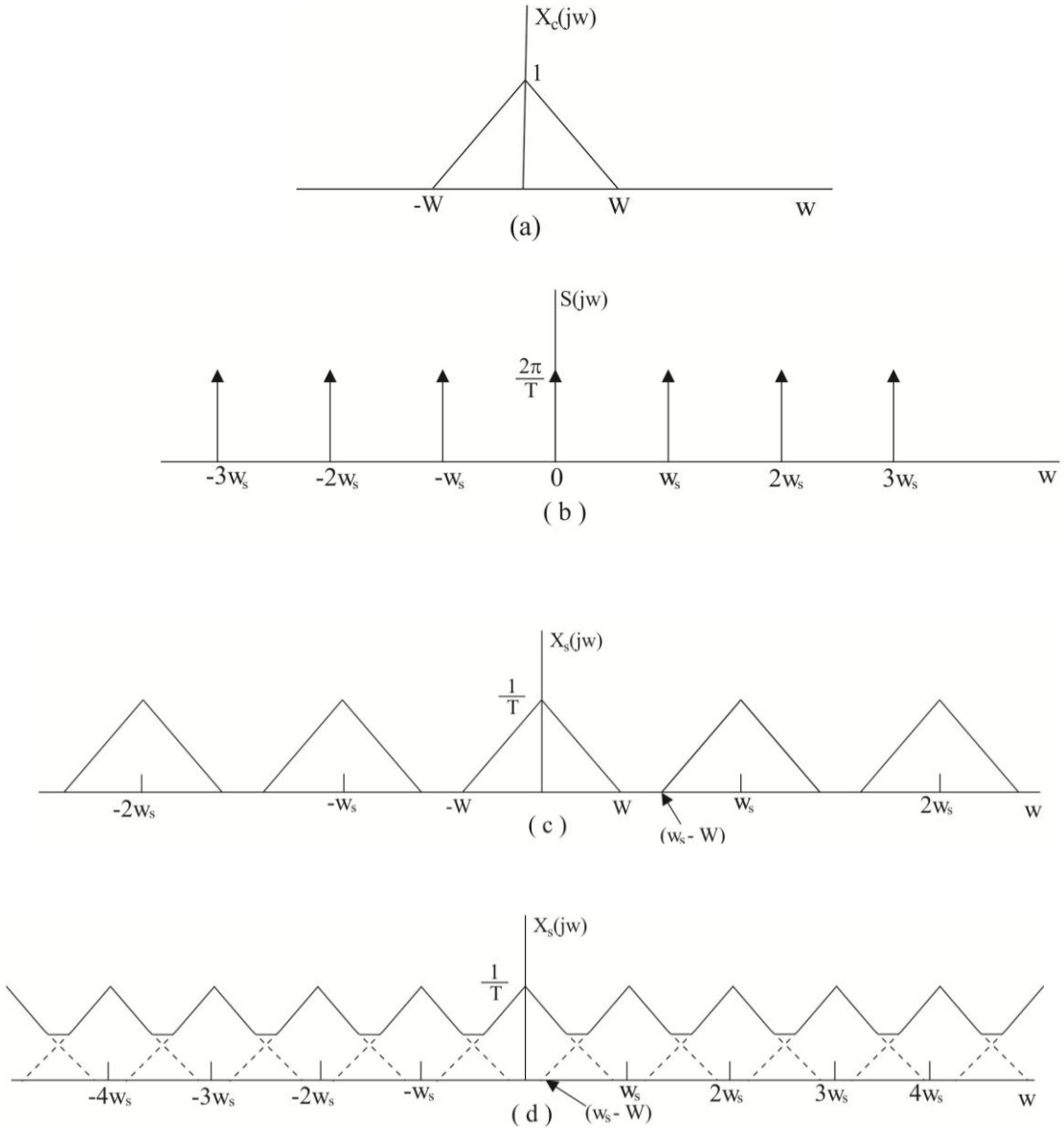


Figura 3.1: Efeitos da amostragem: (a) espectro de Fourier do sinal original; (b) espectro de Fourier do sinal amostrador; (c) amostragem a uma taxa superior à taxa de Nyquist; (d) amostragem a uma taxa inferior à taxa de Nyquist.

3.1.2.1.ERROS DE QUANTIZAÇÃO

A diferença entre os elementos da sequência original $x[n]$ e os elementos da sequência quantizada $\hat{x}[n]$ é conhecida como erro ou ruído de quantização. Mensurar o erro de quantização introduzido é fundamental para se avaliar criteriosamente a qualidade dos sinais quantizados. A qualidade do processo de quantização é verificada quando se

reduz os erros de quantização. As técnicas de projetos de dicionário para quantização têm como um de seus fundamentais objetivos a redução dos ruídos de quantização inseridos, buscando uma maior fidelidade na representação dos dados (RABINER&SCHAFER, 2007).

Uma métrica padrão para avaliar o erro de quantização introduzido é a Relação Sinal-Ruído (SNR - *Signal to Noise-Ratio*). Sejam $x[n]$ o sinal original, $\hat{x}[n]$o sinal processado, $e[n] = x[n] - \hat{x}[n]$ o erro de quantização da n-ésima componente da sequência. A medida de SNR , expressa em dB, de um quantizador é dada pela Equação 3.4:

$$SNR = 10log_{10}\frac{\sum_n x^2[n]}{\sum_n [x[n]-\hat{x}[n]]^2}, \qquad (3.4)$$

Existem diversas técnicas de quantização (KONDOZ, 2004), porém, as descritas nesta dissertação são:

I. Uniforme – é o método em que se divide a máxima amplitude do sinal em intervalos uniformemente espaçados e cada respectivo segmento obtido recebe uma representação única, por uma palavra-código com comprimento de $n = log_2^M\ bits$, em que M representa o número de níveis distintos (KONDOZ, 2004). Um exemplo de quantização uniforme é visto na Figura 3.2.

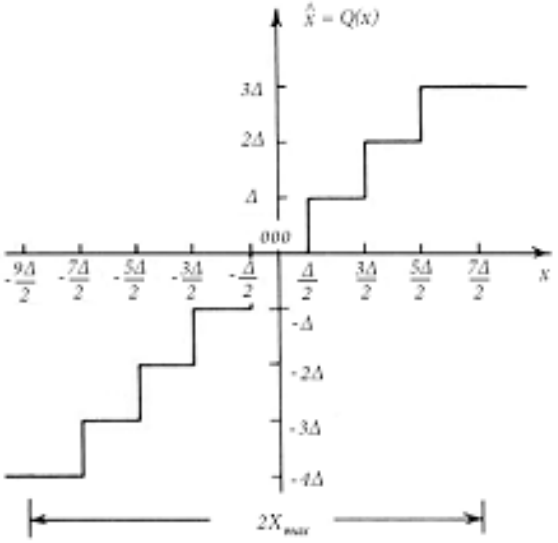


Figura 3.2: *Características de entrada-saída de um quantizador uniforme.* **(**SOTERO, 2009).

II. Não uniforme – é o método em que se divide a amplitude em segmentos de comprimentos diferentes, desta forma a quantização se adapta melhor a distribuição do parâmetro a ser quantizado. Dessa forma o passo do quantizador diminui para valores mais baixos do sinal de entrada e cresce para valores mais altos. A ideia é escolher o tamanho do passo de quantização com comprimento "dependente" da faixa de valores do sinal. O método mais usual é uma quantização logarítmica. A quantização não uniforme é realizada em duas etapas: a primeira realiza a distorção do sinal original, por meio de uma compressão logarítmica realizada pela Equação 3.5 (ilustrada na Figura 3.3); a outra etapa consiste em realizar a quantização uniforme do sinal distorcido (BARBOSA, 2009).

$$y[n] = h + k \times log_{10} x[n], \qquad (3.5)$$

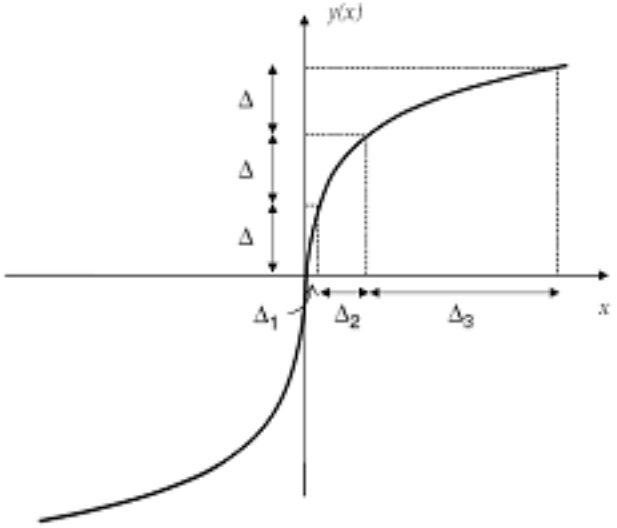


Figura 3.3: *Função de compressão de um quantizador não uniforme típico* **(**SOTERO, 2009).

Os parâmetros da curva de compressão são ajustados em padrões distintos, sendo os mais utilizados a Lei A e a Lei μ (KONDOZ, 2004).

3.2. MODELO DO SISTEMA DE PRODUÇÃO VOCAL

Devido à complexidade de muitos sistemas, torna-se necessário a elaboração de modelos ou padrões de estudos. Um dos modelos que representa o sistema de produção de voz é o modelo fonte-filtro, que pode ser visto na Figura 3.4. Esse modelo é composto por:

I. Um gerador de excitação: componente que simula o elemento produtor de harmônicos, no nosso sistema representado pela glote.

II. Um filtro linear variante no tempo: simulador dos formantes, que são os máximos do espectro de frequência ressonante, obtidos quando o ar ressoa através do trato vocal e suas respectivas cavidades (VASEGHI, 2007).

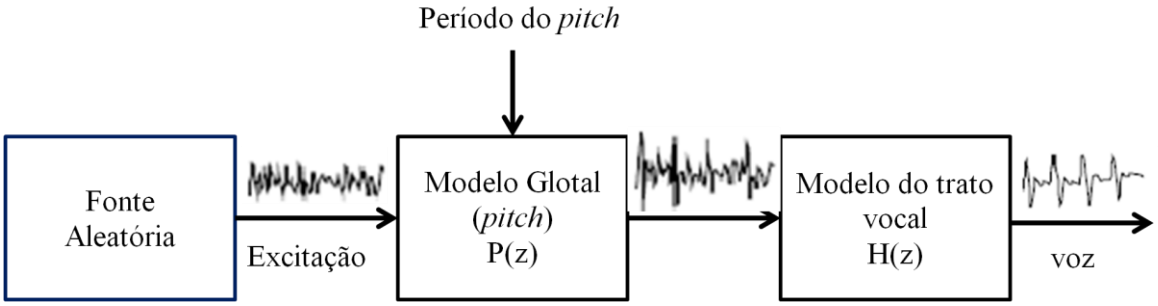


Figura 3.4: *Modelo de um gerador de voz (fonte-filtro) (VASEGHI, 2007).*

Torna-se comum representar o modelo do trato vocal para pequenos intervalos de tempo pela função de transferência:

$$H(z) := \frac{\sum_{k=o}^{M} b_k z^{-k}}{1-\sum_{k=1}^{N} a_k z^{-k}} = \frac{b_0 \prod_{k=1}^{M}(1-d_k z^{-1})}{\prod_{k=1}^{N}(1-c_k z^{-1})}, \qquad (3.6)$$

em que a_k e b_k são os coeficientes do filtro, responsáveis por modelar o trato vocal; c_k modela os formantes e d_k os sons nasais e fricativos (RABINER&SCHAFER, 2007). Esses coeficientes são os mesmos utilizados no modelo da Figura 3.4.

As alterações efetuadas nos coeficientes da Equação (3.6) ocasionam modificações no modelo do filtro. Em algumas aplicações, para tornar mais simples as estimativas dos parâmetros vocais, assume-se apenas a função de transferência com seus respectivos pólos.

Como o sinal de voz não sofre variações rápidas, devido às características do trato vocal, tem-se que, para pequenos intervalos de tempo, entre 10 a 40 *ms*, pequenos segmentos de onda sonora podem ser considerados estacionários (OPPENHEIM&SCHAFER, 2010). Um sinal é dito estacionário, quando suas características estatísticas não variam de forma considerável com o tempo (LATHI, 1998).

O método por codificação preditiva linear, LPC *(Linear Predictive Coding*) consiste basicamente no fato de que amostras futuras são determinadas pela combinação linear de amostras passadas (RABINER&JUANG, 1993). O método por predição linear

estima a frequência fundamental (*pitch*), os formantes, o espectro, a função área do trato vocal e representa o sinal de voz (SOTERO, 2009).

Os sistemas discretos também podem ser representados por equações de diferenças (OPPENHEIM&WILLSKY, 2010). Para o método LPC, a equação de diferença é dada por:

$$s(n) := \sum_{k=1}^{p} a_k \, s(n-k) + \, Gx(n), \qquad (3.7)$$

em que $s(n)$ é o sinal de voz amostrado e G é o ganho da excitação normalizada $x(n)$.

A maior dificuldade existente na elaboração da Equação (3.7) está em determinar os coeficientes a_k, de modo a fornecer uma boa estimativa para os parâmetros do sinal de voz. Além disso, em razão do caráter estacionário das ondas sonoras, os coeficientes deste modelo devem ser calculados para pequenos intervalos de tempo. Deve-se ter em mente, também, que a escolha destes coeficientes deve minimizar o erro médio quadrático no respectivo intervalo de tempo. Os parâmetros obtidos na Equação (3.7) são os mesmos da função de transferência da Equação (3.6) para o modelo de produção de voz.

3.3. ALGORITMOS DE DETECÇÃO DE *PITCH*

Diversos métodos têm sido utilizados para os algoritmos de detecção de *pitch* (CHENG *et* al., 1976; RABINER *et* al., 1976a, 1976b; KADAME&BARTELS, 1991; MARTIN, 1982; OH&HU, 1984). Na maioria dos casos, o desempenho destes algoritmos é validado empregando um banco de dados contendo gravações de voz masculina e feminina. Algumas técnicas propõem a validação dos experimentos com sinais de voz obtidos por gravações previamente armazenadas, no entanto, existem outras que se valem de sinais de voz em tempo real. Como visto, algumas dessas técnicas propõem uma análise parcial ou totalmente automática. Dentre estes métodos, pode-se citar:

a) Método de autocorrelação (SAMAD *et* al., 2000, BRANDÃO *et* al., 2007);
b) Método cepstral (AHMADI, 1999);
c) Método ST - curto prazo (BOERSMA, 1993, CHARPENTIER, 1986);
d) Método com base em *wavelets* (KADAME&BARTELS, 1992, JANER *et* al., 1996);
e) Método baseado em Classificadores de redes neuronais (BARNARD, 1991, HARBECK, 1995);
f) Método AMDF (YING, 1996, LI *et* al., 2006).

O primeiro método propõe a detecção do pico de amplitude em função da estimativa do período de *pitch,* e usualmente este método conduz a menores erros. No segundo método, a informação da frequência de *pitch* é extraída pela modificação do método baseado em *cepstrum* (RABINER, 1978) e então cuidadosamente refinada, usando um rastreamento de *pitch,* seguido da correção e aplicação de algoritmos de suavização.

No terceiro método apresenta-se uma técnica para a detecção de periodicidade no domínio da "autocorrelação". Esse método é capaz de medir a relação harmônico-ruído nesse domínio, com uma precisão e confiabilidade ainda maior do que com métodos usuais de domínio da frequência.

No quarto método, computa-se a Transformada Discreta *Wavelet* (DWT) do sinal e compara-se o desempenho relativo à fase linear com a fase mínima, para estimação dos períodos de *pitch*. A DWT é então utilizada para detectar o fechamento da glote e estimar o período de *pitch,* medido em um período. O quinto método para detecção de *pitch* consiste em um classificador de rede neuronal que opera com características invariantes, baseadas nas propriedades da forma de onda dos picos. Verifica-se que o melhor rastreador de *pitch* de rede neuronal se aproxima do nível de aproximação feito por pessoas para o mesmo conjunto de dados, e tem um desempenho competitivo em relação aos sofisticados rastreadores baseados em características da fala.

No sexto método são introduzidos novos algoritmos de detecção de *pitch* baseados na função diferença de magnitude média de curto tempo (AMDF) e na função de autocorrelação de curto tempo (ACF). Os quadros dos sinais de voz nesses algoritmos são considerados em uma representação binária para reduzir o custo computacional, facilitando a implementação desses algoritmos em sistemas de processamento de sinais em tempo real. O método se propõe também a diminuir os efeitos de amplitude e de formantes (em certos momentos prejudiciais à estimação) do sinal de voz para detecção de *pitch*.

O método adotado para a estimativa de *pitch,* nesta dissertação considerando-se os sons vocálicos investigados foi o método de XUEJING (2002b). De fato, entre a cornucópia de métodos descritos, optou-se por selecionar este método tendo em vista a facilidade de implementação e a disponibilidade de um código fonte aberto (também em Matlab®) para a implementação do mesmo. Um programa Matlab® para a sua implementação foi disponibilizado gratuitamente na Internet na URL:
http://www.speakingx.com/blog/2008/01/02/pitch-determination.

4. ESTIMATIVA DO COMPORTAMENTO VOCÁLICO DE LOCUTORES

O quarto capítulo descreve um dos estudos desenvolvidos nesta dissertação: uma estimativa do comportamento vocálico de locutores. Inicialmente é descrito a forma de aquisição do sinal de voz e o tipo e o tamanho da "janela" utilizada no algoritmo. Em seguida, detalha-se o preenchimento com zeros da ultima janela do sinal, com e a respectiva normalização espectral realizada. Em sequência, tem-se a descrição da seção mais importante deste estudo, que consiste na a descrição da derivada binária e na determinação dos picos de frequência. O capítulo é finalizado com os resultados experimentais obtidos neste estudo.

4.1. AQUISIÇÃO DE VOZ E JANELAMENTO

Os dados foram coletados em uma sala fechada e com baixo nível de ruído (sem ruídos provenientes de outras fontes, tais como, ar-condicionado, aparelhos de multimídia, etc.). A coleta foi realizada com o auxílio de um computador Compaq® Presario CQ40 – 740BR e microfone acoplado a um fone de ouvido. Foram realizadas sete coletas vocálicas para cada um entre seis locutores diferentes, divididos igualmente entre os sexos masculino e feminino (VASEGNHI, 2007, RUMSEY&MCCORMICK, 2007). A idade dos colaboradores variou entre 22 e 35 anos. A taxa de amostragem de frequência foi de 44,1kHz, a mesma utilizada em sistemas de áudio padrão CD (RUMSEY&MCCORMICK, 2007), com quantização de 16 *bits*.

O programa utilizado para a captura da voz dos participantes foi o Audacity 1.3®. O processo consistiu em realizar coletas de sons das vogais, do alfabeto brasileiro, incluindo as variações regionais para as letras "e" e "o", por intervalos de tempo de aproximadamente 7 segundos, sem interrupções. Diferentemente das técnicas usuais para codificadores de voz LPC, em que se requer a estimativa de *pitch* em uma janela curta (tipicamente 30 *ms*), trecho quase-estacionário de voz, (CHU, 2003, OPPENHEIM&SCHAFER, 2003), o procedimento descrito aqui trabalha com uma repetição bastante longa dos sons vocálicos, de forma a melhor estimar as características do locutor. Modela-se *offline* o comportamento do trato vocal em regiões vocálicas, como

um processo de "aprendizado". Pela teoria clássica de Fourier (de OLIVEIRA, 2007), sons periódicos (tais como os sons vocálicos) apresentam essencialmente espectro discreto, caracterizado por picos (raias espectrais).

O pré-processamento também foi realizado pelo Audacity 1.3®, para eliminar os trechos de silêncio no início e no final de cada arquivo e para eliminar trechos ruidosos ocasionados, no início e fim, pelo *click* do mouse. A ilustração do arquivo de áudio pré-processado obtido com o auxílio do Audacity 1.3® com ausência de silêncio e sem trechos ruidosos no início e no fim, para a vogal "a", pode ser vista na Figura 4.1 As gravações geraram arquivos na extensão .wav.

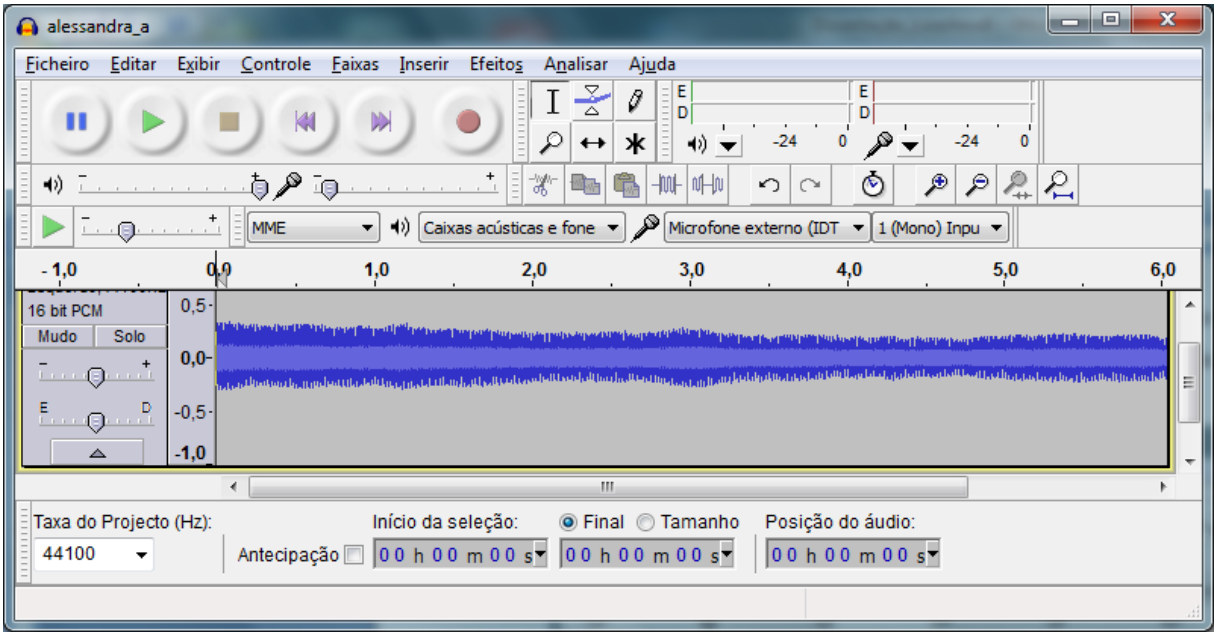


Figura 4.1: *Ilustração do sinal pré-processado correspondente a um sinal de voz da vogal "a" sendo repetida por 7 segundos pelo locutor Alessandra.*

Este mesmo procedimento foi conduzido para cada um dos locutores, para cada uma das vogais, incluindo acentuação (grave/agudo para e, o). A leitura do arquivo é realizada no início do algoritmo e os arquivos são amostrados a uma taxa de 44,1kHz e quantizados em 16 *bits*. O comprimento da janela utilizada no estimador foi estabelecido com base em observações realizadas no Audacity 1.3®. A janela tomada, inicialmente, foi de 128 amostras (comprimento da transformada com a taxa de amostragem fixada). A opção da janela retangular (RABINER&SCHAFER, 2007, OPPENHEIM&SCHAFER, 2010) se dar por esta ser mais simples, o que implica menor complexidade computacional.

A despeito da dessa janela conter descontinuidade e gerar perturbações no espectro, este efeito não é relevante tendo em vista que se trabalhou com uma taxa de amostragem muito elevada (bem acima da taxa de *Nyquist* do sinal de voz). Os espectros foram calculados expressando-os em escalas de frequência linear e logarítmica. A escala de frequência linear não apresenta uma visão tão ampla ao longo de seu eixo como a escala de frequência logarítmica, que foi, por esta razão, a escala adotada. A ilustração dessa janela pode ser vista na Figura 4.2.

Nota-se que, para o comprimento N = 128, não se constata o aparecimento de picos no espectro (como seria de se esperar, uma vez que o som registrado é quase periódico). Isto significa que esta resolução espectral ainda é insuficiente. Como forma de melhorar esta resolução, optou-se por aumentar o número de amostras por janela.

Como pode ser visto na Figura 4.3, para a janela de *N* amostras, a distância entre as amostras não sofre variação no domínio do tempo, pois a taxa de amostragem não se altera 44100 amostras/s. Entretanto, a distância entre as amostras no domínio da frequência depende do comprimento *N* adotado. A fim de se obter uma melhor visualização dos picos de frequências, variou-se então o comprimento *N* da janela.

Analisando ainda a Figura 4.3, observa-se que à medida que o tamanho da janela aumenta, isto é, mais picos de frequência são visualizados (melhora-se o nível de detalhamento no espectro). Para comprimento de janela maior que 2048 não se notou melhora visual, logo este foi o valor adotado.

Como pode ser visto na Figura 4.4, para a janela de 512 amostras, a distância entre as amostras não sofre variação em relação ao tempo, pois a taxa de amostragem foi mantida constante em 44100 amostras/s. Definindo, portanto a DFT, tem-se:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2k\pi}{N}}, \quad com \quad 0 \leq k \leq N-1 \qquad (4.1)$$

em que $x[n]$ são as amostras discretas do sinal no domínio do tempo, $X[k]$ são as raias espectrais definidas pela DFT do sinal $x[n]$ e N o número de amostras.

No entanto, quando essas mesmas amostras são convertidas para o domínio da frequência, avaliando o conteúdo harmônico nas frequências $k/(NT_S)$, com $k = 0, 1, \ldots, N-1$ e sendo N o comprimento da DFT (Transformada Discreta de Fourier), a distância entre as raias espectrais varia com relação ao número de amostras (de

OLIVEIRA, 2007). A Equação (4.2), vista a seguir, descreve a variação das raias espectrais com a distância:

$$\Delta f = 1/\,(NT_S), \qquad (4.2)$$

em que Δf é o valor da resolução das raias espectrais em Hz/ amostras e T_S é o período de amostragem.

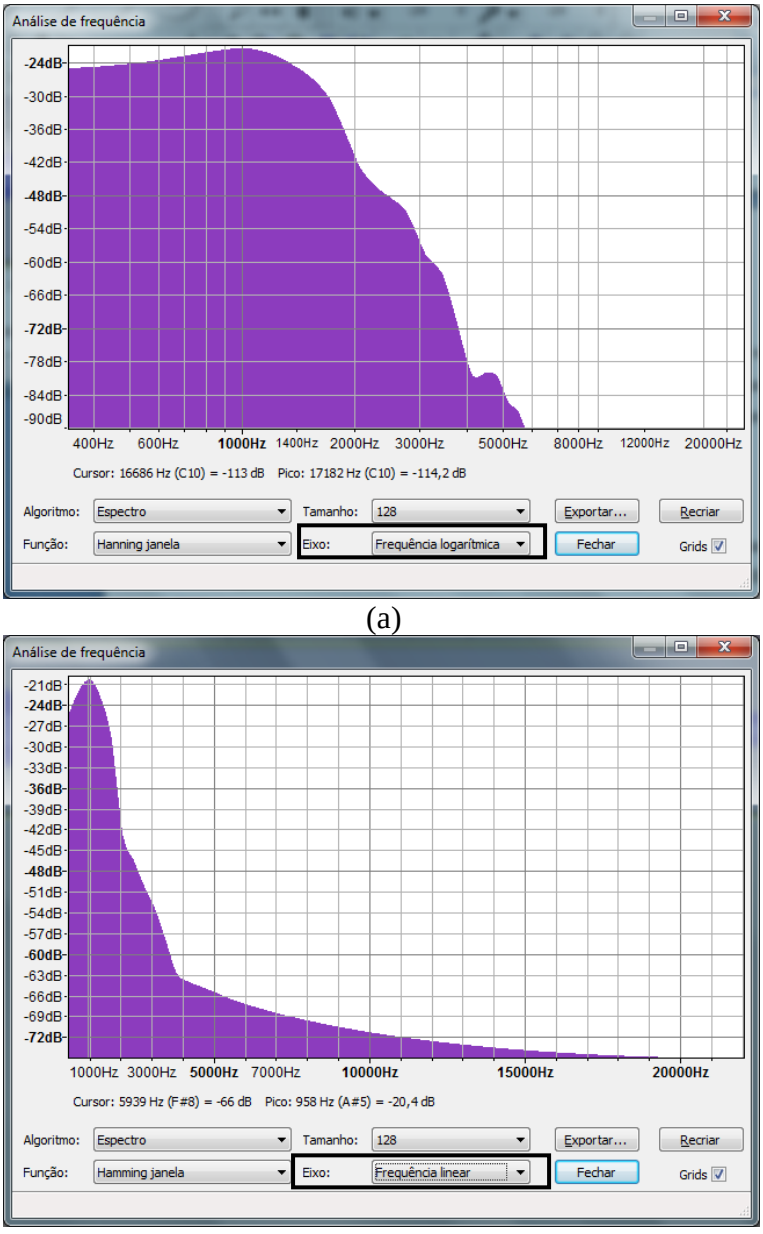
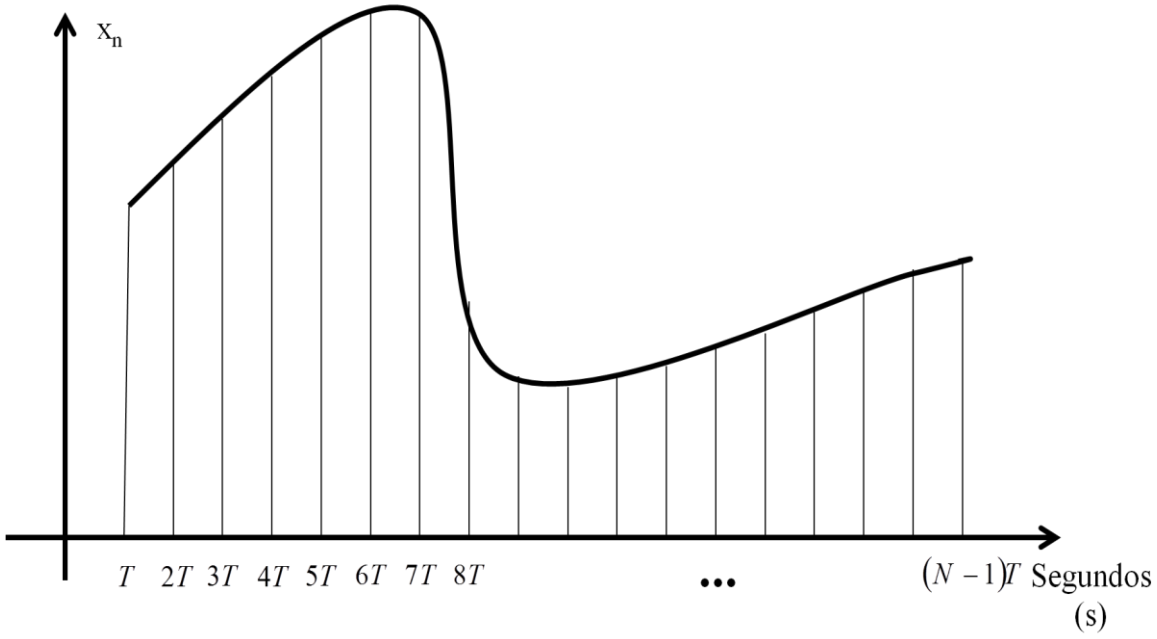
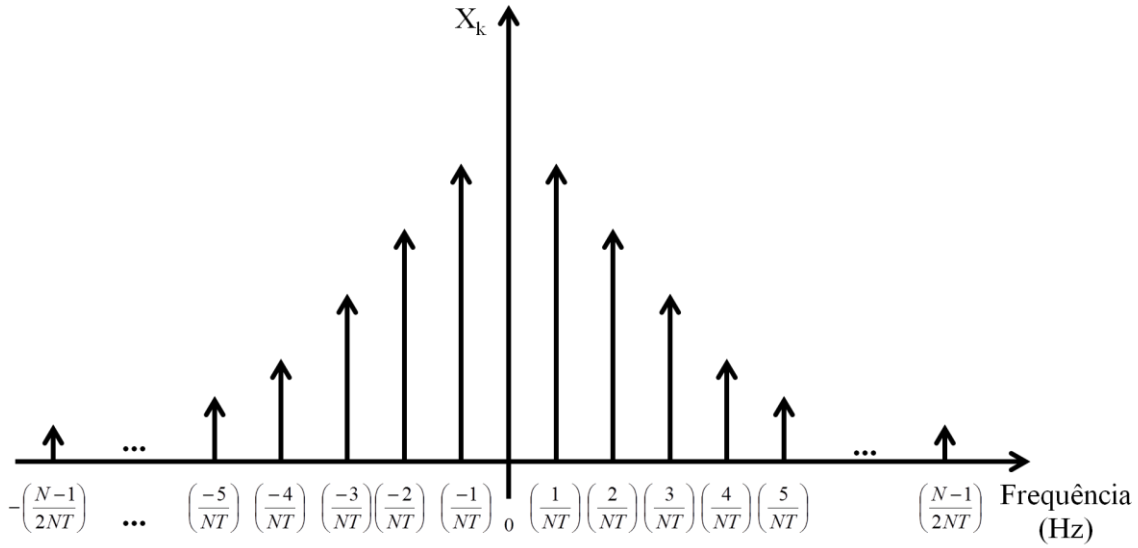


Figura 4.2: *Interface do Audacity 1.3® para análise de frequência com 128 amostras para a vogal a, sendo repetida por 7 segundos pelo locutor Alessandra.(a) Espectro em escala frequencial linear. (b) Espectro em escala frequencial logarítmica.*



(a)



(b)

Figura 4.3: *Ilustração para um sinal amostrado com N amostras. a) Amostras no domínio do tempo. b) Amostras no domínio da frequência.*

Por isso, a fim de se obter uma melhor visualização dos picos de frequência, variou-se o comprimento da janela utilizada para amostragem. As Figuras 4.4, 4.5 e 4.6 ilustram o espectro obtido para janelas com 512, 1024 e 2048 amostras, respectivamente.

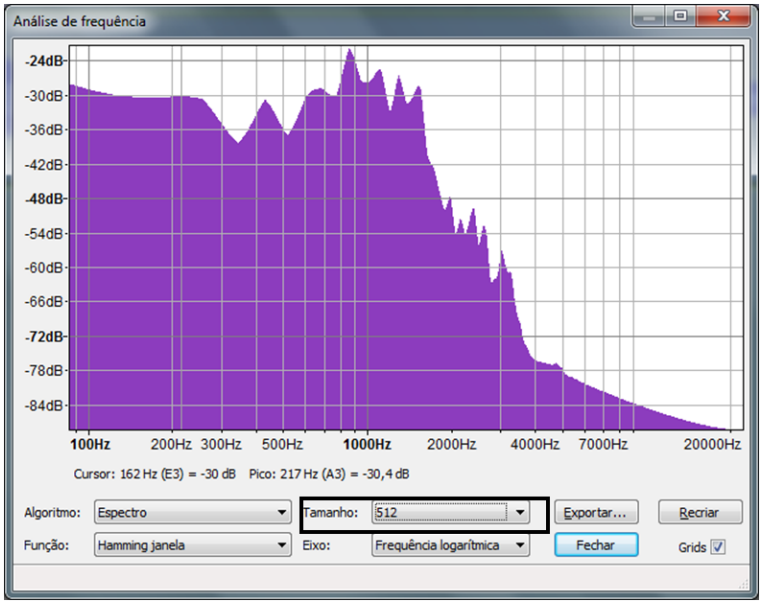


Figura 4.4: *Interface do Audacity 1.3® para análise de frequência com 512 amostras para a vogal "a", sendo repetida por 7 segundos pelo locutor Alessandra. O aumento do comprimento da DFT conduz a explicitar os picos relacionados com os sons harmônicos.*

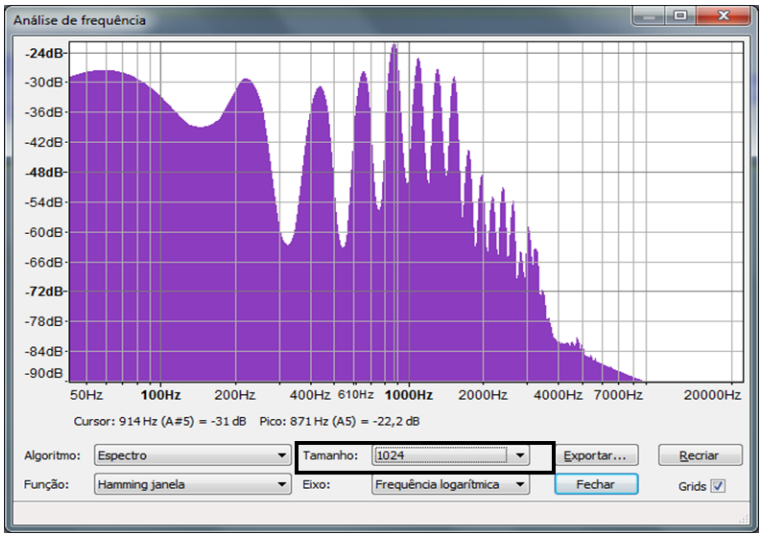


Figura 4.5: *Interface do Audacity 1.3 ® para análise de frequência com 1024 amostras para a vogal "a' sendo repetida por 7 segundos pelo locutor Alessandra".*

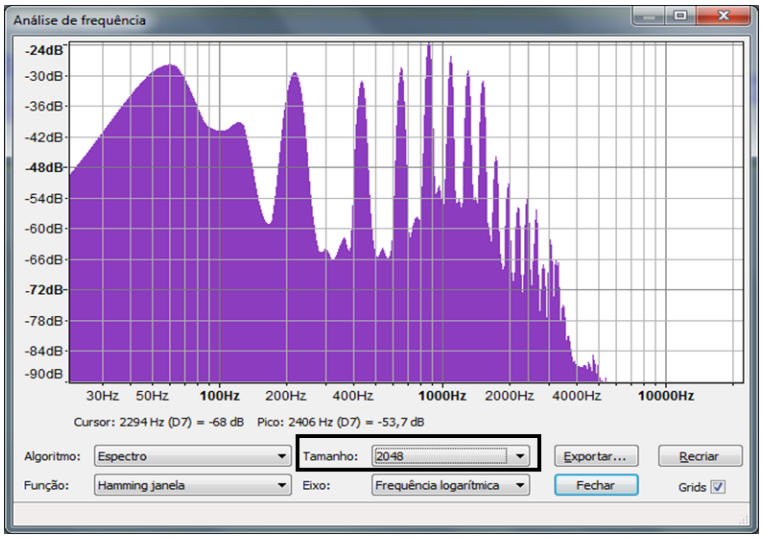


Figura 4.6: *Interface do Audacity 1.3® para análise de frequência com 2048 amostras para a vogal "a" sendo repetida por 7 segundos pelo locutor Alessandra.Verifica-se uma estabilização no formato espectral.*

Analisando as figuras, observa-se que, à medida que o comprimento da janela aumenta, o número de amostras também aumenta, e mais picos de frequência são visualizados (melhora-se o nível de detalhamento no espectro). Isto pode ser visto na Tabela 4.1, que ilustra a variação da resolução (DFT) de uma função de comprimento N. Para o comprimento de janela de 2048 amostras, a resolução é de 21,5 Hz, sendo, portanto satisfatória e por isso adotada neste trabalho.

A duração do trecho de áudio selecionado (cerca de 7 segundos) fornece cerca de 150 janelas e a média aritmética da magnitude do espectro normalizada em energia é tomada.

Tabela 4.1:: DFT com relação à respectiva resolução.

Comprimento da DFT	**Resolução (*Hz*)**
128	344
256	172
512	86
1024	43
2048	21,5

4.2. PREENCHIMENTO COM ZEROS E NORMALIZAÇÃO ESPECTRAL

Nesta etapa, torna-se a cardinalidade do "conjunto de amostras" um múltiplo de 2048. Para tornar o conjunto de amostras um múltiplo deste comprimento, foi realizado um pequeno ajuste, que consiste em dividir o número total de amostras por 2048 e usar o teto deste valor. Em seguida, esse número de janelas é utilizado no cálculo do comprimento do vetor nulo (preenchimento com zeros*)*, que realiza a complementação exata de elementos necessários no vetor de amostras totais, para que esse vetor se torne um múltiplo de 2048 (OPPENHEIM&SCHAFER, 2010).

Após o cálculo da DFT, uma normalização da janela final, obtida pelo somatório acumulativo é realizada. Esse cálculo é necessário para o estabelecimento de critérios de verificação estatísticos, utilizados na localização dos elementos de *pitch*/formantes em etapas posteriores. O cálculo da normalização consiste em dividir os elementos de amplitude espectrais da janela final, obtida depois do cálculo da DFT, pela soma dos quadrados de cada elemento da janela (energia desta janela). O cálculo da normalização é realizado com o auxílio do valor "$norm$", definido a seguir:

$$norm: = \sum_{i=1}^{L} \sum_{n=1}^{N} |fft(x(n + (i-1) * N))|^2, \qquad (4.3)$$

em que N = 2048 é o comprimento da janela , em amostras, em que a fft é calculada e L é o número de janelas contidas no trecho de voz analisado, na Equação (4.3). A normalização se faz dividindo os coeficientes da DFT pelo valor "$norm$", definido na Equação (4.3).

A razão da normalização advém do fato de que há níveis de áudio distintos de gravação e trechos com energia diferente. Após os valores nesta janela serem normalizados realiza-se uma busca, dentro do vetor, pelo máximo valor de amplitude. Esse valor é utilizado em etapas subsequentes.

4.3. A DERIVADA BINÁRIA E IDENTIFICAÇÃO DE PICOS

O identificador espectral proposto busca realizar a extração dos elementos do sinal de voz de maneira simplificada. Para isto, utiliza um critério de análise, no qual cada amostra obtida até a etapa de normalização deve ser comparada com a respectiva amostra

posterior adjacente. Este critério baseia-se na comparação entre o valor da amostra atual e o valor da amostra posterior. A estimativa de picos é, naturalmente, feita no domínio frequencial, obtendo-se a DFT $x_n \leftrightarrow X_k$ e examinando-se o comportamento de $|X_k|$, $k = 0, 1, \dots, {}^{N}/_{2}$. Se o valor da amostra atual for menor do que o valor da amostra posterior adjacente e se o valor absoluto dessa amostra for maior do que 0,1% do máximo valor obtido na etapa de normalização, é associado o valor 1 a essa amostra; caso contrário, se o valor da amostra analisada for maior do que o valor da amostra posterior e se o valor dessa amostra for superior a 0,1% do máximo valor obtido na etapa de normalização, é associado o valor -1 para essa amostra. Este procedimento é computacionalmente atrativo, pois as amostras espectrais são essencialmente "reduzidas" (i.e., mapeadas) em ±1. Esse procedimento é bem mais simples que adotar a magnitude das raias. Ao término desse processo, tem-se um vetor com o mesmo comprimento de 2048 amostras, porém, agora esse vetor só possui dois valores, -1 ou 1. O processo corresponde a uma indicação do sinal, (função sgn) da derivada do espectro:

+1 indica áudio em região crescente do espectro

-1 indica áudio em região decrescente do espectro

A existência de picos espectrais está associada a pontos de máximo, logo de derivada nula. Se há inversão no sinal da derivada, segue-se que o espectro sai de uma região crescente para decrescente. Finalmente, sobre o vetor obtido, com valores -1 ou +1, são extraídos os picos de frequência. O pico é localizado na região de transição de +1 para -1. Por exemplo, em uma sequência I_k como pode ser visto na Tabela 4.2, um pico é localizado na raia 3 (transição +1 → -1).

Tabela 4.2: Exemplo para o vetor derivada binária

I_k	+1	+1	+1	-1	-1	-1	+1	+1
K	1	2	3	4	5	6	7	8

A coleta é realizada apenas para os N^* = 150 primeiros elementos, pois os espectros típicos de voz encontram-se localizados neste intervalo (300 a 3300 Hz). Como visto na Tabela 4.2, para a janela (2048) a resolução é 21,5 Hz, ou seja, são necessárias apenas 150 amostras para cobrir a parte de maior conteúdo de energia do espectro de áudio tornando-se portanto desnecessário avaliar raias espectrais com valores de frequências mais elevadas.

O identificador proposto lida tão somente com $sgn(.)$ da grandeza $\Delta X_k := |X_{n+1}| - |X_n|,\ n = 0, 1, \ldots, N^* \ll {}^{N}/_{2}$, em que N = 2048. Para finalizar o processo, realiza-se a contagem dos picos de frequência. Essa contagem se dá da seguinte forma: se o elemento correspondente à amostra analisada tiver o valor numérico 1 e se a amostra subsequente possuir o valor -1, cria-se um novo vetor no qual a posição dessa amostra receberá o valor 1, que indica que nessa determinada posição existe um pico; caso contrário, o valor numérico atribuído é 0 e isso indicará a ausência de pico espectral nessa posição.

A seguir é apresentado o algoritmo para a derivada binária:

```
% Classifica os picos de frequência segundo o critério anterior
sinal = zeros(NN,1);
%
for i=1:NN-1
  if (X(i+1,1)> X(i,1)) && (abs(X(i,1)> tol))
    sinal(i,1) = 1;
  else if (X(i+1,1)<X(i,1)) && (abs(X(i,1)>tol))
      sinal(i,1) = -1;
    end
  end
end
```

A operação mais complexa de todo o procedimento é o cálculo da DFT a análise de picos relevantes do sinal vocálico se faz apenas através de contadores e comparações. Métodos alternativos utilizam estratégias bem mais sofisticadas, como por exemplo, métodos cepstrais, redes neuronais, wavelets, autocorrelação modificada, etc. Na maioria desses métodos a estimação do espectro com base na DFT também é realizada como etapa prévia. Em termos de implementação, tanto em hardware como em software, os cálculos de derivada binária são bem mais simples que as demais alternativas.

4.4. RESULTADOS EXPERIMENTAIS

Como visto, os experimentos foram realizados para 42 arquivos de áudio, pois cada locutor realizou sete locuções, relativas a cada vogal do alfabeto brasileiro mais as duas variações tônicas (ô e ê). Com o auxílio do programa Audacity 1.3® foram obtidas também, as raias espectrais relevantes para os mesmos 42 arquivos de áudio citados anteriormente. O objetivo dessa análise foi comparar as posições dos picos de frequência obtidos pelo identificador espectral e pelo Audacity 1.3®.

Observa-se que as amostras coletadas no Audacity 1.3® foram obtidas por método de coleta manual, ou seja, as posições de cada pico de frequência eram observadas na

janela da interface do aplicativo, contadas uma a uma. Já no identificador espectral, a contagem é realizada de maneira automática pelo próprio algoritmo.

A ilustração da interface do identificador espectral com a extração de parâmetros da vogal a pode ser vista na Figura 4.7. Os resultados obtidos pelo identificador de espectro vocálico e pelo Audacity 1.3® podem ser vistos na Tabela 4.2 para o experimento em que a vogal "a" é pronunciada por um locutor do sexo feminino. O código do estimador de *pitch* pode ser visto no ANEXO A.

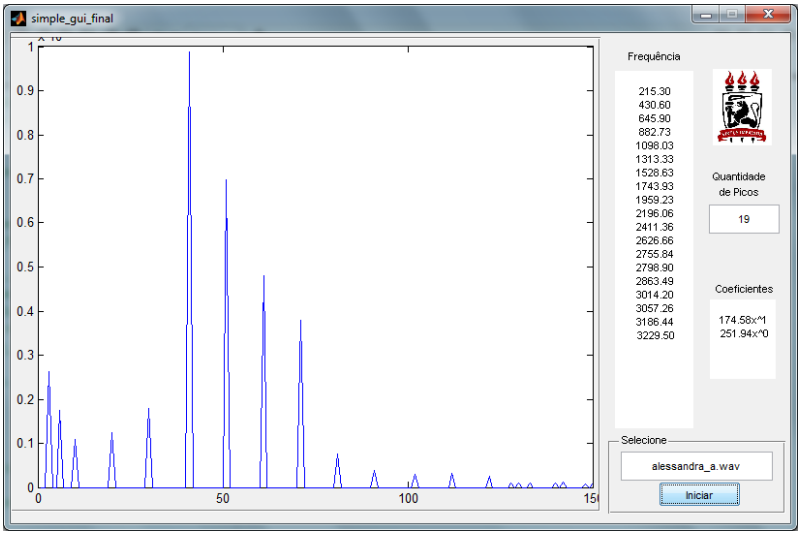


Figura 4.7: *Ilustração da Interface gráfica em Matlab®, do estimador de Pitch para a vogal "a" (Alessandra). Seleciona-se o arquivo extensão .wav (canto direito da tela) 19 picos são mostrados para esta vogal.*

A concordância dos resultados obtidos manualmente com o Audacity 1.3®, selecionando visualmente os picos, e com o programa Matlab® desenvolvido, é notável, com r^2 = 0,9997 e EMQ = 30 Hz, em que r^2 é o "fator de correlação" da curva de tendência linear com a curva obtida. Para verificar a aderência das raias espectrais identificadas pelo aplicativo e um espectro do sinal vocálico, realizou-se um ajuste linear para cada locutor, como pode ser visto no ANEXO C. Mostra-se (Figura 4.8), a título ilustrativo, o ajuste obtido para locutores selecionados.

O mesmo procedimento foi realizado para cada uma das "vogais" (a, é, ê, i, ó, ô, u) e para cada um dos locutores testados. A estimativa do passo foi estabelecida usando regressão linear (mínimos quadrados) entre a ordem do pico e sua frequência. Para o exemplo descrito (locutor Alessandra, vogal "a"), obteve-se a curva com "coeficiente" inicial 215,3 Hz (vide ANEXO B, Tabela I) e passo médio em harmônicos 174,6 Hz.

Como visto, na determinação de picos, os dois primeiros valores dos elementos de frequência podem ser eventualmente retirados da análise do algoritmo de identificador de raias espectrais.

Na coleta das amostras do Audacity 1.3® foram considerados todos os valores de frequência. A título comparativo, também se empregou um algoritmo de detecção de *pitch* clássico (XUEJING, 2002a) para avaliá-lo em cada fonema vocálico repetido, para cada um dos locutores como pode ser visto no ANEXO B, Tabela II. O código de detecção de *pitch* usado, com base na relação sub-harmônica-harmônica, encontra-se disponibilizado em (XUEJING, 2002b).

Há uma boa concordância (erro inferior a 5%) entre os valores estimados de *pitch* com o método de estimativa de raias vocálicas deste trabalho e as correspondentes estimativas de *pitch* usando um algoritmo construído especificamente para tal função (XUEJING, 2002b). Isto funciona como um indicador (uma validação parcial) que o método de estimação do comportamento vocálico proposto neste trabalho fornece dados coerentes com o esperado. Vale salientar que os resultados contêm mais informação que a extração de *pitch*, tais como, os demais picos existentes no sinal (como visto na Tabela 4.2). Por este motivo, as complexidades dos algoritmos não foram comparadas.

O vetor vocálico assim extraído pode ter múltiplas aplicações, como síntese de voz e reconhecimento de locutor. Outro "vetor de parâmetros" bastante útil na caracterização do locutor, além do vetor de *pitch* de cada som vocálico, pode ser construído. Este vetor *7 – D* contém a distância inter-raias de cada som vocálico (inclinação da regressão ou coeficiente angular):

$$\rho = \big((a), (é), (ê), (i), (ó), (ô), (u)\big), \qquad (4.4)$$

em que $(.)$ indica uso da frequência associada à vogal. Por exemplo, os locutores Ricardo e Lizandra têm perfis vocálicos expressos por:

$$\rho_{Ricardo} = \big((102{,}6), (89{,}0), (102{,}0), (82{,}0), (83{,}3), (94{,}5), (87{,}6)\big)$$

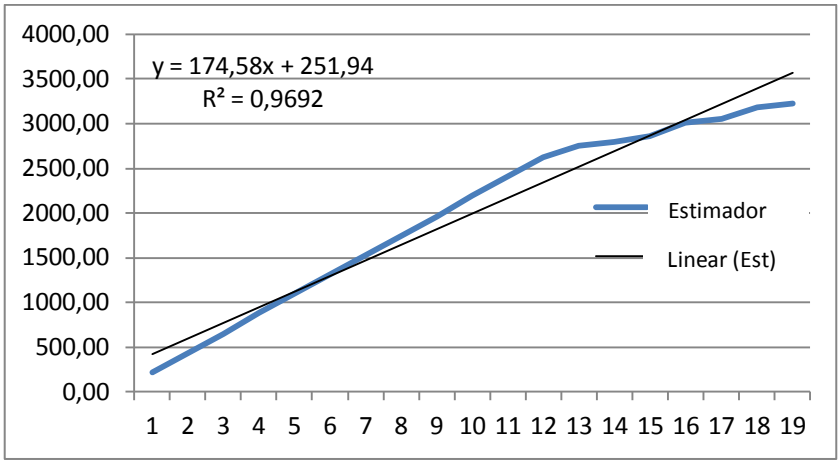
$$\rho_{Lizandra} = \big((210{,}3), (197{,}7), (328{,}0), (203{,}1), (224{,}9), (164{,}4), (103{,}4)\big)$$

Este modelo com base em *template* foi empregado com sucesso em um recente sistema de reconhecimento de locutor (SOTERO& de OLIVEIRA, 2009). Uma análise do comportamento espectral, com base no gênero, conduziu a uma constatação experimental curiosa. Ao se obter as curvas de ajuste de regressão para estimar a periodicidade das raias, observou-se (no espaço amostral investigado) que uma correlação maior é obtida para a

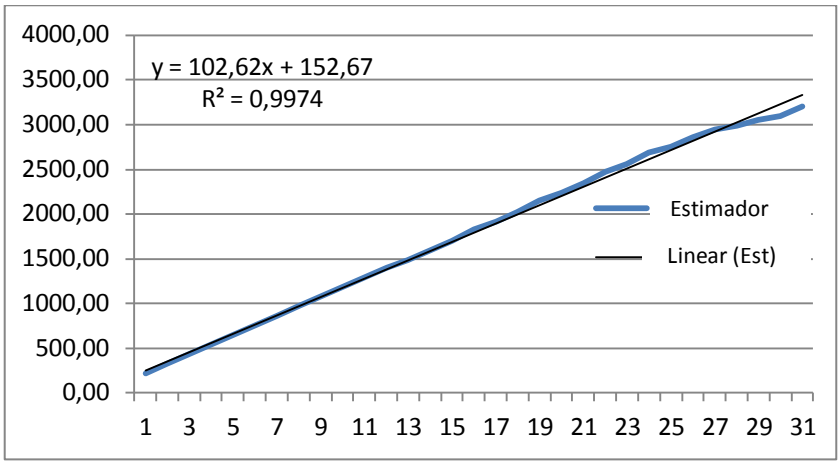
voz masculina. A aderência das curvas ao modelo linear foi sempre mais fraca para a voz feminina, sugerindo que o grau de periodicidade dos sons vocálicos provenientes de locutores femininos é menor. A despeito da ausência de explicação plausível, o fato é, no mínimo, um achado interessante que provoca pesquisar mais aprofundadamente tais efeitos.

Tabela 4.3: *Picos de frequências, em Hz, para os 19 picos significativos, obtidos pelo Audacity 1.3® e pelo Identificador para a vogal "a" pronunciada por um locutor do sexo feminino.*

Picos	Audacity 1.3® (*HZ*)	Identificador de Picos Espectrais (*Hz*)
1°	59,00	-
2°	120,00	-
3°	213,00	215,30
4°	431,00	430,60
5°	653,00	645,90
6°	872,00	882,73
7°	1088,00	1098,03
8°	1301,00	1313,33
9°	1518,00	1528,63
10°	1731,00	1743,93
11°	1970,00	1959,23
12°	2156,00	2196,06
13°	2371,00	-
14°	-	2411,36
15°	2544,00	-
16°	-	2626,66
17°	-	2755,84
18°	-	2798,90
19°	2855,00	2863,49
20°	3008,00	3014,20
21°	-	3057,26
22°	-	3186,44
23°	3280,00	3229,50



(a)



(b)

Figura 4.8: *Correlação entre picos identificados pelo aplicativo (ajuste linear) para: a) locutor Alessandra, pronunciando longamente o som vocálico "a". b) locutor Ricardo, pronunciando longamente o som vocálico "a". Equação de regressão e coeficiente de determinação indicados. Ajustes de regressão linear com n pontos, 9 <n <37, dependendo do som.*

5. IMPLEMENTAÇÃO DE UM ALGORITMO DE DIVISÃO SILÁBICA AUTOMÁTICA EM ARQUIVOS DE VOZ NA LÍNGUA PORTUGUESA

O quinto capítulo descreve o segundo estudo desenvolvido nesta dissertação: implementação de um algoritmo de divisão silábica automática para arquivos de voz na língua portuguesa. Inicialmente é descrita a forma de aquisição e o pré-processamento realizado no sinal. Em seguida, descreve-se o processo de retificação da onda, a obtenção do valor *RMS* e da descarga linear do detector de envoltória. Em sequência, tem-se a descrição do localizador silábico. Na seção seguinte é obtida a identificação de supersílabas e novos conceitos. Logo após, realiza-se a descrição de uma rotina para aglutinação de sílabas separadas indevidamente. E finalmente o capítulo se encerra com uma seção dos resultados obtidos pelo divisor silábico.

5.1. AQUISIÇÃO E PRÉ–PROCESSAMENTO DE VOZ

Os dados foram coletados em uma sala fechada e com baixo nível de ruído (sem ruídos audíveis provenientes de outras fontes, tais como, ar-condicionado, aparelhos de multimídia, etc.). A coleta foi realizada com o auxilio de um computador HP® Intel Atom sem uso de fone de ouvido. Foram realizadas 50 coletas distintas por indivíduos do sexo masculino e feminino. O programa utilizado para a captura da voz do colaborador foi o Audacity 1.3®. As gravações geraram arquivos em extensão *.wav*. A Figura 5.1 ilustra a interface do programa Matlab® para o arquivo de áudio contendo a palavra "departamento".

O pré-processamento também foi realizado com o auxilio do Audacity 1.3®. Esta etapa consiste na eliminação dos trechos de silêncio no início e no final de cada arquivo gravado. Como os arquivos gravados podiam estar em formato estéreo, tornou-se necessário uma conversão para o formato mono. A leitura do separador silábico é realizada logo no início do algoritmo a uma taxa de amostragem de 44,1 kHz e com quantização de 16 *bits*. O comprimento da janela utilizada no algoritmo foi de 2048 amostras. Este valor foi selecionado examinando-se o comportamento vocálico de locutores de língua portuguesa (da SILVA&de OLIVEIRA, 2012).

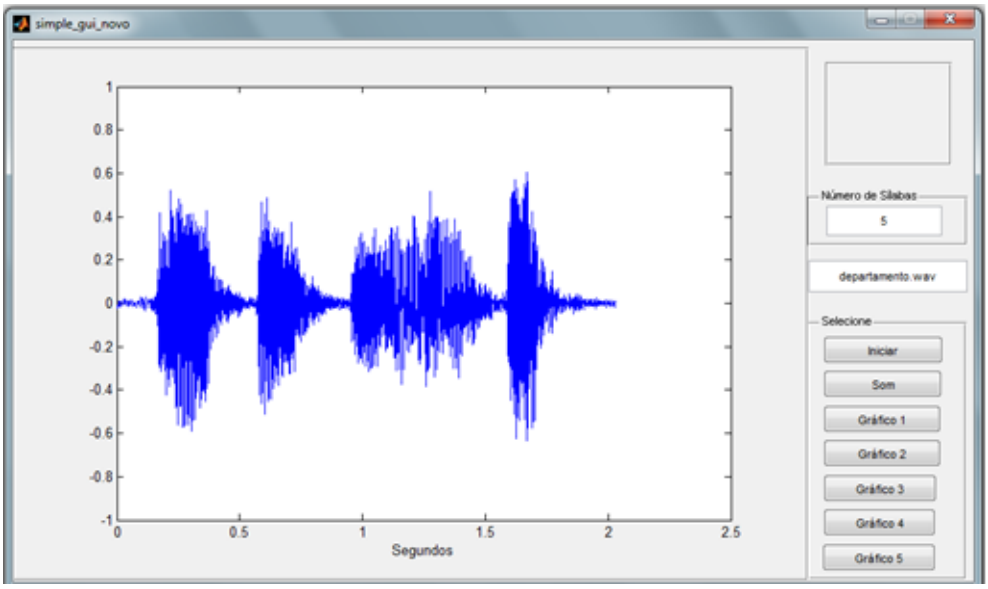


Figura 5.1: *Ilustração da interface do programa Matlab® para a palavra "departamento" após o pré-processamento.*

5.2. RETIFICAÇÃO DA ONDA, VALOR *RMS* E DESCARGA LINEAR

Nesta etapa ocorre, primeiramente, a retificação de meia onda do sinal de entrada. A retificação do sinal de áudio é realizada com o auxilio da função $signal$,

$$sign[x] := \begin{cases} -1, e\ x < 0 \\ 0, se\ x = 0, \\ 1, se\ x > 0 \end{cases}, \qquad (5.1)$$

em que cada valor de amostra $x[n]$ é convertido para $x[n]/2 \times \big(1 + sign(x[n])\big)$. A Figura 5.2 ilustra a interface gráfica do programa Matlab® do sinal de onda retificado, para a palavra "batata".

Após o cálculo da retificação de meia onda, efetua-se o cálculo do valor *RMS* (*Root Mean Square)* ou valor médio quadrático. O valor *RMS*, é utilizado posteriormente no processo de localização silábica, que juntamente com um percentual estatístico, servem como critério de avaliação para a determinação da localização das sílabas na palavra. Em seguida, realiza-se a detecção da envoltória através de uma descarga linear do sinal. Para isso, utilizam-se os valores de amostras obtidos no processo de retificação da onda.

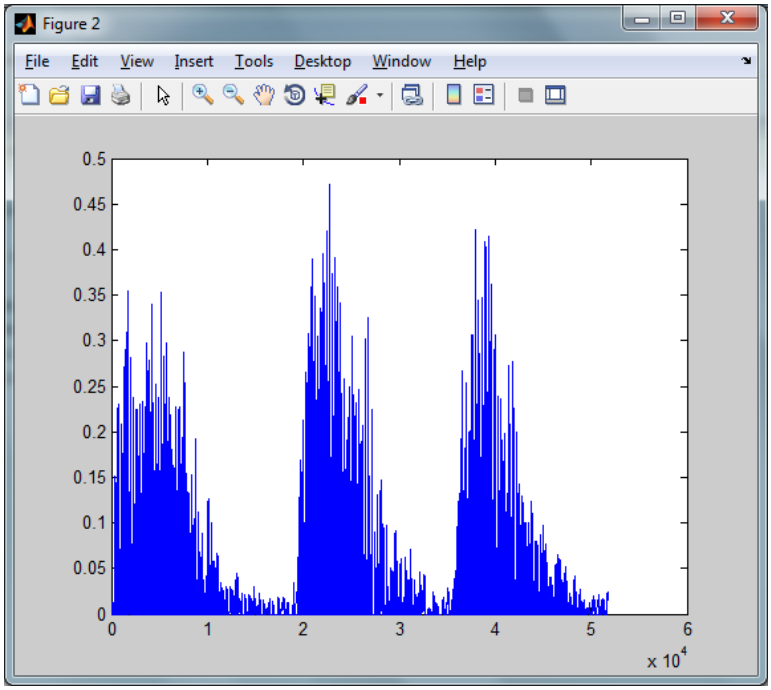


Figura 5.2: *Ilustração da Interface gráfica do Matlab® para o sinal de áudio da palavra "batata" após retificação de meia-onda.*

A identificação das silabas é baseada no sinal da envoltória do sinal de voz. A envoltória foi obtida inicialmente com a retificação (diodo), anulando-se todas as amostras negativas. O ajuste da taxa de descarga do detector linear de envoltória é realizado através do parâmetro Δ (contado em termos do número de amostras requeridas até a descarga total da envoltória) que representa, no caso de detecção de AM (de Oliveira, 2012), a constante de tempo do demodulador. Este valor foi ajustado empiricamente, após diversas tentativas.

Para valores muito grandes de Δ, por exemplo, Δ = 9702 amostras, obtém-se uma envoltória com perdas, como ilustrado na Figura 5.3. A redução para o valor de Δ = 970 amostras, resulta na forma de onda ilustrada na Figura 5.4 e descreve melhor a envoltória do sinal. A notar que esta escolha corresponde a um tempo de descarga (4RC) de 22ms. Isto coincide com os tempos típicos envolvidos em janelas de diferentes *vocoders* para sinais de voz (PROAKIS, 1989). Para uma taxa de descarga muito pequena, e.g., Δ = 97 amostras, como pode ser visto na Figura 5.5, a envoltória torna-se imperceptível.

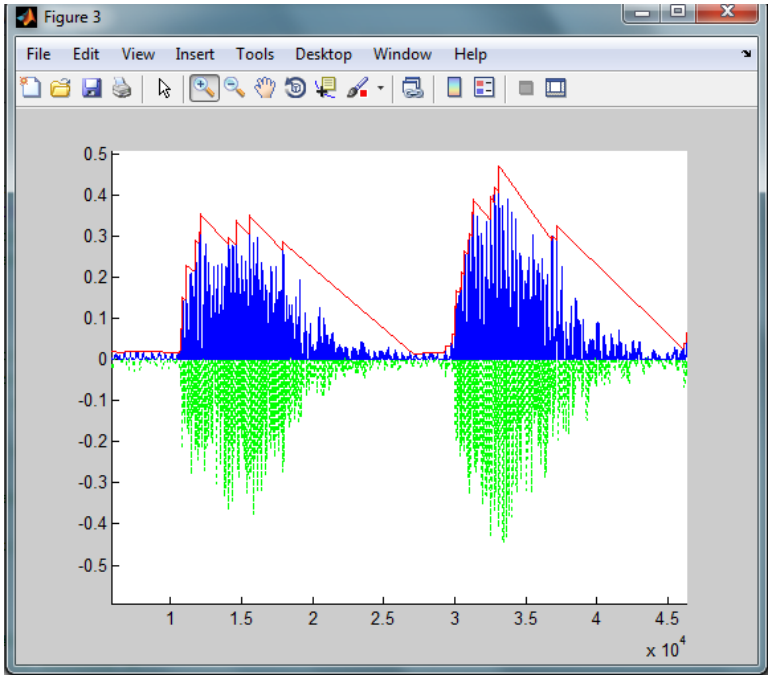


Figura 5.3: *Envoltória mal ajustada, por uso de taxa de descarga muito elevada (9702amostras, o que excede até a janela padrão de 2048 amostras). O sinal de voz corresponde a um trecho do sinal "batata".*

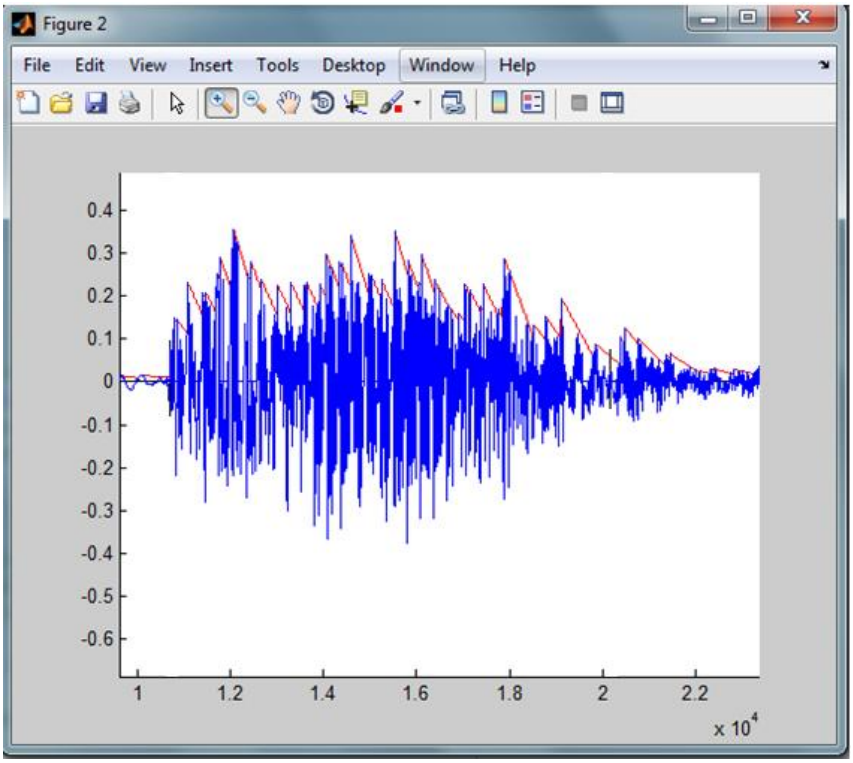


Figura 5.4: *Forma de onda da envoltória demodulada corretamente para a sílaba "BA" da palavra "batata". A descarga completa da envoltória pode ocorrer em até 22 ms.*

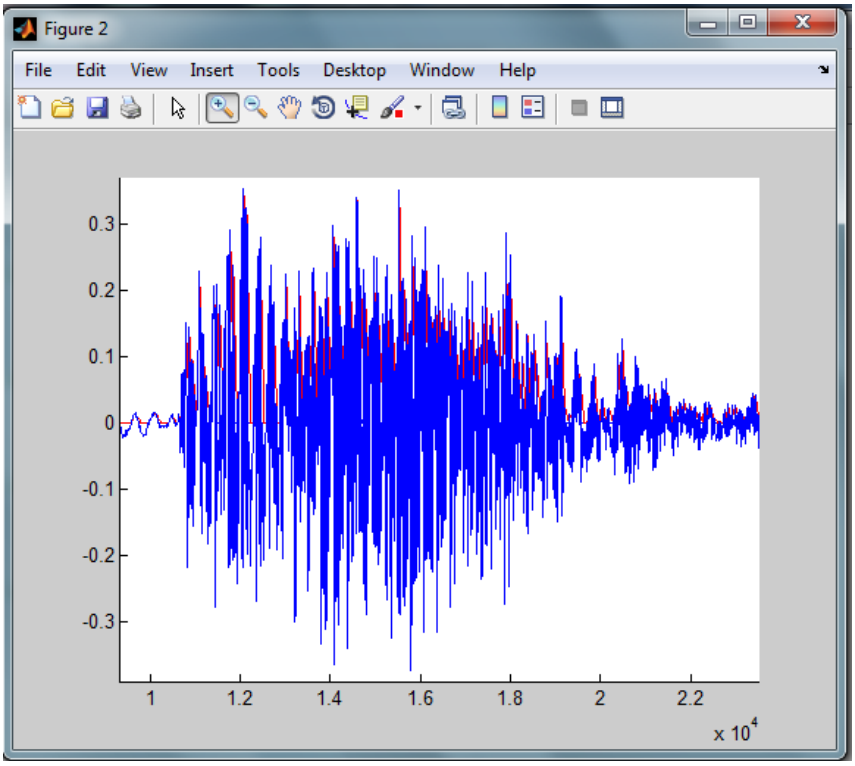


Figura 5.5: *Forma de onda da envoltória recuperada usando uma constante de tempo muito pequena, 2,2 ms, para a silaba "BA" da palavra "batata" ilustrada na Figura 5.2.*

O processo consiste em descrever o envelope da onda retificada através de um método preditivo linear. Nesse método, amostras futuras podem ser determinadas através da observação de amostras passadas (RABINER&JUANG, 1993). As amostras são analisadas dentro de uma subjanela de descarga linear (chamada de delta) dentro da janela maior de 2048 amostras. Esse delta é um valor obtido pelo produto da taxa de amostragem (44,1 kHz) pelo valor da janela de tempo utilizada em codificadores de voz, 22 ms (PROAKIS, 1989). O resultado desse produto fornece uma pequena janela com 970 amostras, que descreve a trajetória da descarga (aproximadamente linearmente) do pico, visando acompanhar a envoltória do sinal. O valor do delta corresponde ao ajuste da constante de tempo do detector AM (de OLIVEIRA, 2012). A ilustração da envoltória do sinal de onda retificado da Figura 5.2 pode ser vista na Figura 5.6.

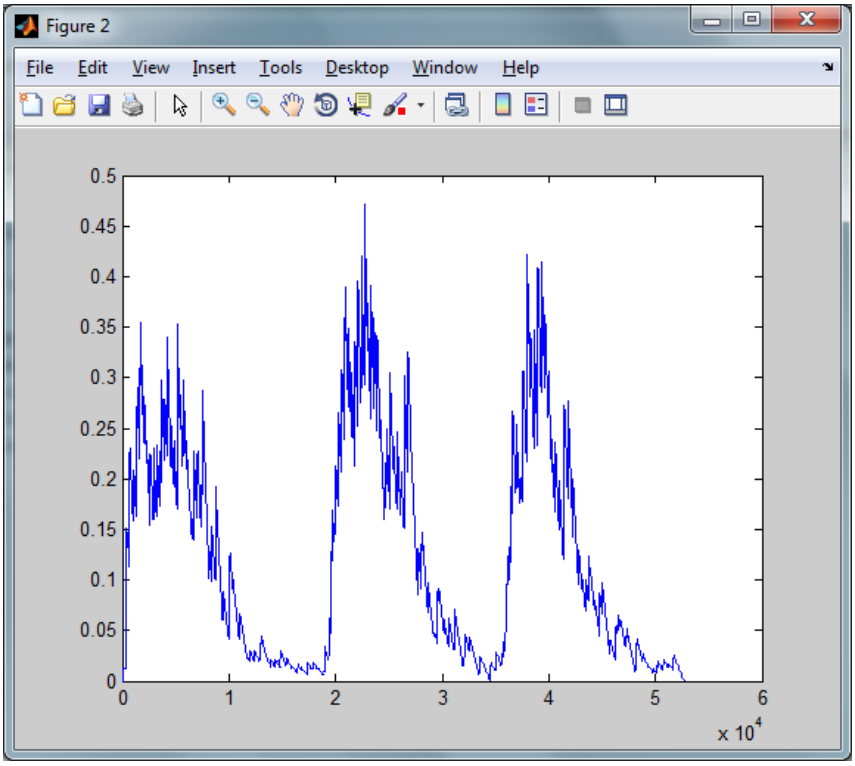


Figura 5.6: *Ilustração da Interface gráfica do Matlab® para a envoltória do sinal de áudio da palavra "batata"* referente a Figura 5.2.

5.3. LOCALIZADOR SILÁBICO

A localização das sílabas no sinal de áudio é realizada pela comparação do valor de cada amostra, obtida na etapa de detecção da envoltória, com um valor limite. Esse limiar foi obtido experimentalmente, após diversos ajustes em sucessivas análises. Ele foi expresso percentualmente (valor ajustado, Perc = 1,2) em termos do valor *RMS*, obtido na etapa de retificação de meia onda. O algoritmo consiste em atribuir o valor "1", em um novo vetor (denominado de "sílaba") na respectiva posição do valor da amostra, se o valor da amostra da envoltória for maior do que o valor limite, ou atribuir o valor "0" em caso contrário. Em seguida, realiza-se a contagem dos números de "zeros" e "uns" na sequência do vetor-sílaba.

A contagem das sequências consecutivas de amostras é realizada e, em seguida, armazenada na matriz denominada de matriz de contagem. Essa matriz armazena a quantidade de elementos encontrados na sequência (segunda coluna) e o respectivo valor associado a esta sequência (primeira coluna). A matriz contagem pode ser vista na Tabela 5.1, para a separação da palavra "batata".

A etapa seguinte consiste em eliminar as sequências muito curtas de amostras representadas pelo valor 1, com o intuito de que essas pequenas sequências não interfiram na localização das sílabas dentro de cada palavra. As amostras das sequências curtas são eliminadas por comparação com um percentual.

Esse percentual também foi obtido por ajustes *ad hoc* e corresponde a 1,8 do comprimento da janela (970 amostras), utilizada para descrever a trajetória da envoltória. Uma representação mais compacta pode ser utilizada, simplificando os dados para uma notação 0_n1_m em que os índices *n* e *m* denotam os comprimentos das sequencias de "zeros" e "uns", respectivamente. A matriz contagem exibida da Tabela 5.1 pode ser representada por uma sequência (*run length*) alternada de 0's e 1's:

$0_{10702}1_{9472}0_{204}1_{788}0_{185}1_{40}0_{55}1_{92}0_{8238}1_{29}0_{192}1_{9432}0_{523}1_{484}0_{128}1_{10}0_{423}1_{46}0_{512}1_{156}0_{4633}1_{8631}0_{189}1_{611}$
$0_{1137}1_{79}0_{68}1_{24}0_{19717}$

Depois da eliminação das sequências curtas de elementos "1" (falsa localização de sílaba), ela é transformada em:

$0_{10702}1_{9472}0_{204}0_{788}0_{185}0_{40}0_{55}0_{92}0_{8238}0_{29}0_{192}1_{9432}0_{523}0_{484}0_{128}0_{10}0_{423}0_{46}0_{512}0_{156}0_{4633}1_{8631}0_{189}0_{611}$
$0_{1137}0_{79}0_{68}0_{24}0_{19717}$

ou seja, $0_{10702}1_{9472}0_{9823}1_{9432}0_{6915}1_{8631}0_{21825}$

Ao final dessa etapa, realiza-se o somatório da quantidade de posições que possuem apenas valores unitários, a fim de se obter o número total de sílabas. Para o exemplo visto acima, na sequência *run length* para a palavra "batata", tem-se a quantidade de sílabas (3 sílabas), o início e o fim de cada sílaba.

Como visto, os elementos da sequência de contagem determinam que amostras contenham as sílabas da palavra analisada. Esses mesmos valores unitários também são responsáveis por determinar o início de cada sílaba, contida na palavra.

Para determinar a localização de cada sílaba, realiza-se uma rotina em que primeiramente calcula-se a soma acumulativa dos elementos da sequência de contagem, em um novo vetor (denominado de vetor posições) e, em seguida, verifica-se a posição em

que ocorre a mudança desse valor. Para a palavra “batata”, tem-se o vetor posições com 76800 amostras (i.e., duração de ~ 1,75 segundos), representado aqui da seguinte maneira:

$$0_{10702}1_{9472}1_{29997}2_{39429}2_{46344}3_{54975}3_{76800}.$$

Tabela 5.1: Matriz de Contagem Obtida Via Matlab® para o Arquivo de Trecho de Áudio Pré-processado Referente à Palavra “batata”.

Valores na Sequência	Quantidade de elementos	Valores na Sequência	Quantidade de elementos
0	10702	1	10
1	9472	0	423
0	204	1	46
1	788	0	512
0	185	1	156
1	40	0	4633
0	55	1	8631
1	92	0	189
0	8238	1	611
1	29	0	1137
0	192	1	79
1	9432	0	68
0	523	1	24
1	484	0	19717
0	128	-	-

5.4. IDENTIFICAÇÃO DE SUPERSÍLABAS E QUEBRA

Em geral, a média de amostras contidas em uma sílaba é de 6543 amostras, com desvio padrão de 5467 amostras (valores obtidos restritos à análise estatística das palavras analisadas, porém representativo para uma linha geral de análise na língua portuguesa, ao menos para os propósitos deste algoritmo). Portanto, sílabas longas cujos comprimentos excedem a média em um desvio padrão (ou seja, 12010 amostras) são consideradas supersílabas. Com base em experimentos realizados, foi observado que sequências com um número igual ou superior a esse de amostras, possuíam *de facto* mais do que uma sílaba em sua composição. Além disso, tornou-se necessária a utilização de outro critério para a localização do início e término de cada sílaba, dentro da sílaba com mais de 12010 amostras.

O critério utilizado para separar sílabas que contenham número igual ou superior a 12010 amostras consiste em comparar os valores de amostras desta sílaba com um valor

pré-estabelecido. Esse valor é o resultado do produto do limiar utilizado na localização da sílaba, por um fator *épsilon*. O processo é bastante similar àquele usado para detecção de sílabas por limiar, agora em outro nível (padrão do tipo *Matrioshka*).

A Tabela 5.2 mostra a separação efetuada pelo algoritmo para a palavra "departamento". Foram identificadas S = 4 sílabas, correspondentes a de/par/tamen/to, cuja forma de onda é mostrada na Figura 5.7. Pode ser observado que apenas a sílaba "tamen" tem duração maior do que 12010 amostras e, portanto, é identificada como uma supersílaba. O procedimento empregado para a separação de supersílabas resulta nos dados indicados na Tabela 5.3 e a forma de onda correspondente é mostrada na Figura 5.8.

A seguir é apresentado o algoritmo para a supersílaba:

Passo1:

Se tamanho da sílaba detectado $> 0{,}25 f_s$, uma supersílaba é identificada. No estado, isto corresponde a silabas com mais de 12010 amostras.

Passo2:

Caso haja supersílaba, não se registra o limiar usado na identificação silábica, aumentando-o gradativamente em passos de 10%, até que mais de uma sílaba seja detectada usando este novo limiar.

Tabela 5.2: *Dados do separador silábico para a palavra "departamento". Notar a identificação de uma supersílaba (terceira sílaba).*

Sílaba	Amostra inicial	Amostra final	Quantidade de amostras	Supersílaba?
De	7371	17913	10542	Não
par	25364	33320	7956	Não
tamen	42188	64029	**21841>12010**	Sim
to	70141	76111	5970	Não

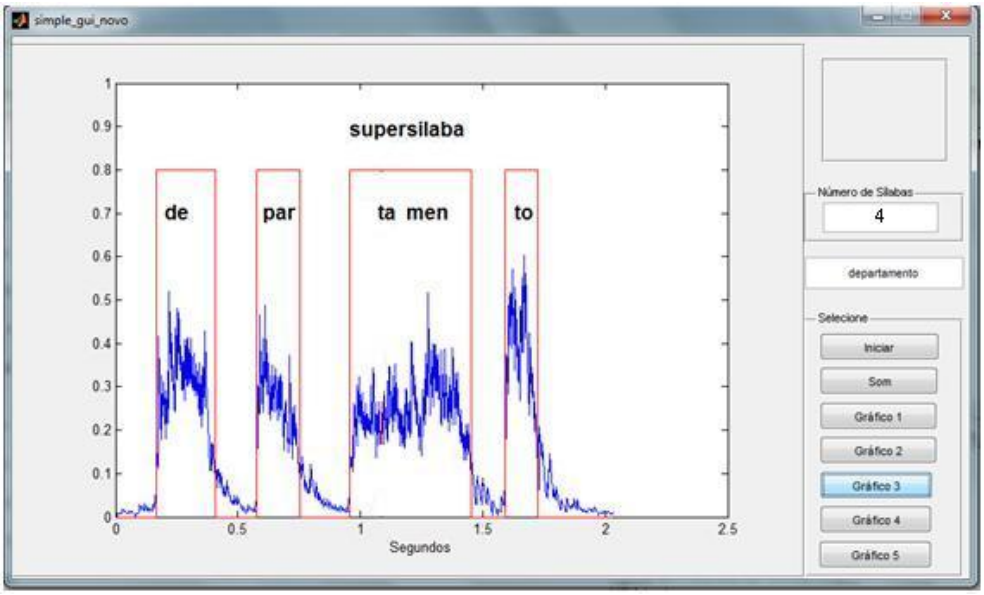


Figura 5.7: *Forma de onda correspondente à identificação de sílabas para a palavra "departamento". São identificadas 4 sílabas sendo a terceira uma supersílaba.*

Tabela 5.3: *Dados do separador silábico para a palavra "departamento", após a correta divisão da supersílaba identificada.*

Sílaba	Amostra inicial	Amostra final	Quantidade de amostras	Supersílaba?
De	7371	17913	10542	Não
par	25364	33320	7956	Não
ta	42188	47975	5787	Não
men	47975	64029	**16054>12010**	Sim ?
to	70141	76111	5970	Não

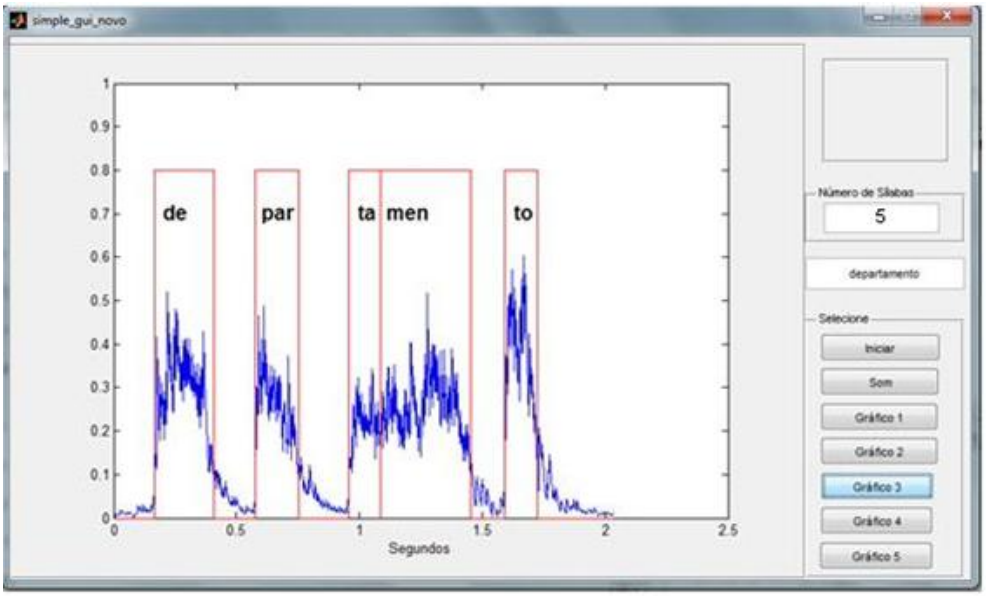


Figura 5.8: *Forma de onda correspondente à identificação de sílabas para a palavra "departamento". Após a "quebra" de uma supersílaba são identificadas corretamente as 5 sílabas.*

5.5. AGLUTINAÇÃO DE SÍLABAS SEPARADAS INDEVIDAMENTE

Para uma palavra com SS sílabas tem-se SS vetores correspondentes à envoltória dados por:

$$\{e_1(n), n = amostra\_inicial_i \text{ até } amostra\_final_i\}_{1 \le i \le S}.$$

A ideia de realizar aglutinação é baseada no comportamento médio da envoltória: sílabas consecutivas com uma pequena diferença entre os valores médios da envoltória são associadas em uma única sílaba. Os valores médios da envoltória são então calculados:

$$\{E_1\}_{1 \le i \le S} = \{E_1, E_2, \dots, E_S\},$$

em que *i* denota a sílaba detectada.

Para verificar a diferença do comportamento médio da envoltória intersílaba trabalha-se com o vetor normalizado.

$$\left\{100 * \frac{E_1}{E_{máx}}\right\}_{1 \le i \le S},$$

em que *i* novamente denota a sílaba detectada.

O vetor diferença para as envoltórias das diferentes sílabas é definido por

$$\{\Delta E_i\}_{1\leq i\leq S}$$

em $E_0 = 0$ e $\Delta E_i = \left|\frac{E_i - E_{i-1}}{E_{máx}}\right| * 100$.

A regra de aglutinação é implementada verificando quais as sílabas consecutivas que possuem pequenas (arbitra-se inferior a 3%, sendo o ajuste empírico) diferenças entre as envoltórias médias. Assim,

Se $\Delta E_i \leq 3\%$ aglutinar as sílabas $i-1$ e i.

A seguir descreve-se o exemplo da implementação desta técnica para a palavra "vale" que inicialmente foi classificada contendo S = 3 sílabas, a saber, "va-l-le". As "sílabas" identificadas correspondem as amostras ilustradas na Tabela 5.4. A forma de onda correspondente é mostrada na Figura 5.9.

O vetor de média obtido da envoltória das sílabas foi:

$$(E_i : E_2 : E_3) = (0{,}4953 : 0{,}2157 : 0{,}2072)$$

Tabela 5.4: *Sílabas, amostra inicial e amostra final para a palavra "vale" antes da aglutinação*

Sílaba	Amostra inicial	Amostra final	Quantidade de amostras	Duração (ms)
Va	12362	21601	9239	209,5
l	21601	25522	3921	88,9
le	25596	29830	4234	96

O vetor de diferença entre envoltórias consecutivas é:

$$(\Delta E_1 : \Delta E_2 : \Delta E_3) = (100 : 56{,}45 : 1{,}727).$$

Como $\Delta E_3 \leq 3$, aglutinam-se as sílabas 3 e 2, resultando na divisão silábica indicada na Tabela 5.5 e na Figura 5.10.

Este mesmo procedimento aplicado à palavra "hoje", cuja separação inicial teve três "sílabas", a saber, "ho-j-je", como ilustrado nas Figuras 5.11 e 5.12 a seguir, resulta na aglutinação das sílabas 2 e 3. Vê-se claramente que o comportamento da média da envoltória para as pseudo-sílabas 2 e 3 é bastante semelhante.

Tabela 5.5: Sílabas, amostra inicial, amostra final e durações da palavra "vale" após aglutinação

Sílaba	Amostra inicial	Amostra final	Quantidade de amostras	Duração (ms)
Va	12362	21601	9239	209,5
le	21601	29830	8229	186,6

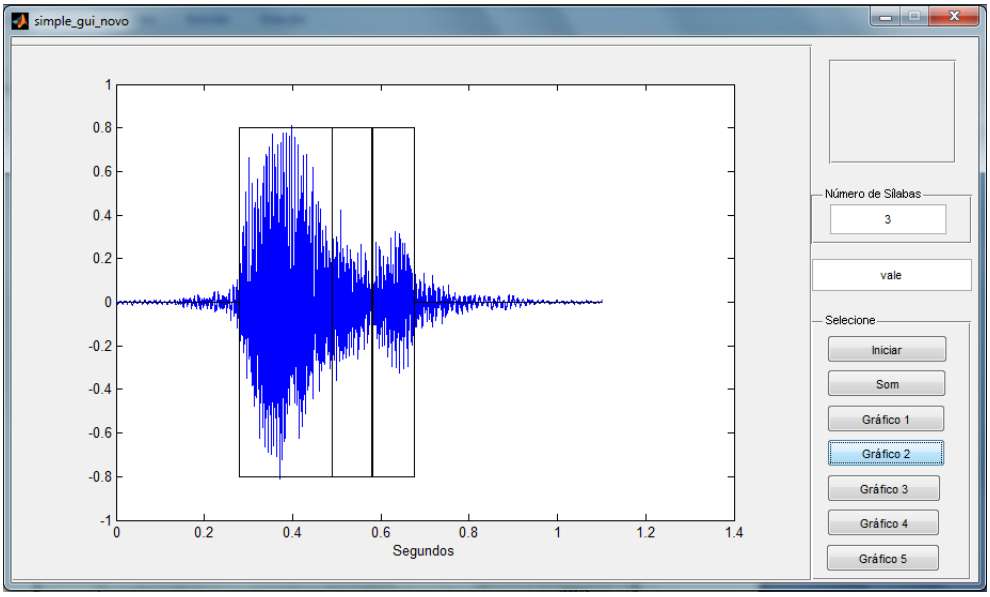


Figura 5.9: *Forma de onda da palavra vale com a separação inicial composta por três "sílabas": va-l-le.*

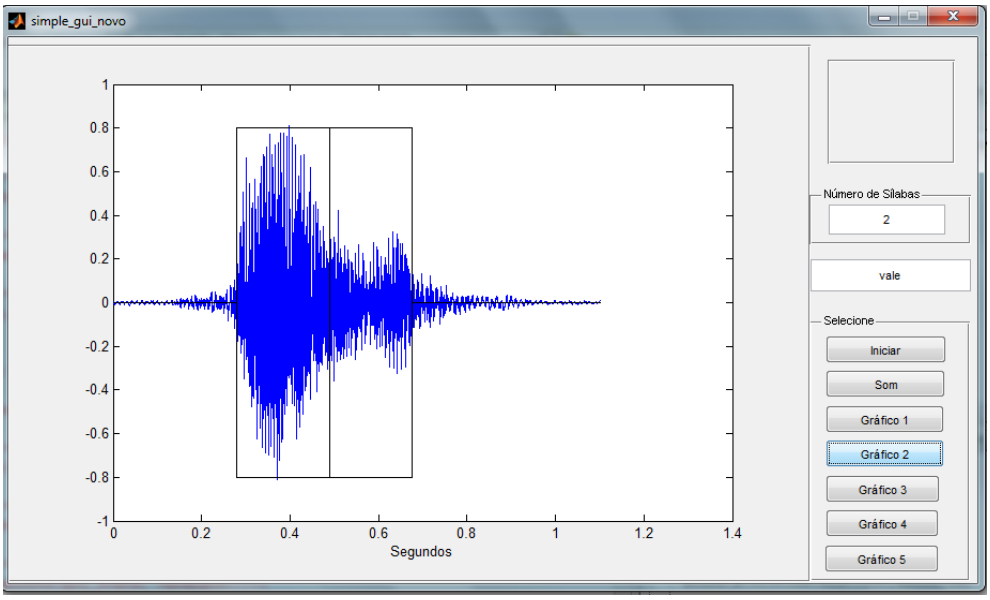


Figura 5.10: *Divisor silábico para a palavra vale após a aplicação do procedimento de aglutinação: as sílabas identificadas correspondem a "va-le".*

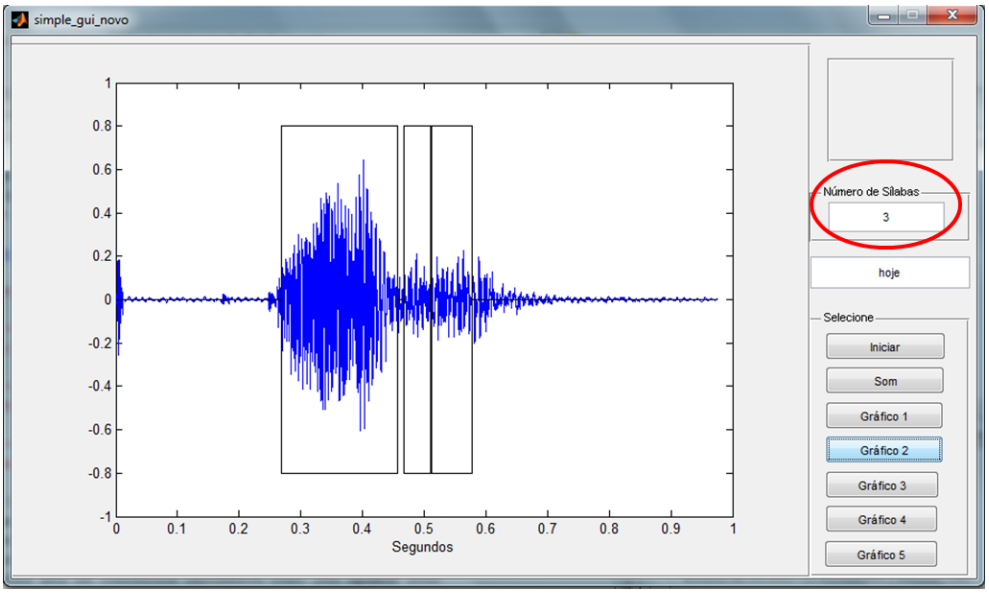


Figura 5.11: ***Forma de onda para a palavra hoje com a separação inicial composta por três "sílabas": "ho-j-je".***

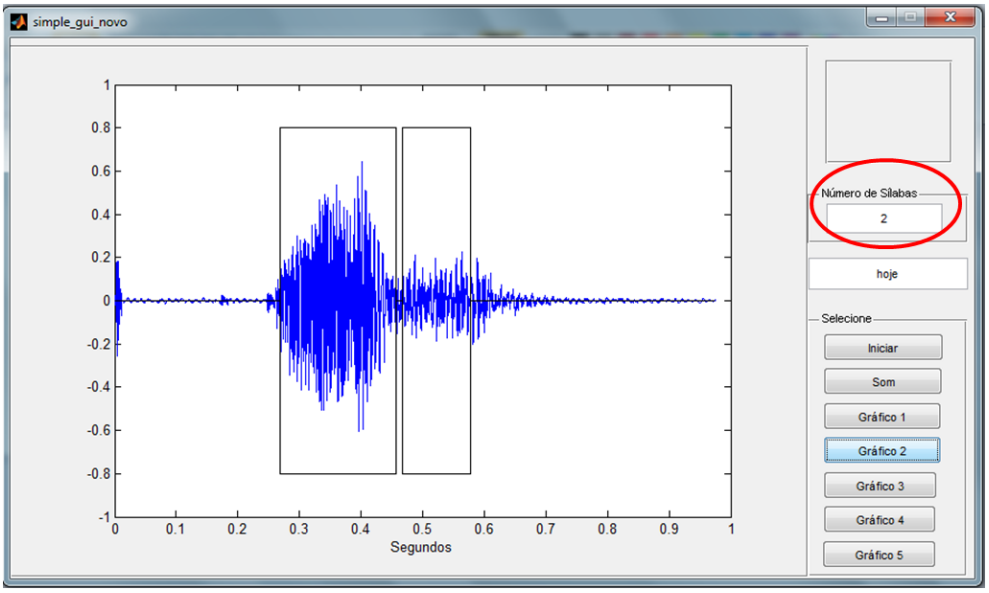


Figura 5.12: *Divisor silábico para a palavra hoje após a aplicação do procedimento de aglutinação: as sílabas identificadas correspondem a "ho-je".*

5.6. DESEMPENHO DO DIVISOR SILÁBICO

O algoritmo proposto conseguiu realizar adequadamente a separação silábica automática, de 43 das 49 palavras gravadas. Dessas 43 palavras, 23 foram separadas por completo (Tabela 5.7) e 20 obtiveram uma separação parcial (Tabela 5.8). Mostra-se a seguir, ilustrações do processo para a separação da palavra "departamento", incluindo detalhes da interface gráfica (veja URL citada no resumo).

A palavra analisada, “departamento”, correspondeu a um total de 76800 amostras na gravação, com duração total de 1,74 segundos. A Tabela 5.6 identifica a saída do divisor silábico, com os respectivos delineamentos das sílabas (amostras inicial e final) e duração estimada. A análise para as palavras descritas nas Tabelas 5.7 e 5.8 encontra-se no ANEXO D. Para algumas palavras, a divisão silábica conduziu a um resíduo de uma consoante na separação (e.g. música = mú-si-ca, semicondutor = sem-mi-com-du-tor), porém isto não caracteriza uma separação silábica incorreta e o resultado foi considerado integralmente correto e incorporado na Tabela 5.7. Por outro lado, houve a ocorrência de sílabas “perdidas”, como é o caso da palavra “duzentos” (vide Tabela 5.8).

Tabela 5.6: *Sílabas Separadas no Arquivo de Voz Contendo a Palavra “departamento”. Índice da Amostra Inicial e da Amostra Final e Duração Estimada da Sílaba (exemplo de saída de dados).*

Sílaba	Início	Fim	Duração (ms)
De	7371	17913	239
par	25364	33320	180
ta	42188	47975	131
men	47975	64029	364
to	70141	76111	135

Como pode ser visto, as palavras da Tabela 5.8 não puderam ser inteiramente separadas (e.g. vião, silei). Embora alguns trechos de voz possuíssem mais do que uma sílaba, estas sílabas não atenderam ao critério usado nas supersílabas e, portanto, ainda necessitariam sofrer uma nova quebra. Noutros casos, letras isoladas com duração maior que o usual (e.g. b, m, s) foram indicadas como pseudossílabas (isto é, não puderam ser eliminadas alterando-se o limiar, sob pena de “perder” algumas sílabas muito curtas).

Em algumas palavras, não foi possível realizar a separação das sílabas, como por exemplo, oito, roxo e banana. As Figuras de 5.13 a 5.16 ilustram a interface, desenvolvida na plataforma Matlab®, para as etapas da análise da palavra “departamento”. Testes preliminares com diferentes locutores e diversas palavras também indicaram que um aumento na velocidade com que o locutor pronuncia a(s) palavra(s) (ou fala) não teve efeito sobre o desempenho do sistema de reconhecimento de sílabas para as palavras testadas.

Tabela 5.7: Lista de palavras com as respectivas sílabas separadas corretamente pelo algoritmo de divisão silábica (23 Palavras).

Palavras	**Sílabas**				
abacate	a	ba	ca	te	
batata	ba	ta	ta		
berimboca	be	rim	bo	ca	
bonita	bo	ni	ta		
butantã	bu	tan	tã		
café	ca	fé			
campus	cam	pus			
complexo	com	ple	xo		
computador	com	pu	ta	dor	
corpo	cor	po			
departamento	de	par	ta	men	to
história	his	to	ria		
hoje	ho	je			
música	mus	si	ca		
pitoco	pi	to	co		
recife	rec	ci	fe		
roupa	rou	pa			
semicondutor	sem	mi	con	du	tor
siri	si	ri			
solteiro	sol	tei	ro		
uva	u	va			
vale	va	le			
zebra	ze	bra			

Um teste adicional consistiu na execução do algoritmo para um arquivo de áudio obtido pela leitura fluente (em voz contínua) de um texto escrito. O texto selecionado foi o poema “Vou-me embora para Pasárgada”, do poeta Manuel Bandeira, gravado em .wav 16 bits mono, com duração total de 1 minuto e meio, o qual contém um total de 316 silabas. O identificador /separador silábico proposto encontrou um total de **360** sílabas (já incluindo, entre elas, a detecção de um total de 62 supersílabas, ANEXO E). O tempo de resposta para o poema proposto é de aproximadamente 2 minutos.

Há ainda entre 5% de trechos identificados como sílabas, mas que se constituem de meras letras isoladas. A inclusão de critérios adicionais para reduzir estas ocorrências pode melhorar o desempenho do separador.

Tabela 5.8: Lista de palavras com as respectivas sílabas separadas de forma parcial pelo algoritmo de divisão silábica (20 palavras)

Palavras	Sílabas					
abacaxi	a	ba	cax	i		
assado	as	as	d	do		
avião	av	vião	o			
brasileiro	bra	silei	ro			
cabelo	ca	belo				
cabeça	ca	be	ç	ça		
circunferência	cir	cun	fe	rên	c	ia
duzentos	duz	zen				
economia	e	cono	mia			
eletrônica	ele	tro	ni	ca		
engenharia	em	g	genhar	ria		
farmácia	far	r	ma	cia		
matemática	ma	tema	a	ti	ca	
minuto	m	m	inu	to		
mistério	mi	s	te	rio		
oficina	o	fic	c	ina		
pernambuco	per	nam	b	bu	co	
televisão	tele	v	visão			
universidade	uni	ver	si	dade		
vestibular	ves	ti	bula	ar		

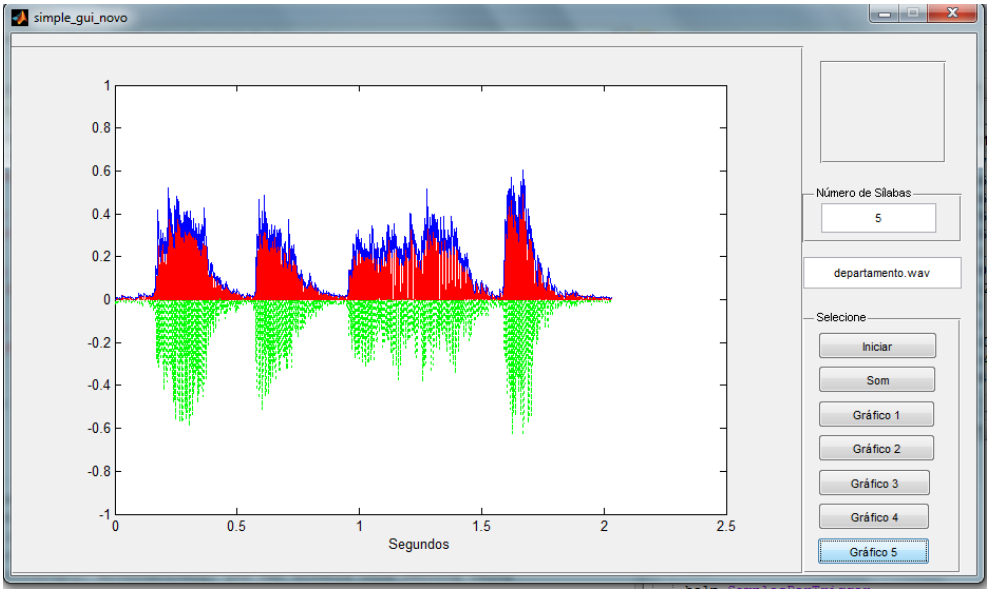


Figura 5.13: *Formas de onda envolvida (as cores são disponibilizadas na versão eletrônica)s: a) em cor verde, sinal de áudio referente à palavra "departamento"; b) em cor vermelha, a onda retificada em meia-onda; c) em cor azul, o envelope do sinal obtido com envoltória de descarga linear.*

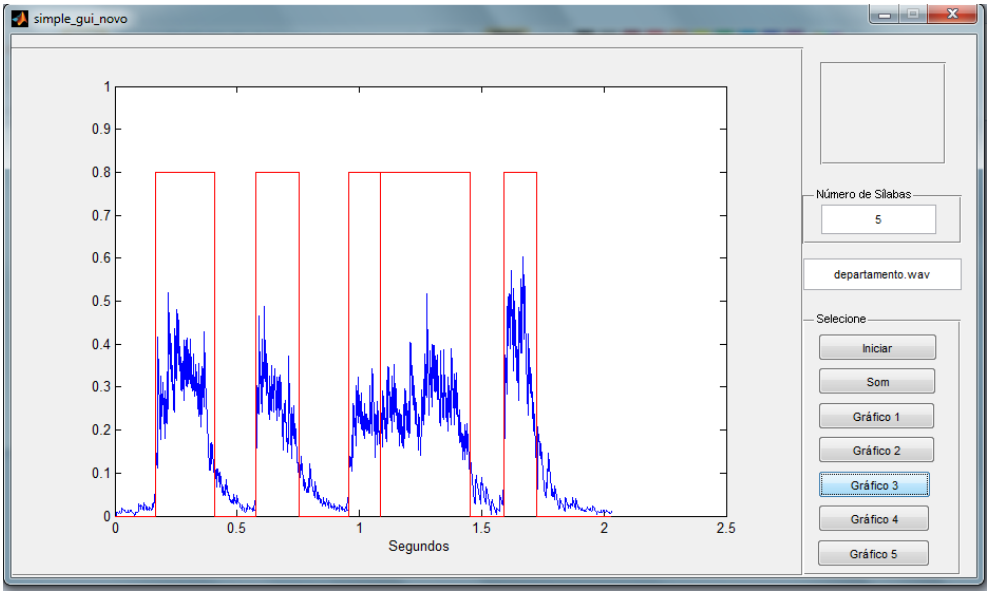


Figura 5.14: *Forma de onda com a respectiva separação silábica obtida por limiar aplicado ao envelope do sinal de voz para a palavra "departamento".*

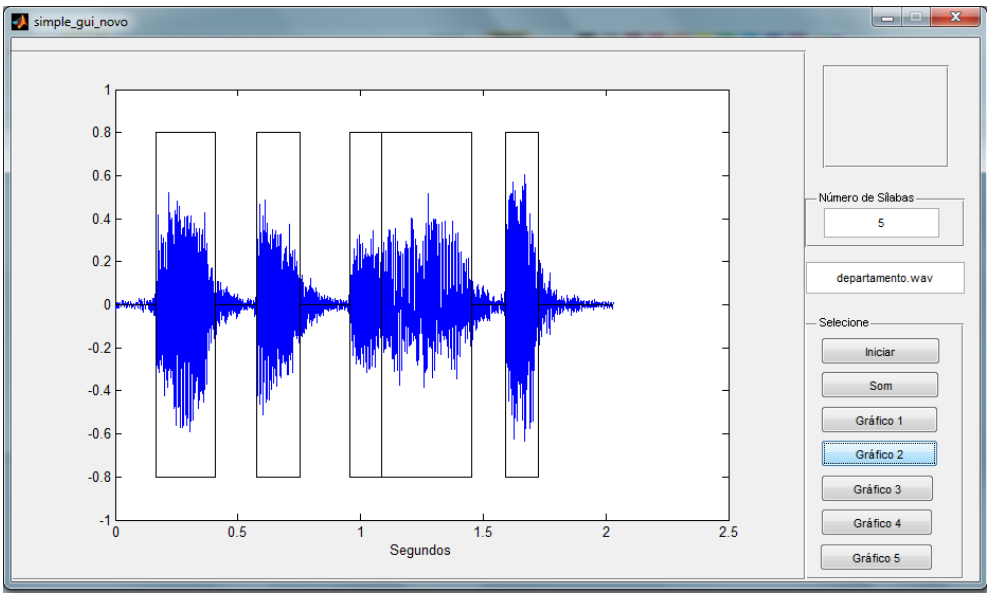


Figura 5.15*: Separação silábica no sinal original: note a separação clara dos fonemas "DE-PAR-TAMEN-TO". A supersílaba "TAMEN" foi corretamente subdividida.*

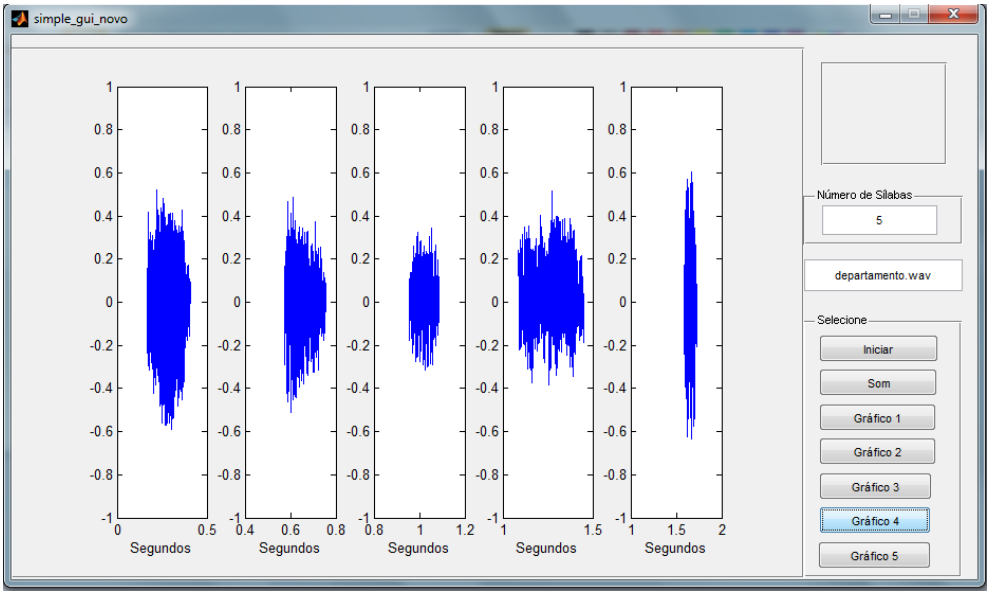


Figura 5.16: *Divisão silábica completada. Trechos são indicados; há acesso ao comprimento em amostras e duração das sílabas e ao trecho de áudio isolado de cada silaba para a palavra "departamento".*

Do ponto de vista de complexidade a proposta do algoritmo de divisão silábica com base em envoltória envolve pouco uso de operações aritméticas quando comparado a estratégias como HMM ou redes neuronais.

Certamente a incorporação de técnicas de descontaminação/remoção de ruído deve melhorar o desempenho do divisor silábico proposto. Isto poderia ser de especial valia em aplicações comerciais (iPhone, etc.), propõe-se, portanto, em trabalhos futuros, investigar os efeitos da remoção do ruído no desempenho do método em ambientes sem controle de ruído.

6. CONCLUSÕES

Neste capítulo são descritas as principais conclusões de ambos os trabalhos. Primeiramente são apresentadas as conclusões para o estimador de raias espectrais e em seguida para o separador silábico.

6.1. SOBRE O ESTIMADOR VOCÁLICO

- O estimador de raias espectrais proposto nesta dissertação possui como atrativo principal a simplicidade. Por isso, o método em questão pode viabilizar o seu uso em sistemas em tempo real bem como em sistemas embarcados. Mais especificamente, em sistemas para reconhecimento e identificação de locutores.
- Como visto, o método proposto possui valores de picos de frequência muito similares, para as amostras coletadas, daqueles obtidos com o auxílio do programa Audacity 1.3®. Além disso, constatou-se, em praticamente todos os casos, que há uma flutuação na cadencia das raias, de forma que o espectro apresenta "pequenos desvios" em torno dos "harmônicos teóricos" para um sinal periódico.
- Também é possível observar, no âmbito do universo estudado, um "achado" no mínimo curioso. A correlação entre a curva teórica e a obtida pelo estimador para um sinal de voz é maior para voz masculina do que para a voz feminina.

6.2. SOBRE O DIVISOR SILÁBICO

- Com relação ao separador silábico conclui-se que o método proposto, por ser um método bastante simples e que não seja intensivo em recursos computacionais, pode ser utilizado em sistemas de tempo real, tais como, na etapa inicial em sistemas conversão automática de "fala-para-texto" (STT), adaptados a língua portuguesa, com qualidade aceitável.
- Além disso, valendo-se ainda da baixa complexidade computacional do método no processamento da fala, o sistema proposto pode ser utilizado em sistema de plataforma embarcada, tais como FPGA (*field-programmable gate arrays*).

- O separado silábico, também, proporciona sua aplicação em sistemas automáticos com comandos de voz específicos e de reconhecimento de locutor. Outras contribuições potenciais do método são:
    - — O auxílio na educação infantil e para estrangeiros, em que estes podem aprender e exercitar a forma correta da separação das palavras da língua portuguesa;
    - — O apoio para pacientes em tratamento fonoaudiólogo, onde estes podem não só exercitar a correta separação das palavras, como também, ouvir cada sílaba separada quantas vezes forem necessário.

7. TRABALHOS FUTUROS

Finalmente têm-se as descrições sobre possíveis linhas de trabalhos futuros: estimador de raias espectrais e separador silábico, na ordem descrita.

7.1. POTENCIAIS MELHORIAS NO ESTIMADOR VOCÁLICO

Para o estimador de raias espectrais, propõe-se como possíveis investigações:

- A aplicação e verificação do método em trechos de voz natural, nos quais existam trechos alternados vocálicos e não vocálicos;
- Um aprofundamento na análise do comportamento baseado em gênero, já que como foi visto, uma correlação maior para voz masculina do que para voz feminina;
- Comparação da complexidade com diversos algoritmos de detecção de *pitch*, pois o método proposto realizou uma única estimação com base no método sub-harmônico-harmônico do XUEJING (2002a), como maneira de validar que o estimador também determina o elemento de *pitch;*
- Finalmente, uma investigação do sinal de voz utilizando as séries *quase-harmônicas* de Fourier, propostas por VERMEHREN *et al*. (2010).

7.2. POTENCIAIS MELHORIAS NO DIVISOR SILÁBICO

Com relação a separador silábico, propõe-se:

- A inclusão de critérios adicionais para melhorar o desempenho do método, tanto no processo de separação, quanto no processo de aglutinação de algumas sílabas.
- A combinação com outros algoritmos de identificação silábicos mais complexos, tais como, aqueles constituídos por HMM, redes neuronais, método Cepstral, etc., em etapa inicial de segmentação da fala.
- E na transconversão de voz. A transconversão de voz foi o “pontapé” inicial para os trabalhos desenvolvidos nesta dissertação. A ideia consiste na extração dos parâmetros vocais de um locutor, para serem inseridos em um arquivo de áudio de uma palavra pronunciado por outro locutor. A

combinação desses elementos pode gerar uma saída em que se tem um locutor falando com as características do outro locutor? A abordagem pode auxiliar na concepção de um logaritmo de “troca de falante”.

ANEXO A – CÓDIGO FONTE DOS ALGORITMOS

A Figura A.1 apresenta o diagrama em blocos do estimador vocálico. Na sequência é apresentado o código fonte em linguagem Matlab® do estimador.

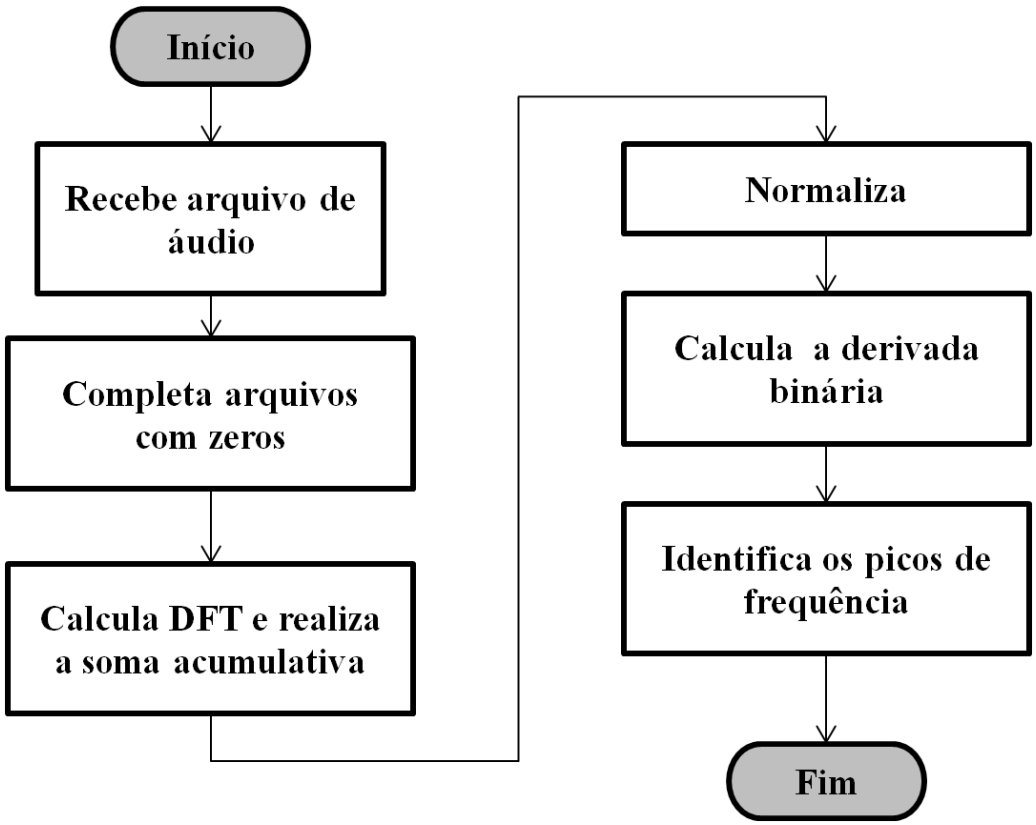


Figura A.1: *Diagrama em blocos do código fonte do estimador vocálico.*

```
%%%%%%%%%%%%%%%%%%%%%%
% ESTIMADOR VOCÁLICO %
%%%%%%%%%%%%%%%%%%%%%%
function [F1,num_picos,saida] = identificador_pitch(trecho_voz)
NN=2048;

% Completa o arquivo com zeros
auxi = trecho_voz;
mm = length(auxi);
L=ceil(mm/NN); % numero de janelas (teto)
paddingzero = zeros(NN*L-mm,1)';
arquivo_vetor_car = [auxi paddingzero]';

% Calculo da transformada de Fourier
m = length(arquivo_vetor_car);
L = m/NN; % numero de janelas (teto);
X = zeros(NN,1);
for i=1:L
  X = X+abs(fft(arquivo_vetor_car(1+(i-1)*NN:NN+NN*(i-1))));
end

% Normaliza o vetor X.
norm = sum(sum(X.^2,2),1);
X = X/norm;
maximo = max(X,[],1);

% Critério de localização dos picos de frequência
tol = 0.001* maximo;

% Classifica os picos de frequência segundo o critério anterior
sinal = zeros(NN,1);
%
for i=1:NN-1
  if (X(i+1,1)> X(i,1)) && (abs(X(i,1)> tol))
    sinal(i,1) = 1;
  else if (X(i+1,1)<X(i,1)) && (abs(X(i,1)>tol))
      sinal(i,1) = -1;
    end
  end
end

% Determina os picos de frequência
picos = zeros(150,1);
contador =0;
Aux = zeros(150,1);
for i=1:150
  if (sinal(i+1,1)== -1 && sinal (i,1)==1 )
    picos(i,1)= 1;
    contador = contador +1;
    Aux(i)= X(i);
  else
    picos(i,1)= 0;
  end
end
% Lista os picos de frequência em uma matriz (Posição, amostra)
F=find(picos);
num_picos = length(F)-2;
F1 = F(1:length(F))*(21.53);% Freq/Janela (44100/2048)
F(1:length(F));
A =zeros(contador-2,2);
```

```
j=0;
for i=1:150
  if Aux(i)~= 0
    j=j+1;
    if j>2
    A(j-2,1) = Aux(i);
    A(j-2,2) = F(j);
    end
  end
end
```

A Figura A.2 apresenta o diagrama em blocos do separador silábico. Na sequência é apresentado o código fonte em linguagem Matlab® do separador.

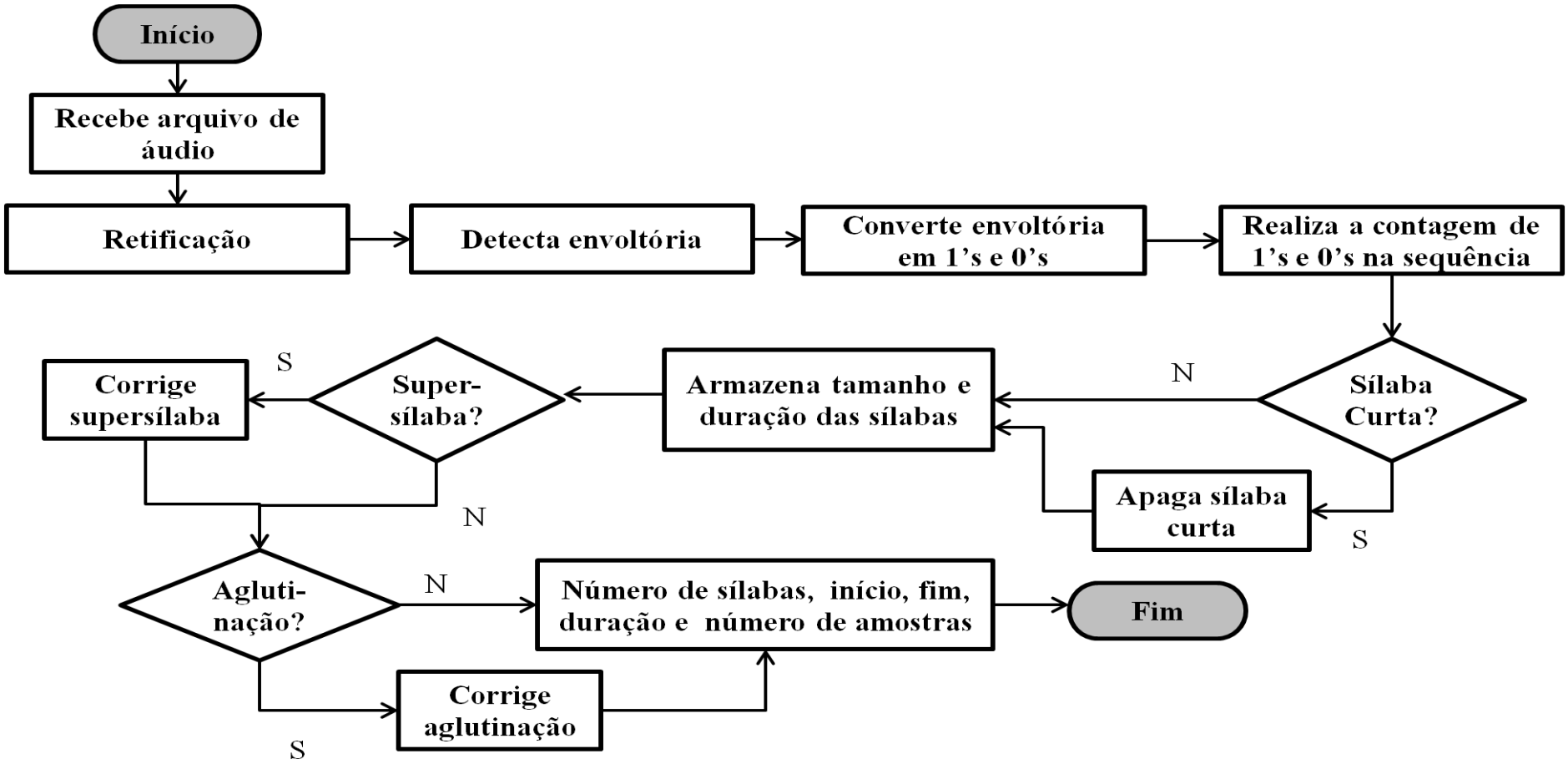


Figura A.2: *Diagrama em blocos do código fonte do separador silábico.*

```
%%%%%%%%%%%%%%%%%%%%%%
% SEPARADOR SILÁBICO %
%%%%%%%%%%%%%%%%%%%%%%
function [x,silabas,silabas1,QTD_SILABAS,LOCAO,e,rect] =
separador_teste_novo(arq)
PERC=1.2;
N=2048;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ENTRADA DO SINAL DE ÁUDIO %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% leitura do arquivo de áudio extensão wav %
[x,fs,NBITS]=wavread(arq);
% os arquivos são, em princípio, mono, amostragem fs=44,1 kHz)
% análise janelada: adotadas janelas de 22 mseg
window = 0.022;
% xs1 = silencio(xs);
% x = xs1';
% com taxa de amostragem fs, isso corresponde a
% um número de amostras DELTA dentro de cada janela
% DELTA=fs*window=44.100*22.05=970 amostras
DELTA = floor(fs*window);
% transformando o arquivo em mono, se necessário
rect = x(:,1);
% tamanho do arquivo de áudio, em amostras
N = length(rect);
% total de janelas de 22 ms no arquivo
N_windows = ceil(N/DELTA);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% DETECTANDO A ENVOLTÓRIA DO SINAL DE ENTRADA %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%inicializando a envoltória
e = zeros(N,1);
% retificação de meia-onda
for n=1:N
  rect(n)=(1+sign(rect(n)))*rect(n)/2;
end
% valor rms total do arquivo
rms = sqrt(var(rect,0,1));
var_e= rms;
% descarga do detector de envoltória, com descarga LINEAR
% a constante de tempo é regulada por DELTA, i.e.
% Após uma janela, o sinal deve estar descarregado
for n=1:N-1
  if rect(n) > e(n)
    for j=n:(n+DELTA)
      e(j) = (n-j+DELTA)/DELTA*rect(n);
    end
  end
  e(n) = max(rect(n),e(n));
end

%%%%%%%%%%%%%%%%%%%%%%%%%%
% LOCALIZADOR DE SILABAS %
%%%%%%%%%%%%%%%%%%%%%%%%%%
% localizando as silabas
Vlim = PERC*var_e;
silaba = zeros(N,1);
for n=1:N
```

```
  if e(n) > Vlim
    silaba(n) = 1;
  else
    silaba(n)=0;
  end
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% CONTAGEM DAS SEQUÊNCIAS DE AMOSTRAS %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% fazendo a contagem da sequência de amostras
% silabas=1 e não-silabas=0
n=1;
Qde = 0;
i = 1;
contagem(i,1)=0;
contagem(i,2)=0;
while (n<=N)
  caracter = silaba(n);
  while (n<=N) && (silaba(n) == caracter)
    Qde = Qde + 1;
    n = n + 1;
  end
  contagem(i,1)=caracter;
  contagem(i,2)=Qde;
  Qde = 0;
  i = i +1;
end
% apagando as (pseudo)silabas de duração muito curta
larguraMinima = 1.8*DELTA;
for n=1:length(contagem)
  if (contagem(n,1) == 1) && (contagem(n,2) < larguraMinima)
    contagem(n,1)=0;
  end;
end;

%%%%%%%%%%%%%%%%%%%%%%
% SÍLABAS E DURAÇÕES %
%%%%%%%%%%%%%%%%%%%%%%
%
index = find(contagem(:,1));
totalsilabas = length(index);
duracoes = zeros(totalsilabas,1);
posicoes = cumsum(contagem);
t = 1;
inicio_silabas = zeros(totalsilabas,1);
for k=1:(length(contagem)-1)
  if posicoes(k,1) ~= posicoes(k+1,1)
    inicio_silabas(t) = posicoes(k,2);
    t = t +1;
  end
end
for k=1:totalsilabas
  duracoes(k) = contagem(index(k),2);
end
LOCA = zeros(totalsilabas,2);
limiar=zeros(totalsilabas,1);
for k=1:totalsilabas
  LOCA(k,1)=inicio_silabas(k);
```

```
  LOCA(k,2)=duracoes(k);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%
% IMPRIMINDO AS SILABAS...%
%%%%%%%%%%%%%%%%%%%%%%%%%%%
 LOCA;
  [Y,tonica] = max(duracoes);
  provavel_TONICA = tonica;

%%%%%%%%%%%%%%%%%%%%%%%%%%
% CORRIGINDO AS SSILABAS %
%%%%%%%%%%%%%%%%%%%%%%%%%%
% Considerado particionar supersilaba quando tempo>250 ms
% NB para fs=44100 amostras/s, N_STAR=11.025 amostras
% Comprimento medio de silabas = 6543 amostras
% Desvio padrão do comprimento de silabas = 5467 amostras
% Definição de supersilaba, quando o comprimento da silaba excede um desvio
% padrão da média isto corresponde a N_STAR = 6543+5467 = 12010 amostras
N_STAR = 0.25*fs;
ssilabas = 0;
inicio_ssilaba = zeros(totalsilabas,1);
quebra_silabao = zeros(totalsilabas,1);
for k=1:totalsilabas
if LOCA(k,2)>N_STAR
  ssilabas = ssilabas+1;
  inicio_ssilaba(k)=k;
% Corrigindo as ssilabas
  silabao = ones(LOCA(k,1)+LOCA(k,2)+1,1);
  produtorio=1;
  epslon=0;
  while produtorio ~=0
  for n= LOCA(k,1)+2000: LOCA(k,1)+LOCA(k,2)-2000
    if e(n) < Vlim+epslon
      silabao(n)=0;
      quebra_silabao(k)=n-LOCA(k,1);
    end
    produtorio=produtorio*silabao(n);
  end
  epslon=epslon+0.1*Vlim;
  end

end
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% SAÍDA CASO OCORRA SUPERSÍLABAS %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
LOCAO =zeros(2*size(LOCA,1),2*size(LOCA,2));
j = 1;
for i = 1:length(LOCA)
if quebra_silabao(i)~=0
  LOCAO(j,1) = LOCA(i,1);
  LOCAO(j,2)= quebra_silabao(i);
  j = j+1;
  LOCAO(j,1) = LOCA(i,1)+ quebra_silabao(i);
  LOCAO(j,2)= LOCA(i,2) - quebra_silabao(i);
else

  LOCAO(j,1) = LOCA(i,1);
```

```
  LOCAO(j,2) = LOCA(i,2);
end
 j = j+1;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                    %
% IDENTIFICADOR DE SÍLABAS ISOLADAS %
%                    %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

media_env = zeros (1,length(LOCAO));
delta_env = zeros (1,length(LOCAO));
for i=1:length(LOCAO)
  media =0;
  for ii=LOCAO(i,1):(LOCAO(i,1)+LOCAO(i,2))
    if ii~= 0
    media = media + abs(e(ii));
    end
  end
  media_env(i)= media/ LOCAO(i,2);
end
resultado = 100*media_env./max(media_env);
delta_env(1)= resultado(1);
for i= 2: length(LOCAO)
delta_env (i) = abs(resultado(i)- resultado(i-1));
end
[L,C]= find(delta_env<3);
LOCAO(C-1,2)=LOCAO(C,1)- LOCAO(C-1,1)+LOCAO(C,2);
LOCAO(C,1)=0;
LOCAO(C,2)=0;
disp('silaba   inicio     final     (amostras)   (mseg)');
format short;
for i=1:length(LOCAO)
  if LOCAO(i,2)>12010
    NB = ' Dividir??';
  else
    NB = '';
  end
  if LOCAO(i,1)~=0
    disp([num2str([i LOCAO(i,1) (LOCAO(i,1)+LOCAO(i,2)) LOCAO(i,2)
LOCAO(i,2)/44.1]) NB]);
  end
end

QTD_SILABAS = j-length(C);
silabas1 = zeros(size(x));
silabas = zeros(size(x));
for jota = 1: (QTD_SILABAS -1)
  for i = LOCAO(jota,1)+1:LOCAO(jota,2)+ LOCAO(jota,1)-1
    silabas(i)= 0.8;
    silabas1(i)= -0.8;
  end
end
```

ANEXO B – TABELAS E GRÁFICOS

Tabela B.1: *Raias espectrais iniciais e respectivos passos para os seis locutores, para as vogais "a","e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

Locutor	Freq. Inicial **a** (passo)	Freq. Inicial **ê** (passo)	Freq. Inicial **é** (passo)	Freq. Inicial **i** (passo)	Freq. Inicial **ô** (passo)	Freq. Inicial **ó** (passo)	Freq. Inicial **u** (passo)
Alessandra	215,3 (174,6)	215,3 (220,3)	215,3 (213,6)	258,4 (193,4)	236,8 (203,3)	236,8 (188,9)	258,4 (214,5)
Lidiane	215,3 (197,8)	193,8 (98,8)	193,8 (184,2)	236,8 (126,2)	193,8 (74,4)	193,8 (160,5)	280,0 (162,3)
Lizandra	301,4 (210,3)	279,9 (197,7)	323,0 (328,0)	366,0 (203,1)	323,0 (224,9)	279,9 (174,9)	323,0 (103,4)
Paulo Freitas	129,20 (117,3)	150,71 (115,9)	129,20 (133,5)	150,71 (143,8)	150,71 (101,6)	150,71 (141,1)	150,71 (121,3)
Paulo Martins	172,2 (167,8)	172,2 (144,8)	322,9 (151,8)	301,4 (118,6)	236,8 (114,5)	236,8 (138,6)	322,9 (84,1)
Ricardo	107,65 (102,6)	107,65 (89,0)	107,65 (102,0)	129,20 (82,0)	129,20 (83,3)	107,65 (94,5)	129,20 (87,6)

Tabela B.2: *Comparação da estimativa de pitch para os seis locutores, para as vogais a, e , incluindo acentuação (grave/agudo), de acordo com o programa de estimativa espectral vocálica (Nesta Dissertação) e algoritmo de identificação de pitch (XUEJING, 2002) (valores em Hz).*

Locutor	(a)	(ê)	(é)	(i)	(ô)	(ó)	(u)
Alessandra	215,3 *218,3*	215,3 *219,2*	215,3 *212,4*	258,4 *247,1*	236,8 *233,5*	236,8 *233,3*	258,4 *250,4*
Lidiane	215,3 *220,9*	409,1 *375,3*	193,8 194,2	236,8 *226,6*	193,8 *193,6*	193,8 *203,0*	279,9 *274,9*
Lizandra	301,4 *293,9*	279,9 *285,4*	322,9 *327,8*	366,0 *371,6*	322,9 *193,7*	279,9 *284,6*	323,0 *329,5*
Paulo Freitas	129,2 *130,6*	150,7 *141,3*	129,2 *133,5*	150,7 *145,4*	150,7 *149,8*	150,7 *142,2*	150,7 *157,3*
Paulo Martins	172,2 172,3	172,2 *169,9*	172,2 *162,5*	150,7 *156,1*	150,7 *160,5*	150,7 *159,4*	323,0 *317,9*
Ricardo	107,6 109,7	107,6 *103,0*	107,6 *107,9*	129,2 *129,5*	129,2 *123,3*	107,7 ***115,6***	129,2 *134,3*

Tabela B.3: *Raias espectrais de Alessandra, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	59 -	59 -	59 -	59 -	59 -	59 -	59 -
2	120 -	120 -	120 -	118 -	120 -	120 -	121 -
3	213 (215,3)	213 (215,3)	211 (215,3)	248 (258,36)	229 (236,83)	231 (236,83)	251 (258,36)
4	431 (430,6)	436 (430,6)	424 (430,6)	492 (495,19)	452 (452,13)	466 (473,66)	497 (495,16)
5	653 (645,9)	659 (667,43)	638 (645,9)	744 (753,55)	683 (688,96)	700 (710,49)	749 (753,55)
6	872 (882,73)	877 (882,73)	850 (861,2)	989 (990,38)	916 (925,79)	931 (925,79)	1003 (1011,91)
7	1088 (1098,03)	1094 (1098,03)	1066 (1054,97)	- (1141,09)	1139 (1141,09)	1166 (1162,62)	1257 (1248,74)
8	1301 (1313,33)	1314 (1313,33)	1278 (1270,27)	1238 (1248,74)	1371 (1377,92)	1395 (1399,45)	- (1507,1)
9	1518 (1528,63)	1539 (1528,63)	1494 (1485,57)	- (1377,92)	1596 (1593,22)	1632 (1636,28)	- (1765,46)
10	1731 (1743,93)	1758 (1765,46)	1707 (1700,87)	1486 (1485,57)	1826 (1830,05)	1867 (1873,11)	- (2002,29)
11	1970 (1959,23)	1979 (1980,76)	1922 (1916,17)	1733 (1743,93)	2060 (2066,88)	2102 (2109,94)	- (2260,65)
12	2156 (2196,06)	2202 (2196,06)	2132 (2131,47)	1976 (1980,76)	2290 (2282,18)	2337 (2346,77)	- (2519,01)
13	2371 (2411,36)	2414 (2411,36)	2352 (2346,77)	2226 (2217,59)	2516 (2519,01)	- (2454,42)	- (2691,25)
14	2632 (2626,66)	2640 (2648,19)	2565 (2562,07)	2471 (2475,95)	- (2648,19)	2571 (2583,6)	- (2755,84)
15	- (2755,84)	2863 (2863,49)	2780 (2777,37)	2724 (2712,78)	2747 (2755,84)	2757 (2755,84)	- (2820,43)
16	- (2798,9)	3076 (3078,79)	2987 (2992,67)	- (2841,96)	2975 (2971,14)	- (2820,43)	- (3014,2)
17	- (2863,49)	-	3203 (3207,94)	2968 (2971,14)	- (3057,26)	- (2992,67)	- (3229,5)
18	- (3014,2)	-	-	3220 (3229,5)	3199 (3186,44)	- (3035,73)	-
19	- (3057,26)	-	-	-	3434 -	- (3229,5)	-
20	- (3186,44)	-	-	-	3669 -	-	-
21	- (3229,5)	-	-	-	-	-	-

Tabela B.4: *Raias espectrais de Lidiane, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

Picos (Hz)	A	Ê	É	I	Ô	Ó	U
	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	59 -	59 -	59 -	59 -	59 -	59 -	59 -
2	119 -	120 -	120 -	120 -	120 -	120 -	120 -
3	215 (215,3)	195 (193,8)	194 (193,8)	228 (236,8)	189 (193,8)	198 (193,8)	274 (279,9)
4	441 (430,6)	402 (409,1)	402 (409,1)	451 (452,1)	229 (236,8)	403 (409,1)	550 (559,8)
5	661 (667,4)	597 -	600 (602,8)	683 (689)	382 (387,5)	604 (602,8)	822 (818,1)
6	883 (882,7)	- (602,8)	- (796,6)	- (839,7)	575 (581,3)	809 (818,1)	1093 (1098)
7	1098 (1098)	796 (796,6)	804 -	905 (904,3)	769 (775,1)	1012 (1011,9)	- (1205,7)
8	1318 (1313,3)	997 -	1005 (1011,9)	- (1076,5)	960 (968,8)	1217 (1227,2)	- (1356,4)
9	1538 (1550,2)	- (1011,9)	1205 (1205,7)	1135 (1141,1)	- (1076,5)	1416 (1421)	- (1571,7)
10	1757 (1765,5)	- (1076,5)	1411 (1421)	- (1205,7)	- (1141,1)	- (1614,7)	- (1614,7)
11	- (1894,6)	- (1141,1)	1611 (1614,7)	- (1248,7)	- (1270,3)	- (1830,1)	- (1743,9)
12	1972 (1980,8)	- (1205,7)	- (1700,9)	- (1291,8)	- (1313,3)	- (2023,8)	- (1873,1)
13	2192 (2196,1)	- (1313,3)	1820 (1808,5)	1368 (1356,4)	- (1377,9)	- (2174,5)	- (2153)
14	2416 (2411,4)	- (1377,9)	2017 (2023,8)	1597 (1593,2)	- (1507,1)	- (2217,6)	- (2196,1)
15	2638 (2626,7)	- (1421)	2209 (2217,6)	- (1679,3)	- (1550,2)	- (2325,2)	- (2454,4)
16	2852 (2863,5)	- (1507,1)	2422 (2411,4)	1824 (1830,1)	- (1614,7)	- (2432,9)	- (2583,6)
17	3071 (3078,8)	- (1550,2)	2620 (2626,7)	- (1980,8)	- (1679,3)	- (2583,6)	- (2755,8)
18	- (3229,5)	- (1614,7)	2829 (2820,4)	2059 (2066,9)	- (1722,4)	- (2626,7)	- (2906,6)
19	3306 -	- (1743,9)	3028 (3035,7)	2282 (2282,2)	- (1808,5)	- (2820,4)	- (3035,7)

Tabela B.4 (Continuação): *Raias espectrais de Lidiane, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
20	-	- (1787)	3227 (3229,5)	2522 (2519)	- (1851,6)	- (3014,2)	- (3208)
21	-	- (1830,1)	-	- (2605,1)	- (1916,2)	- (3229,5)	-
22	-	- (1916,2)	-	2714 (2712,8)	- (1980,8)	-	-
23	-	- (2023,8)	-	2983 -	- (2023,8)	-	-
24	-	- (2239,1)	-	3213 (3208)	- (2088,4)	-	-
25	-	- (2389,8)	-	-	- (2153)	-	-
26	-	- (2432,9)	-	-	- (2217,6)	-	-
27	-	- (2583,6)	-	-	- (2346,8)	-	-
28	-	- (2798,9)	-	-	- (2389,8)	-	-
29	-	- (2842)	-	-	- (2454,4)	-	-
30	-	- (2992,7)	-	-	- (2519)	-	-
31	-	- (3057,3)	-	-	- (2583,6)	-	-
32	-	- (3186,4)	-	-	- (2626,7)	-	-
33	-	-	-	-	- (2691,3)	-	-
34	-	-	-	-	- (2755,8)	-	-
35	-	-	-	-	- (2820,4)	-	-
36	-	-	-	-	- (2863,5)	-	-
37	-	-	-	-	- (2928,1)	-	-
38	-	-	-	-	- (3057,3)	-	-
39	-	-	-	-	- (3164,9)	-	-

Tabela B.5: *Raias espectrais de Lizandra, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	59 -	59 -	59 -	59 -	59 -	59 -	59 -
2	119 -	121 -	120 -	120 -	119 -	120 -	120 -
3	296 (301,4)	280 (279,9)	325 (323)	369 (366)	317 (323)	279 (279,9)	326 (323)
4	598 (602,8)	571 (581,3)	658 (667,4)	744 (753,6)	637 (645,9)	570 (581,3)	660 (667,4)
5	902 (904,3)	855 (861,2)	989 (990,4)	1117 (1119,6)	956 (947,3)	854 (861,2)	- (839,7)
6	- (1055)	1142 (1141,1)	1315 (1313,3)	- (1313,3)	1268 (1270,3)	1139 (1141,1)	987 (990,4)
7	1201 (1205,7)	- (1291,8)	1643 (1636,3)	1497 (1485,6)	1589 (1593,2)	1423 (-)	- (1205,7)
8	1509 (1507,1)	1433 (1442,5)	1974 (1980,8)	- (1679,3)	- (1787)	1712 (1722,4)	1316 (1313,3)
9	1814 (1808,5)	- (1528,6)	2302 (2303,7)	- (1722,4)	1905 (1916,2)	- (2002,3)	- (1507,1)
10	2113 (2109,9)	- (1571,7)	2630 (2626,7)	1857 (1851,6)	2228 (2217,6)	- (2088,4)	- (1571,7)
11	- (2282,2)	1716 (1722,4)	2954 (2949,6)	- (2023,8)	- (2432,9)	- (2153)	- (1657,8)
12	2414 (2411,4)	2010 (2002,3)	3274 -	2235 (2239,1)	2552 (2562,1)	- (2196,1)	- (1787)
13	- (2540,5)	2287 (2282,2)	-	2600 (2605,1)	- (2648,2)	- (2282,2)	- (1851,6)

Tabela B.5 (Continuação): *Raias espectrais de Lizandra, para as vogais "a", "e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
14	2711 (2712,8)	2570 (2583,6)	-	2912 (2992,7)	2858 (2863,5)	- (2432,9)	- (1894,6)
15	- (2885)	2876 (2863,5)	-	3319 -	- (3164,9)	- (2583,6)	- (2002,3)
16	3010 (3014,2)	3126 (3121,9)	-	-	-	- (2712,8)	- (2045,4)
17	3310 -	-	-	-	-	- (2863,5)	- (2088,4)
18	-	-	-	-	-	- (3014,2)	- (2260,7)
19	-	-	-	-	-	- (3121,9)	- (2303,7)
20	-	-	-	-	-	-	- (2368,3)
21	-	-	-	-	-	-	- (2454,4)
22	-	-	-	-	-	-	- (2497,5)
23	-	-	-	-	-	-	- (2648,2)
24	-	-	-	-	-	-	- (2820,4)
25	-	-	-	-	-	-	- (2971,1)
26	-	-	-	-	-	-	- (3229,5)

Tabela B.6: *Raias espectrais de Paulo Freitas, para as vogais "a", "e" ,incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

Picos (Hz)	A	Ê	É	I	Ô	Ó	U
	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	59 -	59 -	59 -	59 -	59 -	59 -	59 -
2	123 (129,2)	141 (150,71)	125 (129,2)	144 (150,71)	144 (150,71)	142 (150,71)	149 (150,71)
3	253 (258,4)	278 (279,9)	261 (258,4)	291 (301,4)	296 (301,4)	277 (279,9)	314 (323)
4	385 (387,5)	422 (430,6)	400 (409,1)	426 (430,6)	445 (452,1)	424 (430,6)	467 (473,7)
5	515 (516,7)	559 (559,8)	532 (538,3)	575 (581,3)	595 (602,8)	564 (559,8)	622 (624,4)
6	646 (645,9)	704 (710,5)	662 (667,4)	725 (732)	745 (753,6)	705 (710,5)	780 (775,1)
7	776 (775,1)	- (775,1)	795 (796,6)	873 (882,7)	897 (904,3)	850 (839,7)	940 (947,3)
8	901 (904,3)	840 (839,7)	931 (925,8)	1009 (1011,9)	1046 (1055)	990 (990,4)	1092 (1098)
9	1033 (1033,4)	984 (990,4)	1066 (1076,5)	1158 (1162,6)	- (1141,1)	1130 (1141,1)	1247 (1248,7)
10	1164 (1162,6)	- (1076,5)	1198 (1205,7)	1306 (1313,3)	- (1205,7)	- (1270,3)	- (1421)
11	1296 (1291,8)	1126 (1141,1)	1330 (1334,9)	1455 (1464)	- (1313,3)	- (1399,5)	- (1550,2)
12	1428 (1421)	1263 (1205,7)	1464 (1464)	1606 (1614,7)	- (1356,4)	- (1550,2)	- (1851,6)
13	- (1550,2)	- (1270,3)	1604 (1593,2)	1748 (1743,9)	- (1507,1)	- (1700,9)	- (1894,6)
14	- (1700,9)	1407 (1421)	1735 (1743,9)	- (1894,6)	- (1614,7)	- (1830,1)	- (1980,8)
15	- (1808,5)	1551 (1550,2)	1864 (1873,1)	- (2045,4)	- (1657,8)	- (1980,8)	- (2023,8)

Tabela B.6 (Continuação): *Raias espectrais de Paulo Freitas, para as vogais "a", "e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
16	- (1894,6)	1696 (1700,9)	1998 (2002,3)	- (2196,1)	- (1743,9)	- (2131,5)	- (2174,5)
17	- (1937,7)	- (1851,6)	- (2131,5)	- (2346,8)	- (1808,5)	- (2260,7)	- (2217,6)
18	- (2066,9)	- (1959,2)	- (2260,7)	- (2497,5)	- (1916,2)	- (2411,4)	- (2325,2)
19	- (2196,1)	- (2109,9)	- (2411,4)	- (2648,2)	- (1959,2)	- (2540,5)	- (2368,3)
20	- (2346,8)	- (2260,7)	- (2540,5)	- (2777,4)	- (2023,8)	- (2669,7)	- (2519)
21	- (2454,4)	- (2389,8)	- (2669,7)	- (2928,1)	- (2109,9)	- (2842)	- (2669,7)
22	- (2583,6)	- (2519)	- (2798,9)	- (3057,3)	- (2260,7)	- (2971,1)	- (2798,9)
23	- (2712,8)	- (2562,1)	- (2928,1)	- (3143,4)	- (2389,8)	- (3100,3)	- (2842)
24	- (2863,5)	- (2691,3)	- (3078,8)	- (3208)	- (2562,1)	- (3229,5)	- (2992,7)
25	- (2928,1)	- (2842)	- (3208)	-	- (2691,3)	-	- (3100,3)
26	- (2992,7)	- (2928,1)	-	-	- (2842)	-	- (3143,4)
27	- (3057,3)	- (3078,8)	-	-	- (2992,7)	-	-
28	- (3100,3)	- (3229,5)	-	-	- (3057,3)	-	-
29	-	-	-	-	- (3100,3)	-	-
30	-	-	-	-	- (3143,4)	-	-

Tabela B.7: *Raias espectrais de Paulo Martins, para as vogais "a", "e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	59 -	59 -	59 -	59 -	59 -	59 -	59 -
2	119 -	120 -	162 -	145 -	156 -	151 -	165 -
3	167 (172,2)	167 (172,2)	319 (323)	303 (301,4)	229 (236,8)	229 (236,8)	323 (323)
4	231 (236,8)	234 (236,8)	486 (495,2)	458 (452,1)	318 (323)	317 (323)	487 (495,2)
5	340 (344,5)	338 (344,5)	644 (645,9)	616 (624,4)	475 (473,7)	472 (473,7)	656 (645,9)
6	512 (516,7)	508 (516,7)	809 (818,1)	768 (775,1)	639 (645,9)	636 (645,9)	819 (818,1)
7	683 (689)	678 (667,4)	968 (968,8)	- (925,8)	799 (796,6)	792 (796,6)	985 (990,4)
8	855 (861,2)	842 (839,7)	- (1055)	- (1076,5)	961 (968,8)	951 (947,3)	- (1141,1)
9	1028 (1033,4)	1012 (1011,9)	1133 (1141,1)	- (1141,1)	1121 (1119,6)	1112 (1119,6)	- (1270,3)
10	1198 (1205,7)	- (1098)	1294 (1291,8)	- (1248,7)	- (1270,3)	1268 (1270,3)	- (1313,3)
11	1371 (1377,9)	1180 (1184,2)	1457 (1464)	- (1291,8)	- (1442,5)	- (1442,5)	- (1442,5)
12	1546 (1550,2)	1352 (1356,4)	1621 (1614,7)	- (1377,9)	- (1507,1)	- (1593,2)	- (1485,6)
13	1715 (1722,4)	- (1442,5)	1781 (1787)	- (1550,2)	- (1593,2)	- (1657,8)	- (1550,2)
14	1889 (1894,6)	1524 (1528,6)	1939 (1937,7)	- (1614,7)	- (1743,9)	- (1743,9)	- (1614,7)
15	2059 (2066,9)	1697 (1700,9)	- (2109,9)	- (1700,9)	- (1787)	- (1916,2)	- (1808,5)
16	- (2239,1)	1866 (1873,1)	- (2260,7)	- (1787)	- (1937,7)	- (2088,4)	- (1894,6)

Tabela B.7 (Continuação): *Raias espectrais de Paulo Martins, para as vogais "a", "e",incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
17	- (2411,4)	- (2045,4)	- (2432,9)	- (1851,6)	- (2066,9)	- (2239,1)	- (1959,2)
18	- (2583,6)	- (2196,1)	- (2583,6)	- (2002,3)	- (2109,9)	- (2282,2)	- (2045,4)
19	- (2755,8)	- (2368,3)	- (2755,8)	- (2153)	- (2153)	- (2389,8)	- (2088,4)
20	- (2928,1)	- (2540,5)	- (2906,6)	- (2303,7)	- (2260,7)	- (2562,1)	- (2153)
21	- (3078,8)	- (2712,8)	- (3078,8)	- (2454,4)	- (2411,4)	- (2712,8)	- (2217,6)
22	- (3208)	- (2885)	- (3229,5)	- (2626,7)	- (2540,5)	- (2863,5)	- (2282,2)
23	-	- (3057,3)	-	- (2777,4)	- (2712,8)	- (3035,7)	- (2325,2)
24	-	- (3208)	-	- (2928,1)	- (2755,8)	- (3186,4)	- (2411,4)
25	-	-	-	- (3078,8)	- (2863,5)	-	- (2454,4)
26	-	-	-	- (3229,5)	- (2906,6)	-	- (2519)
27	-	-	-	-	- (3014,2)	-	- (2626,7)
28	-	-	-	-	- (3186,4)	-	- (2734,3)
29	-	-	-	-	-	-	- (2777,4)
30	-	-	-	-	-	-	- (2928,1)
31	-	-	-	-	-	-	- (2992,7)
32	-	-	-	-	-	-	- (3035,7)
33	-	-	-	-	-	-	- (3121,9)
34	-	-	-	-	-	-	- (3164,9)
35	-	-	-	-	-	-	- (3229,5)

Tabela B.8: *Raias espectrais de Ricardo, para as vogais "a", "e", incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
1	58 -	58 -	58 -	59 -	59 -	59 -	59 -
2	106 (107,65)	106 (107,65)	105 (107,65)	123 (129,2)	122 (129,2)	112 (107,65)	124 (129,2)
3	213 (215,3)	212 (215,3)	212 (215,3)	254 (258,4)	242 (236,8)	230 (236,8)	265 (258,4)
4	319 (323)	320 (323)	318 (323)	383 (387,5)	364 (366)	341 (344,5)	402 (409,1)
5	427 (430,6)	426 (430,6)	425 (430,6)	512 (516,7)	489 (495,2)	461 (452,1)	533 (538,3)
6	535 (538,3)	534 (538,3)	531 (538,3)	596 (602,8)	614 (624,4)	574 (581,3)	664 (667,4)
7	642 (645,9)	595 (602,8)	639 (645,9)	641 (645,9)	733 (732)	686 (689)	803 (796,6)
8	747 (753,6)	643 (645,9)	744 (753,6)	705 (710,5)	858 (861,2)	805 (796,6)	940 (947,3)
9	854 (861,2)	703 (710,5)	852 (861,2)	770 (775,1)	983 (990,4)	920 (925,8)	- (1011,9)
10	961 (968,8)	750 (753,6)	959 (968,8)	833 (839,7)	- (1098)	1028 (1033,4)	- (1076,5)
11	1066 (1076,5)	852 (861,2)	1064 (1011,9)	899 (904,3)	- (1227,2)	1151 (1162,6)	- (1141,1)
12	1176 (1184,2)	- (904,3)	- (1055)	- (968,8)	- (1356,4)	- (1270,3)	- (1205,7)
13	1282 (1291,8)	- (968,8)	1171 (1162,6)	- (1033,4)	- (1464)	- (1377,9)	- (1270,3)
14	1387 (1399,5)	- (1011,9)	- (1270,3)	- (1076,5)	- (1507,1)	- (1507,1)	- (1313,3)
15	1493 (1485,6)	- (1076,5)	- (1377,9)	- (1162,6)	- (1550,2)	- (1614,7)	- (1377,9)
16	- (1593,2)	- (1141,1)	- (1485,6)	- (1205,7)	- (1614,7)	- (1722,4)	- (1507,1)
17	- (1700,9)	- (1184,2)	- (1593,2)	- (1291,8)	- (1722,4)	- (1830,1)	- (1550,2)
18	- (1830,1)	- (1291,8)	- (1700,9)	- (1377,9)	- (1808,5)	- (1937,7)	- (1614,7)

Tabela B.8 (Continuação): *Raias espectrais de Ricardo, para as vogais "a", "e" , incluindo acentuação (grave/agudo), "i", "o", "u" (valores em Hz).*

	A	Ê	É	I	Ô	Ó	U
Picos (Hz)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)	Audacity (Estimador)
19	- (1916,2)	- (1399,5)	- (1808,5)	- (1421)	- (1851,6)	- (1980,8)	- (1873,1)
20	- (2023,8)	- (1507,1)	- (1916,2)	- (1507,1)	- (1916,2)	- (2045,4)	- (1980,8)
21	- (2153)	- (1614,7)	- (2023,8)	- (1550,2)	- (1980,8)	- (2153)	- (2045,4)
22	- (2239,1)	- (1722,4)	- (2131,5)	- (1614,7)	- (2045,4)	- (2217,6)	- (2088,4)
23	- (2346,8)	- (1830,1)	- (2239,1)	- (1679,3)	- (2088,4)	- (2282,2)	- (2217,6)
24	- (2476)	- (1937,7)	- (2346,8)	- (1743,9)	- (2153)	- (2325,2)	- (2282,2)
25	- (2562,1)	- (2045,4)	- (2454,4)	- (1808,5)	- (2217,6)	- (2411,4)	- (2389,8)
26	- (2691,3)	- (2153)	- (2562,1)	- (1937,7)	- (2346,8)	- (2454,4)	- (2454,4)
27	- (2755,8)	- (2260,7)	- (2669,7)	- (2066,9)	- (2389,8)	- (2540,5)	- (2562,1)
28	- (2863,5)	- (2368,3)	- (2777,4)	- (2196,1)	- (2476)	- (2669,7)	- (2648,2)
29	- (2949,6)	- (2476)	- (2885)	- (2303,7)	- (2583,6)	- (2798,9)	- (2691,3)
30	- (2992,7)	- (2583,6)	- (2971,1)	- (2432,9)	- (2691,3)	- (2885)	- (2755,8)
31	- (3057,3)	- (2691,3)	- (3078,8)	- (2562,1)	- (2820,4)	- (2992,7)	- (2820,4)
32	- (3100,3)	- (2798,9)	- (3186,4)	- (2691,3)	- (2885)	- (3100,3)	- (2885)
33	- (3208)	- (2906,6)	-	- (2820,4)	- (2928,1)	- (3143,4)	- (2928,1)
34	-	- (3014,2)	-	- (2949,6)	- (2992,7)	- (3229,5)	- (3057,3)
35	-	- (3057,3)	-	- (3078,8)	- (3057,3)	-	- (3186,4)
36	-	- (3121,9)	-	- (3208)	- (3186,4)	-	-

ANEXO C – CURVAS DE CORRELAÇÃO DE CADA VOGAL PARA CADA LOCUTOR OBTIDOS PELO AUDACITY 1.3® E PELO ESTIMADOR.

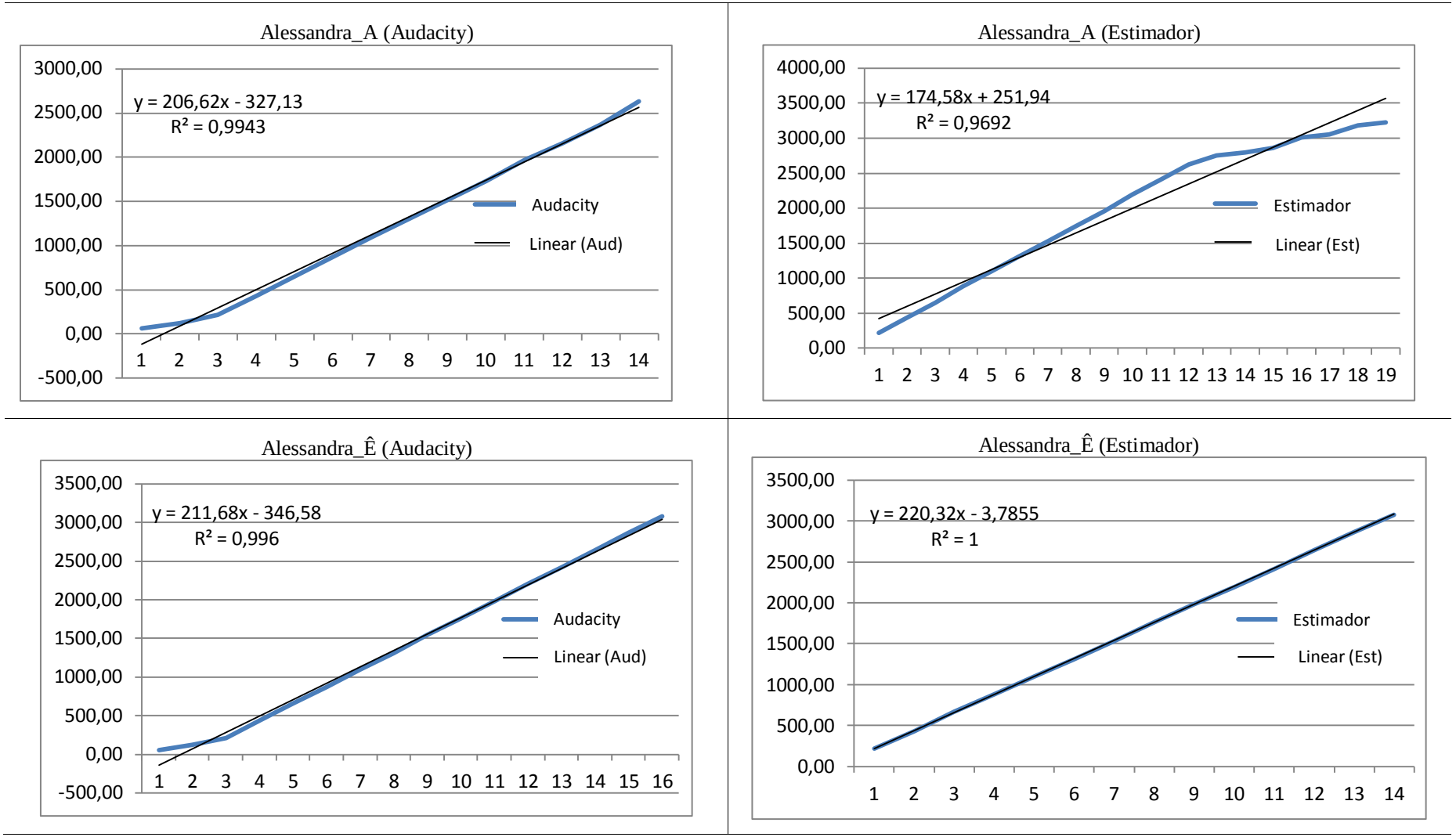


Figura C.1: *Curvas de correlação das vogais "A" e "Ê" para a locutora Alessandra.*

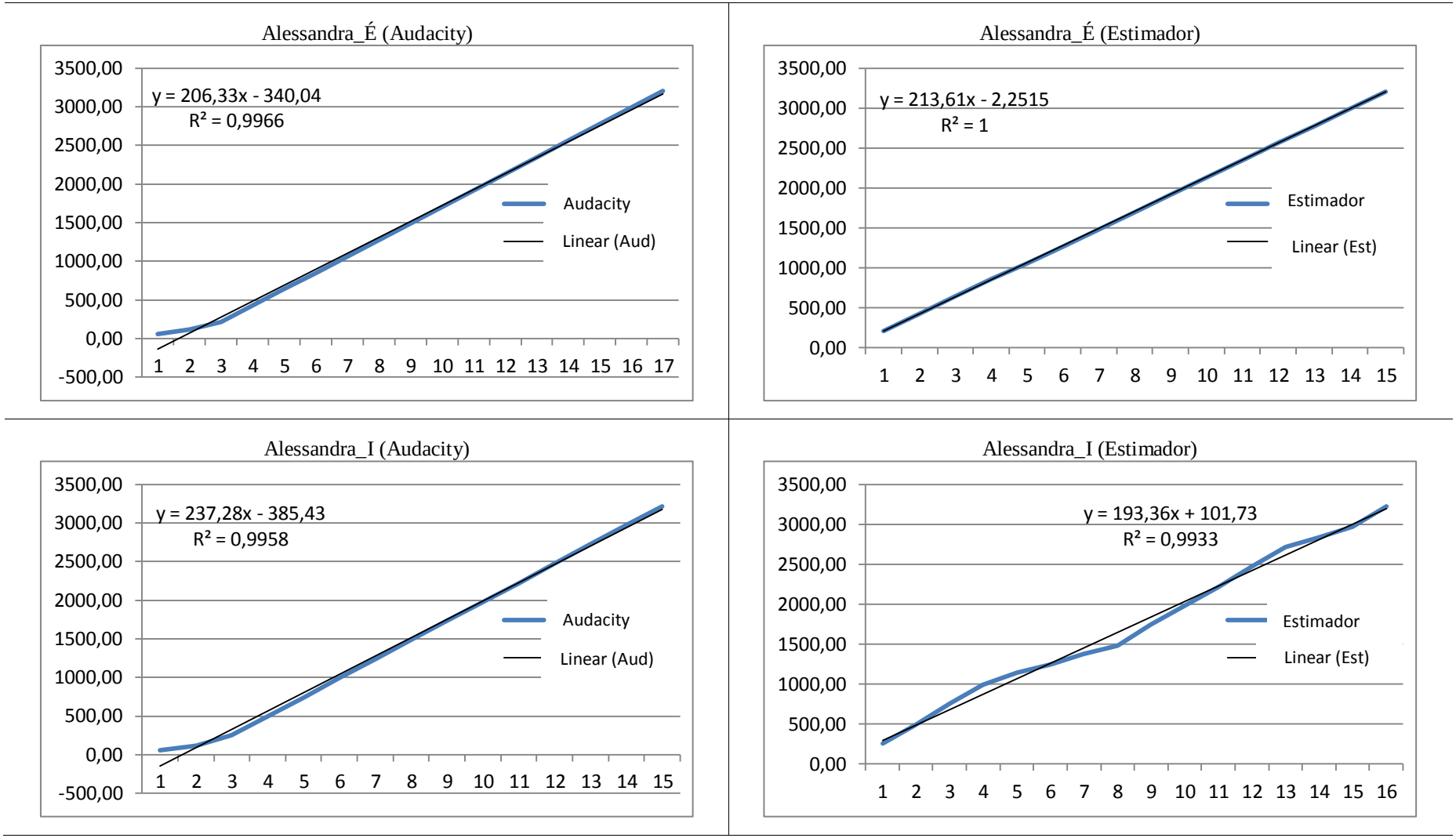


Figura C.2: *Curvas de correlação das vogais "É" e "I" para a locutora Alessandra.*

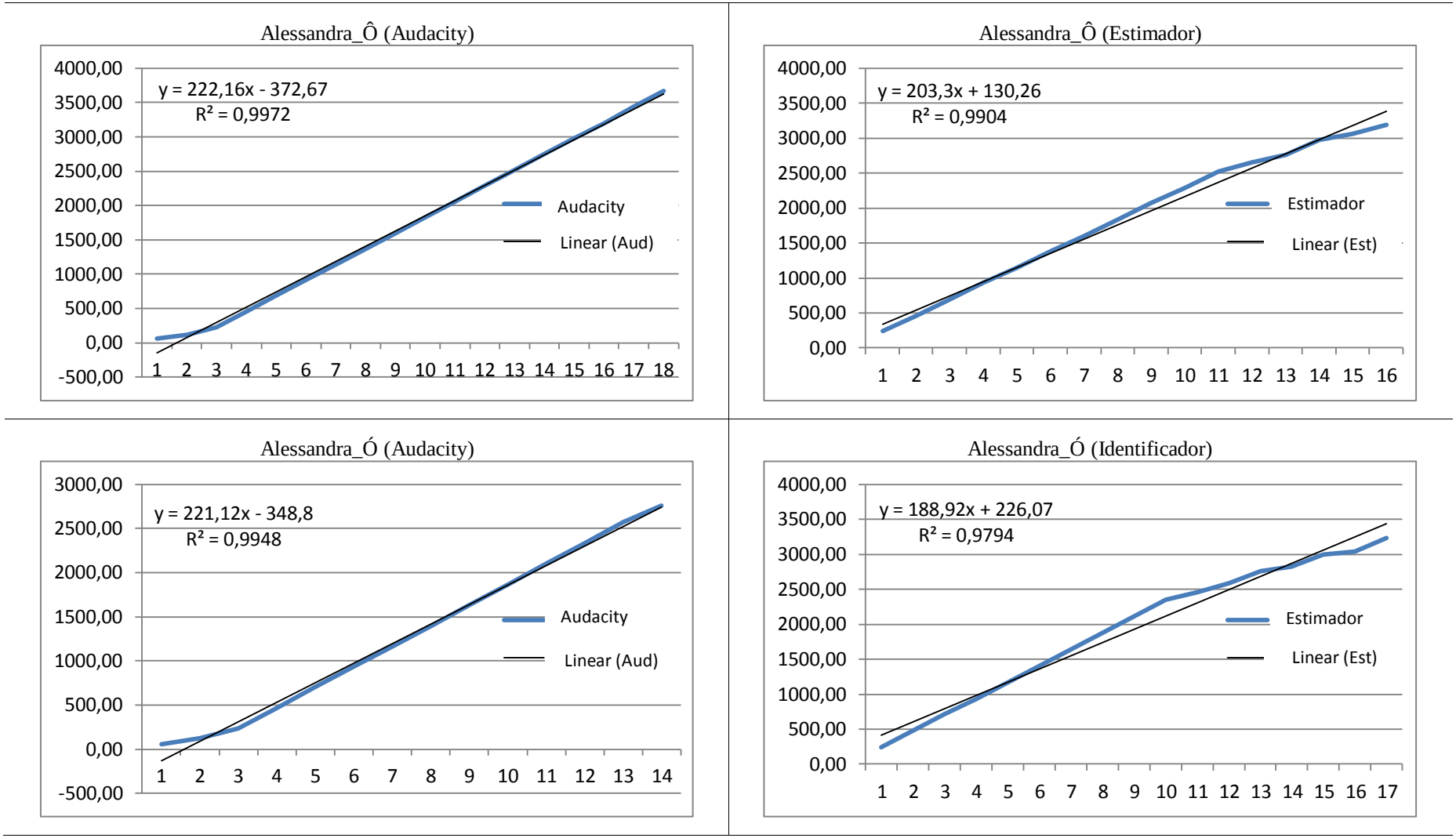
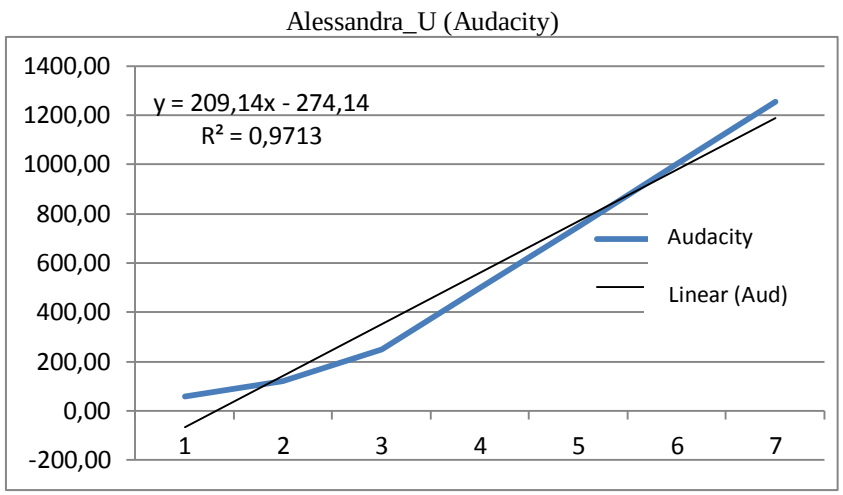


Figura C.3: *Curvas de correlação das vogais "Ô" e "Ó" para a locutora Alessandra.*



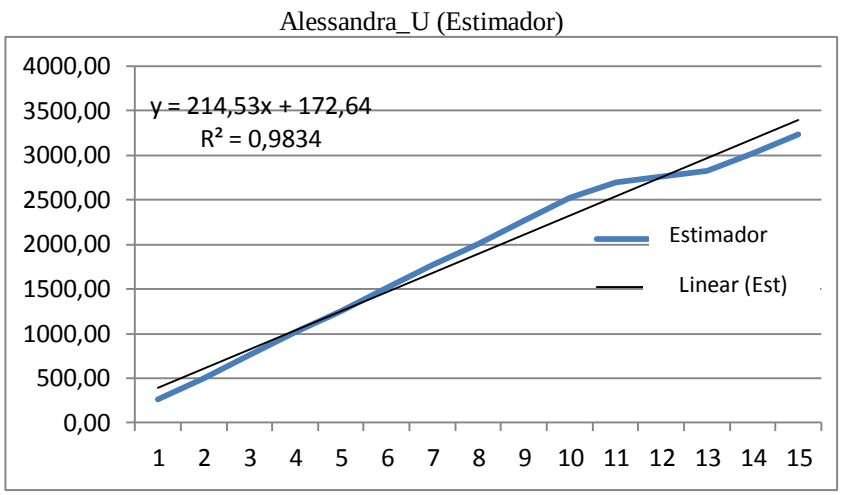


Figura C.4: *Curvas de correlação da vogal "U" para a locutora Alessandra.*

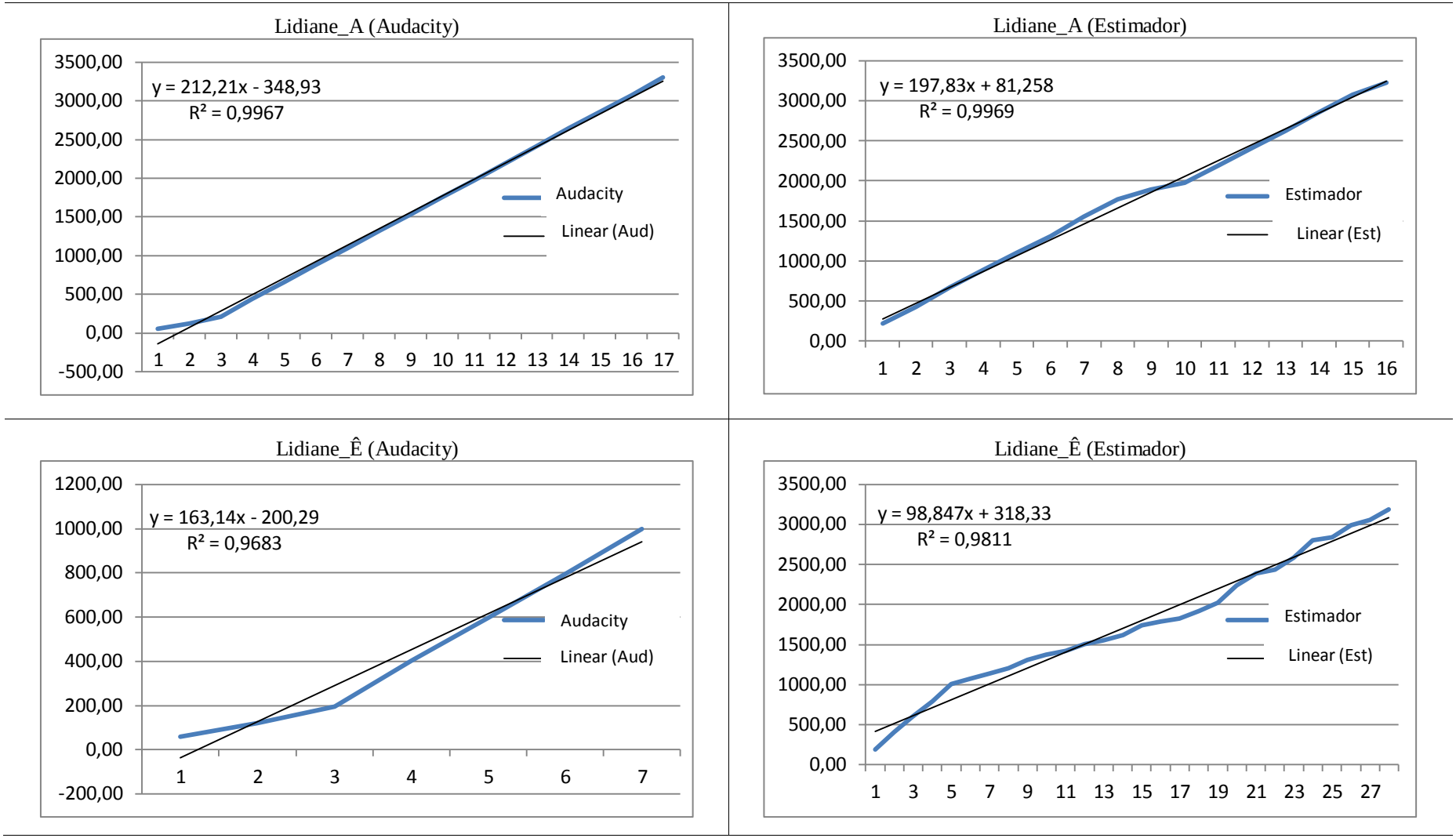


Figura C.5: *Curvas de correlação das vogais "A" e "Ê" para a locutora Lidiane.*

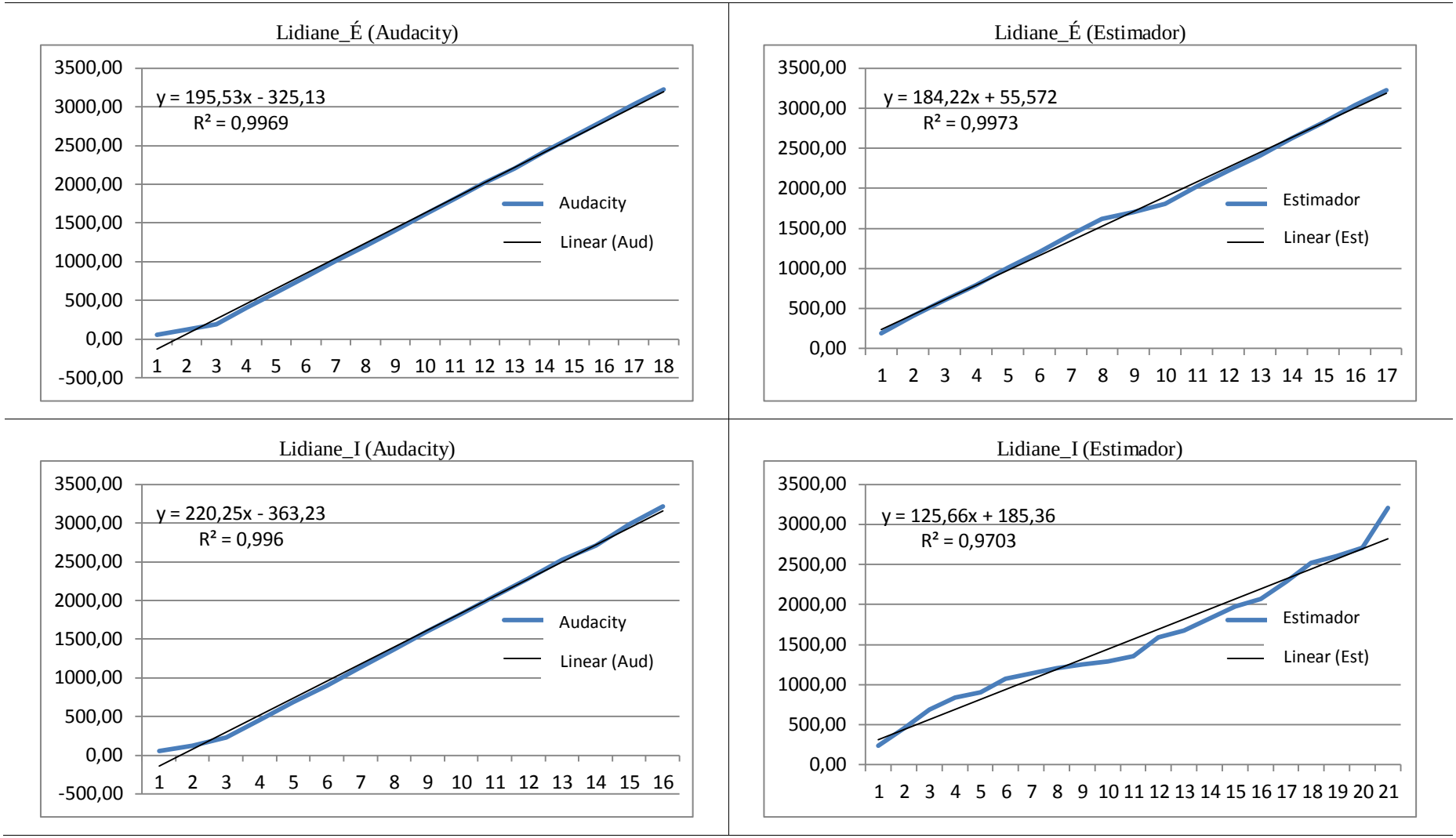


Figura C.6: *Curvas de correlação das vogais "É" e "I" para a locutora Lidiane.*

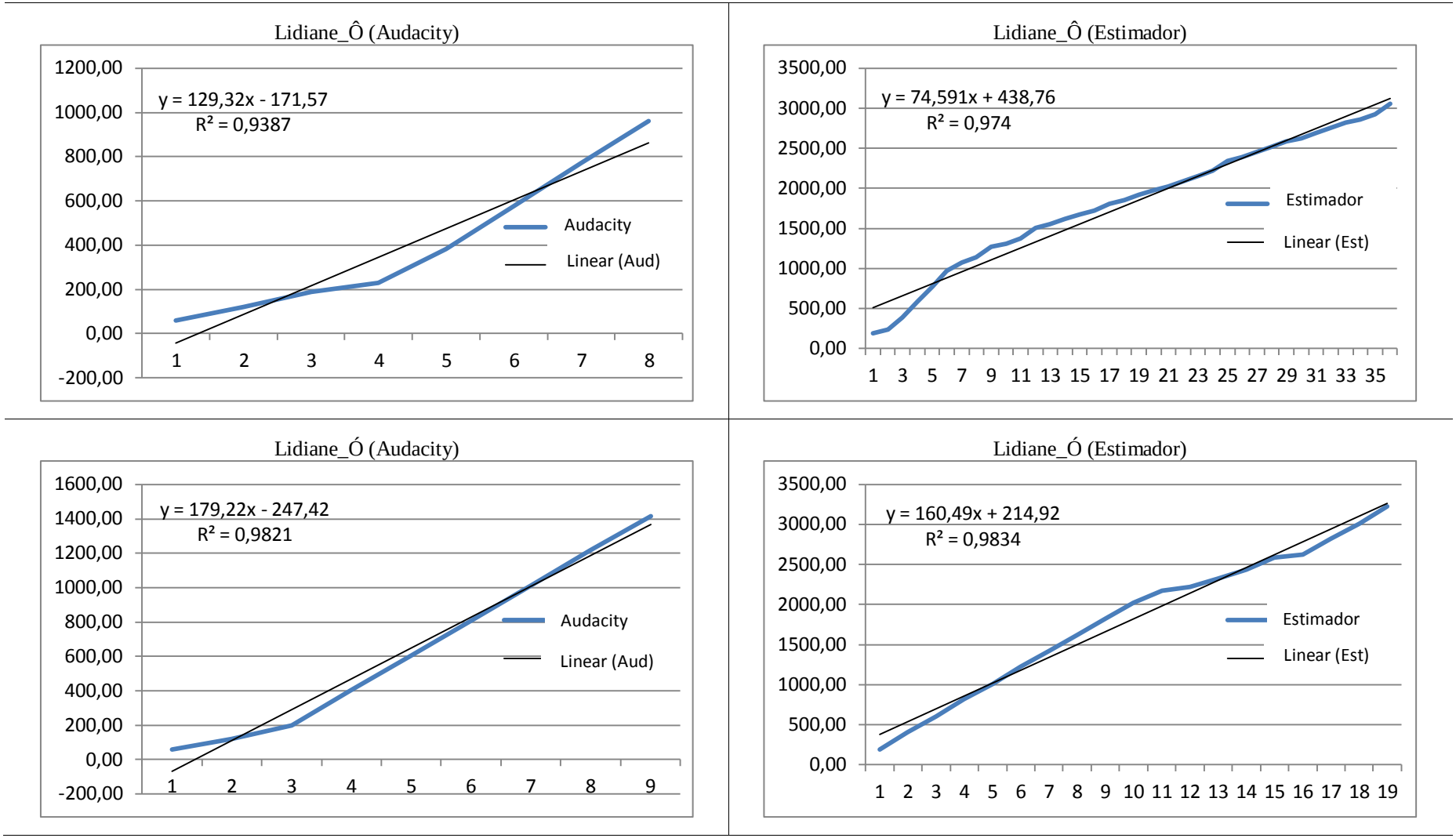


Figura C.7: *Curvas de correlação das vogais "Ô" e "Ó" para a locutora Lidiane.*

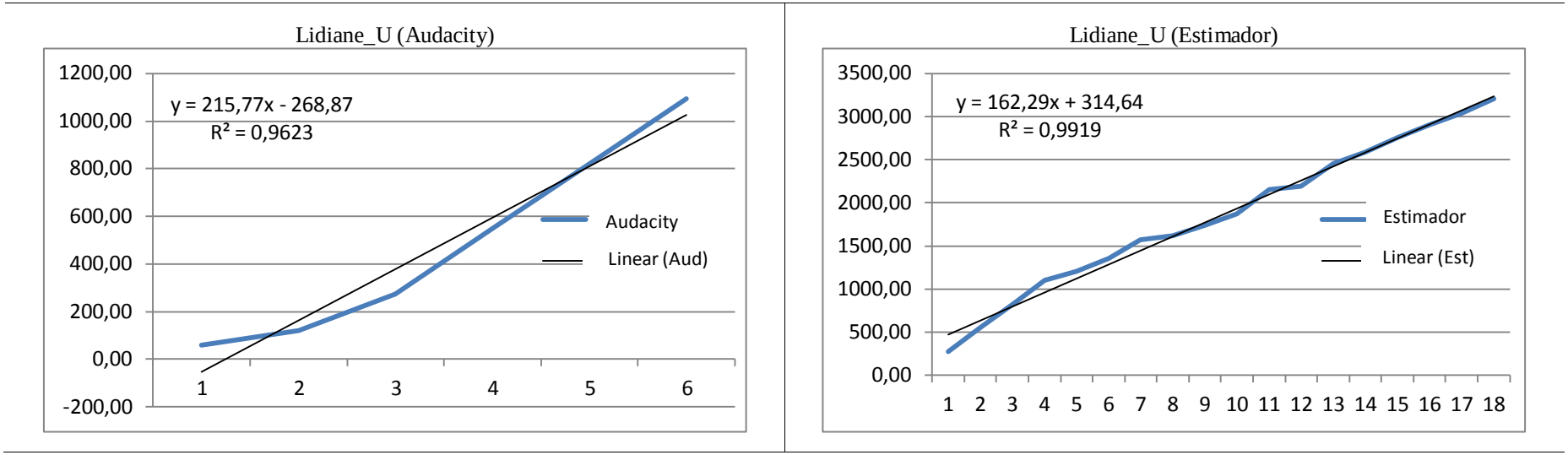


Figura C.8: *Curvas de correlação da vogal "U" para a locutora Lidiane.*

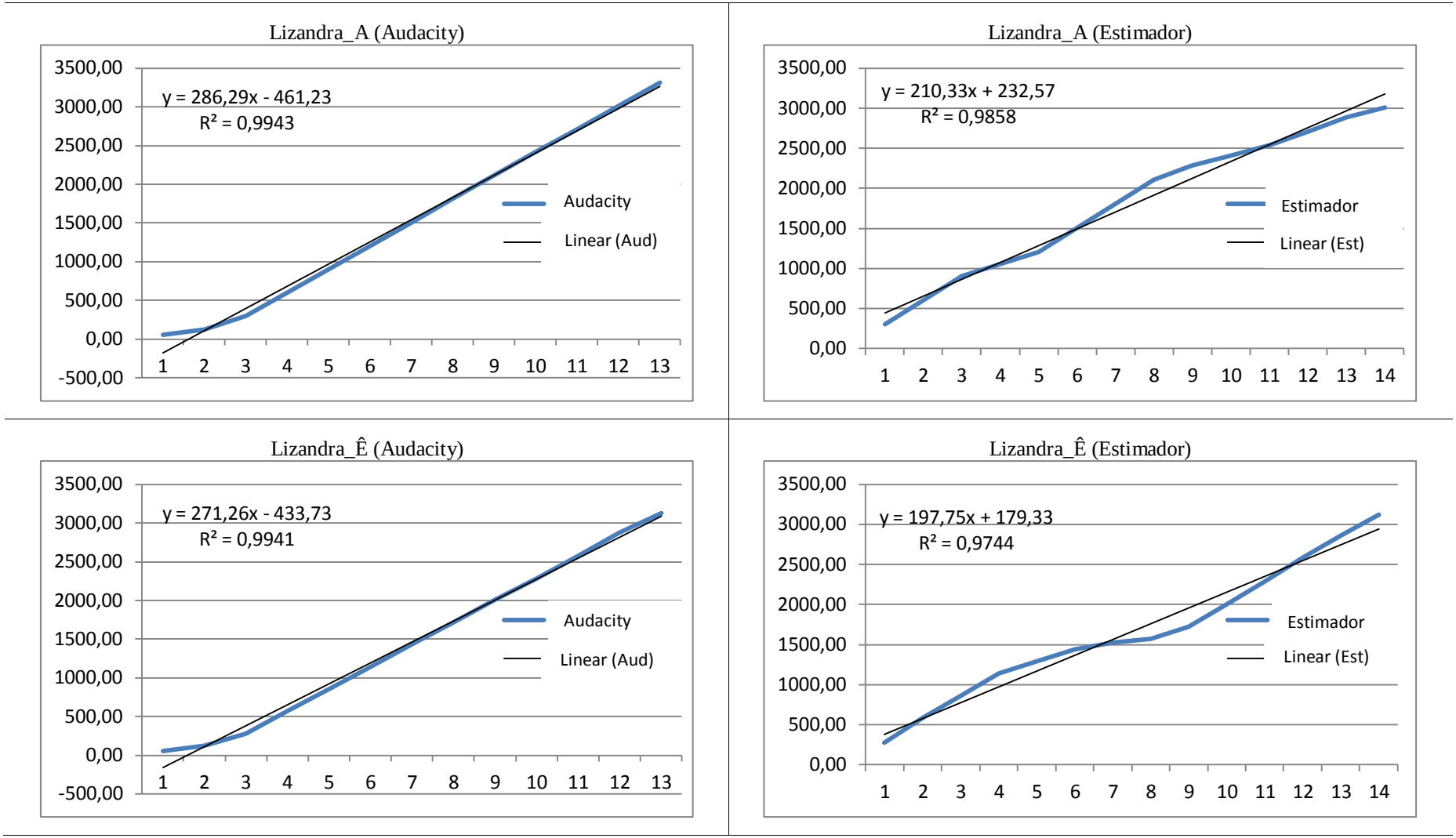


Figura C.9: *Curvas de correlação das vogais "A" e "Ê" para a locutora Lizandra.*

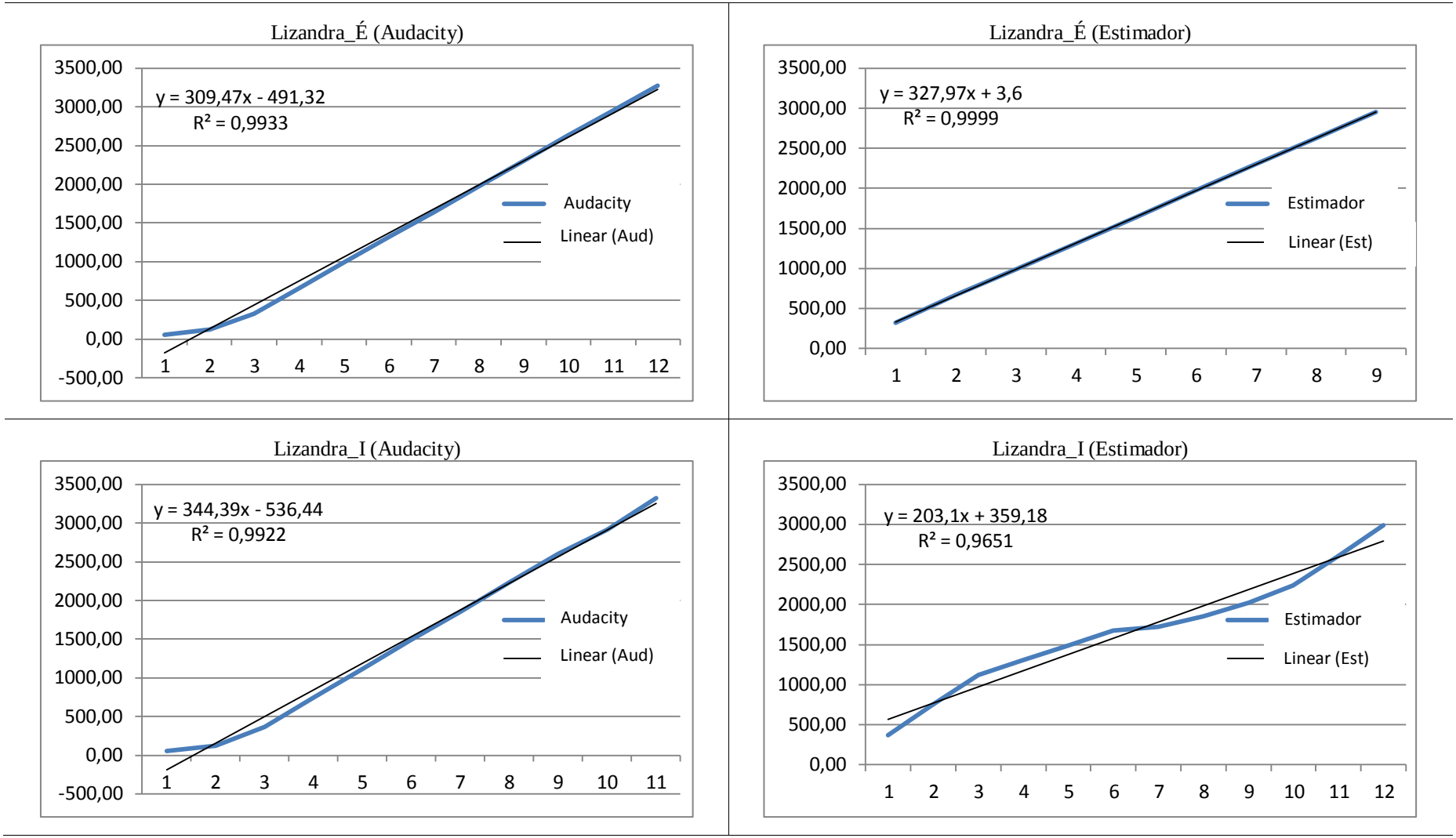


Figura C.10: *Curvas de correlação das vogais "É" e "I" para a locutora Lizandra.*

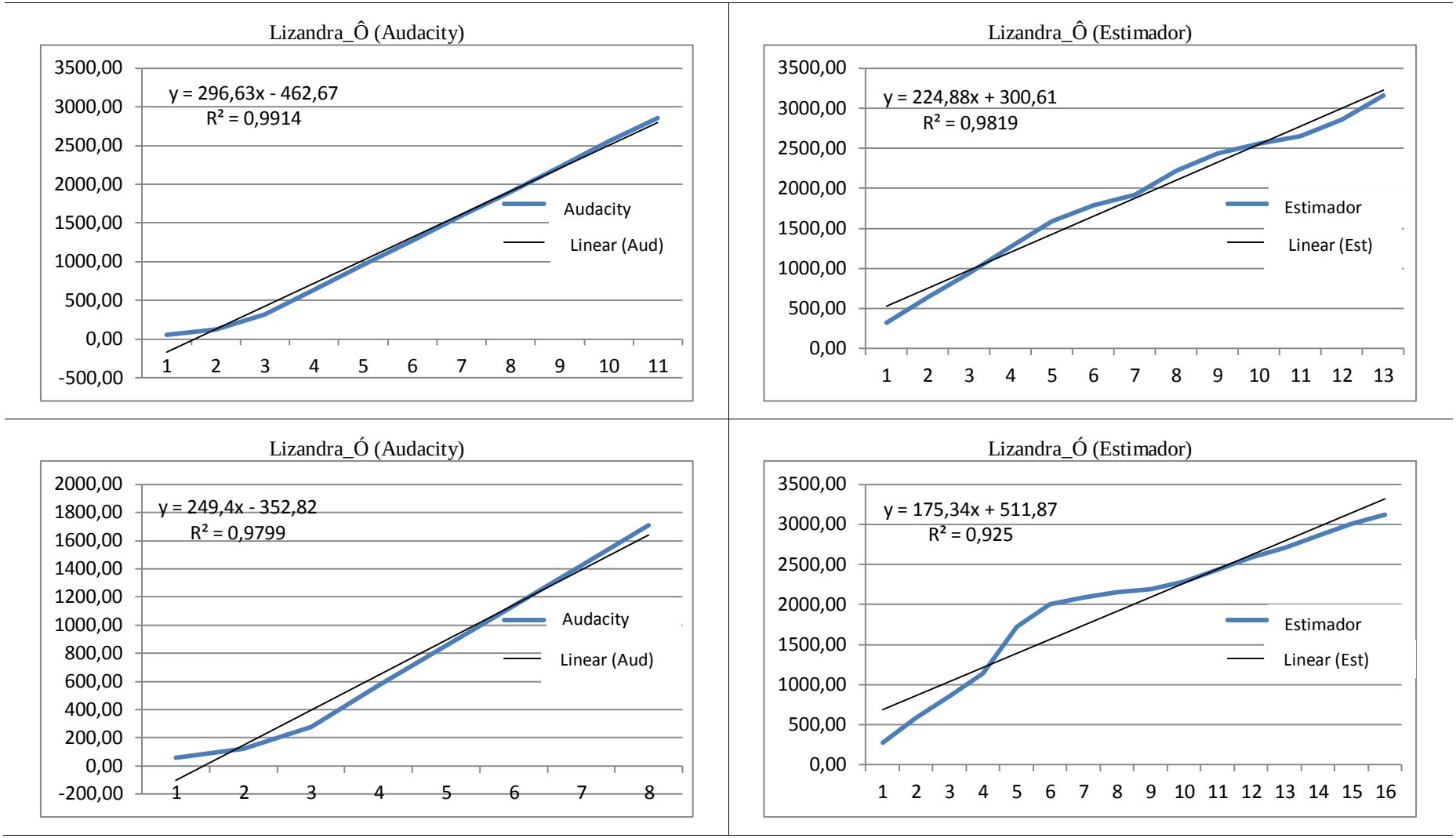


Figura C.11: *Curvas de correlação das vogais "Ô" e "Ó" para a locutora Lizandra.*

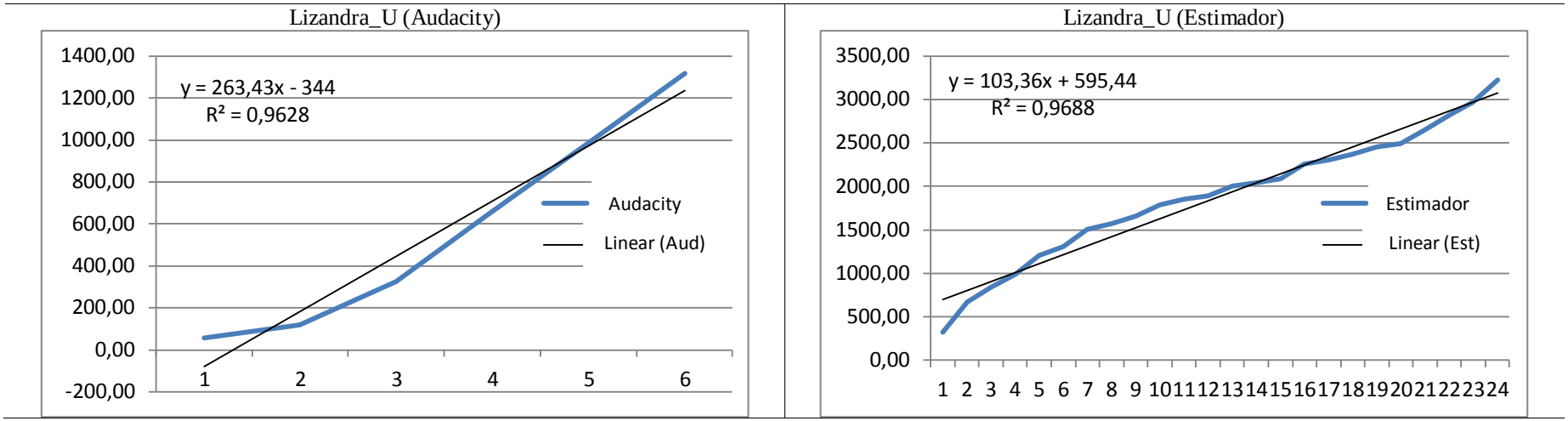


Figura C.12: *Curvas de correlação da vogal "U" para a locutora Lizandra.*

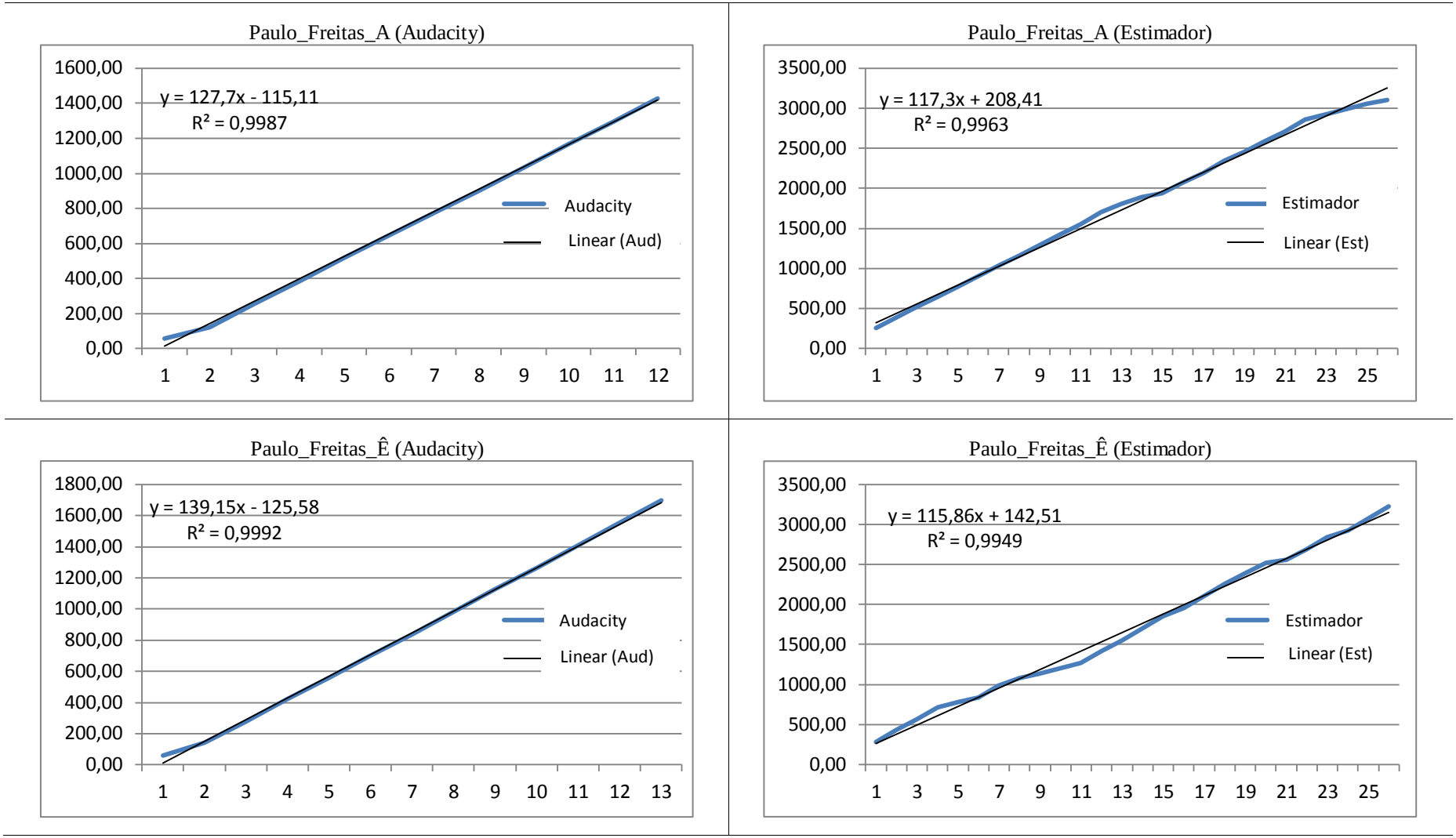


Figura C.13: *Curvas de correlação das vogais "A" e "Ê" para o locutor Paulo Freitas.*

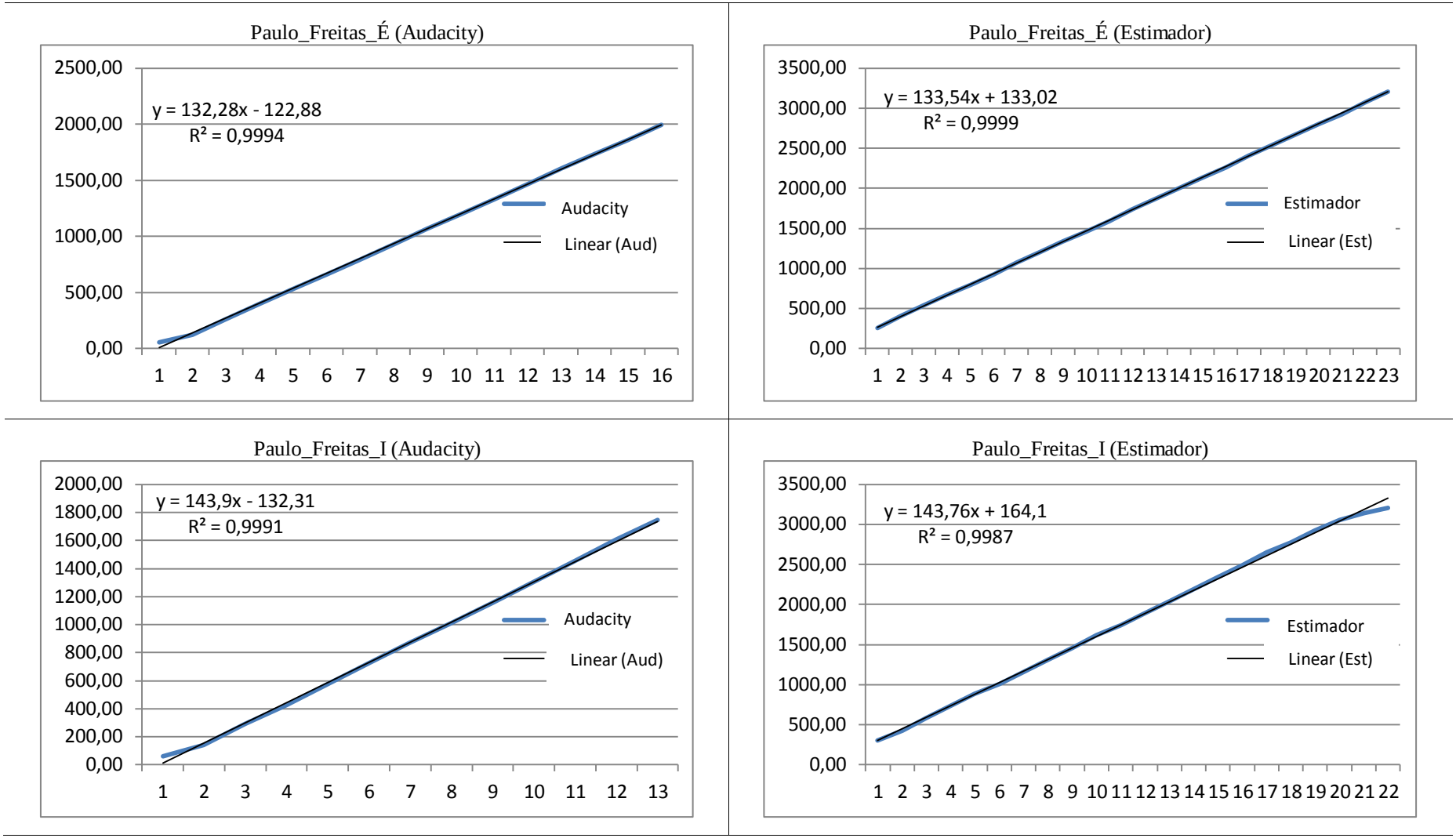


Figura C.14: *Curvas de correlação das vogais "É" e "I" para o locutor Paulo Freitas.*

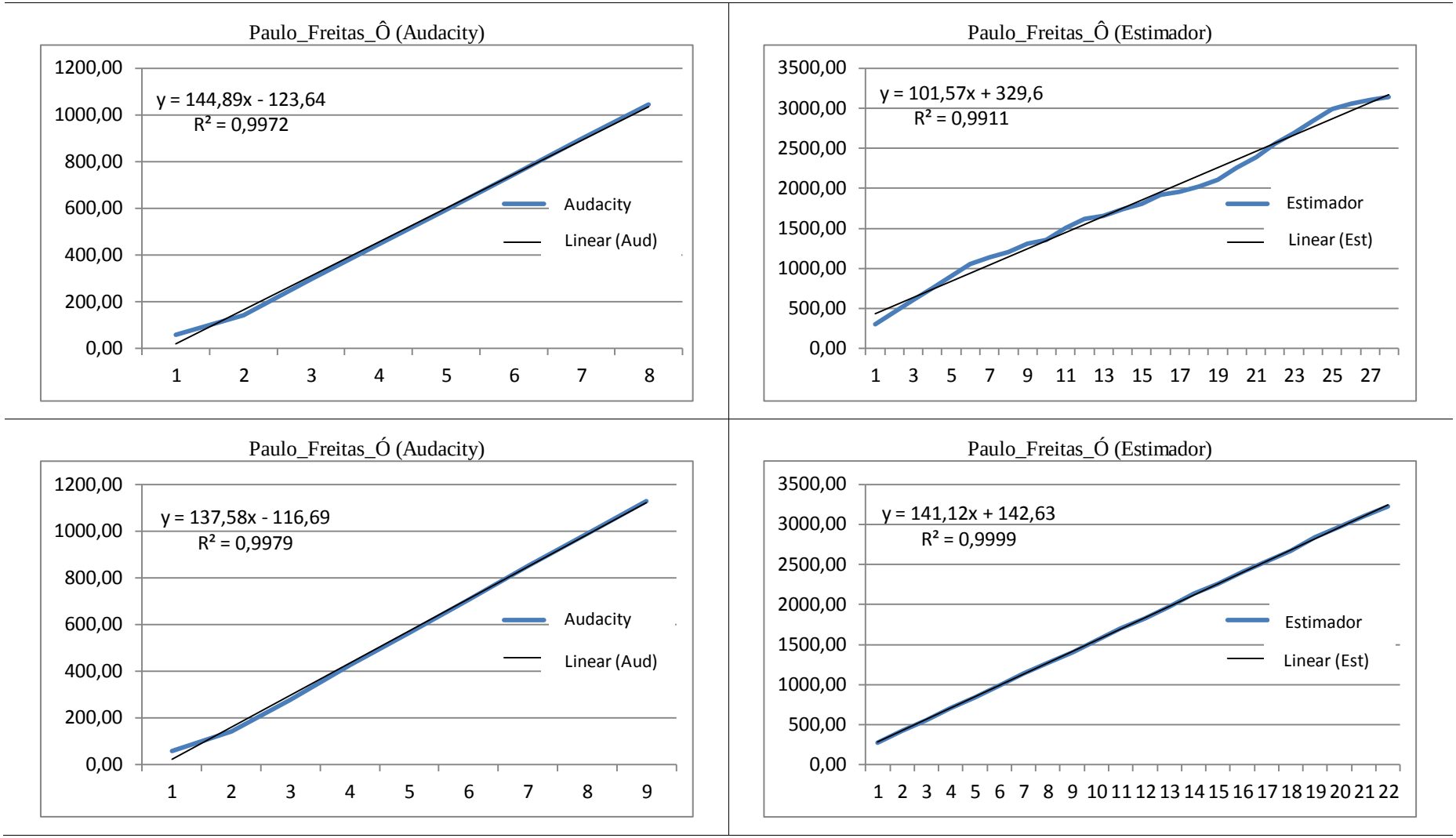


Figura C.15: *Curvas de correlação das vogais "Ô" e "Ó" para o locutor Paulo Freitas.*

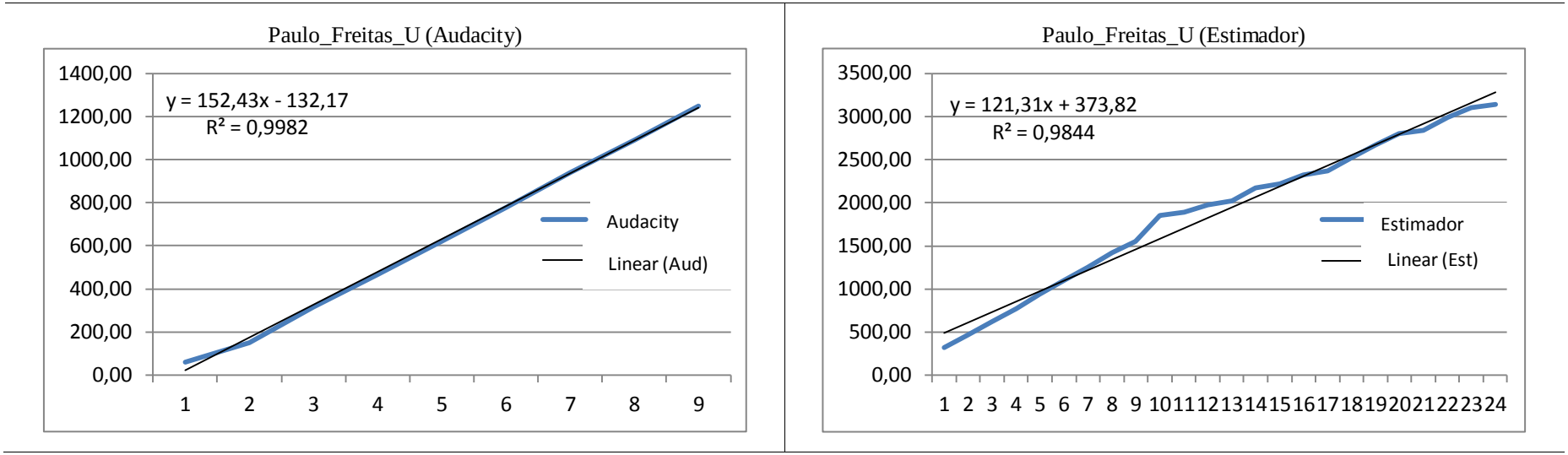


Figura C.16: *Curvas de correlação da vogal "U" para o locutor Paulo Freitas.*

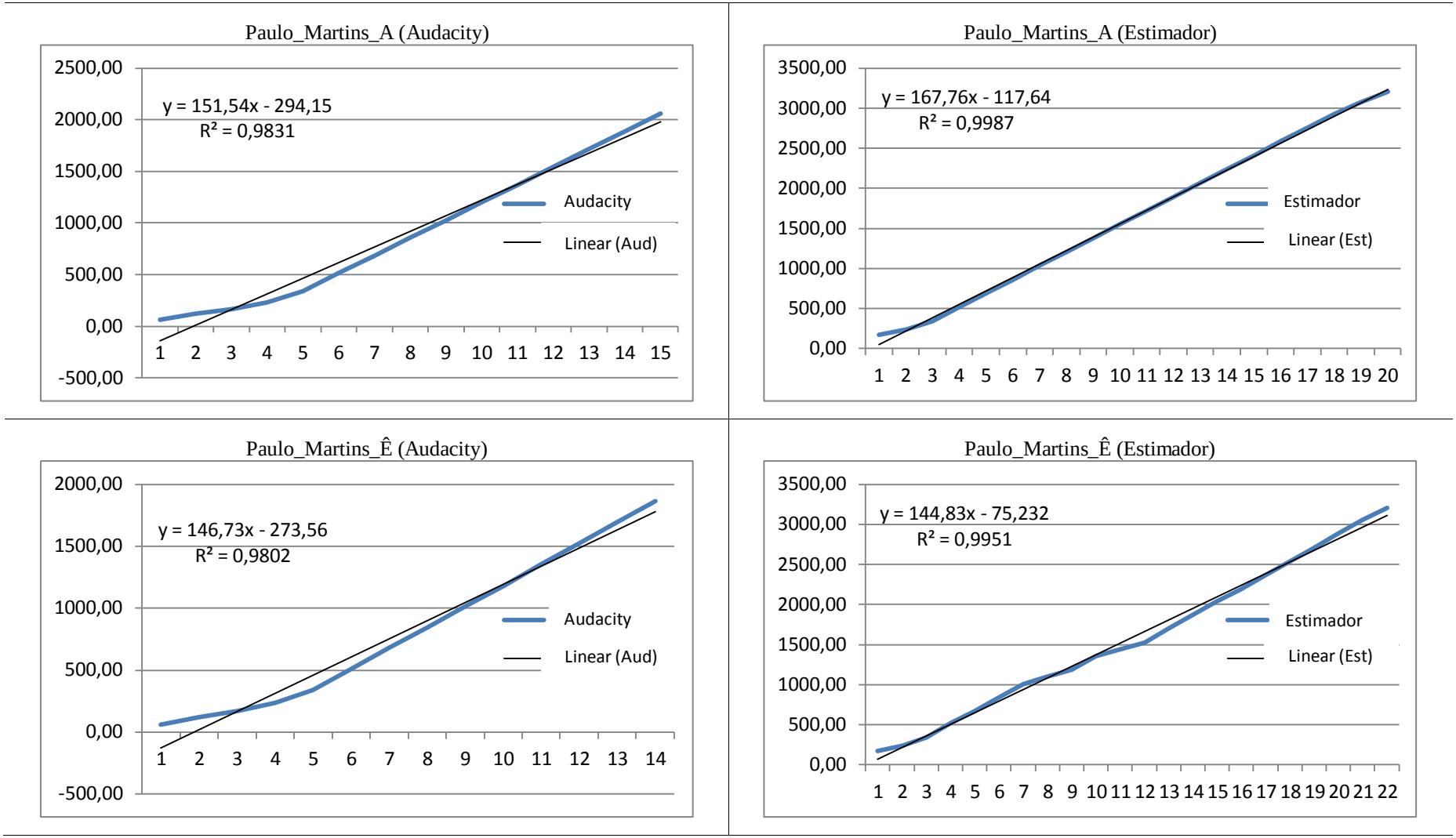


Figura C.17: *Curvas de correlação das vogais "A" e "Ê" para o locutor Paulo Martins.*

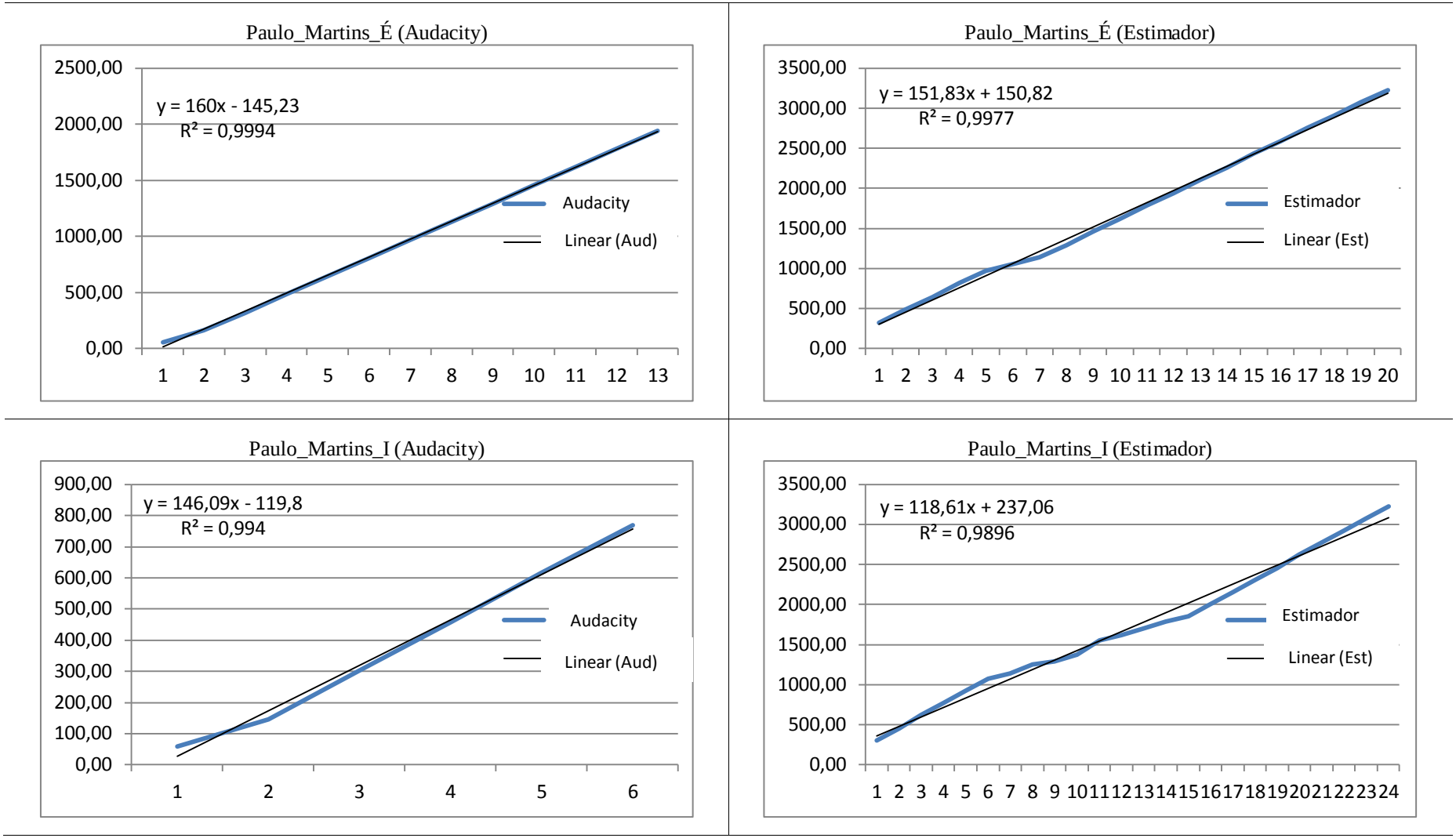


Figura C.18: *Curvas de correlação das vogais "É" e "I" para o locutor Paulo Martins.*

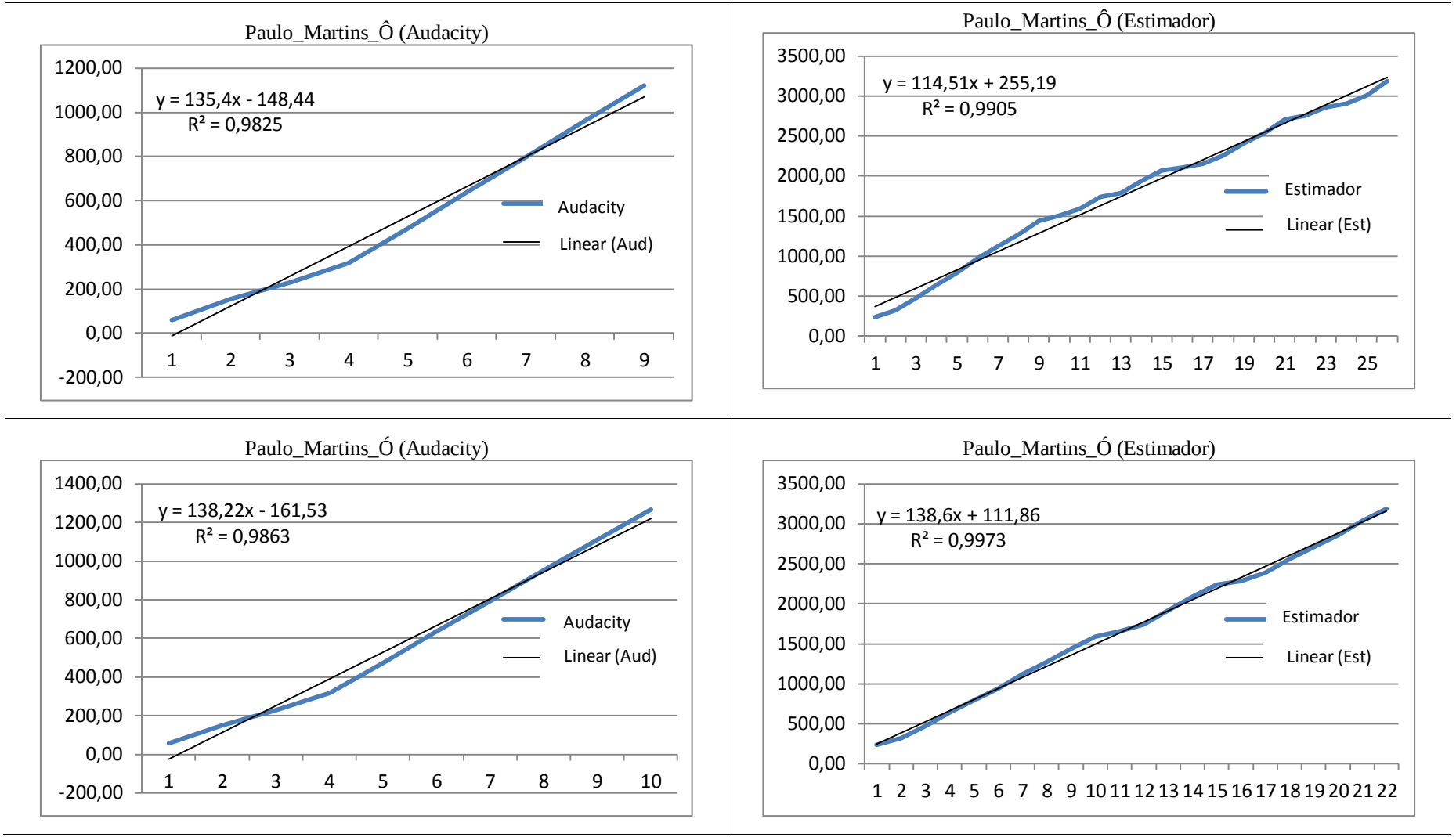


Figura C.19: *Curvas de correlação das vogais "Ô" e "Ó" para o locutor Paulo Martins.*

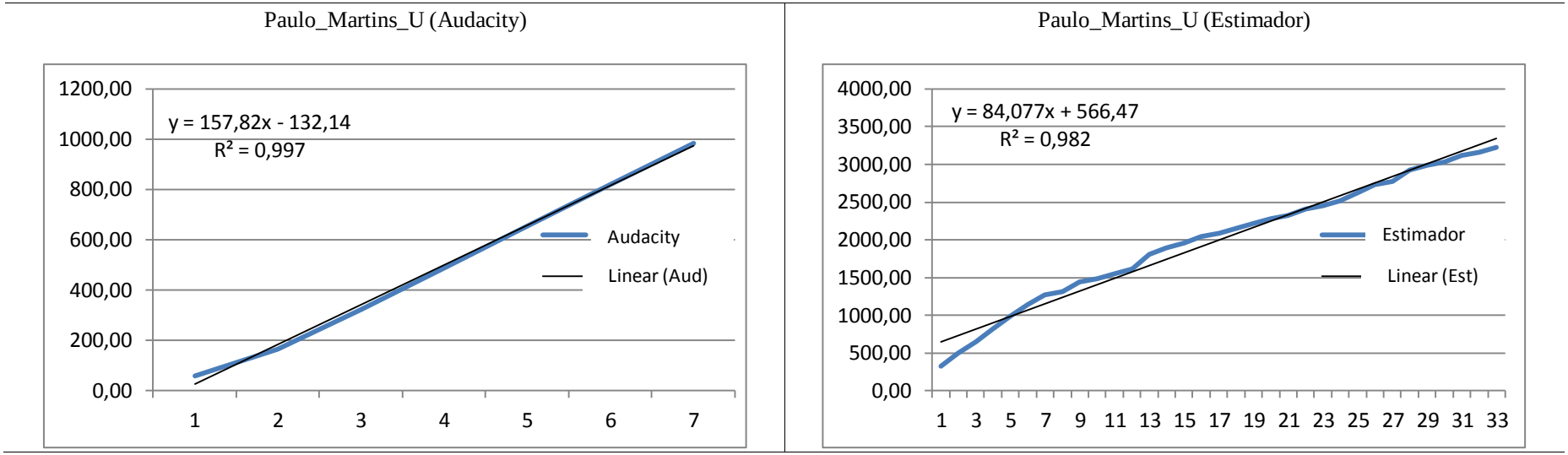


Figura C.20: *Curvas de correlação da vogal "U" para o locutor Paulo Martins.*

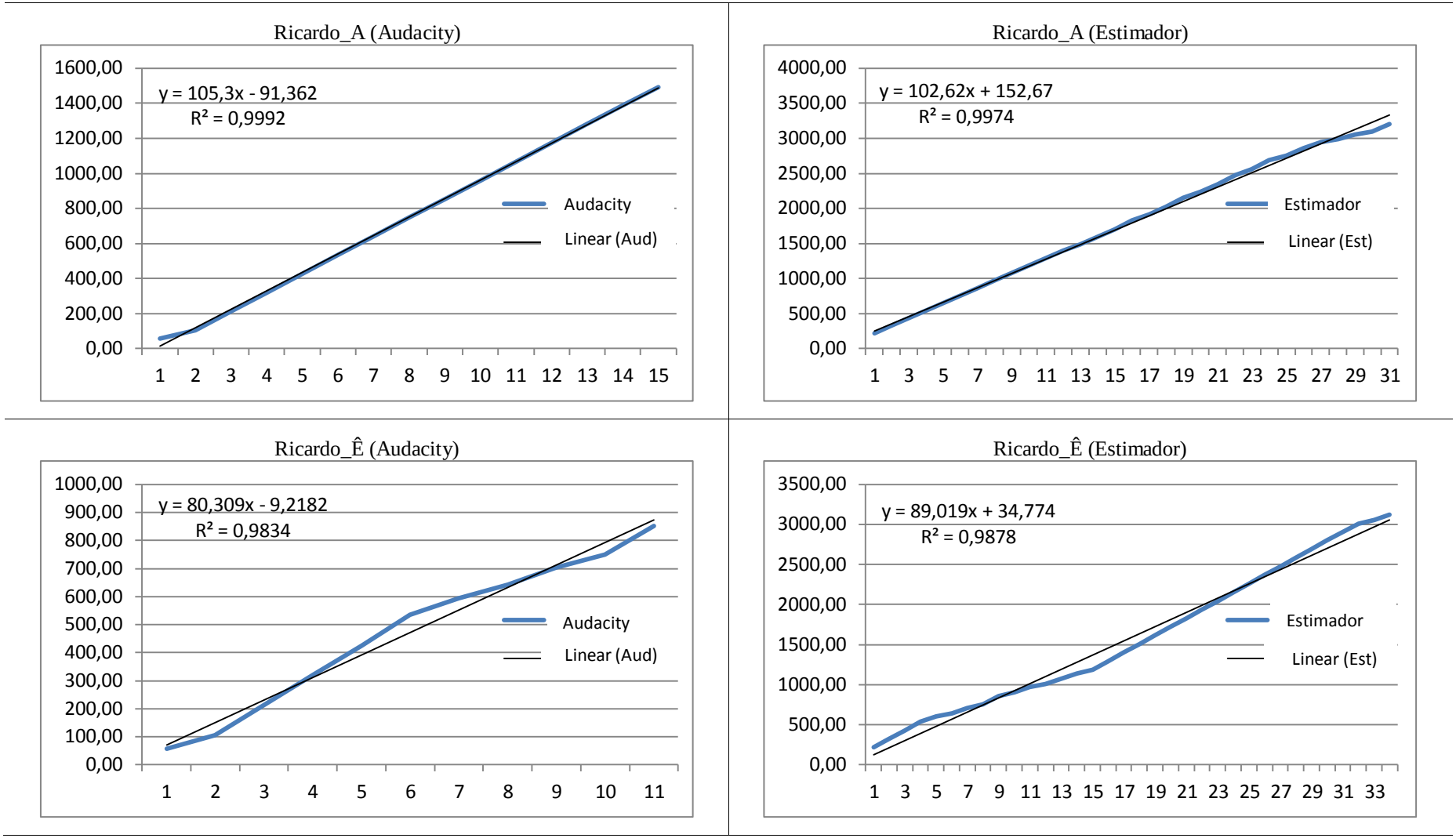


Figura C.21: *Curvas de correlação das vogais "A" e "Ê" para o locutor Ricardo.*

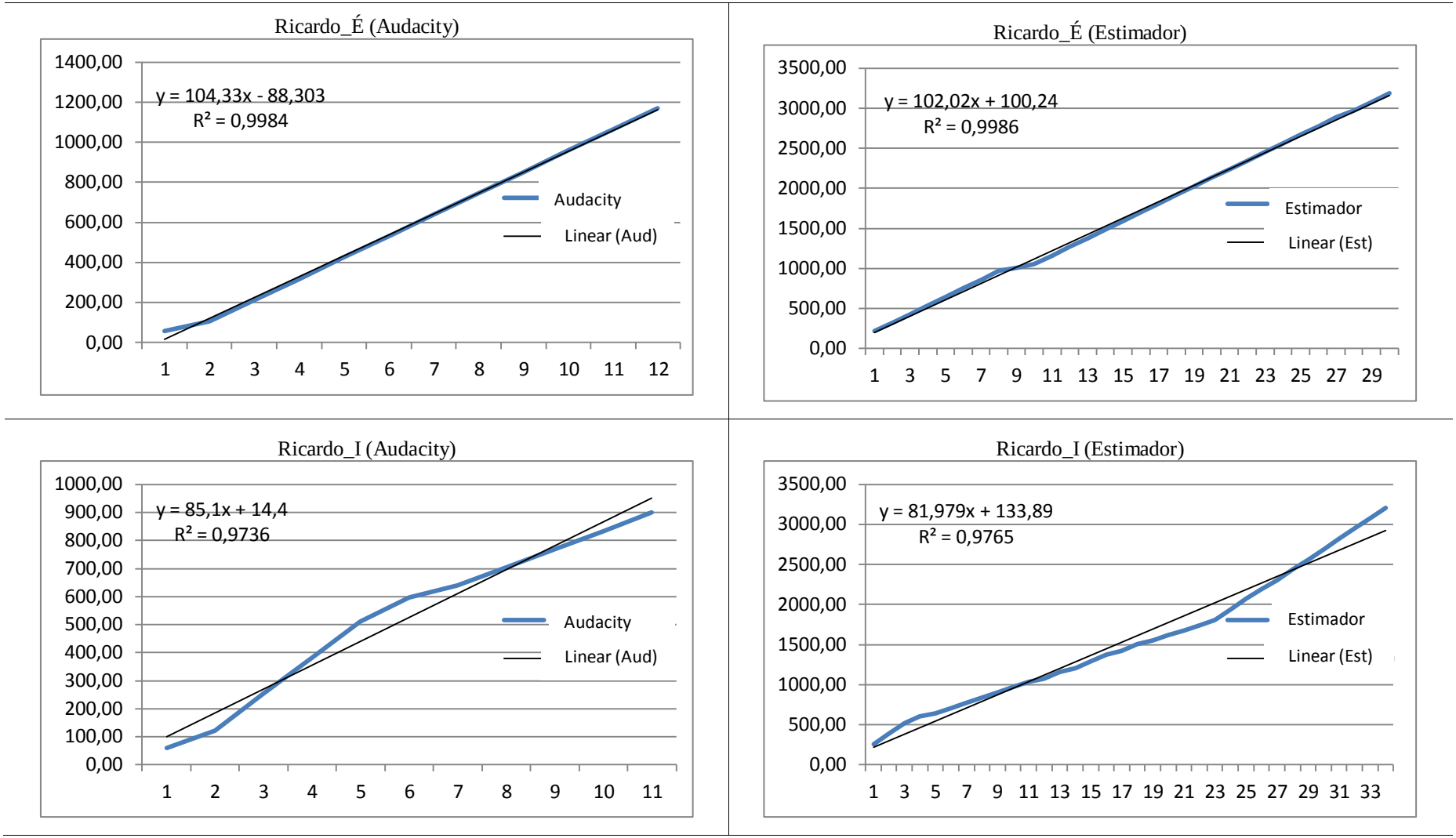


Figura C.22: *Curvas de correlação das vogais "É" e "I" para o locutor Ricardo.*

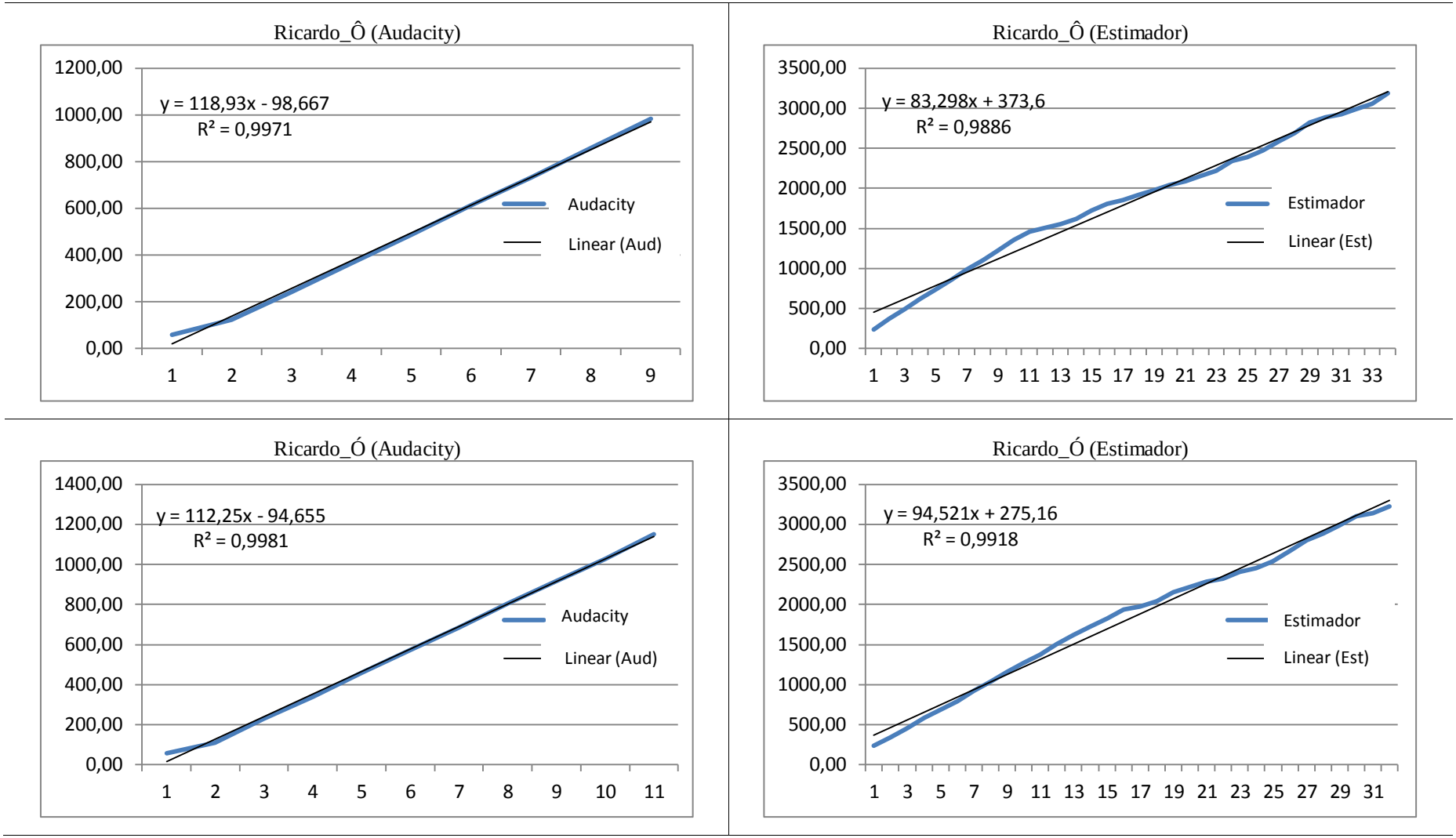


Figura C.23: *Curvas de correlação das vogais "Ô" e "Ó" para o locutor Ricardo.*

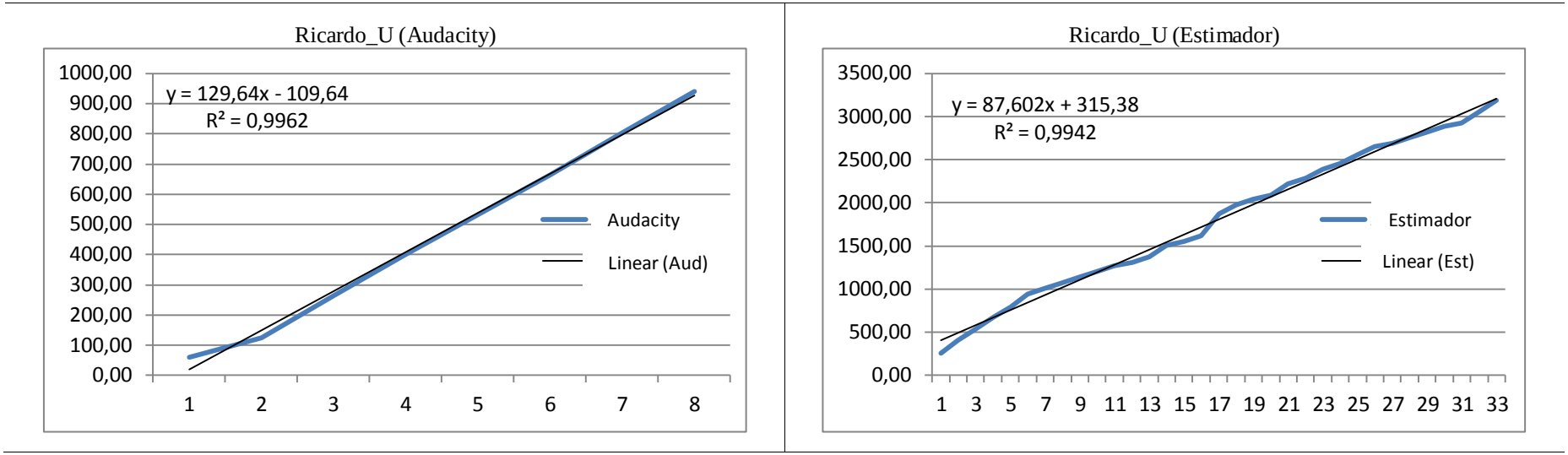


Figura C.24: *Curvas de correlação da vogal "U" para o locutor Ricardo.*

ANEXO D – SEPARAÇÃO SILÁBICA DAS PALAVRAS TESTADAS

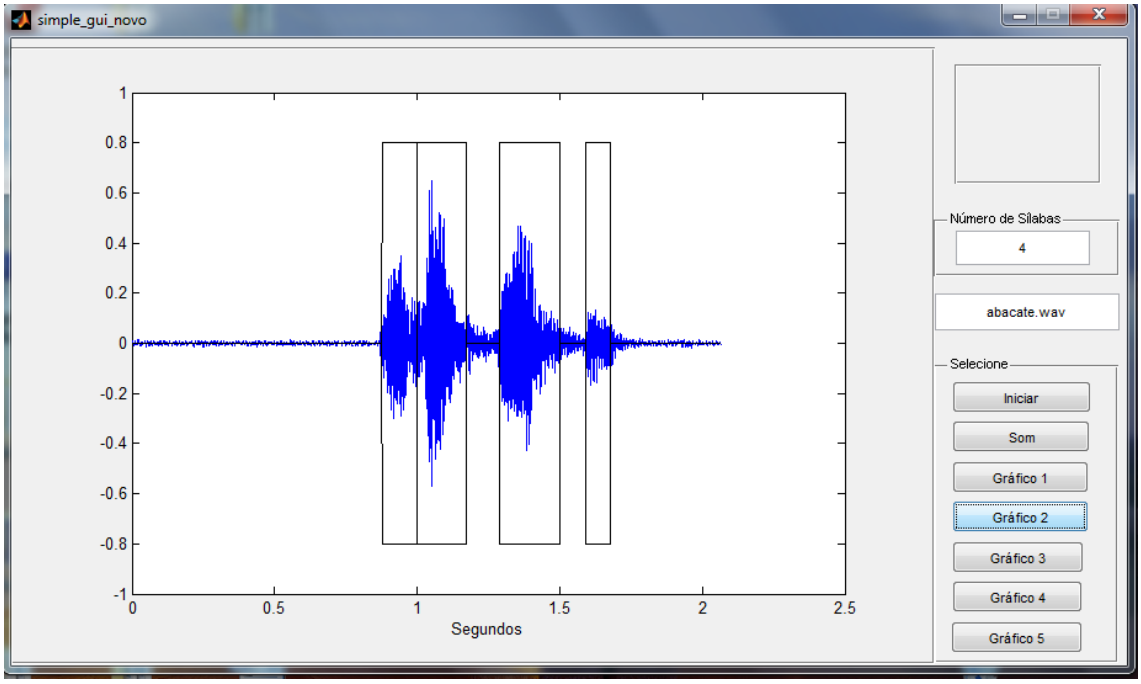


Figura D.1: Separação silábica da palavra “ABACATE”.

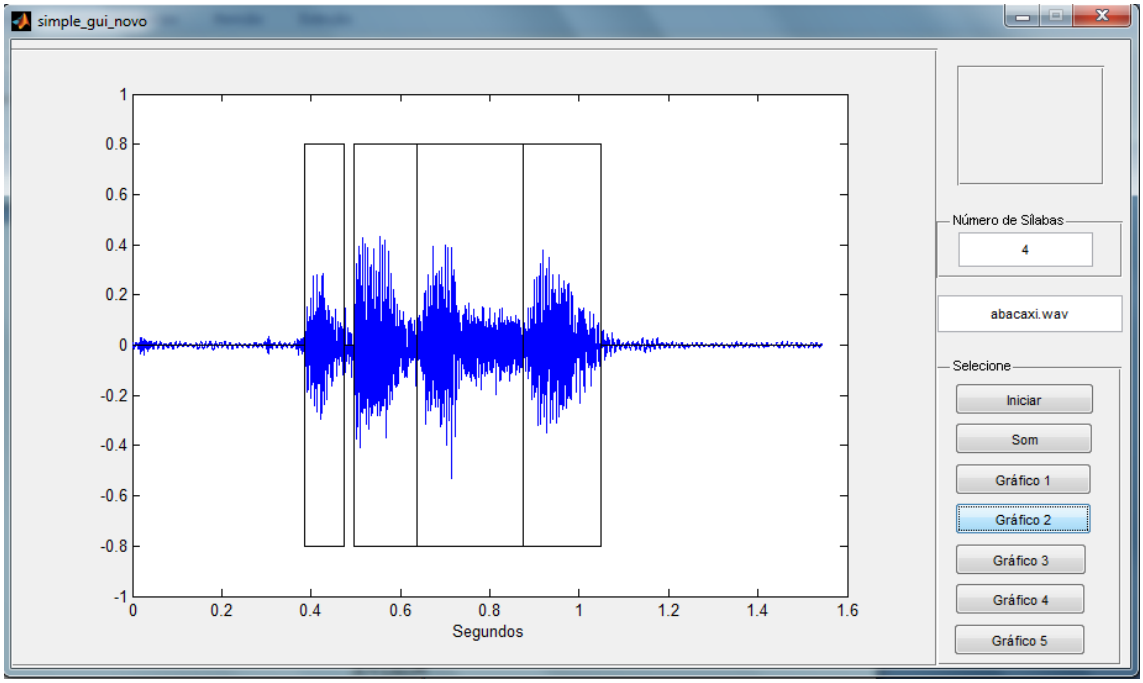


Figura D.2: Separação silábica da palavra “ABACAXI”.

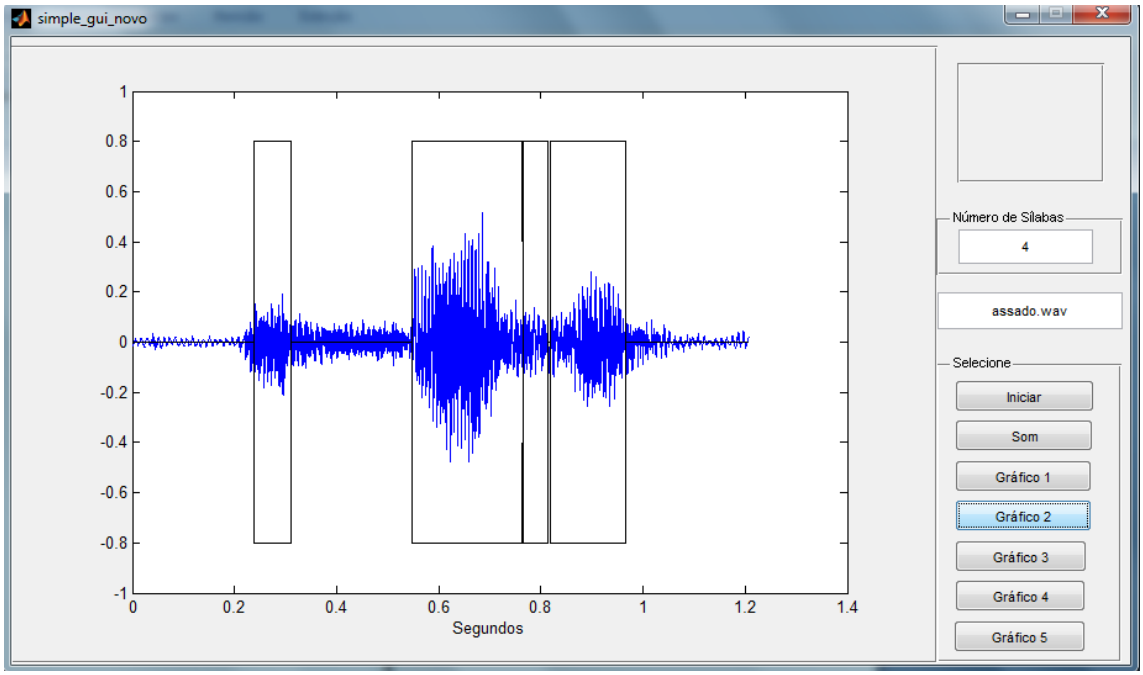


Figura D.3: Separação silábica da palavra “ASSADO”.

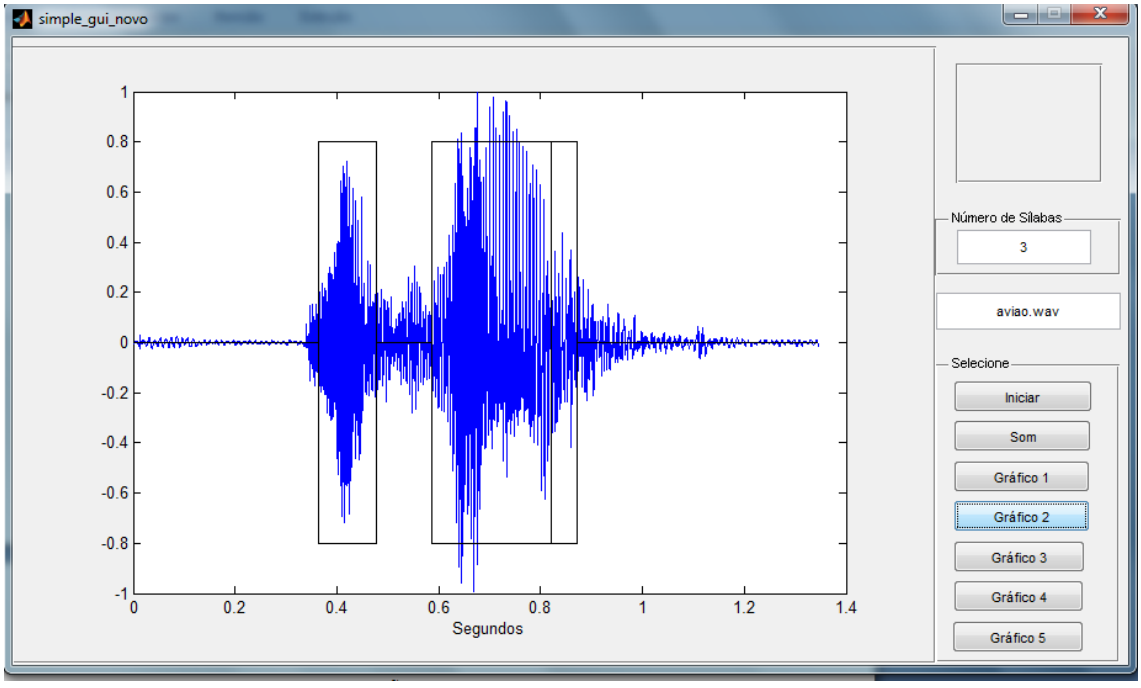


Figura D.4: Separação silábica da palavra “AVIÃO”.

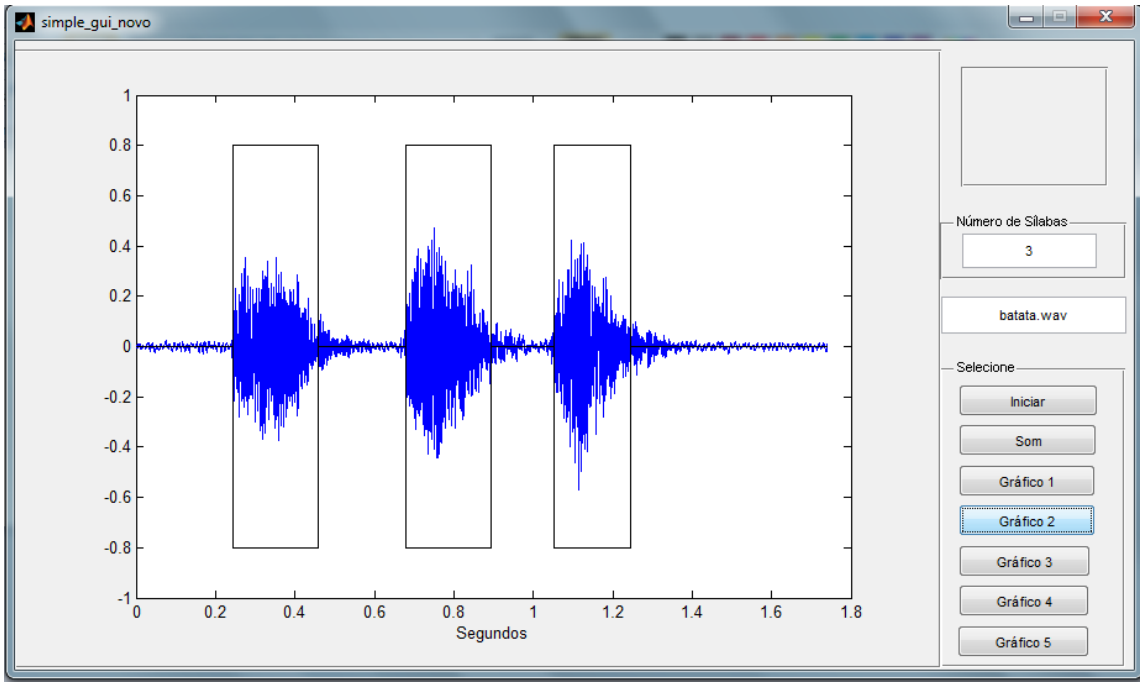


Figura D.5: Separação silábica da palavra “BATATA”.

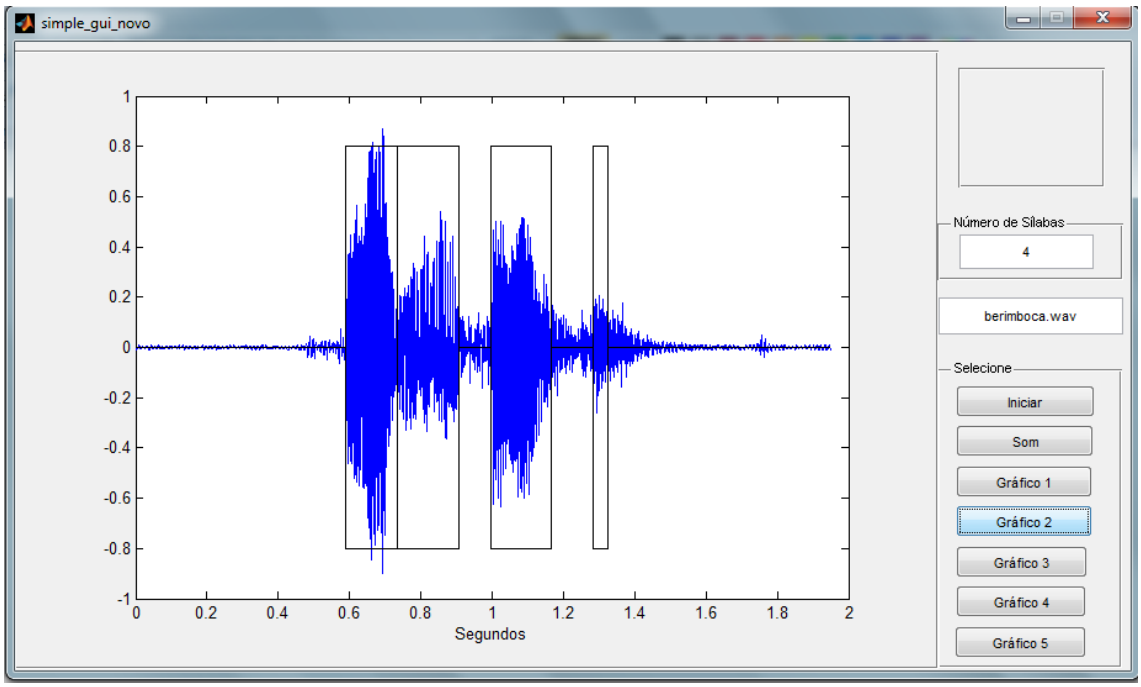


Figura D.6: Separação silábica da palavra “BERIMBOCA”.

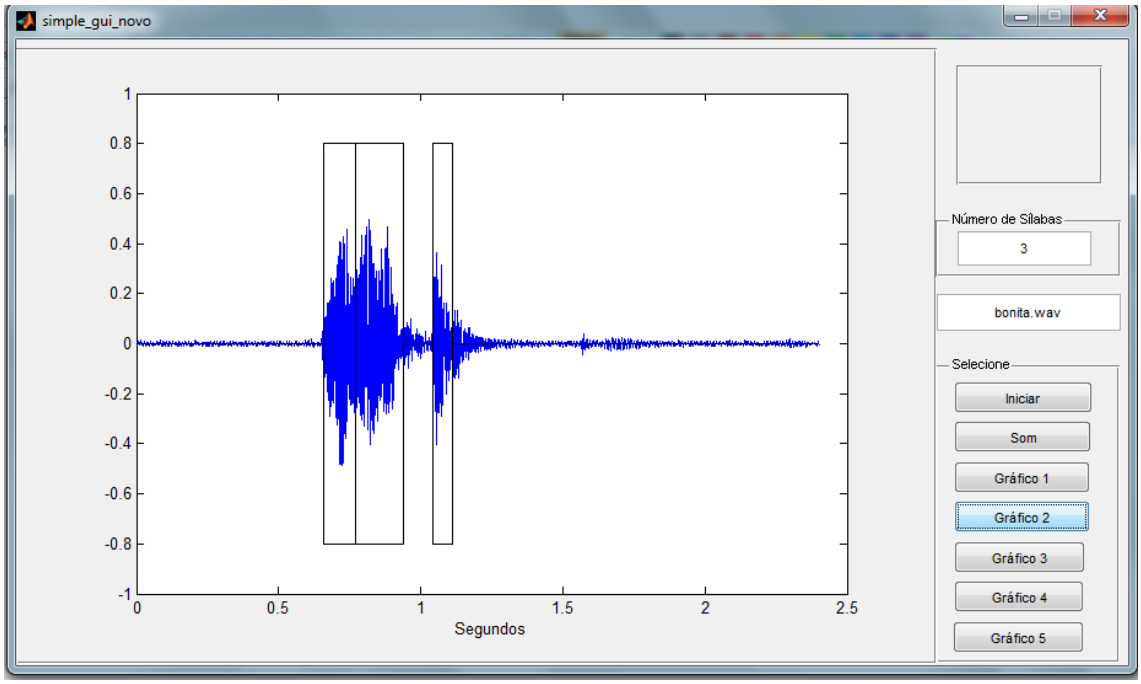


Figura D.7: Separação silábica da palavra "BONITA".

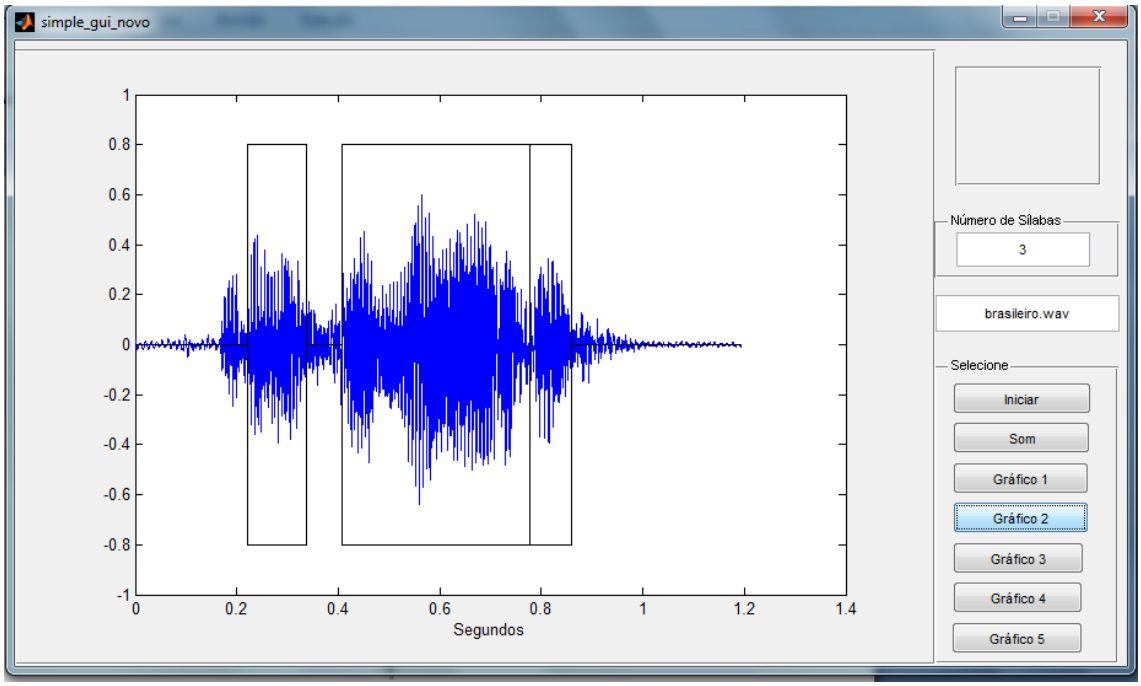


Figura D.8: Separação silábica da palavra "BRASILEIRO[1]".

[1]Veja um exemplo claro em que a envoltória não é um critério atrativo na separação silábica.

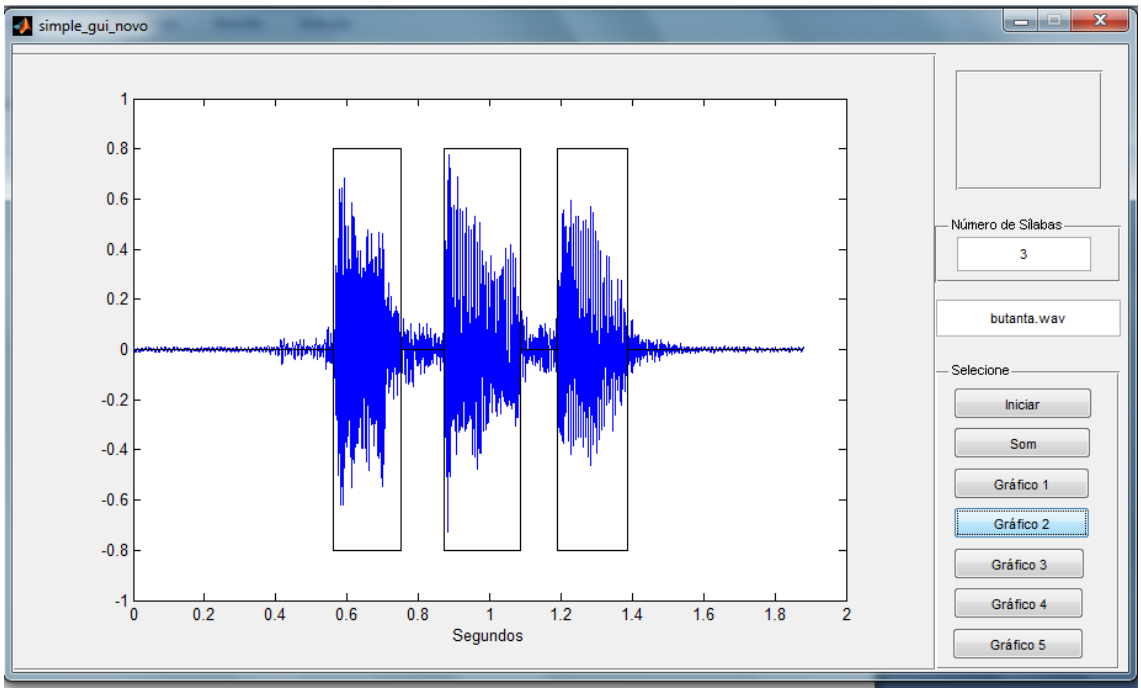


Figura D.9: Separação silábica da palavra “BUTANTÃ”.

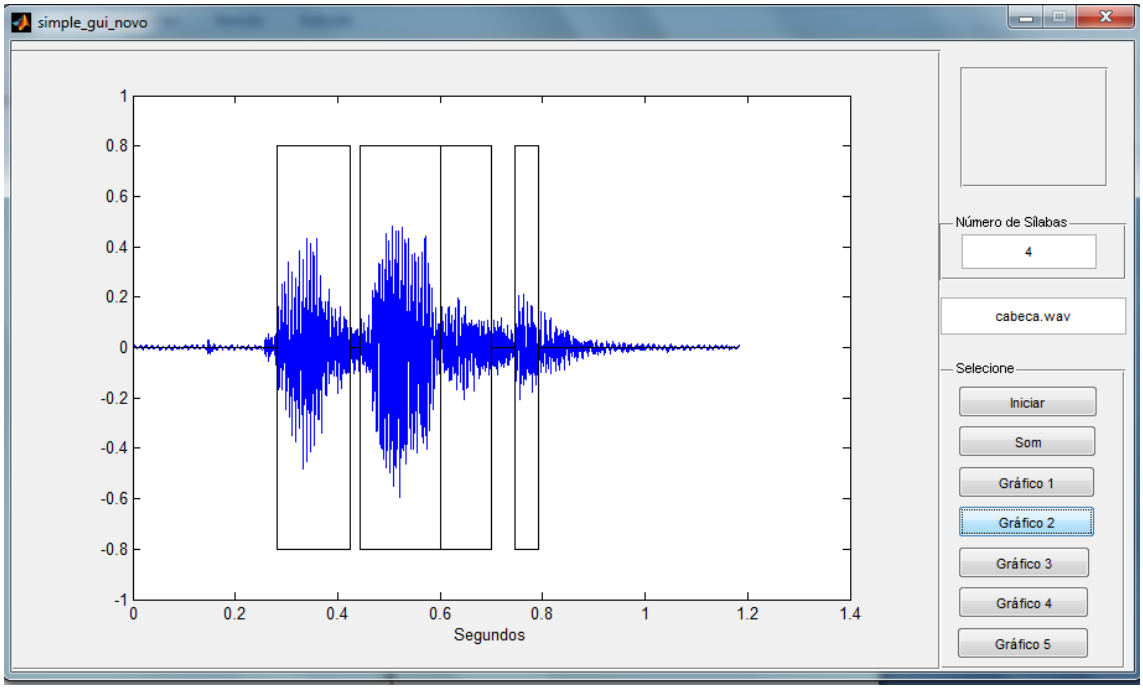


Figura D.10: Separação silábica da palavra “CABEÇA”.

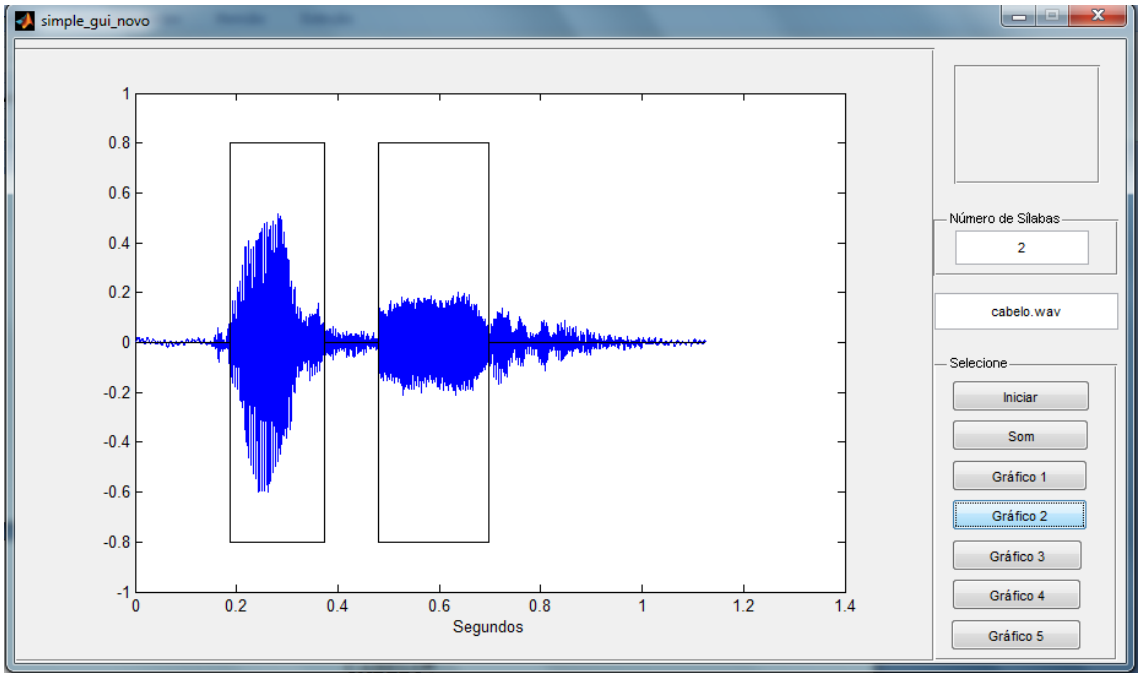


Figura D.11: Separação silábica da palavra “CABELO”.

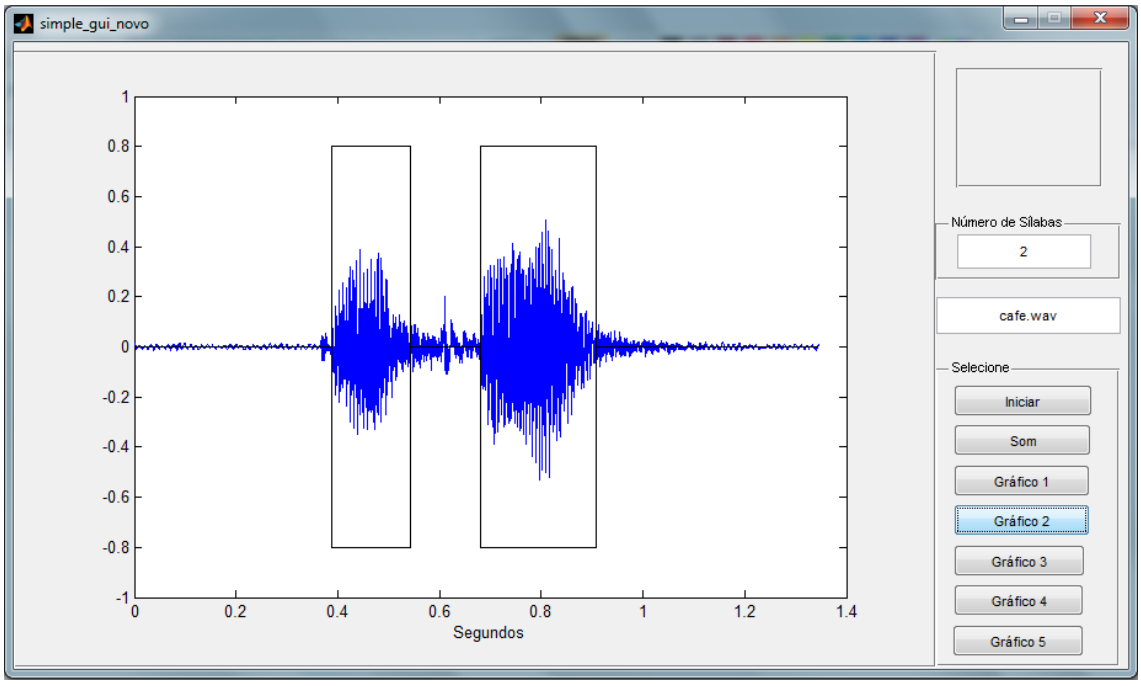


Figura D.12: Separação silábica da palavra “CAFÉ”.

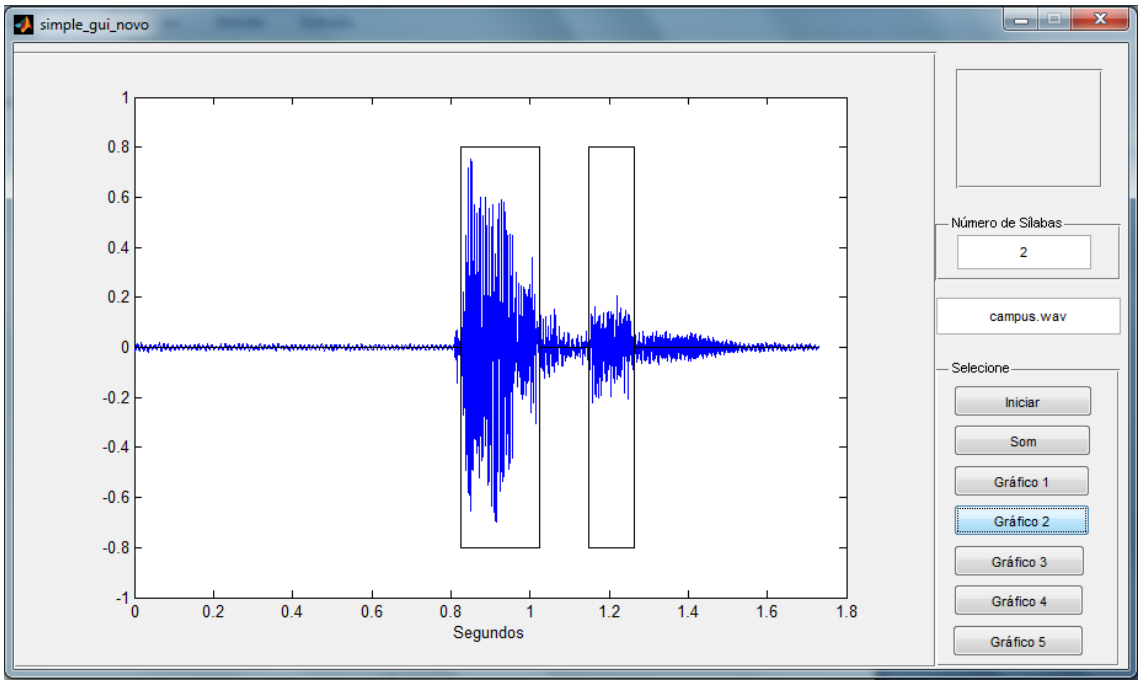


Figura D.13: Separação silábica da palavra "CAMPUS".

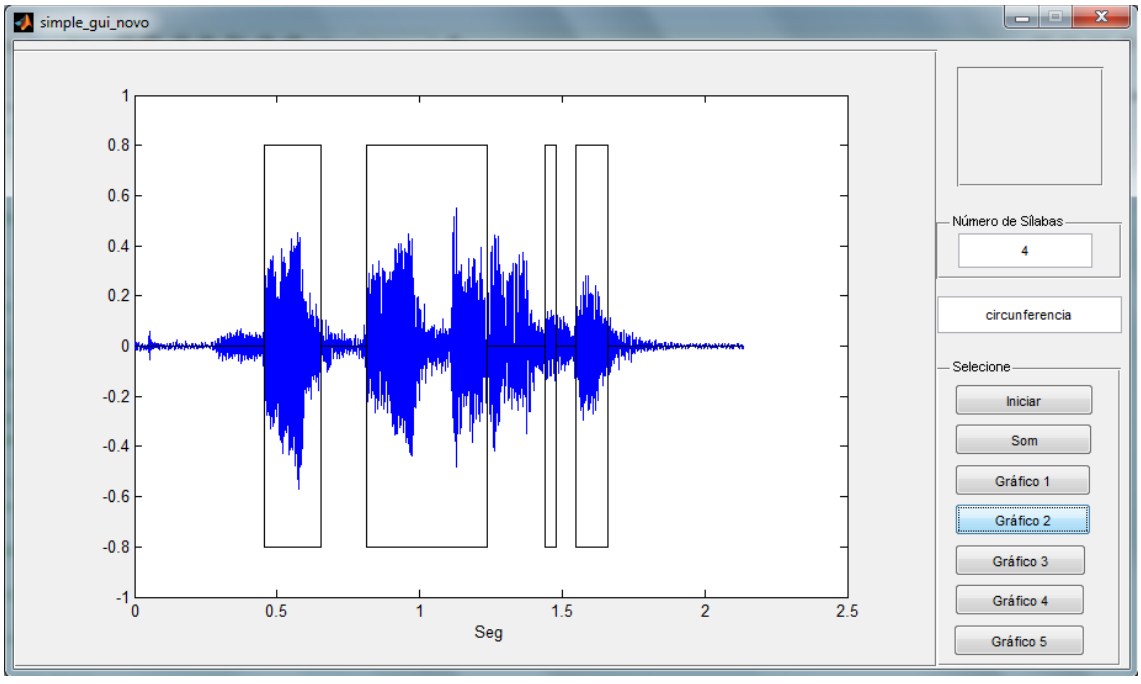


Figura D.14: Separação silábica da palavra "CIRCUNFERÊNCIA".

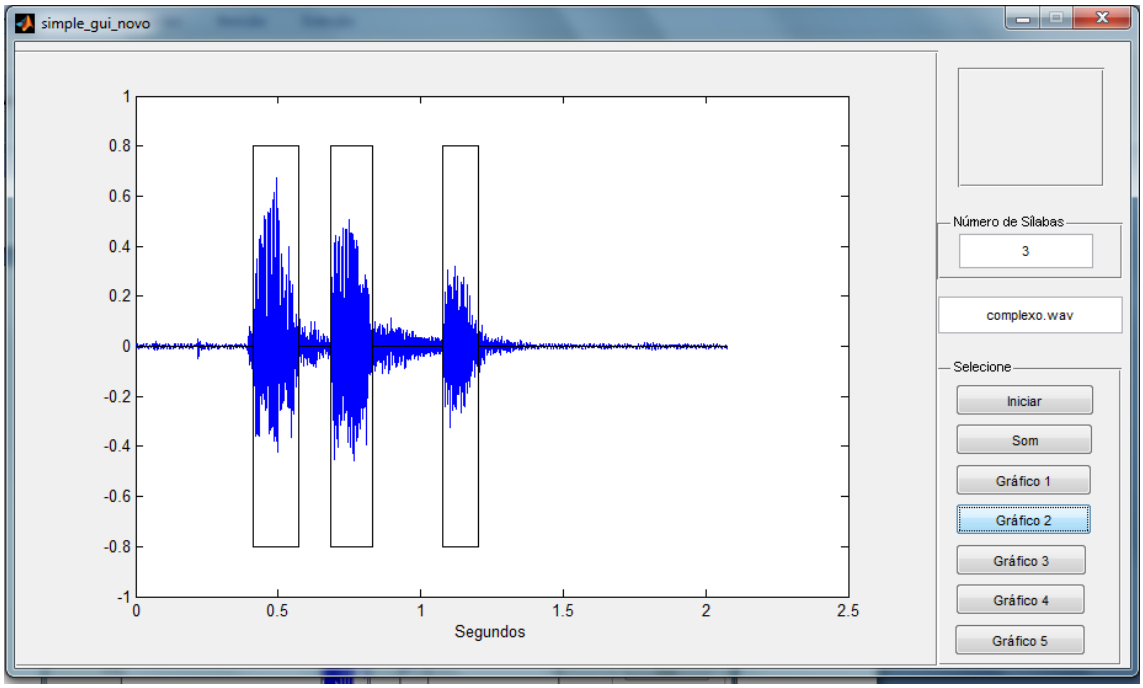


Figura D.15: Separação silábica da palavra “COMPLEXO”.

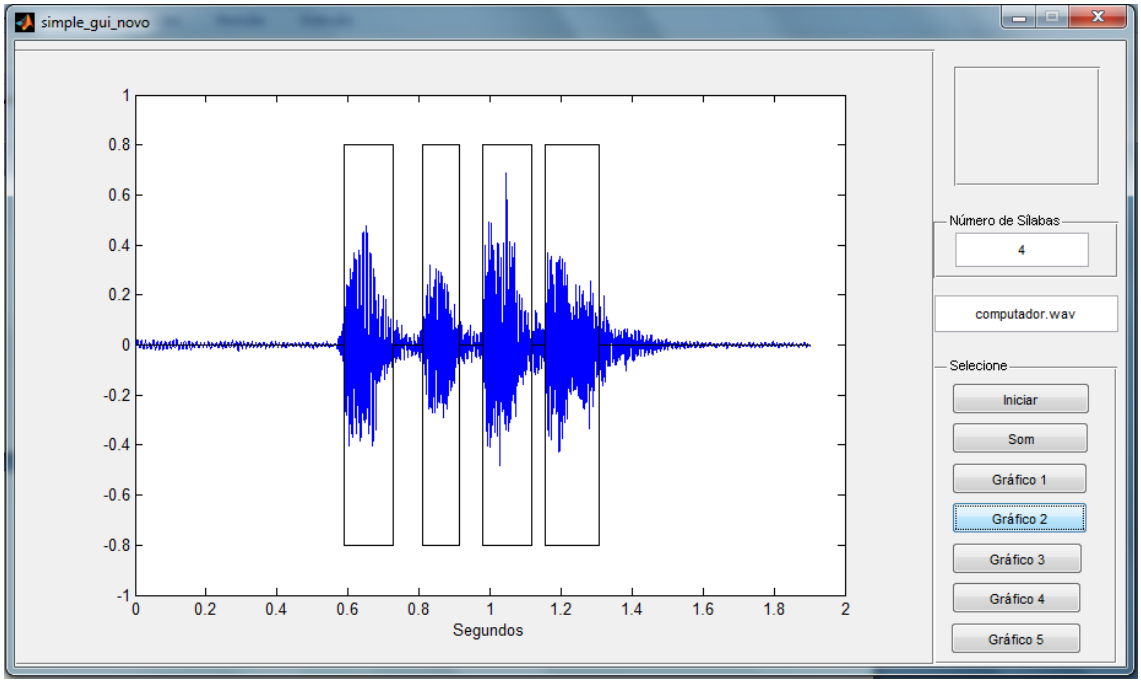


Figura D.16: Separação silábica da palavra “COMPUTADOR”.

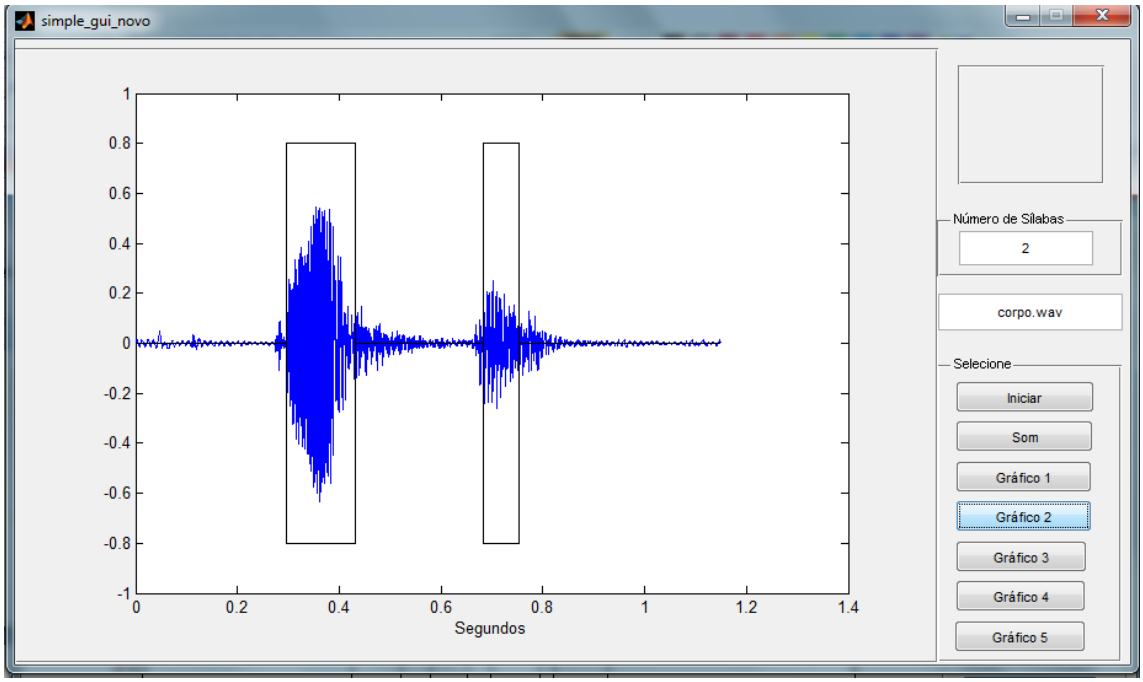


Figura D.17: Separação silábica da palavra "CORPO".

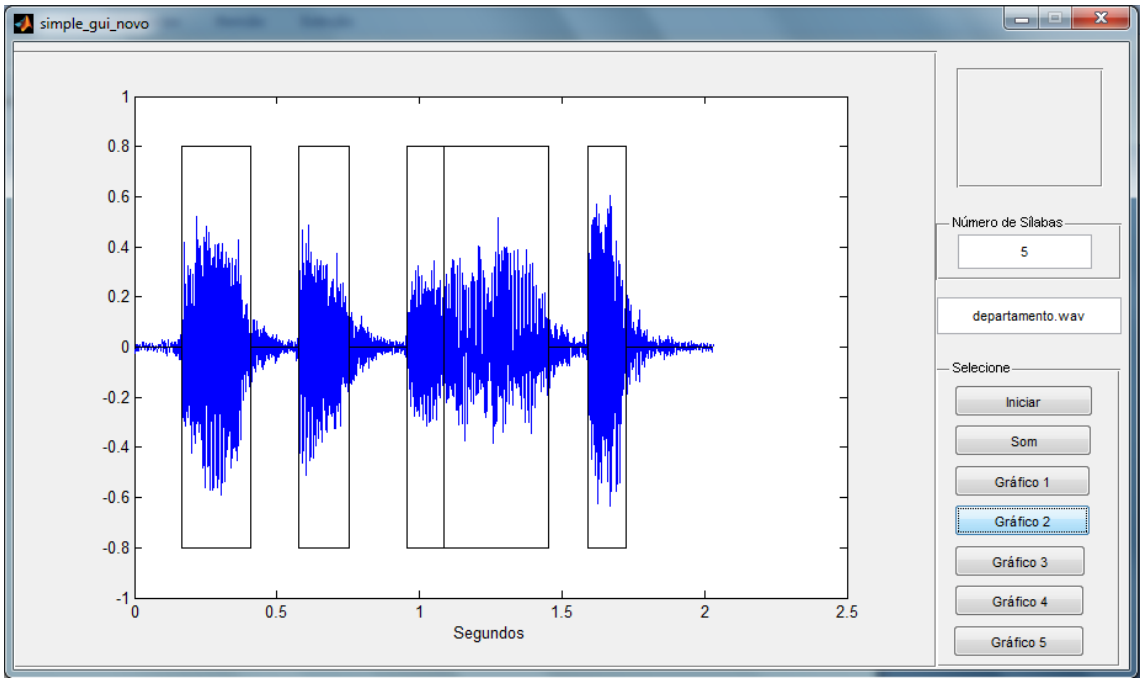


Figura D.18: Separação silábica da palavra "DEPARTAMENTO".

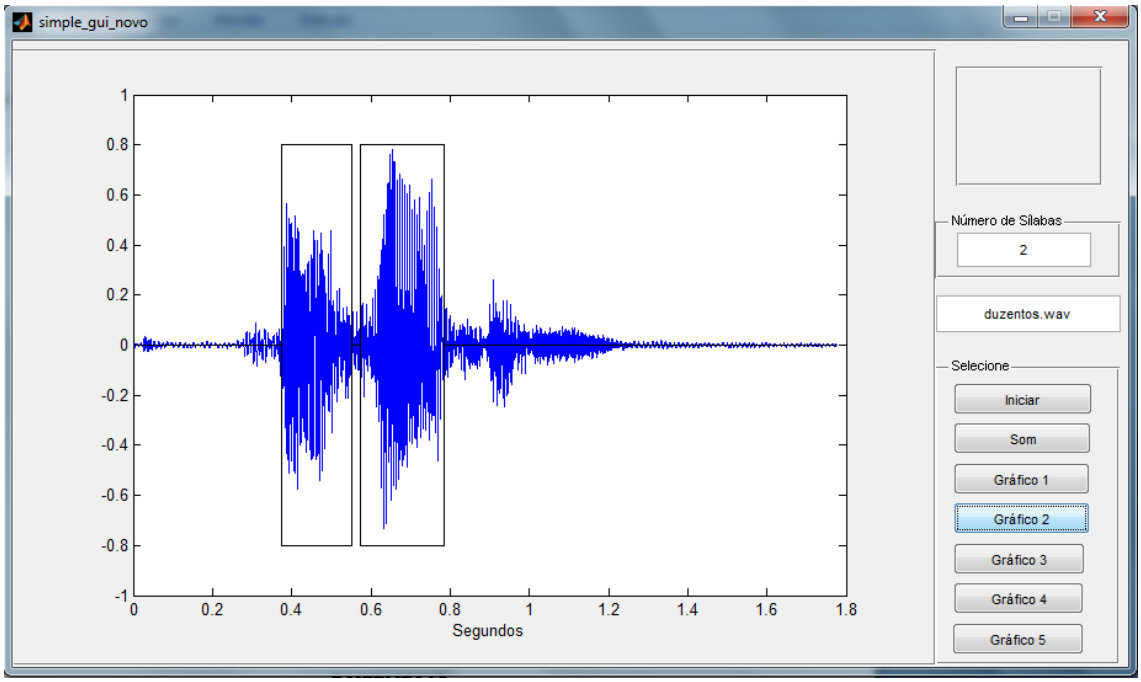


Figura D.19: Separação silábica da palavra "DUZENTOS"[2].

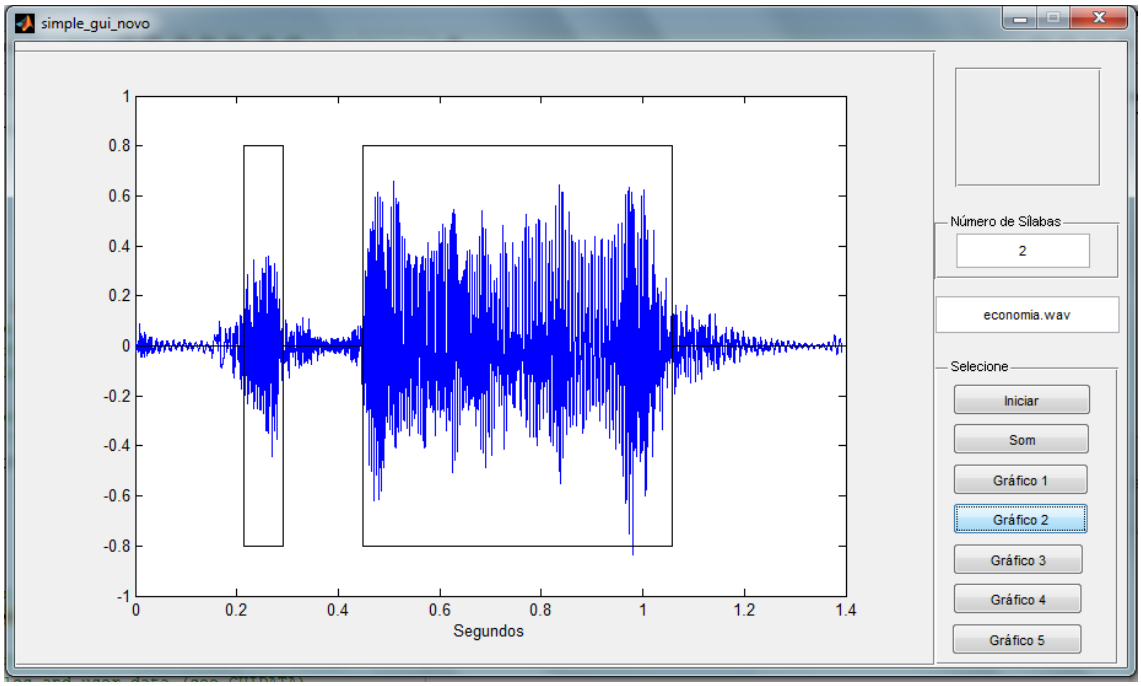


Figura D.20: Separação silábica da palavra "ECONOMIA".

[2] Veja um exemplo claro em que a envoltória não é um critério atrativo na separação

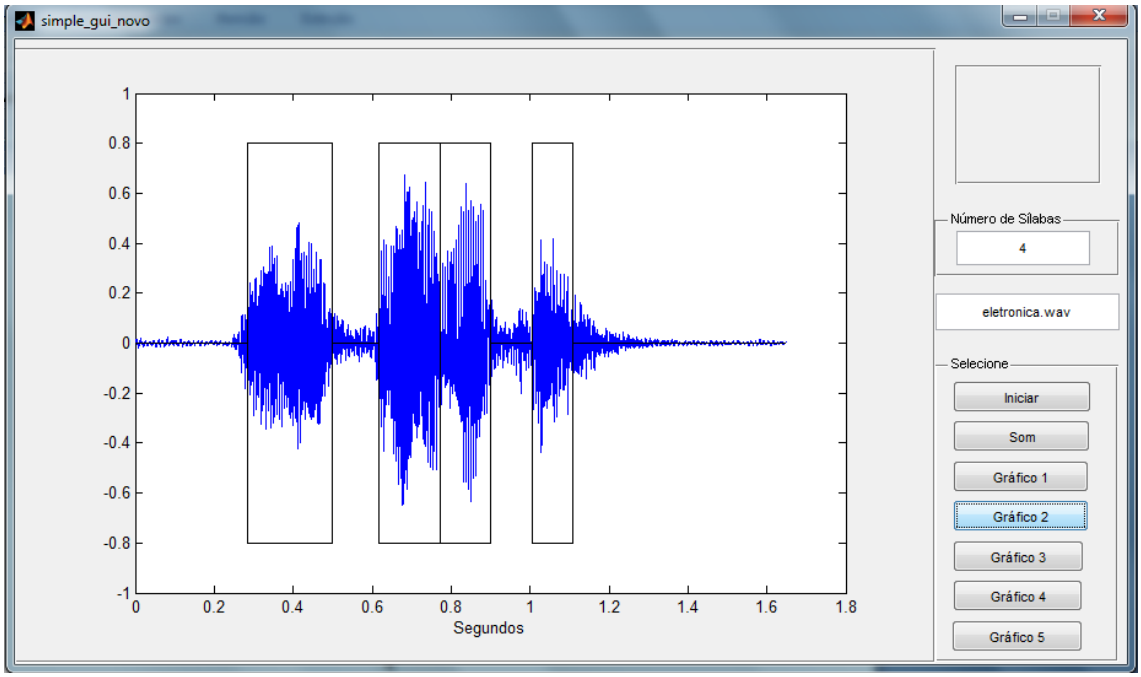


Figura D.21: Separação silábica da palavra “ELETRÔNICA”.

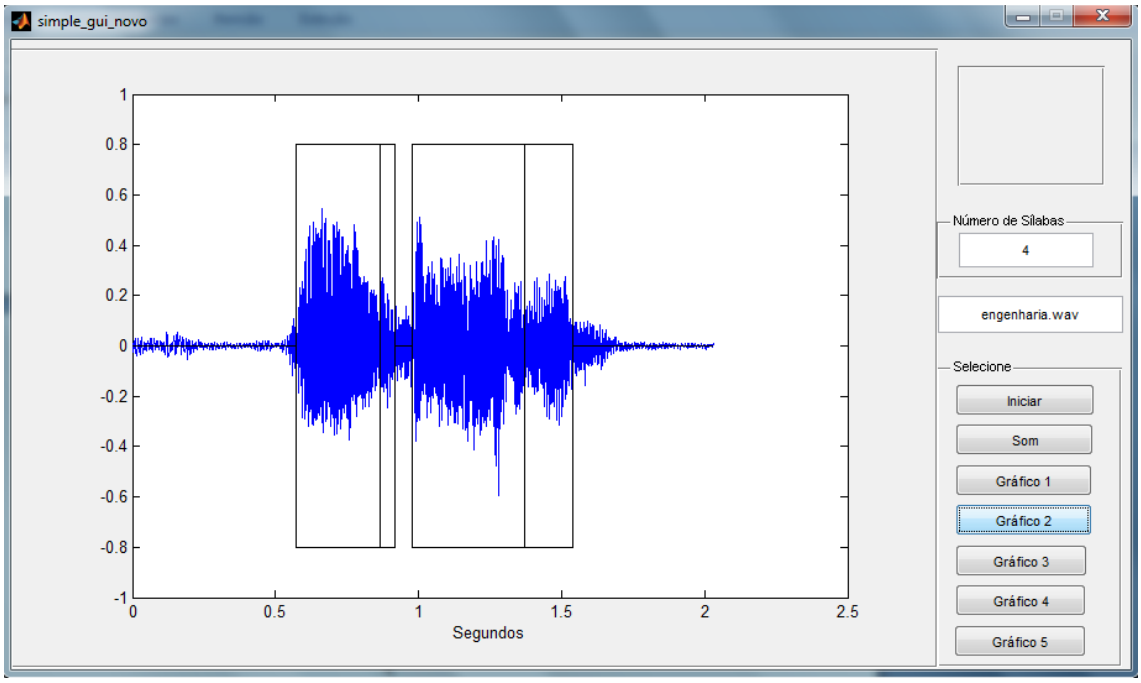


Figura D.22: Separação silábica da palavra “ENGENHARIA”.

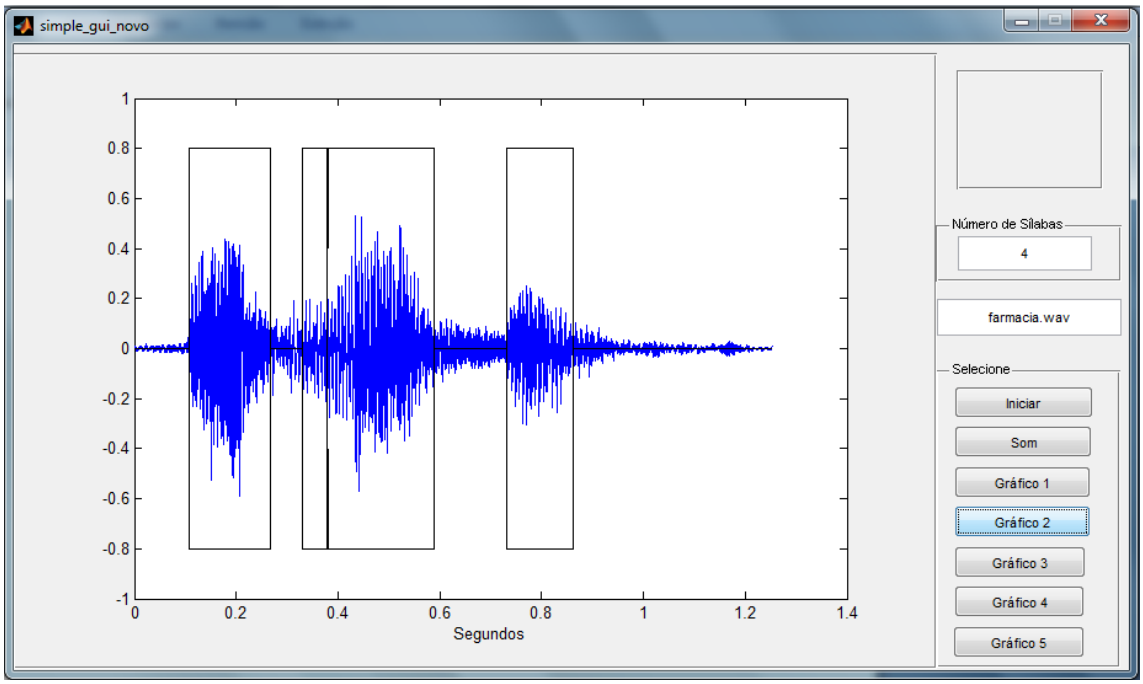


Figura D.23: Separação silábica da palavra "FARMÁCIA".

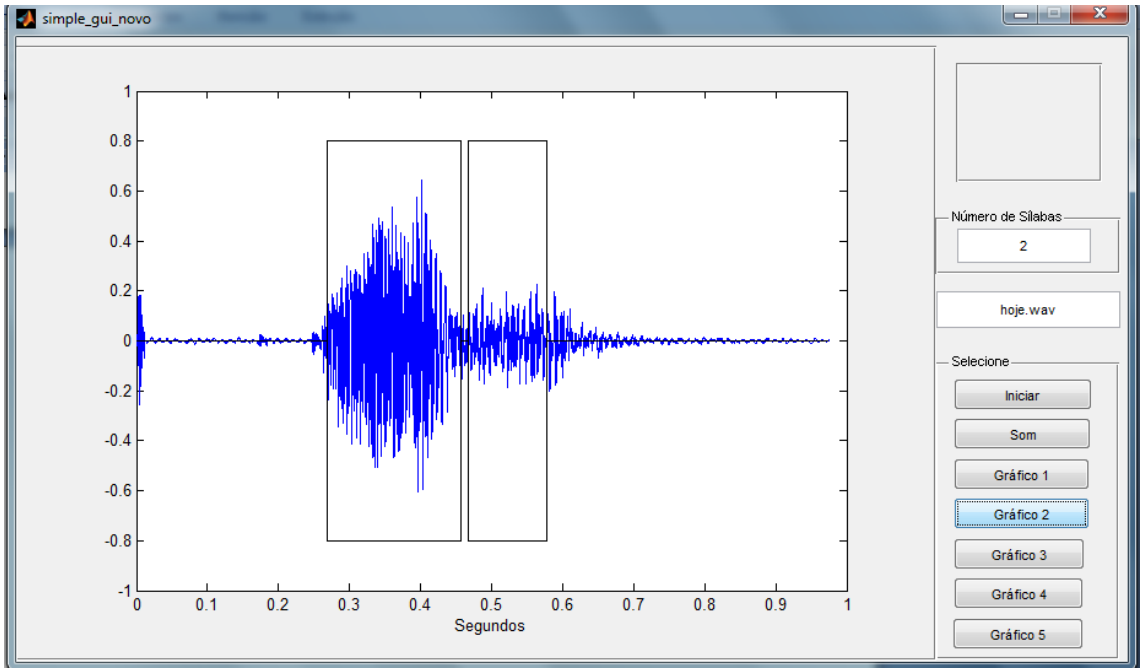


Figura D.24: Separação silábica da palavra "HOJE".

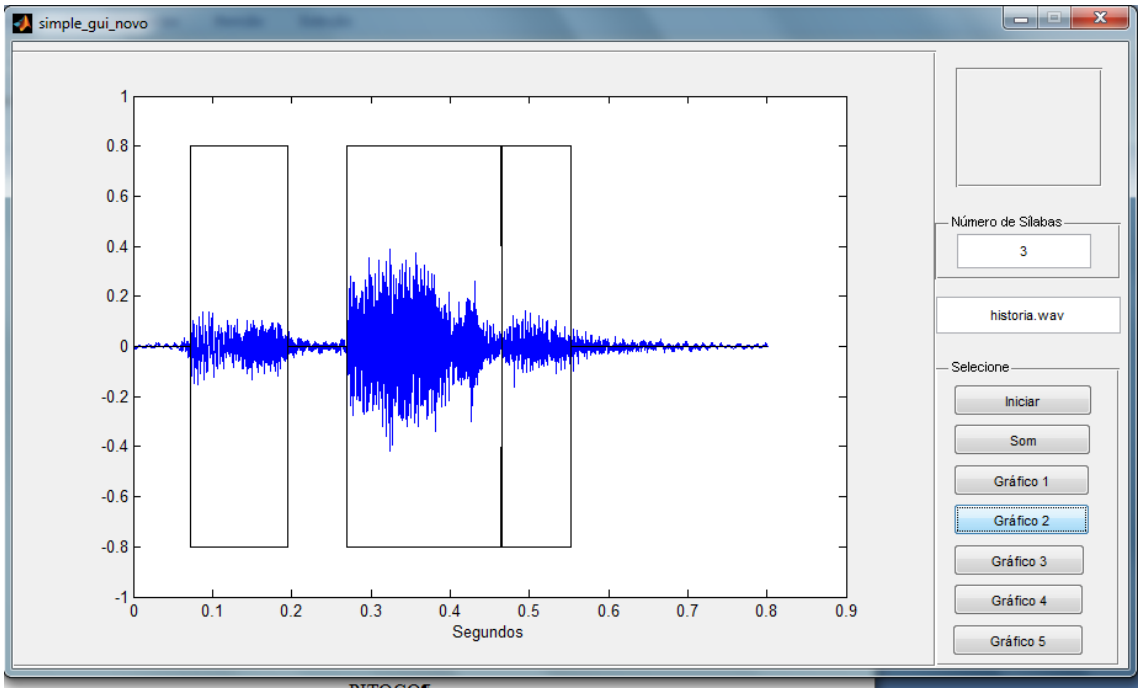


Figura D.25: Separação silábica da palavra “HISTÓRIA”.

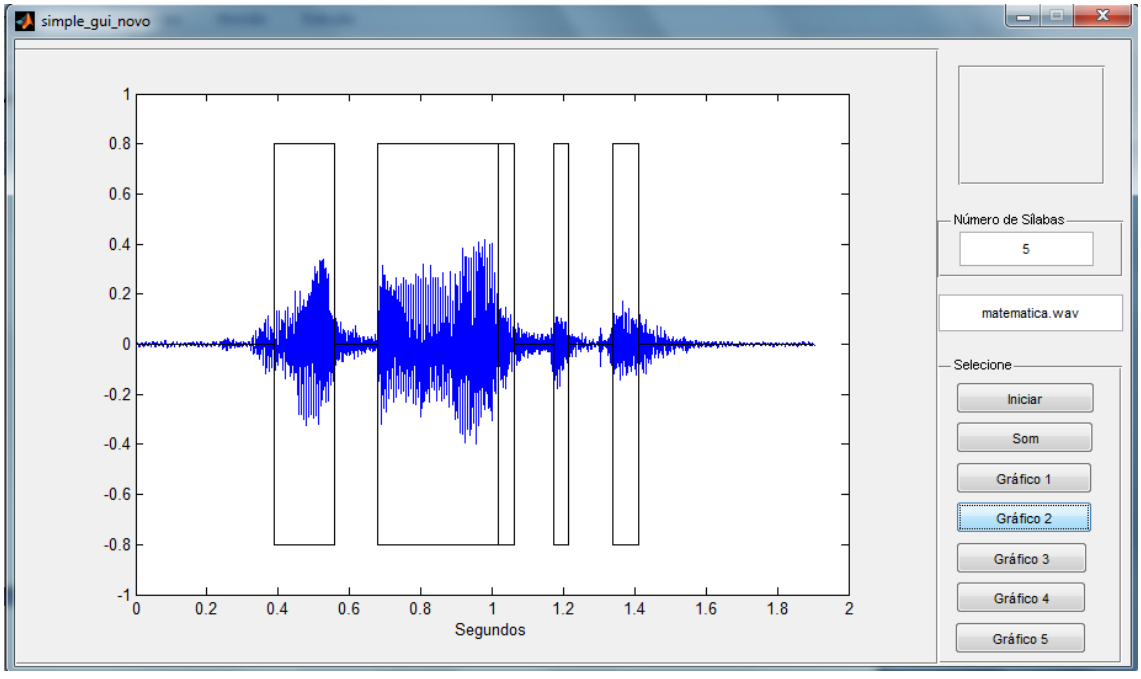


Figura D.26: Separação silábica da palavra “MATEMÁTICA”.

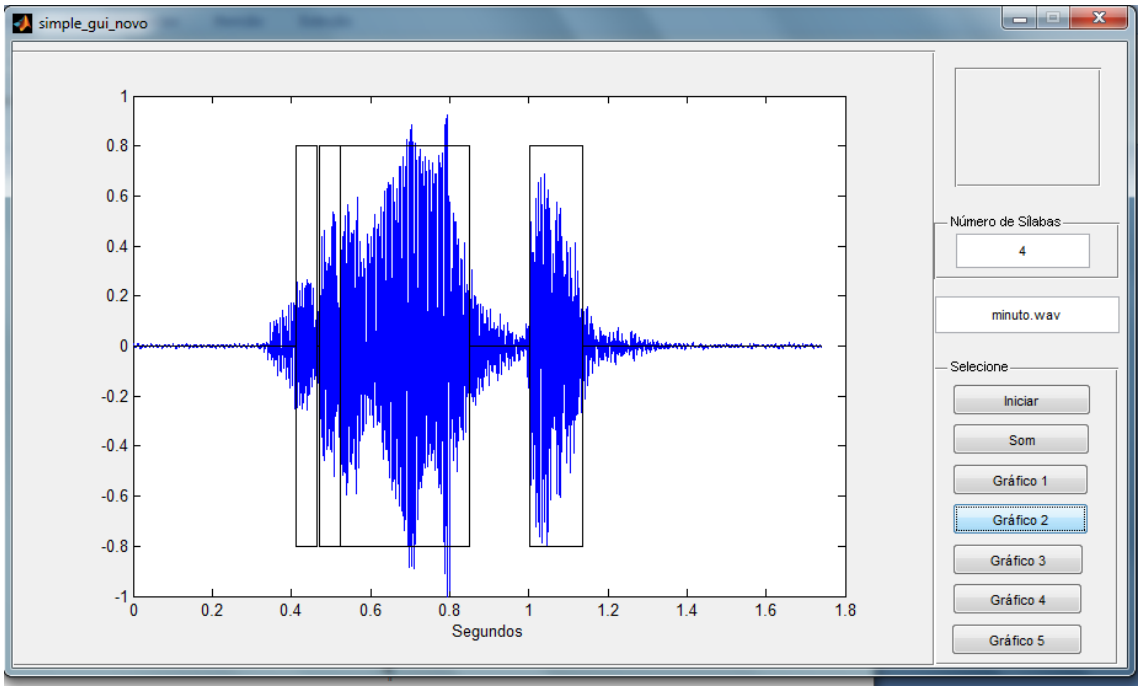


Figura D.27: Separação silábica da palavra “MINUTO”.

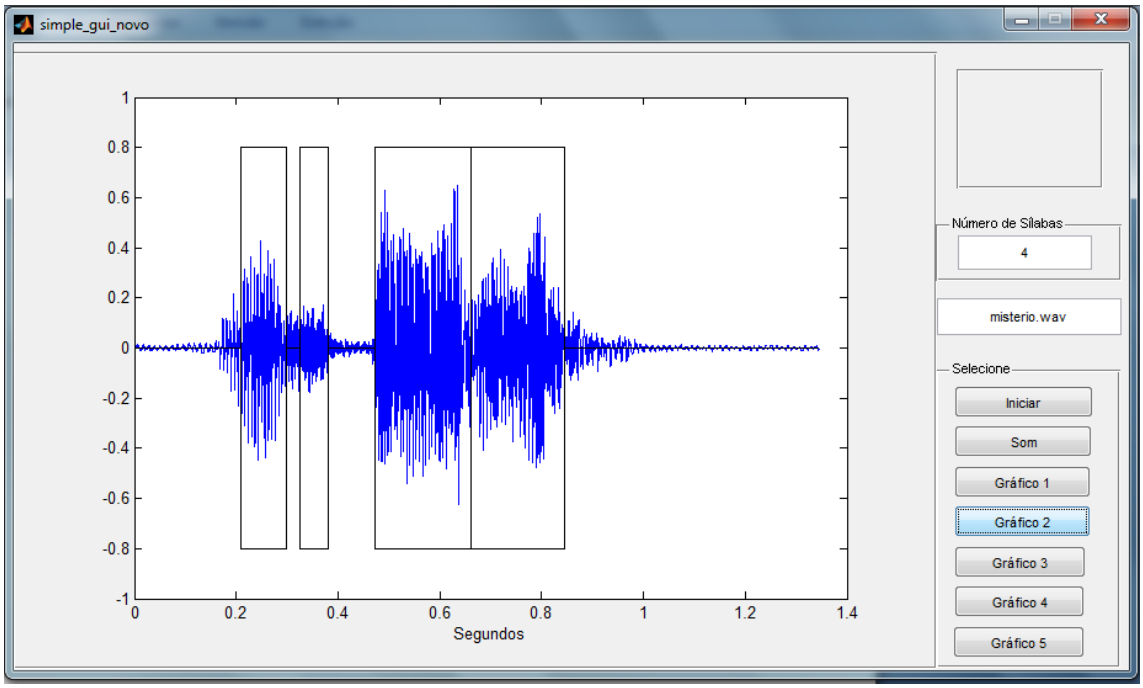


Figura D.28: Separação silábica da palavra “MISTÉRIO

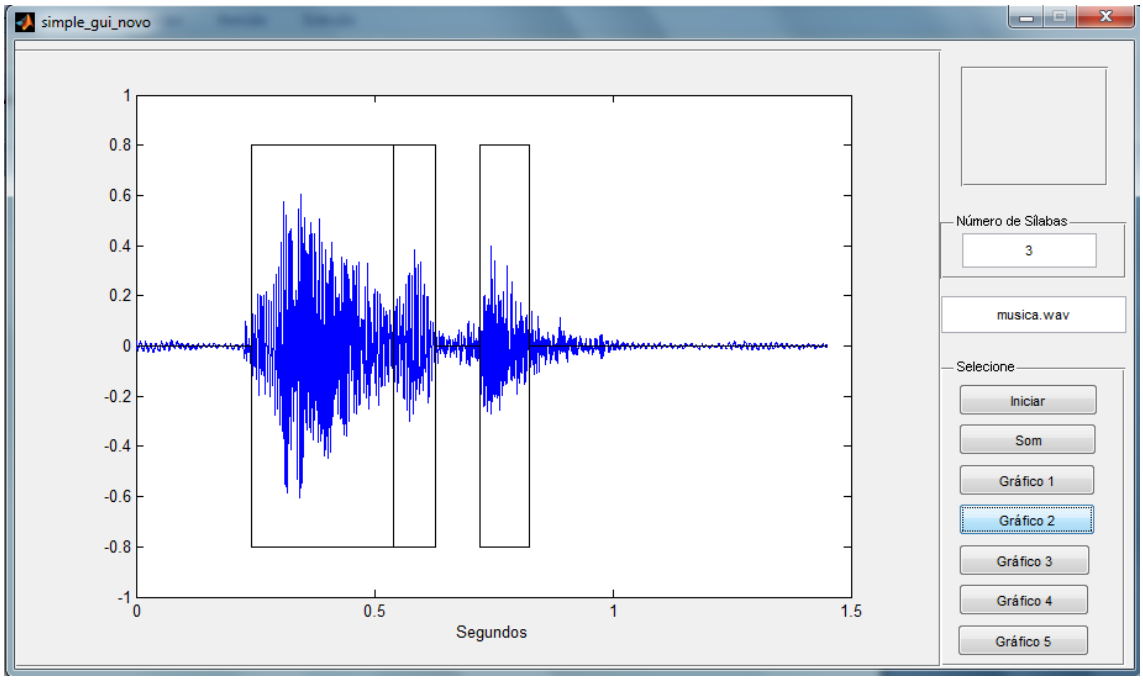


Figura D.29: Separação silábica da palavra “MÚSICA”.

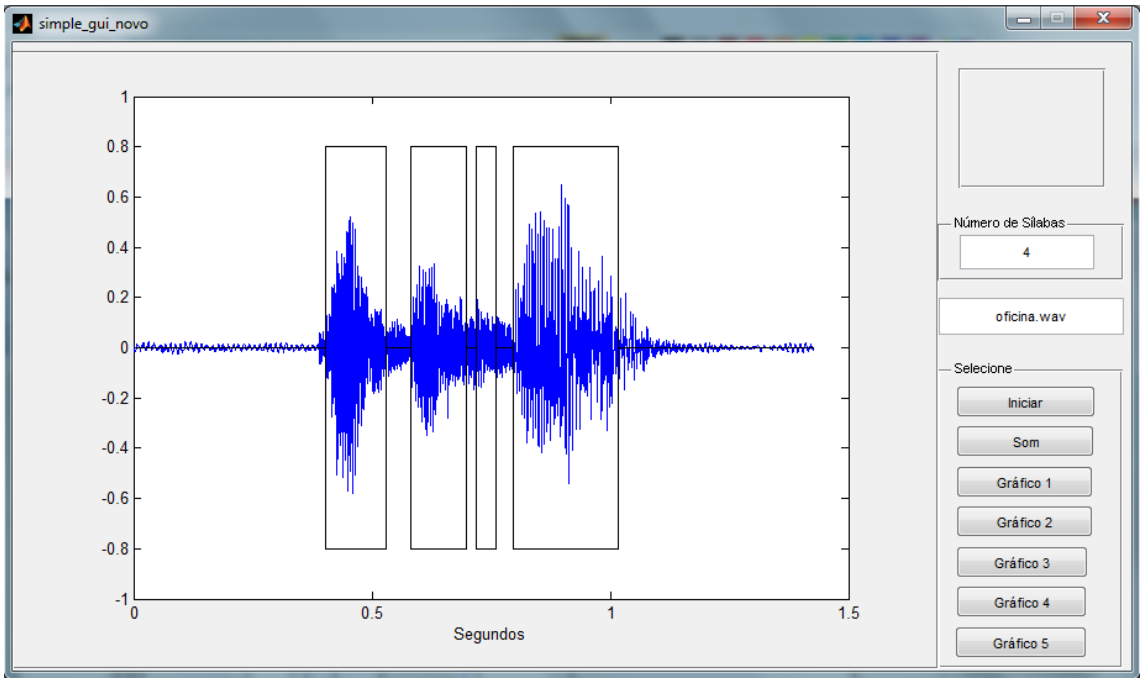


Figura D.30: Separação silábica da palavra “OFICINA”.

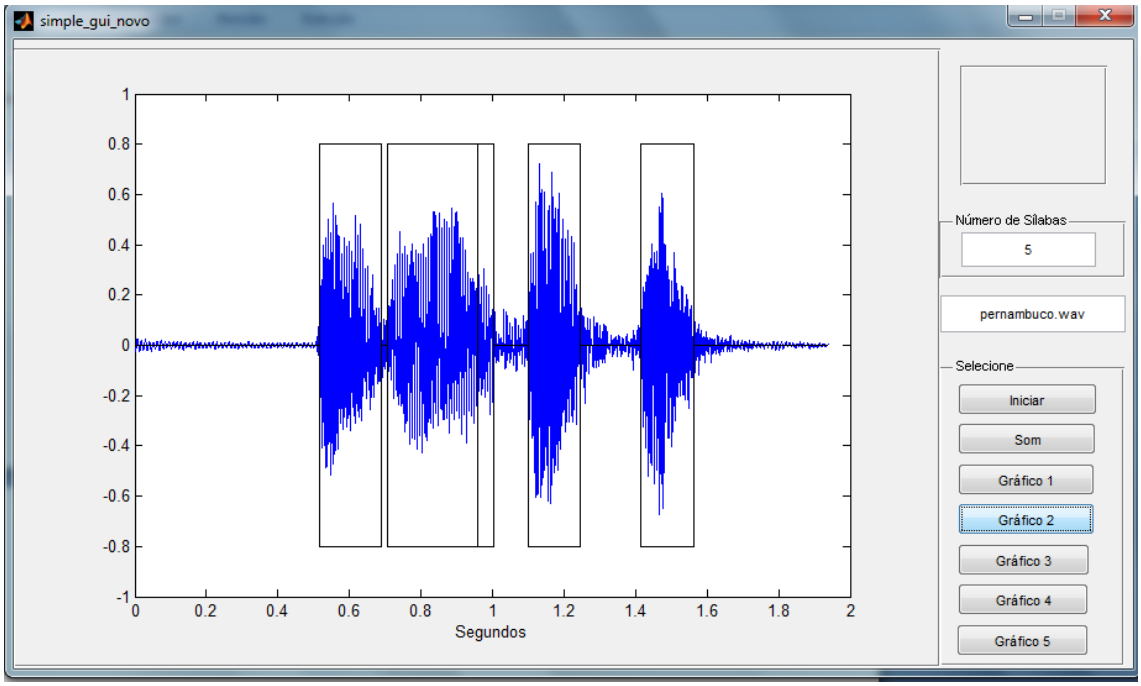


Figura D.31: Separação silábica da palavra "PERNAMBUCO".

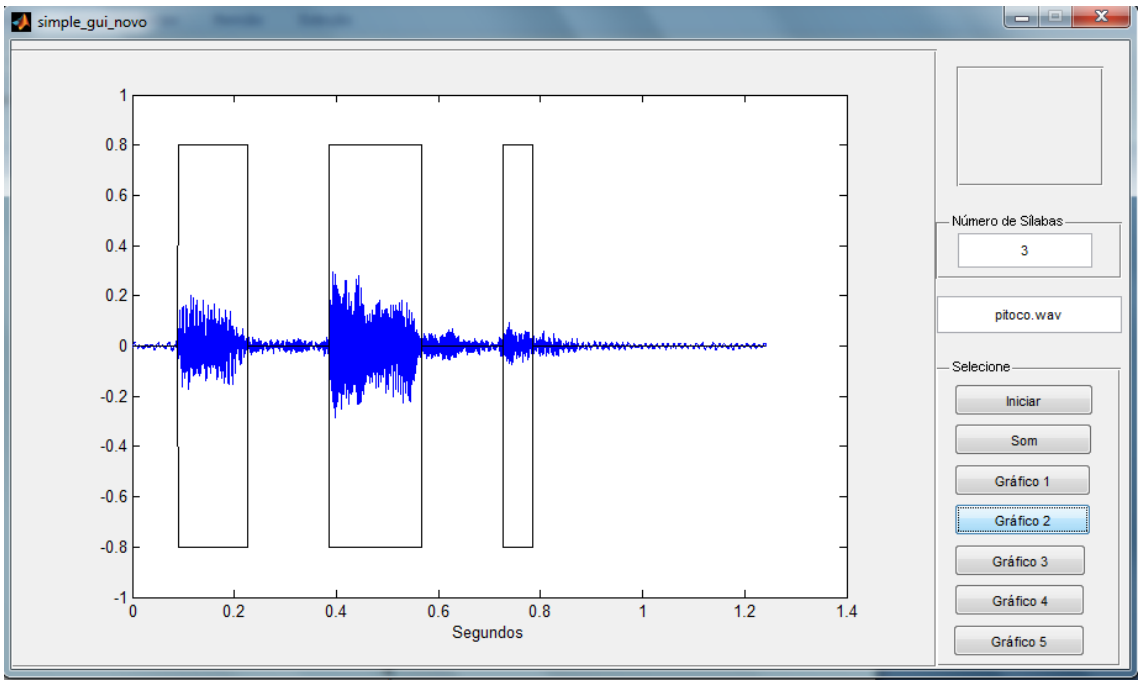


Figura D.32: Separação silábica da palavra "PITOCO".

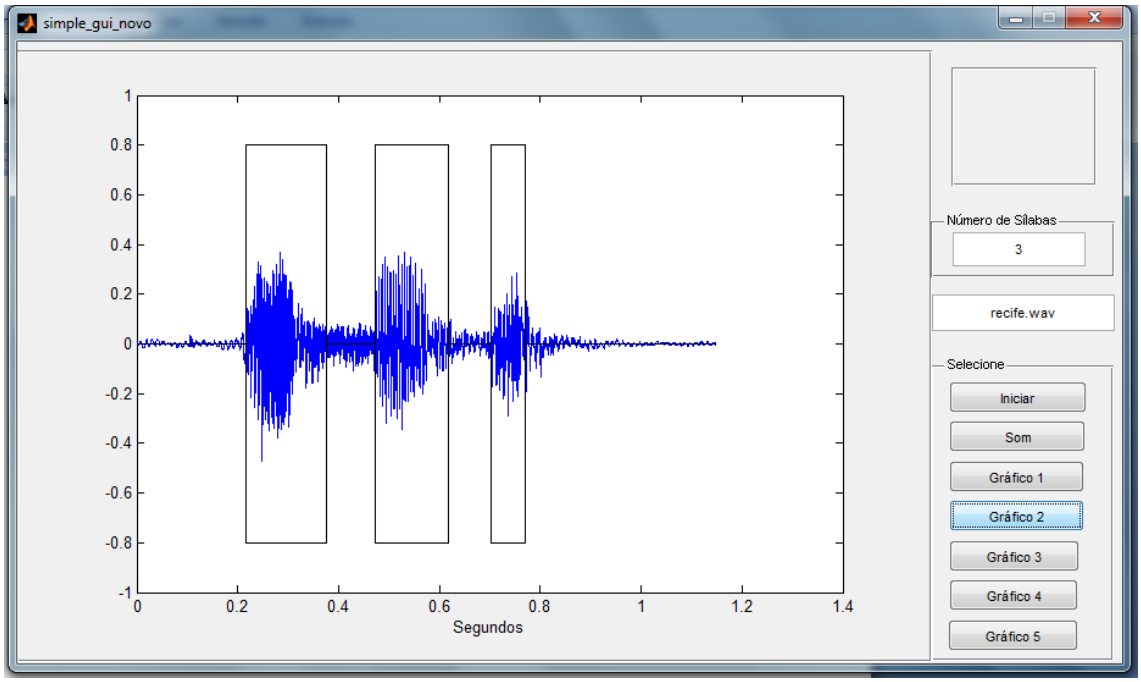


Figura D.33: Separação silábica da palavra “RECIFE”.

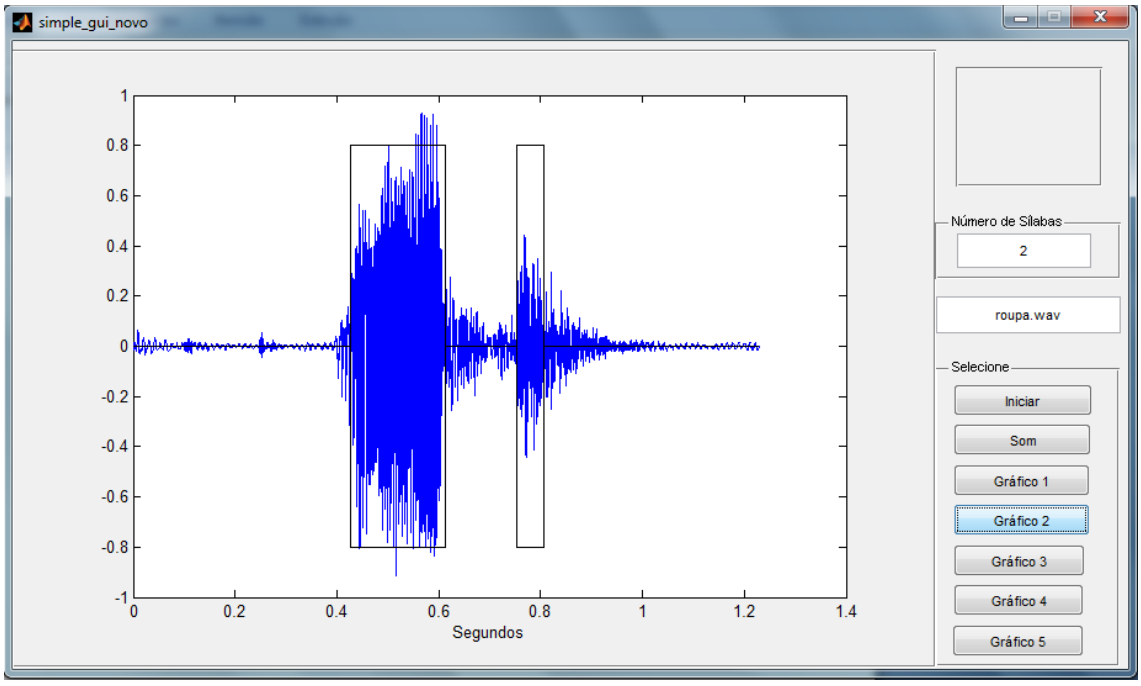


Figura D.34: Separação silábica da palavra “ROUPA”.

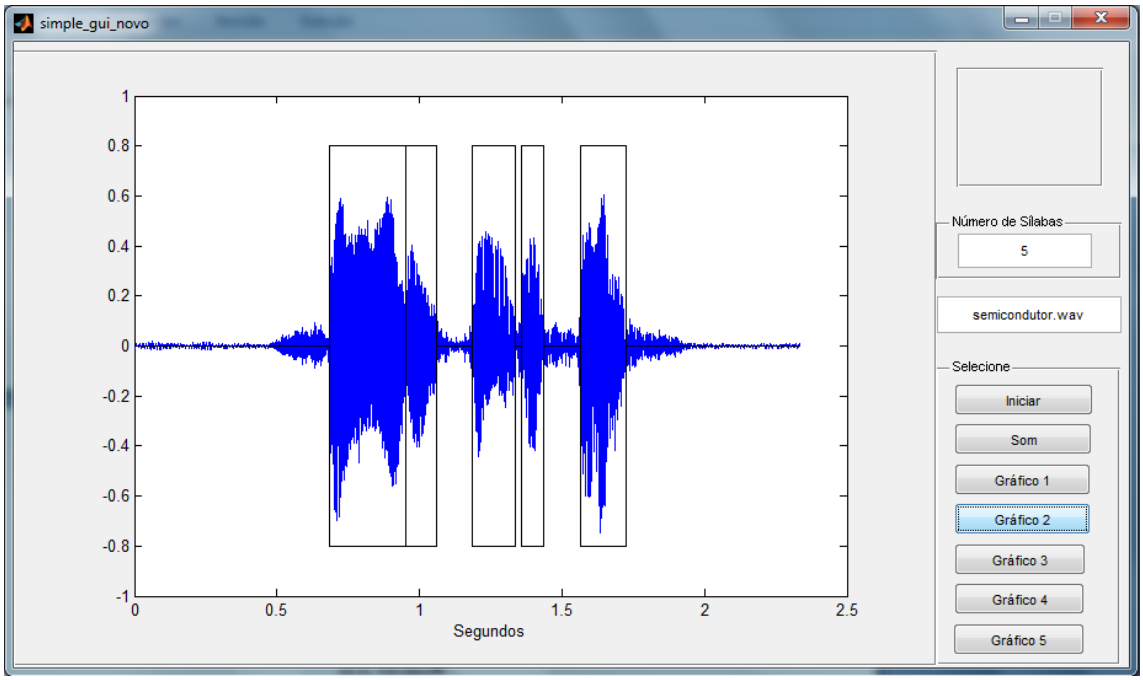


Figura D.35: Separação silábica da palavra “SEMICONDUTOR”.

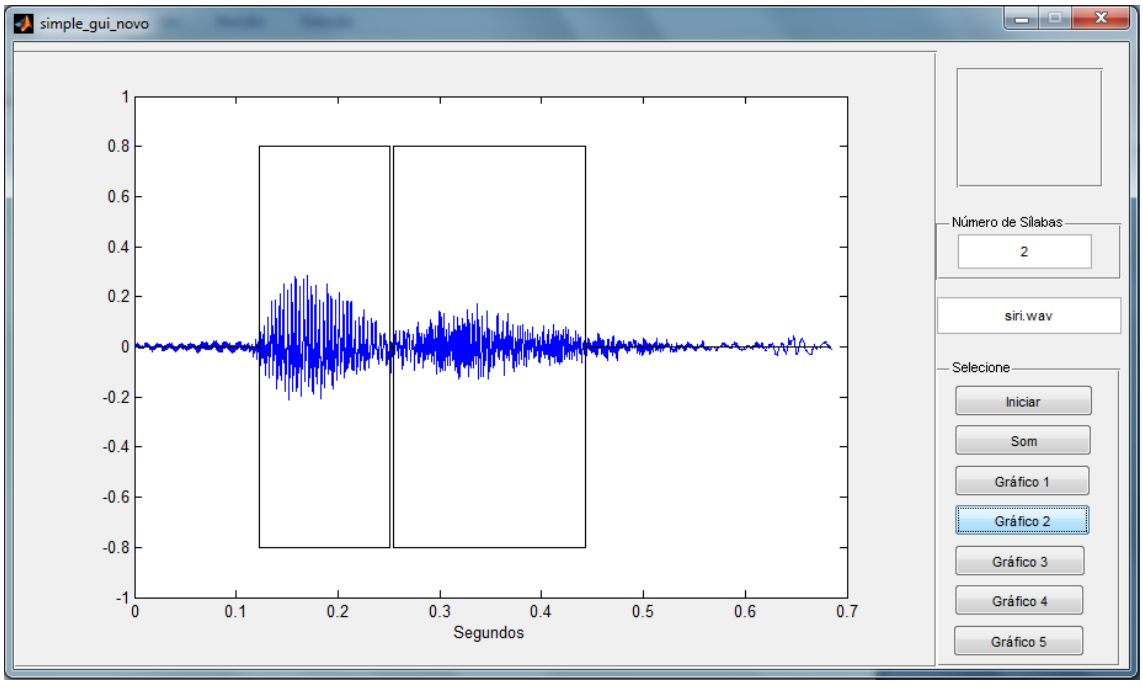


Figura D.36: Separação silábica da palavra “SIRI”.

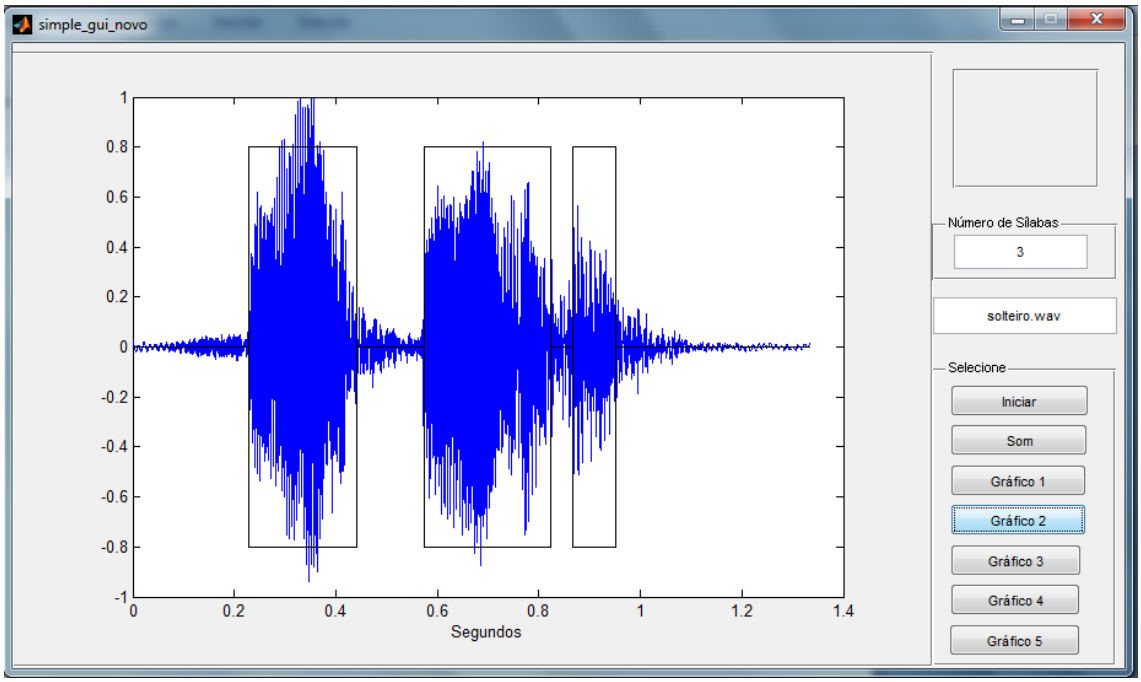


Figura D.37: Separação silábica da palavra “SOLTEIRO”.

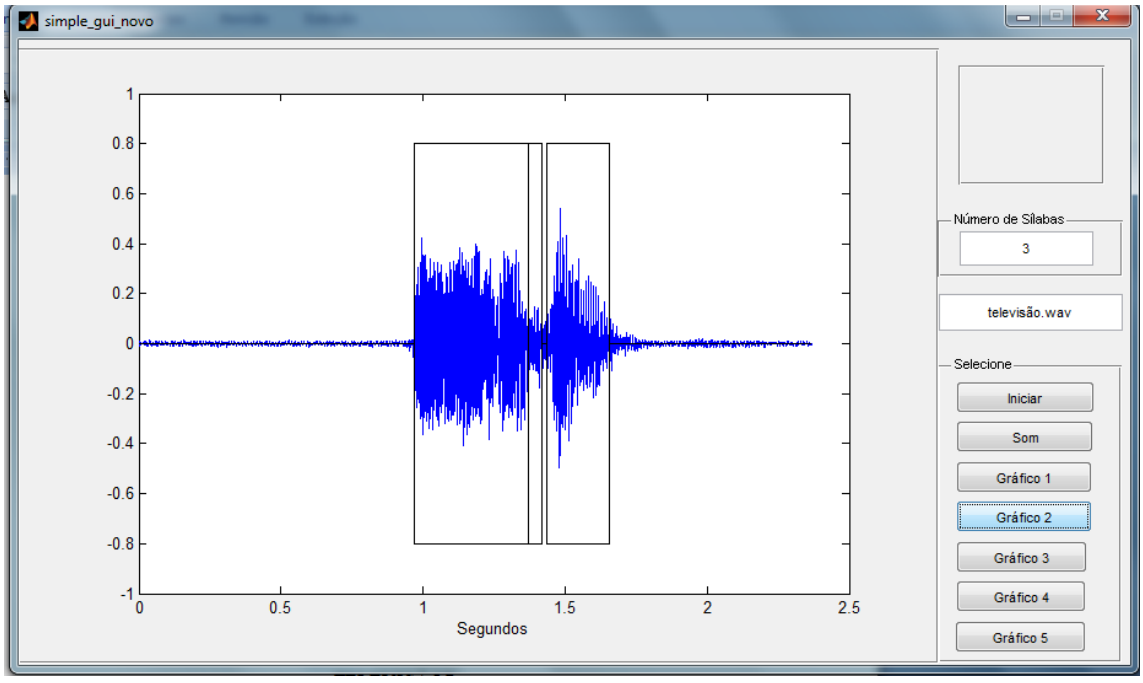


Figura D.38: Separação silábica da palavra “TELEVISÃO”.

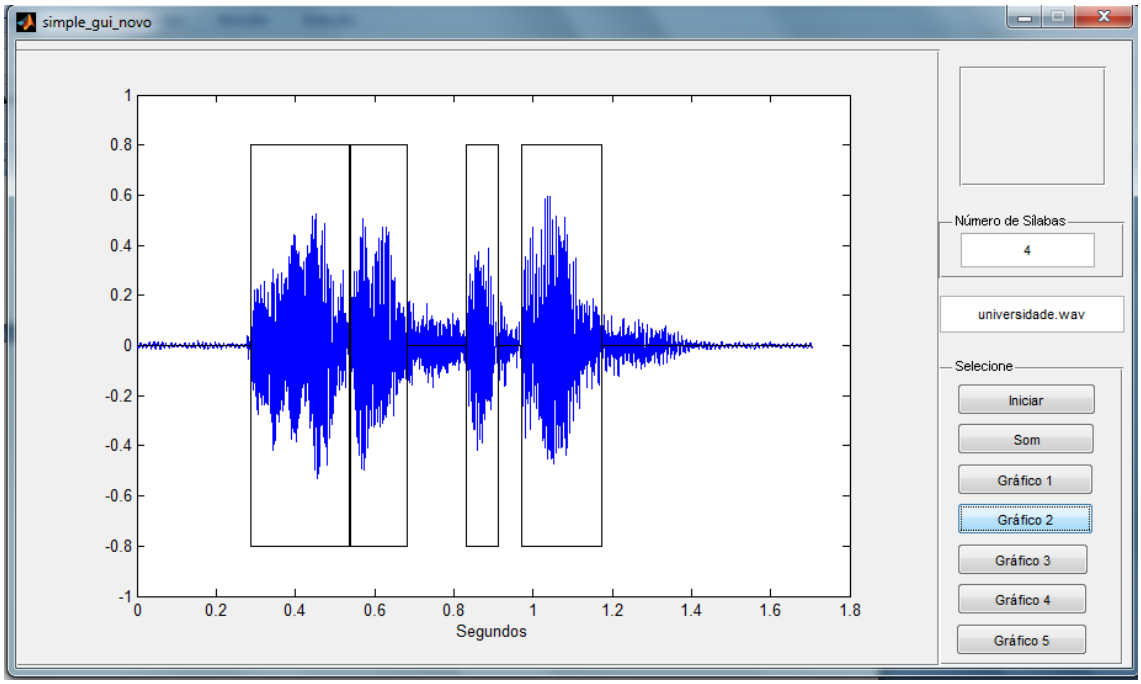


Figura D.39: Separação silábica da palavra "UNIVERSIDADE".

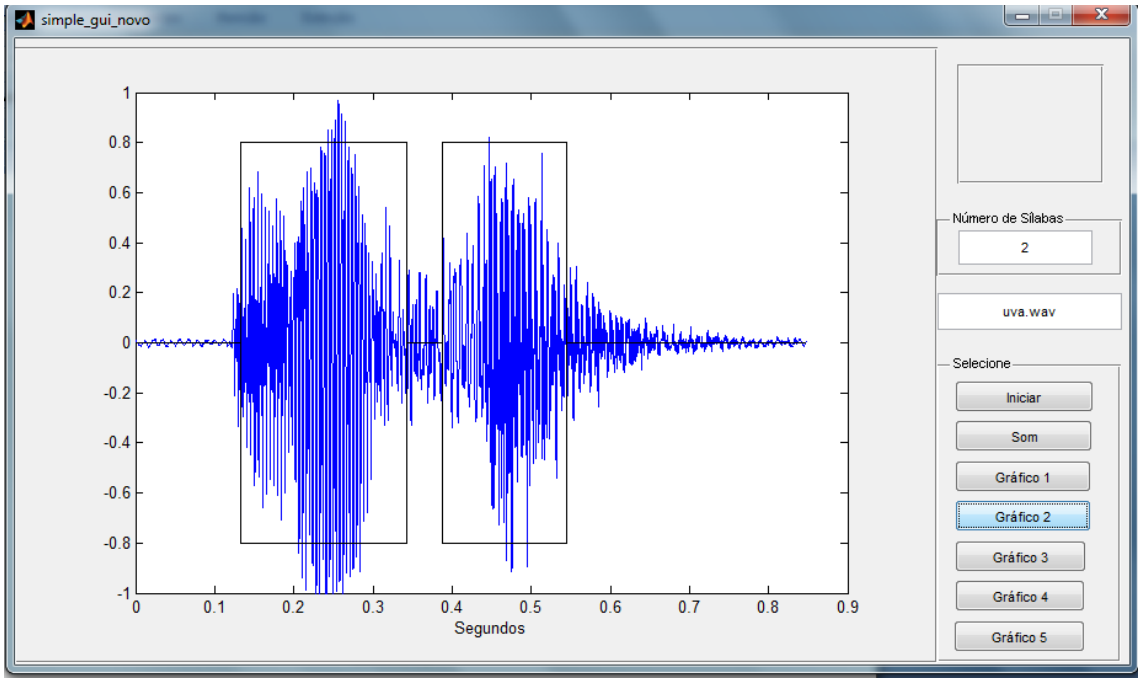


Figura D.40: Separação silábica da palavra "UVA".

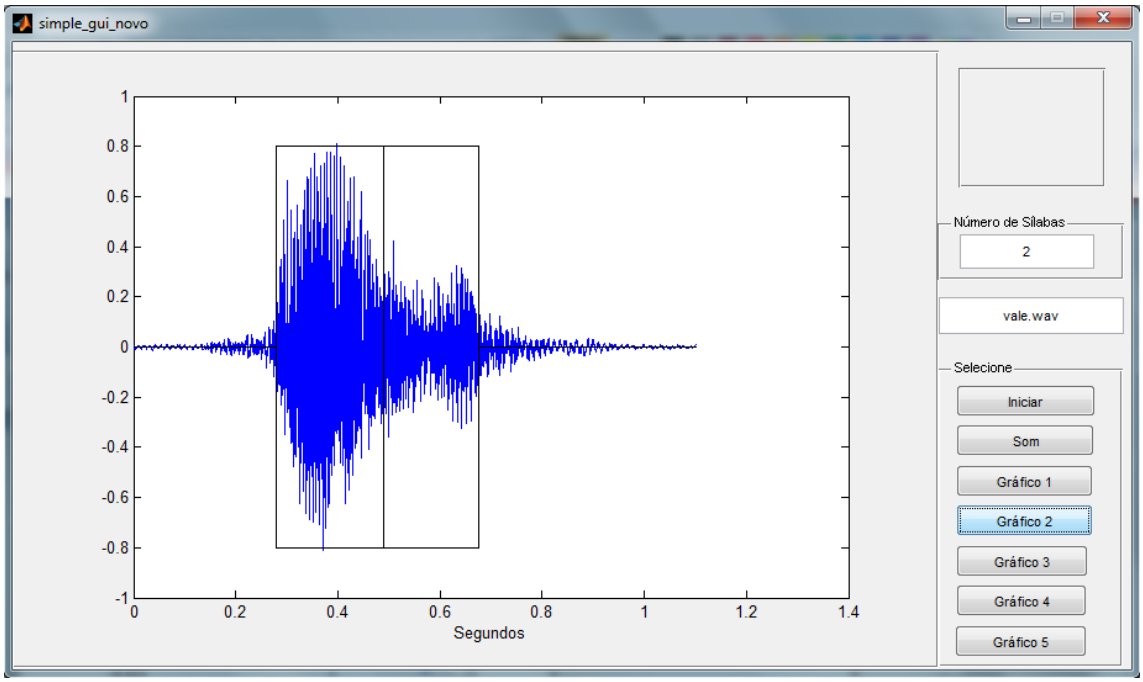


Figura D.41: Separação silábica da palavra “VALE”.

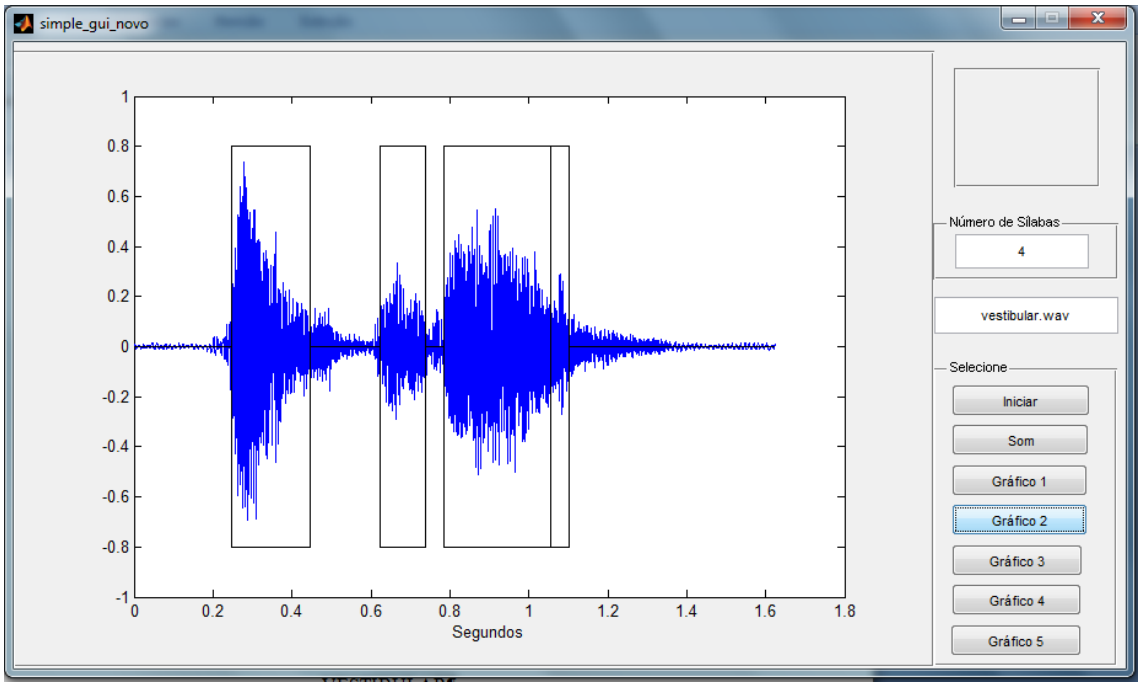


Figura D.42: Separação silábica da palavra “VESTIBULAR”.

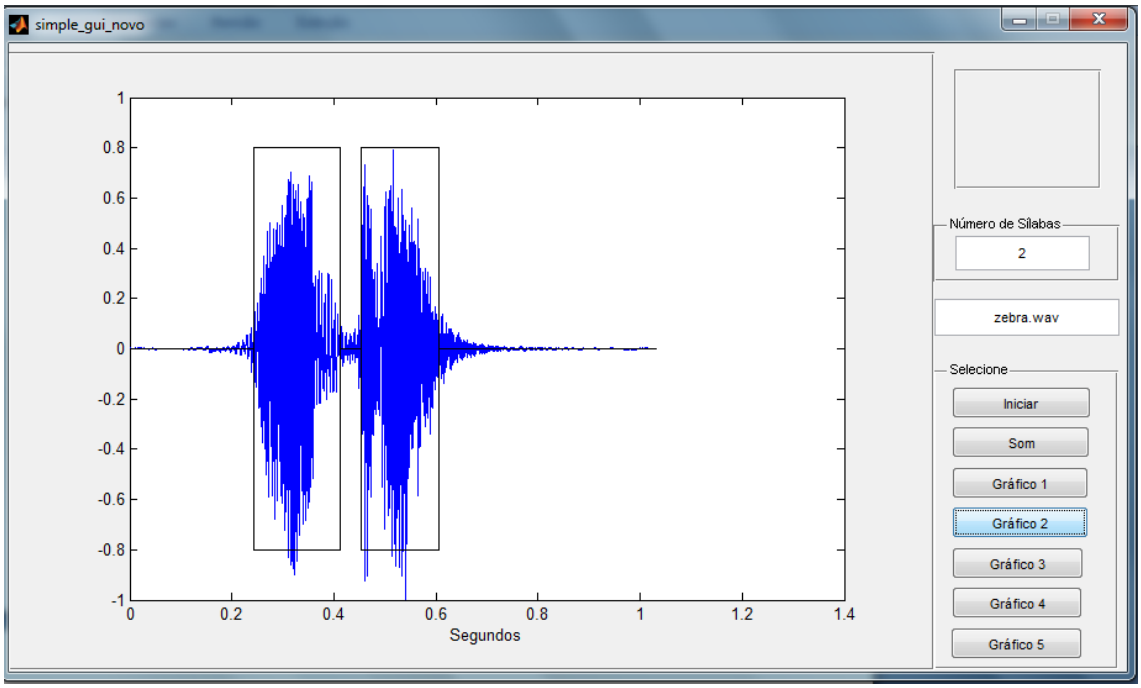


Figura D.43: Separação silábica da palavra “ZEBRA”.

ANEXO E – ANÁLISE DA DIVISÃO SILÁBICA EFETUADA PELO ALGORITMO DE SEPARAÇÃO SÍLABICA PARA POESIA DE MANUEL BANDEIRA

SEPARAÇÃO SILÁBICA REALIZADA DE FORMA MANUAL

SEPARAÇÃO SILÁBICA REALIZADA PELO SEPARADOR AUTOMÁTICO PROPOSTO

/Vou/me/ /em/bo/ra/ /pra/ /Pa/sár/ga/da/

/Vou me/ /-/ /b/ /bo/ /r/ /ra/ /pra/ /Pa/ /s/ /sar/ /gada/

/Lá/ /sou/ /a/mi/go/ /do/ /rei/

/Lá s/ /sou amigo/ /do rei/

/Lá/ /te/nho/ /a/ /mu/lher/ /que/ eu/ /que/ro/

/Lá/ /te/ /nho/ /-/ /mulher/ que /-/ /quer/ /ro/

/Na/ /ca/ma/ /que/ /es/co/lhe/rei/

/Na/ /cam/ /ma/ /que/ /-/ /co/ /lhe/rei

/Vou/me/ /em/bo/ra/ /pra/ /Pa/sár/ga/da/

/Vou me/ /bora/ /pra/ /Pas/ /sárg/ /gada/

/Vou/me/ /em/bo/ra/ /pra/ /Pa/sár/ga/da/

/-/ /bora/ /pra/ /Pas/ /árgad/ /da/

/A/qui/ /eu/ /não/ /sou/ /fe/liz/

/A/ /qui/ /-/ /não/ /sou/ feliz/

/Lá/ /a/ /e/xis/tên/cia/ /é/ uma /a/ven/tu/ra/

/L/ /Lá/ /a existên/ /-/ /é/ /uma/ /-/ /ven/ /tu/ /ra/

/De/ /tal/ /mo/do/ /in/con/se/qüen/te/

/-/ /tal modo/ /inconse/ /quen/ /-/

/Que/ /Jo/a/na/ /a/ /Lou/ca/ /de/ /Es/pa/nha/

/Que Joa/ /na/ /a lou/ /ou/ /ca/ /de/ /-/ /panha/

/Ra/i/nha/ /e/ /fal/sa/ /de/men/te/

/Rai/ /inha e/ /fal/ /sa/ /-/

/Vem/ /a/ /ser/ /con/tra/pa/ren/te/

/Vem/ /m a/ /ser com/ /tra/ /pare/ /te/

/Da/ /no/ra/ /que/ /nun/ca/ /ti/ve/

/-a/ /nor/ /que/ /numc/ /ca/ /ti/ /v-/

/E/ /co/mo/ /fa/rei/ /gi/nás/ti/ca/

/-/ /como/ /f/ /farei/ /-/ /nas/ /-/

/An/da/rei/ /de/ /bi/ci/cle/ta/

/An/ /darei/ /de/ /bi/ /c/ /ci/ /cle/ /ta/

/Mon/ta/rei/ /em/ /bur/ro/ /bra/bo/

/Mon/ /tarei/ /em/ /bur/ /-/ /brab-/

/Su/bi/rei/ /no/ /pau/-/de/-/se/bo/

/Su/ /b/ /bi/ /rei/ /num/ /pau/ /de/ /seb-/

/To/ma/rei/ /ba/nhos/ /de/ /mar!/

/Tomar/ /ei/ /ban/ /nhos/ /de/ /mar/

/E/ /quan/do/ /es/ti/ver/ /can/sa/do/

/-/ /quand/ /d-/ /-s/ /ti/ /ver can/ /sad/ /do/

/Dei/to/ /na/ /bei/ra/ /do/ /rio/

/Deito na/ /beir/ /ra/ /do rio/

/Man/do/ /cha/mar/ /a/ /mãe/-/d'á/gua/

/Man/ /-/ /ch-/ /mar/ /ra/ /a/ /-/ /d'ag/ /gua/

/Pra/ /me/ /con/tar/ /as/ /his/tó/ri/as/

/Pra/ /me/ /con/ /-/ /hi/ /s/ /tori/ /-/

/Que/ /no/ /tem/po/ /de/ /eu/ /me/ni/no/

/-/ /e no tem/ /pó/ /d/ /deu m/ /me/ /nin/ /no/

/Ro/sa/ /vi/nha/ /me/ /con/tar/

/Ros/ /a/ /vinha/ /me/ /con/ /tar/

/Vou/-/me/ /em/bo/ra/ /pra/ /Pa/sár/ga/da/

/Vou me/ /-/ /bora/ /pra/ /P/ /Pas/ /-/ /ga/ /d/ /a/

/Em/ /Pa/sár/ga/da/ /tem/ /tu/do/

/Em/ /Pas/ /-/ /ga/ /d/ /da/ /t/ /em/ /-/ /udo/

É/ /ou/tra/ /ci/vi/li/za/ção/

/É ou/ /tra/ /c/ /civili/ /za/ /ç/ /cão/

/Tem/ /um/ /pro/ces/so/ /se/gu/ro/

/Tem um/ /proc/ /ce/ /esso/ /se/ /-/

/De/ /im/pe/dir/ /a/ /con/cep/ção/

/-/ /im/ /pe/ /dir a/ /a/ /con/ /cep/ /ç/ /cão/

/Tem/ /te/le/fo/ne/ /au/to/má/ti/co/

/Tem/ /tele/ /f/ /on/ /-/ /-/ /to/ /mátc/ /tico/

/Tem/ /al/ca/lói/de/ /à/ /von/ta/de/

/T/ /em al/ /c/ /calói/ /d/ /de/ /-/ /tad/ /de/

/Tem/ /pros/ti/tu/tas/ /bo/ni/tas/

/-/ /pros/ /ti/ /tutas/ /b-/

/Pa/ra/ /a/ /gen/te/ /na/mo/rar/

/Pra g/ /ge/ /nte/ /namor/ /ar e quan/

/E/ /quan/do/ /eu/ /es/ti/ver/ /mais/ /tris/te/

/quando eu-s/ /ti/ /ver/ /mais/ /tris/ /-/

/Mas/ /tris/te/ /de/ /não/ /ter/ /jei/to/

/Mas/ /trist/ /-/ /não/ /ter jei/ /to/

/Quan/do/ /de/ /noi/te/ /me/ /der/

/Quando/ /-/ /noi/ /t-/ /-/ /der/

/Von/ta/de/ /de/ /me/ /ma/tar/

/Von/ /tad/ /de/ /me/ /ma/ /tar/ /r/

/Lá/ /sou/ /a/mi/go/ /do/ /rei/

/Lá/ /s/ /sou ami/ /migo/ /d/ /do rei/

/Te/rei/ /a/ /mu/lher/ /que/ /eu/ /que/ro/

/Terei a/ /um/ /lher/ /que eu/ /quer/ /ro/

/Na/ /ca/ma/ /que/ /es/co/lhe/rei/

/Na/ /cama/ /que/ /-s/ /colhe/ /rei/

/Vou/-/me/ /em/bo/ra/ /pra/ /Pa/sár/ga/da/

/Vou me/ /b/ /borá/ /pra/ /Pasá/ /arga/ /da/

REFERÊNCIAS BIBLIOGRÁFICAS

AHMADI, S.; SPANIAS, A. S., 1999. Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm. In: Transactions on Acoustics, Speech, and Signal Processing, V. 7, n^0. 3, pp. 333 - 338.

BARBOSA, D.C.P. *Análise de Sistemas de Telefonia IP em Redes Par-a-Par Sobrepostas*. Dissertação de Mestrado, Comunicações - Programa de Pós-Graduação em Engenharia Elétrica, UFPE, 2009.

BARNARD, E; COLE, R.A; VEA, M.P.; ALLEVA, F.A., 1991. *Pitch Detection with a Neural-Net Classifier*. IEEE Transactions on Signal Processing, V. 39, n^0. 2, pp. 298 – 307.

BEIGE, H. *Fundamentals of Speaker Recognition*. Springer, 2011.

BISTAFA, B. S. *Acústica Aplicada Ao Controle do Ruído*. São Paulo, Editora Blucher, 2006.

BOERSMA, P., 1993. Accurate *Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound.* In: Institute of Phonetic Sciences, University of Amsterdam. Anais, pp. 97 – 110.

BOUMAN, C. A. *CONNEXIONS CONSORTIUM - Inside Collection (Course): Purdue Digital Signal Processing Labs (ECE 438). Lab 9a - Speech Processing (part 1).* Disponível em: http://cnx.org/content/m18086/latest/?collection=col10593/latest. Acesso em: 23/01/2012.

BRANDÃO, A. S.; CATALDO, E.; LETA, F.R., 2007. *Um Novo Método Usando Autocorrelação para Extração da Freqüência Fundamental em Sinais de Voz.* TEMA. Tendências em Matemática Aplicada e Computacional, V. 8, n^0. 2, pp. 191-200.

CASIERRA, J. P. C. *Implementação de um Sistema Esteganográfico para Inserção de Textos em Sinais de Áudio*, Brasil. Dissertação de Mestrado, Comunicações - Programa de Pós-Graduação em Engenharia Elétrica, UFPE, 2009.

CHARPENTIER, F.J., 1986. *Pitch Detection Using The Short-Term Phase Spectrum*. In: International Conference on Acoustics, Speech, and Signal Processing, on ICASSP' 86, anais, pp. 113 - 116.

CHU, W. C. *Speech coding algorithms – Foundation and Evolution of Standardized Coders*. First Edition, United States of America, John Wiley & Sons, Inc., 2003.

CIPRO, P. N; INFANTE, U. *Gramática da Língua Portuguesa*. Terceira Edição, Brasil, Scipione, 2009.

COLEMAN, J. S. Material do curso: "Advanced Linguistics: Biological Foundations of

Language". Módulo 2, Speaking and hearing. Disponível em: http://www.phon.ox.ac.uk/jcoleman/speaking_hearing.htm. Acesso em: 07/02/2012.

da SILVA, E. L. F. e de OLIVEIRA, H. M. de Oliveira, 2012. *Estimativa do comportamento vocálico de locutores.* In: Submetido ao Simpósio Brasileiro de Telecomunicações, Brasília.

de OLIVEIRA, H. M. *Análise de Fourier e Wavelets: Sinais Estacionários e não Estacionários*. Brasil, Editora Universitária da UFPE, 2007.

de OLIVEIRA, H. M. *Fundamentos de Engenharia de Telecomunicações*. 2012. Disponível em: http://www2.ee.ufpe.br/codec/engenharia_telecomunicacoes.pdf.

dos SANTOS, S. C. B. e ALCAIM, A, 2001. *Sílabas Como Unidades Fonéticas para o Reconhecimento Automático de Voz Contínua em Português*. In: SBA Controle & Automação. Anais, pp. 64-70.

dos SANTOS, S. C. B. *Reconhecimento de Voz Contínua Para o Português Utilizando Modelos de Markov Escondidos*. Tese de doutorado, Engenharia Elétrica – Departamento de Engenharia Elétrica – PUC, 1997.

FAGUNDES, R. D. R., ZWETSCH, I. C. e SCOLARI, D, 2008. *Scolari. Técnicas de processamento de áudio em sinais de voz, para auxílio diagnóstico de doenças laríngeas.* In: 6^0 Congresso de Engenharia de Áudio. Anais, pp.24-28.

FRAGA, F. J, 2001. *Conversão Fala-texto para o Português com Segmentação Sub-silábica e Vocabulário Ilimitado*. Revista de Telecomunicações, INATEL, V. 4, n^0. 2, pp.44-50, 2001.

GOUVEIA, P. D. F.; TEIXEIRA, J. P. R. e FREITAS D, 2000. *Divisão Silábica Automática do Texto Escrito e Falado.* In: Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. Anais, pp.65-74.

HARBECK, S.; KIEβLING, A.; KOMPE, R.; NIEMANN, H. *Robust Pitch Period Detection Using Dynamic Programming With an ANN Cost Function*, 1995. In: 4 th European Conference on Speech Communication and Technology EUROSPEECH. Anais, pp. 1337 – 1340.

HOLMES, J.; HOLMES, W. *Speech Synthesis and Recognition.* Second Edition, Taylor & Francis, 2001.

HUANG, X.; ACERO, A. e HON, H. W. *Spoken Language Processing*. New Jersey, Prentice-Hall, 2001.

JANER, L.; BONETT, J.J.; SOLANO, E. L., 1996. *Pitch Detection and Voiced/Unvoiced Decision Algorithm based on Wavelet Transforms*. In: International Conference on Spoken

Language ICSLP. Anais, pp. 1209 – 1212.

KADAMBE, S.; BOUDREAUX, G.F.B., 1991. *A Comparison of a Wavelet functions for Pitch Detection of Speech Signals*. In: International Conference on Acoustics, Speech, and Signal Processing ICASSP. Anais, pp. 449 - 452.

KADAMBE, S.; BOUDREAUX, G.F.B., 1992. *Application of the Wavelet Transform for Pitch Detection of Speech Signals*. In: Transactions on Information Theory, V. 38, n^0. 2, pp. 917 – 924.

KONDOZ, A. M. *Digital Speech – Coding for Low Bit Rate Communication Systems*. Second Edition, United Kingdom, John Wiley & Sons Ltd, 2004.

LATHI, B. P. *Modern Digital and Analog Communications Systems.* Third Edition, Oxford University Press, 1998.

LI, H.; DAI, B.Q.; WEI, L., 2006. *A Pitch Detection Algorithms Based on AMDF and* ACF. In: International Conference on Acoustics, Speech, and Signal Processing on ICASSP. Anais, pp. 377 - 380.

LIPEIKA, A; LIPEIKIENÉ, J. e TELKSNY, L, 2002. *Development of Isolated Word Speech Recognition.* Informatica. Vol. 13, n^0. 1, pp 37-46.

MARTIN, P., 1982. *Comparison of Pitch Detection by Cepstrum and Spectral Comb analysis.* In: International Conference on Acoustics, Speech, and Signal Processing ICASSP. Anais, pp. 180 - 183.

NETO, N.; SILVA, E. e SOUSA, E., 2005. *Software Usando Reconhecimento e Síntese Voz: O Estado da Arte para o Português Brasileiro*. In: Latin American conference on Human-computer interaction. Anais, pp. 326-331.

OH, K. A.; UN, C. K., 1984. *A Performance Comparison of Pitch Extraction Algorithms for Noisy Speech.* In: International Conference on Acoustics, Speech, and Signal Processing ICASSP. Anais, pp. 85 - 88.

OPPENHEIM, A. V.; SCHAFER, R. W. *Discrete-Time Signal Processing*. Third Edition, United States of America, Prentice-Hall, 2010.

OPPENHEIM, A. V.; WILLSKY, A. S. *Sinais e Sistemas*. Segunda Edição, Prentice-Hall, 2010.

PROAKIS, J. B.; *Digital Communications*. Second Edition, McGraw-Hill, 1989, pp. 107.

RABINER, L. R.; CHENG, M. J.; ROSENBERG, A; MCGONEGAL, C. A., 1976 a. *Some Comparisons Among Several Pitch Detection Algorithms*. In: International Conference on Acoustics, Speech, and Signal Processing on ICASSP' 76. Anais, pp. 332 – 335.

RABINER, L. R.; CHENG, M. J.; ROSENBERG, A; MCGONEGAL, C. A, 1976 b. *Comparative Performance Study of Several Pitch Detection Algorithms*. In: Transactions on Acoustics, Speech, and Signal Processing, V. 24, n^0. 5, pp. 399 – 418.

RABINER, L. and JUANG, B. H. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

RABINER, L. R. and SCHAFER, R. W. Digital Processing of Speech Signal. Prentice-Hall, 1978.

RABINER, L. R. and SCHAFER, R. W. *Introduction to Digital Speech Processing*. Now Publishers, 2007.

ROSENBERG, A. E.; RABINER, L.R.; WILPON, J.G. e KAHN, D., 1983. *Demisyllable-Based Isolated Word Recognition System*. In: IEEE Transactions on Acoustics, Speech and Signal Processing. Anais, pp.713-726.

RUMSEY, F.; MCCORMICK, T. *Sound and Recording.* Fifth Edition, Elsevier, 2006.

SAMAD, S.A.; HUSSAIN, A.; FAH, L. K., 2000. *Pitch Detection of Speech Signals using the Cross-Correlation Technique.* In: TENCON 2000. Anais, pp. 283 - 286.

SELMINI, A. M. *Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala*. Brasil, 2008. Tese de doutorado, Engenharia de Telecomunicações – Engenharia Elétrica e da Computação, UNICAMP.

SILVA, P.; NETO, N e KLAUTAU, A, 2009. *Novos Recursos de Utilização de Adaptação de Locutor no Desenvolvimento de um Sistema de Reconhecimento de Voz para o Português*. In: SBrT Simpósio Brasileiro de Telecomunicações. Anais.

SILVA, P.; NETO, N.; KLAUTAU, A.; ADAMI, A. e TRANCOSO, I, 2008. Speech *Recognition for Brazilian Portuguese using the Spoltech and OGI-22 Corpora*. In: SBrT Simpósio Brasileiro de Telecomunicações. Anais.

SMITH, S.W. Digital *Signal Processing – A Practical Guide for Engineers and Scientists*. Newnes, 2003.

SOTERO, R. F. B. *Novas Abordagens para Codificação de Voz e Reconhecimento Automático de Locutor Projetadas Via Mascaramento Pleno em Frequência por Oitava*. Dissertação de Mestrado, Comunicações - Programa de Pós-Graduação em Engenharia Elétrica, UFPE, 2009.

SOTERO, R. F. B.; de OLIVEIRA, H. M, 2009. *Reconhecimento de Locutor baseado em Mascaramento Pleno em Frequência por Oitavas*. In: 7^o Congresso de Engenharia de Áudio. Anais, pp.61-66.

VASEGHI, S.V. *Multimedia Signal Processing - Theory and Applications in Speech, Music and*

Communications. First Edition, John Wiley & Sons Ltd, 2007.

VERLAG DASHÖFER: *Segurança, Higiene e Saúde do trabalho*. Disponível em: http://higieneseguranca trabalho.dashofer.pt/library/8f14e45fceea167a5a36dedd4bea25437/images/. Acesso em: 07/02/2012.

VERMEHREN, V; WESEN, J. E; de OLIVEIRA, H.M, 2010. Close *Approximations for Daublets and their Spectra*. In:SBrT/ITS International Telecommunication Symposium. Anais.

XUEJING, S; 2002 a. *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio*. In: IEEE Internacional Conference Acoustics, Speech, and Signal Processing, ICASSP. Anais, pp.333-336.

XUEJING, S; 2002 b. Código em Matlab do algoritmo *Pitch Determination*. Disponível em**:** http://www.speakingx.com/blog/2008/01/02/pitch-determination. Acesso em: 08/04/2012.

YING, G.S; JAMIESON, L H; MICHELL., C. D, 1996. *A Probabilistic Approach to AMDF Pitch Detection*. In: 4 th International Conference on Spoken Language ICSLP. Anais, pp. 1201 – 1204.

Catalogação na fonte
Bibliotecário Marcos Aurélio Soares da Silva, CRB-4 / 1175

S586e Silva, Elda Lizandra Fernandes da.
Estimativas de comportamento vocálico de locutores e um novo sistema de separação silábica / Elda Lizandra Fernandes da Silva. - Recife: O Autor, 2012.
xii, 150 folhas, il., gráfs., tabs.

Orientador: Profº Drº. Hélio Magalhães de Oliveira.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Elétrica, 2012.
Inclui Referências e Anexos.

1. Engenharia Elétrica. 2. Caracterização de Voz. 3. Processamento da Fala. 4.Sons Vocálicos. I. Oliveira, Hélio Magalhães de (Orientador). II. Título.

621.3 CDD (22. ed.) UFPE
BCTG/2013-003

Universidade Federal de Pernambuco
Pós-Graduação em Engenharia Elétrica

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE DISSERTAÇÃO DO MESTRADO ACADÊMICO DE

ELDA LIZANDRA FERNANDES DA SILVA

TÍTULO

"ESTIMATIVAS DE COMPORTAMENTO VOCÁLICO DE LOCUTORES E UM NOVO SISTEMA DE SEPARAÇÃO SILÁBICA"

A comissão examinadora composta pelos professores: RICARDO MENEZES CAMPELLO DE SUZA, DES/UFPE, JULIANO BANDEIRA LIMA, DM/ UFPE e ADRIÃO DUARTE DÓRIA NETO, EC/UFRN sob a presidência do primeiro, consideram a candidata **ELDA LIZANDRA FERNANDES DA SILVA APROVADA.**

Recife, 28 de maio de 2012.

CECÍLIO JOSÉ LINS PIMENTEL
Coordenador do PPGEE

RICARDO MENEZES CAMPELLO DE SOUZA
Membro Titular Interno

ADRIÃO DUARTE DÓRIA NETO
Membro Titular Externo

JULIANO BANDEIRA LIMA
Membro Titular Externo